Reports in Phonetics
Umeå University

December 2001

8

Focal accent — f_0 movements and beyond

Mattias Heldner

PHONUM

Mattias Heldner Focal accent — f_0 movements and beyond PHONUM 8 Phonetics Umeå

Focal accent – f_0 movements and beyond

Akademisk avhandling

som med tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen offentligen kommer att försvaras i Hörsal F, Humanisthuset, lördagen den 15 december 2001 klockan 10.00

av

Mattias Heldner
Fil. Mag.

Fakultetsopponent
Professor Gösta Bruce,
Institutionen för lingvistik,
Lunds universitet

**INSTITUTIONEN FÖR FILOSOFI OCH LINGVISTIK,
UMEÅ UNIVERSITET, UMEÅ 2001**

Focal accent – f_0 movements and beyond

Heldner, Mattias
Department of Linguistics and Philosophy, Umeå University

PHONUM 8, Reports in Phonetics Umeå University
ISBN 91-7305-133-0 ISSN 1101-2714

Abstract

The seven papers presented in this thesis are all, in one way or another, concerned with focal accent in Swedish. What is being investigated, more specifically, are its acoustic correlates and the relevance of these correlates as perceptual cues. The purpose of the investigations is to add perceptually relevant details to the description of focal accent. Needless to say, fundamental frequency movements – or f_0 movements as they will be referred to in the following – are of a great importance for the signaling of accents in general; this is already a well-established fact. Most of the efforts put into the research presented here have therefore been directed towards acoustic correlates other than f_0 movements.

Paper I provides an important background for the rest of the papers. Although it was primarily meant to be an evaluation of an IPA based system for the transcription of Swedish prosody, it also revealed that prosodic transcription is by no means a trivial task, and especially not the transcription of prominence including focal accents. This observation, in turn, initiated the exploration of the acoustic basis for the variability associated with the labeling of prominence, and especially the acoustic basis for the perception of focally accented words. Papers II and III deal with the importance of the focal accent rise for the perception of prominence.

These studies were followed by two fairly large production experiments. Because a number of previous studies had shown that focally accented words in Swedish, as well as accented words in many other languages, in addition to the f_0 movements also tend to be produced with longer duration than non-focused words, Paper IV deals with the temporal effects of focal accents. Moreover, as some kind of loudness variation is also intuitively felt to be part of the signaling of prominence distinctions, Paper VI deals with two features related to perceived loudness – i.e. overall intensity and spectral emphasis. Both production experiments were followed up by smaller scale perceptual experiments (Papers V and VII).

These studies have confirmed that a focally accented word, in addition to specific f_0 movements and especially the focal accent rise, is characterized by longer segmental durations, higher overall intensity and higher spectral emphasis, and that several acoustic features contribute to the perceived naturalness of focally accented words. Although the studies presented in this thesis were primarily undertaken to gain a better understanding of the acoustic correlates associated with focal accents in Swedish and their perceptual relevance, the results may also be put to use in various speech technology applications.

Key words: Focal accent, Swedish, prosody, acoustic correlates, perceptual cues, f_0 movements, duration, overall intensity, spectral emphasis.

CODEN: UM/PHON/R-01/0008

Reports in Phonetics
Umeå University

December 2001

8

Focal accent – f_0 movements and beyond

Mattias Heldner

PHONUM

Mattias Heldner: Focal accent – f_0 movements and beyond

PHONUM 8
Department of Philosophy and Linguistics
Umeå University
SE-901 87 Umeå, Sweden

ISBN 91-7305-133-0
ISSN 1101-2714

Cover by Alexander Pankow
Nyheternas Tryckeri, Umeå 2001

PHONUM is distributed on the basis of mutual exchange and is also available via the Department's Internet pages: http://www.ling.umu.se/

Acknowledgments

Dear reader: The author would like to express his deep gratitude to all the people who have helped him in his life, and with this thesis, including those who deliberately and frequently interrupted his work in order to improve his life. You know who you are, thanks for all the love and support, all the friendship, all the laughing, all the wine, all the Neapolitan coffee breaks, all the breakfasts, luncheons and dinners etc. It has taken a lot of these things to write a thesis…

As far as those who encouraged his work are concerned, the sincerest thanks for all valuable comments and suggestions, for help with practical and financial details, and, in some cases, for the valid criticism. It is assumed that you know who you are, but in case you have forgotten, a list has been compiled for your convenience.

First of all, the author is endlessly thankful to his supervisor, Eva Strangert, who just happens to belong to both the interrupting and encouraging types of people acknowledged above – Eva has literally been involved in all aspects of this thesis. Thank you and thank you again! Thanks also to all other past and present colleagues at the Department of Philosophy and Linguistics at Umeå University.

A substantial part of the research presented here was carried out while the author was a guest at the Centre for Speech Technology (CTT) at KTH in Stockholm, and he is extremely grateful to Björn Granström, Rolf Carlson and the rest of the people there for welcoming him to this stimulating environment.

Among all of those who have contributed to this work with comments and discussions – and the occasional valid criticism – the author would especially like to mention Linda Bell, Rolf Carlson, Anders Eriksson, Gunnar Fant, Christina Heldner, Inger Karlsson, Béata Megyesi and Hartmut Traunmüller.

The research was supported by grants from HSFR, the Swedish Council for Research in the Humanities and Social Sciences, and from the Faculty of Humanities at Umeå University. This support is gratefully acknowledged.

And finally, before these acknowledgments begin to cut into the cocktail hour, the author wishes to thank the love of his life, Magdalena.

Umeå, December 2001
Mattias Heldner

Is that all there is?
Is that all there is?
If that's all there is, my friends, then let's keep dancing
Let's break out the booze and have a ball
If that's all there is
Peggy Lee

Contents

List of original publications

The thesis is based on the following papers, referred to in text by their Roman numerals.

I. Strangert, E. & Heldner, M. (1995) Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In *PHONUM 3,* pp. 85-109. Umeå: Department of Phonetics, Umeå University.

II. Strangert, E. & Heldner, M. (1995) The labelling of prominence in Swedish by phonetically experienced transcribers. In *Proceedings ICPhS 95,* pp. 204-207. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

III. Heldner, M. & Strangert, E. (1997) To what extent is perceived focus determined by f_0-cues? In *Eurospeech '97 Proceedings*, pp. 875-877. Rhodes, Greece: ESCA.

IV. Heldner, M. & Strangert, E. (2001) Temporal effects of focus in Swedish, *Journal of Phonetics,* **29**(3), pp. 329-361.

V. Heldner, M. (2001) On the non-linear lengthening of focally accented Swedish words. In *Nordic Prosody: Proceedings of the VIIIth Conference, Trondheim 2000* (W. van Dommelen & T. Fretheim, eds.), pp. 103-112. Frankfurt am Main: Peter Lang.

VI. Heldner, M. (forthcoming) On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish, Pending revisions for possible publication in *Journal of Phonetics*.

VII. Heldner, M. (2001) Spectral emphasis as a perceptual cue to prominence, *TMH-QPSR,* **2/2001**, 51-57.

Focal accent – f_0 movements and beyond: Introduction
pp. 9–27

Focal accent – f_0 movements and beyond: Introduction

Mattias Heldner

1. General introduction

The seven papers presented in this thesis are all, in one way or another, concerned with focal accent in Swedish. What is being investigated, more specifically, are its acoustic correlates and the relevance of these correlates as perceptual cues. The purpose of the investigations is to add perceptually relevant details to the description of focal accent. Needless to say, fundamental frequency movements – or f_0 movements as they will be referred to in the following – are of a great importance for the signaling of accents in general; this is already a well-established fact. Most of the efforts put into the research presented here have therefore been directed towards acoustic correlates other than f_0 movements.

1.1. Focal accent and the Swedish model of intonation

'Focal accent' is a term used in the Swedish model of intonation (Bruce, 1977; Bruce & Gårding, 1978; Gårding & Bruce, 1981; Bruce, Granström, Gustafson, Horne, House & Touati, 1997; Bruce, 1999) about an accent signaling that a word (or some other constituent within a phrase which may be smaller or larger) is 'focused' or 'in focus'. 'Focal accent', in this sense, is equivalent to the term 'sentence accent' used in earlier descriptions (e.g. Bruce, 1977). 'Focused' is also the term used for the highest level of prominence in the Swedish intonation model.

There are three phonologically distinct prominence levels apart from unstressed in this model: 'stressed', 'accented', and 'focused'. Furthermore, these prominence levels apply to units of different sizes, and within domains of different sizes. At the lowest level – i.e. 'stressed' – there is an alternation of stressed and unstressed syllables within the foot domain. At the next higher level – i.e. 'accented' – feet (containing at least one stressed syllable) within words may be accented or unaccented. Here, there exists a choice between two word accents, called acute accent and grave accent, or accents I and II. These accents are distinctive in many Swedish dialects, but there is no difference in prominence between them. Finally, words in (prosodic) phrases may or may not be 'focused', which is the highest level of prominence. Word and focal accents thus amplify the prominence brought about by the stressed syllable and a focused word contains at least one stressed syllable, a word accent, and a focal accent.

This model has been the starting point for the investigations which constitute the subject-matter of the present thesis. Typically, 'focused' or focally accented words have been compared to merely 'accented' or non-focused ones in the studies reported.

It deserves to be emphasized at this point that the Swedish model distinguishes between three levels of prominence – 'stressed', 'accented', and 'focused' – whereas models for other

languages typically have two levels of prominence, usually ‘stressed’ and ‘accented’. As a consequence, what is called ‘focused’ in the Swedish model – i.e. the highest level of prominence – is the equivalent of ‘accented’ in many other prominence models. The extra middle level in the Swedish model – i.e. ‘accented’ – is needed to account for the word accent prominence level mentioned above.

1.2. Focus

The notion of focus is used here in the same way as for instance in Ladd (1980), Gussenhoven (1984), Nooteboom & Kruyt (1987), or Ladd (1996). Focus in this sense is taken to mean that single words or other constituents in utterances can be put in focus by the speaker in order to indicate that they convey new information to the listener or should otherwise be felt as salient for some reason. Thus, focusing is a means used by the speaker to facilitate the listener's understanding by directing his or her attention to the most important parts of the message.

The word (or any other constituent) put in focus is usually made more prominent, that is, standing out from its environment. Such prominence can be brought about by different means. There are, for example, structural means of achieving prominence such as topicalization and clefting. However, these topics will not be addressed here. Moreover, prominence may be brought about by prosodic means such as accents (or focal accents) associated with the prosodic head of the focus domain. The acoustic means for achieving this type of prosodic prominence in Swedish and their relevance as perceptual cues will be the main concern of the present thesis.

1.3. Acoustic correlates vs. perceptual cues

The distinction between ‘acoustic correlates’ and ‘perceptual cues’ remains important throughout this thesis. Here, the term ‘acoustic correlates’ will be taken to mean the acoustic signaling as produced by the speaker, while the term ‘perceptual cues’ will be used to refer to the acoustic information that can be relevant for and used by the listener. It should be stressed here that these two notions do not necessarily refer to exactly the same physical reality.

An acoustic correlate as measured by some technique may well be totally irrelevant as a perceptual cue, in spite of its being systematically present. An obvious example is what happens when a certain acoustic change is not perceivable; that is, when it does not exceed the difference limen (DL), the differential threshold, or the just noticeable difference (JND) for the acoustic dimension in question. There is a vast literature on DLs and JNDs for various acoustic features, and it is beyond the scope of this thesis to cover it. Just to mention a few studies, however, one could notice that Klatt & Cooper (1975) and Klatt (1976) studied JNDs for segment duration, Klatt (1973) and t' Hart (1981) differential thresholds for fundamental frequency, Flanagan (1955) DLs for intensity, and Mermelstein (1978) and Kewley-Port & Zheng (1999) DLs for formant frequencies.

A consistent acoustic correlate may also in principle be irrelevant for perception in case the measured effect is just a by-product of, or totally dependent on, some other acoustic correlate. Listeners certainly do not have to use all the acoustic information available. Ultimately, it might also be the case that a given acoustic effect is simply an artefact caused by the measuring technique. So even if acoustic measurements are to be considered as crucial for the modeling of

prosodic phenomena, it must also be shown that any acoustic differences that have been observed do have a perceptual relevance.

It might, however, be worthwhile to contemplate the meaning of the term 'perceptual cue' a little further. What does it mean to be "relevant for and used by the listener"? Precisely what is the listener supposed to use the perceptual cues for? Let us explore whether the necessary/sufficient distinction could be of any help in formulating a definition of the notion of a perceptual cue.

To begin with, it is highly unlikely that any acoustic feature will be 'necessary' in the sense that it has to be present for the listener to perceive a prosodic category. As many authors have noticed in passing, it is quite possible to signal various prosodic categories even when the supposedly most important acoustic feature for signaling these categories – i.e. f_0 – is absent, as for instance in whispered speech. A definition in terms of necessary features is therefore probably too strict.

Instead, it might seem more plausible to define perceptual cues as acoustic features which 'in themselves' are sufficient to signal a prosodic category, and where the expression 'sufficient in themselves' is used in the strict sense, meaning that if a particular feature is present, this is enough for a listener to be able to perceive a given prosodic category, also when conflicting information is present. However, this definition may also turn out to be too strict. As we will see, there are indications that the supposedly most important acoustic feature – i.e. f_0 – is not in itself strong enough to outweigh conflicting information under certain circumstances (cf. Paper III, Heldner & Strangert, 1997). Again, there is a definite risk that no features will qualify as perceptual cues under such a definition.

What about a definition, then, requiring that the feature be sufficient only when no conflicting information is present? Should not in fact a feature that changes the percept from one prosodic category into another against a neutral background qualify as a perceptual cue? Intuitively the answer would be yes, but there is nevertheless a risk that the definition is not restrictive enough. The fact that some piece of information is being used in an experimental situation where no other information is present merely shows that listeners *may* use the perceivable information that is present. It does not show that they also use it under normal circumstances. Coughs or hiccups could probably be shown to be perceptual cues in this meaning. If one considers this to be an unwanted effect, then the proposed definition is too unrestricted.

Furthermore, in the hypothetical definitions of prosodic cues stated above, only the signaling of categorical distinctions (e.g. focused vs. non-focused) was considered. For the purpose of this thesis – which is, as has already been said, to add perceptually relevant details to the description of focal accents in Swedish – interest reaches beyond the signaling of these categorical distinctions. In particular, acoustic features affecting the perceived naturalness of (and within) prosodic categories are also of potential interest.

Therefore, a perceptual cue will tentatively be defined here as an acoustic feature either influencing a categorical distinction or affecting the perceived naturalness within a prosodic category.

1.4. The primacy of f_0 in models of prosody

It has long been recognized that f_0 is of primary importance in models of prosody. If ever there was a candidate for being such a thing as an 'in itself sufficient cue' for the perception of accents,

this candidate would certainly be an f_0 movement or an f_0 target. All other acoustic features have been considered in some sense secondary, either because they may not in themselves affect categorical distinctions, or because they are believed to be totally dependent on f_0. It has, for example, been claimed that f_0 movements are intended acoustic effects with a specific linguistic function, whereas durational adjustments are merely a function of the f_0 changes within the affected segments (Lyberg, 1979; Öhman, Zetterlund, Nordstrand & Engstrand, 1979; Lyberg, 1981b; Lyberg, 1981a).

The reason why f_0 is considered primary for models of prosody is obviously the observation that f_0 is the most important acoustic correlate, and that it is also the most important perceptual cue for various prosodic categories, including for instance accents (e.g. Bolinger, 1958; Fry, 1958; van Katwijk, 1974; Bruce, 1977; Beckman, 1986; t' Hart, Collier & Cohen, 1990), and major prosodic phrase boundaries (Beckman & Pierrehumbert, 1986; t' Hart *et al.*, 1990). The same holds true for the focal accents in Swedish (Bruce, 1977) – the prosodic category of special interest in the present thesis. This awareness is of course reflected in prosodic terminology and in various kinds of prosodic models and descriptions.

One apparent reflection of this regards terminology. It may be noticed that the terms 'prosody' and 'intonation' have often been used interchangeably in the literature (see discussion in Hirst & Di Cristo, 1998b). Since the term 'intonation' has also been used in the more restricted sense of descriptions of "linguistically relevant, suprasegmental, non-lexical aspects of the fundamental frequency (f_0) variation" (cf. Grønnum, 1995), an inattentive reader might get the impression that f_0 is all there is to prosody.

It also seems to be the case that investigations of f_0 patterns in different prosodic categories (see for example the survey of intonation systems in Hirst & Di Cristo, 1998a) constitute quite a common approach to the study of prosody. However, far from all intonation models have been supplemented with descriptions of other physical characteristics of the prosodic categories.

Furthermore, prosodic categories are often described and symbolized in terms of their f_0 contours. There are, for example, several transcription systems (with accompanying intonation models) that describe prosodic phenomena in terms of their f_0 events. Within the ToBI (Tones and Break Indices) framework, categorically distinct intonation patterns and prosodic structures are described in terms of high and low f_0-targets and indices of prosodic groupings (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg, 1992; Beckman & Ayers Elam, 1997). ToBI conventions have been developed for a number of languages (cf. the ToBI homepage at http://www.ohio-state.edu/~tobi/).

Another example is the Swedish model of intonation, where accentuation and phrasing – including boundary signaling and other intonation features – are described in terms of high and low f_0-targets associated with stressed syllables or boundaries (e.g. Bruce, 1977; Bruce *et al.*, 1997; Bruce, 1999). We will return to the Swedish model of intonation in the next section.

In the IPO description of intonation (t' Hart *et al.*, 1990) prosodic phenomena such as accents and phrase boundaries are described in terms of f_0-movements; that is, as falls and rises rather than as f_0-targets. However, they are still described in terms of f_0. The IPO description was originally intended for Dutch intonation, but has later on been extended to several other languages.

There are also transcription systems that attempt to give the equivalent of a narrow phonetic transcription of the actual f_0 contour, rather than of prosodic categories. Two examples of such systems are the International Transcription System for Intonation, or INTSINT (e.g. Hirst & Di Cristo, 1998b), and the Tilt model (e.g. Taylor, 2000).

The prosodic phenomena in a fair number of languages of the world are thus described in terms of f_0. The representation of suprasegmentals in the Handbook of the International Phonetic Association (1999) and in the IPA-based transcription system for Swedish (Bruce & Touati, 1990; see also Paper I, Strangert & Heldner, 1995) should rather be considered exceptions in this respect, as these systems do not use symbols related to f_0 for prosodic transcriptions.

It has furthermore been suggested that f_0 should be of special importance in descriptions of Swedish and a few other 'pitch-accent languages' such as Norwegian, Lithuanian, Slovenian, Serbo-Croatian, and Japanese. Beckman (1986) makes a distinction between stress-accent and non-stress-accent languages (i.e. 'pitch-accent languages') and considers Swedish to be of the latter kind, whereas English and Dutch are treated as examples of stress-accent languages. In her account, stress-accent languages differ from non-stress-accent languages in the degree to which they use phonetic attributes other than pitch patterns to signal accents. Pitch accents in stress-accent languages are not only characterized by a pitch movement but also by other phonetic correlates such as greater duration and loudness, while accents in the non-stress-accent languages are characterized by a pitch movement only. Thus, according to Beckman (1986), accents in Swedish should be no more than just f_0-movements. However, Nagano-Madsen & Bruce (1998) have argued that Swedish is both a stress-accent and a pitch-accent language – that is, non-stress-accent in the terminology used by Beckman (1986) – and that the similarities between Swedish and stress-accent languages such as English and Dutch are far greater than those between Swedish and a typical non-stress-accent language such as Japanese.

Judging only from the selected examples above, one may easily get the impression that prosody is synonymous to f_0 and that, basically, f_0 is all there is to it. Clearly, there is a widespread belief that f_0 is very important. However, it would not be a fair description to say that all prosodic models only deal with f_0. Quite a few descriptions centered on linguistic issues incorporate other features than f_0, just as well as many prosodic models for technological applications do. As a matter of fact, f_0 together with duration are the most commonly manipulated features in acoustic models of prosody for speech synthesis (e.g. Klatt, 1987). Furthermore, systems for automatic classification of prosodic categories – including accented words – typically use f_0 features, although often in combination with duration and intensity (Campbell, 1992; Wightman & Ostendorf, 1994; Streefkerk, Pols & Ten Bosch, 1998; Nöth, Batliner, Kießling, Kompe & Niemann, 2000; Shriberg, Stolcke, Hakkani-Tür & Tür, 2000). It should be emphasized at this point that it has by no means been our intention to question the primacy of f_0 as an acoustic correlate or as a perceptual cue to prosodic phenomena. We do, however, wish to put the importance of f_0 in a slightly different perspective by asking questions like: Just how important is f_0 for the signaling of focal accents in Swedish? Is the focal accent f_0 rise really a sufficient cue to perceived focus in Swedish? And further: are there any other important acoustic correlates and perceptual cues to focal accents, and, if so, what are they?

1.5. Acoustic correlates of the prominence levels in the Swedish model of intonation

As we have already seen, the Swedish model of intonation involves categorization of accentuation or prominence, as well as of phrasing and boundary signaling, and these features are described in terms of high and low f_0-targets with association to stressed syllables or boundaries. Figure 1, which is reproduced from Bruce (1977), presents a schematic representation of the f_0 contributions of several of these categories, and Table 1, reproduced from Bruce *et al.* (1997),

their respective tonal turning points. We will, however, leave the phrasing and boundary signaling (i.e. the initial and terminal junctures) aside here and concentrate on the three phonologically distinct prominence levels – 'stressed', 'accented' and 'focused'. This section presents a short overview of the literature on the acoustic correlates of these prominence levels, starting with the lowest level – 'stressed'.

Several authors have observed that the most consistent acoustic correlate for distinguishing stressed syllables from unstressed in Swedish is longer segmental durations (e.g. Fant & Kruckenberg, 1994; Bruce, 1999). Furthermore, overall intensity and spectral contrasts such as vowel quality differences have also been mentioned among the correlates of stress (Bruce, 1999), although these differences may not be as consistent as the durational differences (Fant & Kruckenberg, 1994). Crucially however, the signaling of stress is generally described as non-tonal; stressed syllables belong to the category 'unaccented' in the Swedish intonation model (cf. Table 1). Only the two higher levels of prominence have tonal correlates.

Next, let us address the question of acoustic correlates for distinguishing 'accented' from 'stressed'. According to the proponents of the Swedish intonation model, the most important acoustic correlate for distinguishing an accented from an unaccented foot is the presence of an f_0 fall, referred to as a word accent fall in Figure 1. Thus, accent as a higher prominence level than just stress is signaled mainly by f_0, although an accented foot is usually also longer than an unaccented one (Bruce, 1977; Bruce *et al.*, 1997; Bruce, 1999). There is furthermore a choice between word accents I and II at this level, although there is no difference in prominence between the two. Perceptual experiments have demonstrated that the distinction between accent I and accent II words is primarily a tonal phenomenon and that the difference is a matter of timing of the word accent fall (cf. Bruce, 1977 and references mentioned therein). This is illustrated in Figure 1 and Table 1. The schematized f_0 contours in Figure 1 show that the word accent fall occurs earlier (relative to the stressed vowel) in accent I than in accent II words. Similarly, Table 1 indicates that the association of the turning points in the word accent fall with the stressed vowel (as indicated by the asterisks) differs between accent I and II words. While the low tone is associated with the stressed vowel (HL*) in accent I, it is the high tone that is associated with the stressed vowel in accent II and the low tone occurs after the stressed vowel (H*L).

However, several authors question the view conceiving the word accent fall as the most important acoustic correlate of the 'accented' prominence level, arguing that a word accent fall is not always present in accent I words. This is the case for example in a monosyllabic accent I word in the beginning of a sentence. These authors claim that only accent II has a positively specified f_0 correlate and, accordingly, that there are good reasons for treating accent I as an unmarked member of the word accent contrasts (cf. Engstrand, 1989; Elert, 1994; Fant & Kruckenberg, 1994; Engstrand, 1995). Several reports of decreasing size of word accent falls with increasing prominence in accent I words, on the one hand, and increasing word accent falls with increasing prominence in accent II words, on the other, give further support to this view (Fant & Kruckenberg, 1994; Engstrand, 1995).

Let us finally comment on the acoustic correlates for distinguishing 'focused' from 'accented' words. The primary acoustic correlate of the 'focused' prominence level is again held to be a tonal one – a focal accent or a sentence accent rise following the word accent fall (cf. Figure 1 and Bruce, 1977). In terms of tonal turning points, this rise is indicated by an extra high tone following the low tone of the word accent fall (cf. Table 1). However, this f_0 movement is usually accompanied by an increased duration of the word in focus (Bruce, 1977; Bannert, 1979; Bruce, 1981; Bruce, 1983; Fant, Kruckenberg & Nord, 1991; Fant & Kruckenberg, 1994;

Heldner & Strangert, 2001), by moderate increases in overall intensity and in spectral emphasis (Fant & Kruckenberg, 1994; Fant, Hertegård & Kruckenberg, 1996; Fant, Kruckenberg & Liljencrants, 2000; see also Paper VI, Heldner, forthcoming), and by an increase of the contrast in vowel and consonant spectrum shape and intensity (Fant & Kruckenberg, 1994; Fant *et al.*, 2000).

So, according to the Swedish intonation model, f_0 movements should undoubtedly be considered as an important type of acoustic correlates of prominence – and especially of focal accent – but they are by no means the only existing correlates. Furthermore, f_0 movements are more or less explicitly assumed to be the acoustic features upon which perception of focal accentuation relies. However, experimental studies actually investigating the perceptual aspects of the signaling of focal accent are rare. Consequently, our knowledge of these aspects remains fairly limited.

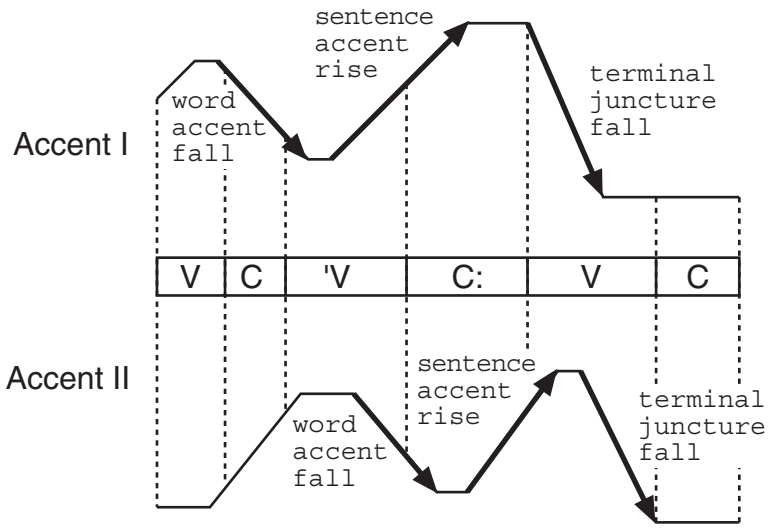

Figure 1. The f0-contributions of word accent, sentence accent and terminal juncture. Schematized f0-contours of one accent I- and one accent II-word. The arrows, drawn in thick lines, indicate word accent fall, sentence accent rise and terminal juncture fall. Note. Picture and caption from Bruce (1977), © Gösta Bruce 1977. Reprinted with permission.

Table 1. The prosodic categories in the Swedish intonation model, and their tonal turning points.

Prosodic category	Tonal turning points
Unaccented	-
Accent I	HL*
Accent II	H*L
Focal accent I	(H)L*H
Focal accent II	H*LH
Focal accent II compound	H*L…L*H
Initial juncture	%L, %H
Terminal juncture	L%, LH%

It deserves to be repeated that the Swedish intonation model has been the point of departure of all the investigations conducted within the research project reported in this thesis. Typically, the acoustic correlates of 'focused' words have been compared to those of merely 'accented' ones. Thus, Papers II and III in this thesis deal with the importance of the focal accent rise. The following papers discuss a few other acoustic correlates of focal accents mentioned in the literature. As for Papers IV and V, they deal with duration, while Papers VI and VII deal with overall intensity and spectral emphasis. Below follows a summary of each of the individual papers.

2. Summaries of the individual papers

2.1. Summary of Paper I: Labeling of boundaries and prominences by phonetically experienced and non-experienced transcribers

Paper I has been incorporated in this thesis because it provides an important background for the rest of the papers and, indeed, for the whole project of describing the acoustic correlates of focal accents in Swedish and their perceptual relevance. Although Paper I was primarily meant to be an evaluation of an IPA based system for the transcription of Swedish prosody, it also revealed that prosodic transcription is by no means a trivial task, and especially not the transcription of prominence. This observation, in turn, initiated the exploration of the acoustic basis for the variability associated with the labeling of prominence, and especially the acoustic basis for the perception of focally accented words.

The IPA based system for transcribing Swedish prosody which was evaluated contains symbols for the prosodic phenomena of boundaries and prominences. The evaluation dealt primarily with various aspects of the system's reliability. Several different indices of interrater reliability and interrater agreement were calculated. In addition, in order to estimate the degree of professionalism required for prosodic labeling, comparisons were made of the labelings by expert phoneticians specializing in prosody and transcribers without any previous experience in

prosodic labeling. Both groups were requested to label a speech material containing samples of read-aloud as well as spontaneous speech.

As might have been expected the experts performed in a more consistent manner than did the transcribers lacking in previous experience. Furthermore, the results showed that the labelings of boundaries were more reliable than those of prominences, and that the read-aloud speech gave more reliable labelings than the spontaneous speech. Moreover, although high interrater reliability figures were achieved, there was still substantial variability in the labelings reflected in the fairly low 'exact agreement between all transcribers' figures, and especially those for the labeling of prominence. Thus, consistent labeling of prosody proved to be a far from straightforward task, even for experts.

2.2. Summary of Paper II: The labeling of prominence in Swedish by phonetically experienced transcribers

The findings from Paper I in turn initiated an exploration of the acoustic basis for the variability associated with the labeling of prominence in the transcription task. As the focal accent f_0 rise is generally believed to be the most important perceptual cue to focus, this is where the exploration began. Paper II contains a summary of the evaluation of the prominence labeling by expert transcribers from Paper I. However, the main issue – at least from the point of view of this thesis – is the exploration of the acoustic basis for the variability associated with the assignment of prosodic categories in the transcription task, and in particular, the extent to which perceived prominence (as reflected by the labeling of the expert transcribers) is dependent on the size of f_0 movements.

This dependence was studied in 115 words taken from the larger speech material used in the labeling study. These words were selected on the ground that they had been judged to be focused by at least two out of the nine expert transcribers. A prominence score based on the labeling of all transcribers was used as an estimate of perceived prominence. Measurements were made of the sizes of the word accent fall and the focal accent rise. These measurements were subsequently used as predictors in regression analyses where the dependent variable was the prominence score.

The analyses revealed a significant positive correlation between perceived prominence (in terms of prominence score) and size of the focal accent rise, while the size of the word accent fall was not significantly correlated with perceived prominence. This means that the perceived prominence on the whole increased with the size of the focal accent rise. However, the explained variance is perhaps of greater interest when it comes to assessing the strength of the relationship. In this case, less than 40% of the variation in perceived prominence was explained by the variation in size of the focal accent rise. This was interpreted as a clear indication that there must be other important perceptual cues to prominence than f_0 movements.

2.3. Summary of Paper III: To what extent is perceived focus determined by f_0 cues

In Paper II it was shown that the size of the focal accent rise could explain only a small proportion of the variation in perceived prominence in natural speech. It was therefore argued that there must exist other important acoustic correlates of focus than the focal accent rise.

Like Paper II, Paper III deals with the importance of the focal accent rise for the perception of prominence, but it does so from a slightly different angle. Here, the importance of the size of the f_0 rise was put in relation to that of other (and sometimes conflicting) local and global acoustic correlates. Furthermore – and in contrast to Paper II, where the size of the f_0 rise was measured in naturally produced speech – this paper involved experimental manipulations of the size of the focal accent rise.

Two experiments were undertaken. The manipulations were performed on phrase-medial words in natural read-aloud Swedish sentences and involved a gradual reduction of the f_0 rise in focally accented words and a gradual addition of a f_0 rise in non-focused words. Thus, the words with a gradual reduction of the f_0 rise contained all other acoustic features inherent in focally accented words, while the words with a gradual addition of an f_0 rise lacked the other acoustic features typical of focally accented words. In other words, the manipulations introduced cases of more or less conflicting information.

In addition, the amount of global information was varied between the first and second experiments. The whole sentence was presented in the first experiment, while, in the second experiment, all that followed the manipulated word was removed. These manipulated sentences were presented to listeners who were asked to determine whether the manipulated target words were focused or not.

The results indicated that the f_0 rise in itself was neither necessary – in the sense that it had to be present for the listener to perceive a word as focused – nor in itself sufficient – in the sense that if it was present, it was enough for the listener to perceive a word as focused also when there was conflicting acoustic information present. This means, firstly, that a word can be perceived as focused even in the absence of an f_0 rise and, secondly, that a word can be perceived to be non-focused even if an f_0 rise is present.

Even though the f_0 might still be the most important perceptual cue to focus, this is another clear indication that also other acoustic correlates than the focal accent rise play a role in the signaling of focus. In the following papers, a few other likely candidates were scrutinized. Thus, Paper IV deals with the effects on durational focal accentuation in Swedish, and Paper V with the perceptual relevance of one specific aspect of the durational adjustments occurring in focused words. Similarly, Paper VI deals with the effects of focus on the overall intensity and the spectral emphasis of the focused word, and Paper VII with the perceptual relevance of increasing spectral emphasis.

2.4. Summary of Paper IV: Temporal effects of focus in Swedish

Paper IV is the first of the papers included in this volume to be actually dealing with acoustic correlates of focal accents in Swedish other than f_0-movements. Because a number of previous studies had shown that focally accented words in Swedish, as well as accented words in many other languages, in addition to the f_0 movements also tend to be produced with longer duration than non-focused words, this is where the exploration of the other acoustic correlates was initiated.

Four experiments concerning the amount and domain of focal accent lengthening in Swedish non-compound words are reported in Paper IV. The duration of words, syllables, and segments in focally accented and non-focused words were measured (i) to estimate the amount of lengthening (that is, how much focally accented words are lengthened as compared to words out of focus),

(ii) to investigate the domain of lengthening (that is, where the lengthening starts and ends) and, (iii) how the lengthening is distributed within this domain. In addition, it was studied if and how the amount and distribution of lengthening was affected by various factors such as the speaker, the position in the phrase, the position of the stressed syllable within the word, the length of the word, and the Swedish word accent distinction. The speech material consisted of short sentences with systematic variation of these features read aloud by a number of speakers.

As expected, these measurements showed that focally accented words were longer than non-focal words in general. Lengthening was thus shown to be a reliable acoustic correlate of focal accent. However, the amount of lengthening varied greatly, primarily due to speaker differences, but also due to position in the phrase. Furthermore, most of the lengthening occurred within the stressed syllable. An analysis of the internal structure of stressed syllables showed that the phonologically long segments – whether they were vowels or consonants – were lengthened most, while the phonologically short vowels were hardly affected at all. Through this non-linear lengthening, the contrast between long and short vowels in stressed syllables was sharpened in focus. We will return to this issue in Paper V.

The fact that most of the lengthening occurred in the stressed syllable shows that the domain of focal accent lengthening includes at least the stressed syllable. Moreover, the unstressed syllable following immediately after the stressed one was also lengthened in focus, while initial unstressed syllables, as well as unstressed syllables to the right of the first unstressed one, were not lengthened. Thus, it is tentatively assumed that the domain of focal accent lengthening in Swedish non-compound words is restricted to the stressed syllable and the immediately following unstressed one.

2.5. Summary of Paper V: On the non-linear lengthening of focally accented Swedish words

One of the findings reported in Paper IV was that the lengthening of focally accented words was non-linearly distributed within the words. It was shown, first, that stressed syllables were lengthened relatively more than unstressed ones and, second, that phonologically long segments within the stressed syllable – whether they were vowels or consonants – were lengthened relatively more than phonologically short segments. Phonologically short vowels were hardly lengthened at all. The two listening experiments described in Paper V concentrate on the non-linear lengthening found within stressed syllables where a phonologically short vowel is followed by a long consonant, i.e. in CVC:-syllables.

The first perceptual experiment was designed to investigate whether a non-linear lengthening pattern in CVC:-syllables is important for the perceived naturalness of focally accented words. Accordingly, the listeners in this experiment were asked to compare pairs of synthesized sentences, where the amount of lengthening of the focally accented word was the same, while the distribution of this lengthening differed.

The second experiment was an attempt to compare the perceptual importance of the two features distinguishing the non-linear from the linear lengthening patterns. These features were: (i) strengthening of temporal contrast, and (ii) short vowel remaining short. Its purpose was consequently to find out whether one of the two features was more important than the other. Again, listeners were asked to compare the naturalness of each member of pairs of synthesized sentences where the segmental durations in the focally accented words differed.

The results of the first perceptual experiment showed that a majority of the listeners was sensitive to the durational differences they were exposed to during the experiment. These listeners preferred a non-linear lengthening of focally accented CVC:-syllables to a linear expansion. Moreover, the second perceptual experiment showed that the most important feature of this non-linear pattern is for the vowel to be maintained short.

2.6. Summary of Paper VI: On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish

Paper VI is the second of the papers in this thesis dealing with the acoustic correlates of focal accents beyond f_0 movements. Whereas Paper IV dealt with temporal aspects of the signaling of focal accents, Paper VI investigates the issue whether two correlates related to the perceived loudness – i.e. overall intensity and spectral emphasis – are to be considered as reliable acoustic correlates of focal accents in Swedish.

Two experiments are reported. The first concerned the reliability of the correlates in a paradigmatic – or between-sentence – sense. More specifically, it had been designed to find an answer to the question whether all focally accented words are characterized by an increase in overall intensity and spectral emphasis as compared to non-focused words, no matter what the word is, or the speaker who utters it, or which position in the sentence the word occurs in.

The second experiment dealt instead with reliability in a within-sentence perspective. The investigation concerned to what extent the highest value in the sentence of the correlates is being found in the focally accented word. The second experiment was run as a detection experiment.

A prerequisite for both experiments, however, is suitable methods for measuring the acoustic correlates in question. In the case of overall intensity, this is fairly unproblematic. For spectral emphasis, however, several different methods have been proposed in the literature. Paper VI evaluates a number of these measures. In addition, it presents a new, continuous, and fully automatic measure of spectral emphasis. Compared to previous implementations, it has the clear advantage of being insensitive to f_0 movements. Again, the speech material consisted of short sentences read aloud by a number of speakers.

The experiments showed that increases in overall intensity and spectral emphasis are in fact reliable acoustic correlates of focal accents in Swedish. Both correlates are reliable in the sense that there are statistically significant differences between focally accented words and non-focal ones for a variety of words, in any position of the sentence, and for all speakers in the analyzed materials. They are also reliable in the sense of their being useful for automatic detection of focal accents.

Moreover, spectral emphasis turned out to be the more reliable correlate, as it proved a better predictor of focal accents in general and for a majority of the speakers. It should also be considered as more reliable because the influence of the factors position in the sentence, word accent and vowel height was less pronounced. As it turns out, this study has resulted in a solid ground of production data for overall intensity and spectral emphasis that might prove important in modeling for speech synthesis. In addition, it indicates that including information about spectral emphasis rather than overall intensity ought to be advantageous in systems for automatic classification of prosodic categories.

2.7. Summary of Paper VII: Spectral emphasis and the perception of prominence

The seventh and last paper of this thesis, Paper VII, represents a first attempt to investigate whether spectral emphasis – the acoustic feature that was shown to be a reliable acoustic correlate of focal accents in Swedish in Paper VI – also has perceptual relevance. In particular, it was examined (i) whether an increase in spectral emphasis would also cause words to be perceived as more prominent, and (ii) whether the modeling of spectral emphasis in connection with focal accent would improve the quality and naturalness of speech synthesis.

To this end, a method was proposed for experimentally increasing spectral emphasis in natural and synthesized speech without degrading speech quality. Essentially, this method involves a uniform amplification of the frequency components above the fundamental by 4 dB, while the fundamental is attenuated by 2 dB. This method for increasing spectral emphasis was subsequently used in two listening experiments.

In the first experiment, spectral emphasis was increased in focally accented words in read-aloud natural speech. Listeners were asked to compare the prominence of the manipulated and the original words. The second experiment involved increased spectral emphasis in focally accented words generated by an mbrola synthesis. This time, the listeners were asked to compare the naturalness of the manipulated and the original words.

The results of these two experiments were on the whole quite negative. Increased spectral emphasis, as implemented here, did not prove to be an unambiguous cue to prominence in the first experiment, and did not improve the naturalness in the second experiment. It would, however, be premature to reject spectral emphasis as a cue to focal accents solely based on these experiments. It might still be the case that other implementations of increased spectral emphasis would produce more salient effects on perceived prominence and on perceived naturalness in speech synthesis. But this will have to be shown in future research.

3. General summary

To summarize, this thesis has presented the results of an evaluation of a prosodic labeling system and has furthermore investigated one supposedly important piece of the acoustic basis for the labeling of prominence in this system. In addition, the results of a perceptual experiment designed to estimate the relative importance of f_0 for the perception of focal accents have been reported. These studies were followed by two fairly large production experiments concerning durational patterns in focally accented words and two loudness related acoustic correlates – overall intensity and spectral emphasis. Both production experiments were followed up by smaller scale perceptual experiments. These studies have confirmed that a focally accented word, in addition to the f_0 movements and especially the focal accent rise, is characterized by longer segmental durations, higher overall intensity and higher spectral emphasis, and that several acoustic features contribute to the perceived naturalness of focally accented words.

3.1. Labeling

Papers I and II showed that the variability in the labeling was substantial also among the expert transcribers, and especially in the labeling of prominence. Thus, also human expert transcribers

find it difficult to rate the level of prominence for a given word. This fact has to be considered by anyone set upon devising an automatic system trying to perform the same task.

3.2. f_0 movements

Regarding the supposedly most important acoustic correlate of focal accents in Swedish, the focal accent f_0 rise, Paper II showed that the variation in size of the focal accent rise explained only a small fraction of the variation in labeling prominence. This was taken as a first indication that also other acoustic correlates than the focal accent rise must be important for the signaling of focus.

Furthermore, the listening test with manipulated f_0 movements in Paper III showed that an f_0 rise is neither necessary nor in itself sufficient to signal that a word is in focus. More exactly, an f_0 rise is not enough to signal that a word is focused when there is conflicting information present. This, in turn, means that combinations of other acoustic features may outweigh f_0 as a perceptual cue to focal accents. And since the supposedly most important acoustic correlate of focus is not in itself sufficient to signal that a given word is focused, it is unlikely that any other acoustic features will be in themselves sufficient to signal focus either. So it seems that even the most important of the acoustic correlates must act together with other correlates in order to signal focus. In themselves, f_0 as well as other correlates such as duration merely contribute to the perceived naturalness within the category.

3.3. Duration

With regard to segmental durations, it has been confirmed by this thesis – and especially by Paper IV – that lengthening is a reliable acoustic correlate of focally accented words in Swedish. As a matter of fact, detailed descriptions have been presented of the lengthening patterns in non-compound, disyllabic and longer focally accented words under various conditions. It was shown in Paper IV that focal accent lengthening occurs in disyllabic as well as in longer words, irrespective of the position of the focally accented word in the sentence, and irrespective of the position of the stressed syllable in the word. The lengthening, however, is not linearly distributed over the words. Stressed syllables are lengthened more than unstressed ones, phonologically long segments within the stressed syllable are lengthened more than short segments, and short vowels within stressed syllables are not lengthened at all. Furthermore, regarding the domain within which lengthening occurs it was shown that the lengthening starts with the onset of the stressed syllable and ends after a following unstressed syllable. Unstressed syllables preceding the stressed syllable, as well as those following the first unstressed one, are not lengthened by the presence of focal accents. Consequently, it is not always the case that the whole word in focus should be lengthened.

Although the experiments provide a fairly broad coverage of Swedish words, there are still details missing. And one of those missing details is whether the domain of focal accent lengthening established for non-compound words is also valid for compound words. It thus remains to be investigated whether the domain extends beyond the stressed syllable and a following unstressed one in words that have both a primary stressed syllable and a secondary stressed one. Furthermore, no monosyllabic words were included in the investigations. Therefore,

it remains an open question whether the lengthening would extend across word boundaries given short enough words.

Regarding the perceptual relevance of durational correlates it was observed (Paper III) that words may be perceived as focused also in the absence of f_0 movements. That is, duration in combination with other non-f_0 correlates may be sufficient to signal that a word is focused, at least in the absence of conflicting information. Furthermore, and far from surprising, duration in combination with f_0 was sufficient to signal that a word is focused (Papers III, V and VII). After all, f_0 and duration are the most commonly manipulated prosodic parameters in speech synthesis.

More importantly, perhaps, it has been shown that the distribution of the lengthening within the focally accented word is important for the perceived naturalness (Paper V). So if a word is to be lengthened, it matters how this lengthening is applied. Paper V showed that perceived naturalness increased when focal accent lengthening in stressed syllables containing a phonologically short vowel (a CVC:-syllable) was modeled using a non-linear lengthening pattern as compared to a linear one. It was moreover shown that the most important feature of this non-linear pattern is for the vowel to be maintained short. A missing detail here which remains to be investigated is whether the non-linear lengthening observed between stressed and unstressed syllables in Paper IV – i.e. that stressed syllables are lengthened more and unstressed syllables less than the word as a whole – is important for perceived naturalness. It seems, however, that focal accent lengthening implemented as a linear lengthening of the whole word is neither an appropriate description of focal accent lengthening from a production perspective, nor from a perception one.

3.4. Spectral emphasis

In Paper VI this thesis deals with spectral emphasis, a measure supposedly related to perceived loudness and capable of capturing the relative high frequency energy. In particular, it has presented a new technique for measuring spectral emphasis by calculating the difference (in dB) between the overall intensity and the intensity of f_0 (or the first partial H1) in each instant. Although this measure bears resemblance to many previous ones, it offers the novelty of accurately describing the intensity of f_0 by the use of a dynamic filter whose cut-off frequency is determined by f_0 in each instant.

This measure of spectral emphasis has furthermore been shown to be a reliable acoustic correlate of focal accents in Swedish. It proved to be more reliable than another acoustic correlate related to perceived loudness, namely overall intensity, and also to be more reliable than several previously published implementations of spectral emphasis. Apart from suggesting that modeling of spectral emphasis might improve the quality and naturalness of speech synthesis, this moreover implies that it ought to be an advantage to include a spectral emphasis measure rather than an overall intensity one in systems for automatic classification of prosodic categories.

This thesis (Paper VII) has also presented a framework for experimental manipulation of spectral emphasis and one particular implementation of it, involving a uniform amplification of all frequency components above f_0. The material obtained was evaluated in listening experiments with natural and synthetic speech. Somewhat discouragingly, however, this implementation proved not to be a very effective perceptual cue for the signaling of focal accents. Increased

spectral emphasis as implemented here neither increased perceived prominence nor perceived naturalness.

These results were unexpected given the results of previous production and perception experiments and they inevitably lead to speculations about the implementation of spectral emphasis not being realistic enough. Although the technique for measuring spectral emphasis presented in Paper VI calculated the total intensity of the frequency components above the fundamental, it is far from obvious that spectral emphasis should also be manipulated using a uniform increase of all frequency components above the fundamental. Maybe the amount of amplification should instead vary with frequency; e.g. a 3 dB increase at 500 Hz, 6 dB at 1000 Hz, and 9 dB at 2000 Hz etc. Fortunately, the framework for experimental manipulation of spectral emphasis presented here is flexible enough to easily accommodate other implementations of spectral emphasis, for example with non-uniform increases in intensity across the different frequency components. However, it will be left for future research to show whether other implementations of increases in spectral emphasis will have a more salient effect on perceived prominence and on perceived naturalness.

3.5. Possible applications of the results

Although the experiments presented in this thesis were primarily undertaken to gain a better understanding of the acoustic correlates associated with focal accents in Swedish and their perceptual relevance, the results may also be put to use in various speech technology applications. For this reason, it is particularly worth noting that there is an increasing awareness among researchers that for speech synthesis to sound natural, all acoustic features must in a sense be combined together. In fact, there seems to be a growing demand for descriptions of the various acoustic features beyond the f_0 and duration features currently used. In the future, therefore, the prosody modules in speech synthesis systems will have to provide modeling not only of f_0 and duration but also of other acoustic features. Incidentally speaking, even systems for speech recognition and understanding, or systems for automatic classification of prosodic categories may benefit from more detailed descriptions of acoustic features associated with prosodic categories. Such systems using combinations of f_0, duration and intensity today might for instance benefit from the inclusion of information about spectral emphasis.

4. References

(1999) *Handbook of the International Phonetic Association: A guide to the use of the international phonetic alphabet.* Cambridge: Cambridge University Press.

Bannert, R. (1979) The effect of sentence accent on quantity. In *Proceedings of the Ninth International Congress of Phonetic Sciences*, pp. 253-259. Copenhagen: Institute of Phonetics, University of Copenhagen.

Beckman, M. E. (1986) *Stress and non-stress accent.* Dordrecht: Foris Publications.

Beckman, M. E. & Ayers Elam, G. (1997) *Guidelines for ToBI labelling, version 3*. Columbus: The Ohio State University Research Foundation.

Beckman, M. E. & Pierrehumbert, J. (1986) Intonational structure in Japanese and English. In *Phonology Yearbook 3* (J. J. Ohala, ed.), pp. 255-309. Cambridge: Cambridge University Press.

Bolinger, D. L. (1958) A theory of pitch accent in English, *Word,* **14**(2-3), 109-149.

Bruce, G. (1977) *Swedish word accents in sentence perspective.* Lund: CWK Gleerup.

Bruce, G. (1981) Tonal and temporal interplay. In *Nordic prosody II* (T. Fretheim, ed.), pp. 63-74. Trondheim: Tapir.

Bruce, G. (1983) Accentuation and timing in Swedish, *Folia Linguistica,* **17**, 221-238.

Bruce, G. (1999) Word tone in Scandinavian languages. In *Word prosodic systems in the languages of Europe* (H. van der Hulst, ed.), pp. 605-633. Berlin, New York: Mouton de Gruyter.

Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D. & Touati, P. (1997) On the analysis of prosody in interaction. In *Computing Prosody* (Y. Sagisaka, N. Campbell & N. Higuchi, eds.), pp. 43-59. New York: Springer-Verlag.

Bruce, G. & Gårding, E. (1978) A prosodic typology for Swedish dialects. In *Nordic Prosody*, Lund, pp. 219-228.

Bruce, G. & Touati, P. (1990) On the analysis of prosody in spontaneous dialogue. In *Working Papers 36*, pp. 37-55. Lund: Department of Linguistics and Phonetics, Lund University.

Campbell, N. (1992) Prosodic encoding of English speech. In *Proceedings ICSLP 92*, pp. 663-666. Alberta: Department of Linguistics, University of Alberta.

Elert, C.-C. (1994) Compounds in a phonology of Swedish, *Acta Linguistica Hafniensia,* **27**, 123-129.

Engstrand, O. (1989) Phonetic features of the acute and grave word accents: data from spontaneous speech, *PERILUS: Phonetic Experimental Research at the Institute of Linguistics University of Stockholm,* **X**, 13-37.

Engstrand, O. (1995) Phonetic interpretation of the word accent contrast in Swedish, *Phonetica,* **52**, 171-179.

Fant, G., Hertegård, S. & Kruckenberg, A. (1996) Focal accent and subglottal pressure, *TMH-QPSR*(2), 29-31.

Fant, G. & Kruckenberg, A. (1994) Notes on stress and word accent in Swedish, *STL-QPSR*(2-3), 125-144.

Fant, G., Kruckenberg, A. & Liljencrants, J. (2000) Acoustic-phonetic analysis of prominence in Swedish. In *Intonation: Analysis, modelling and technology* (A. Botinis, ed.), pp. 55-86. Dordrecht: Kluwer Academic Publishers.

Fant, G., Kruckenberg, A. & Nord, L. (1991) Durational correlates of stress in Swedish, French and English, *Journal of Phonetics,* **19**, 351-365.

Flanagan, J. L. (1955) Difference limen for the intensity of a vowel sound, *Journal of the Acoustical Society of America,* **27**, 1223-1225.

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

Grønnum, N. (1995) Superposition and subordination in intonation: A non-linear approach. In *Proceedings ICPhS 95*, pp. 124-131. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

Gussenhoven, C. (1984) *On the grammar and semantics of sentence accents.* Dordrecht: Foris.

Gårding, E. & Bruce, G. (1981) A presentation of the Lund model for Swedish intonation. In *Nordic Prosody II*, Trondheim, pp. 33-39.

Heldner, M. (forthcoming) On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish, *Submitted to Journal of Phonetics*.

Heldner, M. & Strangert, E. (1997) To what extent is perceived focus determined by F0-cues? In *Eurospeech '97 Proceedings*, pp. 875-877. Rhodes, Greece: ESCA.

Heldner, M. & Strangert, E. (2001) Temporal effects of focus in Swedish, *Journal of Phonetics,* **29**(3), 329-361.

Hirst, D. & Di Cristo, A. eds. (1998a) *Intonation systems: A survey of twenty languages.* Cambrigde: Cambridge University Press.

Hirst, D. & Di Cristo, A. (1998b) A survey of intonation systems. In *Intonation systems: A survey of twenty languages* (D. Hirst & A. Di Cristo, eds.), pp. 1-44. Cambrigde: Cambridge University Press.

Kewley-Port, D. & Zheng, Y. (1999) Vowel formant discrimination: Towards more ordinary listening conditions, *Journal of the Acoustical Society of America,* **106**(5), 2945-2957.

Klatt, D. H. (1973) Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception, *Journal of the Acoustical Society of America,* **53**(1), 8-16.

Klatt, D. H. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, *Journal of the Acoustical Society of America,* **59**(5), 1208-1221.

Klatt, D. H. (1987) Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America,* **82**(3), 737-793.

Klatt, D. H. & Cooper, W. E. (1975) Perception of segment duration in sentence contexts. In *Structure and Process in Speech Perception* (A. Cohen & S. G. Nooteboom, eds.), pp. 69-86. Berlin, Heidelberg, New York: Springer-Verlag.

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

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

Lyberg, B. (1979) Final lengthening — partly a consequence of restrictions on the speed of fundamental frequency change, *Journal of Phonetics,* **7**, 187-196.

Lyberg, B. (1981a) Some consequences of a model for segment duration based on F0-dependence, *Journal of Phonetics,* **9**, 97-103.

Lyberg, B. (1981b) Some observations on the vowel duration and the fundamental frequency contour in Swedish utterances, *Journal of Phonetics,* **9**, 261-272.

Mermelstein, P. (1978) Difference limens for formant frequencies of steady-state and conconant-bound vowels, *Journal of the Acoustical Society of America,* **63**(2), 572-580.

Nagano-Madsen, Y. & Bruce, G. (1998) Comparing pitch accent features in Swedish and Japanese. In *Nordic Prosody: Proceedings of the VIIth Conference, Joensuu 1996* (S. Werner, ed.), pp. 215-224. Frankfurt am Main: Peter Lang.

Nooteboom, S. G. & Kruyt, J. G. (1987) Accents, focus distribution, and the perceived distribution of given and new information: An experiment, *Journal of the Acoustical Society of America,* **82**(5), 1512-1524.

Nöth, E., Batliner, A., Kießling, A., Kompe, R. & Niemann, H. (2000) Verbmobil: The use of prosody in the linguistic components of a speech understanding system, *IEEE Transactions on Speech and Audio Processing,* **8**(5), 519-532.

Shriberg, E., Stolcke, A., Hakkani-Tür, D. & Tür, G. (2000) Prosody-based automatic segmentation of speech into sentences and topics, *Speech Communication,* **32**, 127-154.

Silverman, K. E. A., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B. & Hirschberg, J. (1992) TOBI: A standard for labeling English prosody. In *Proceedings ICSLP 92*, pp. 867-870. Alberta: Department of Linguistics, University of Alberta.

Strangert, E. & Heldner, M. (1995) Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In *PHONUM 3* (R. Bannert & K. Sullivan, eds.), pp. 85-109. Umeå: Department of Phonetics, Umeå University.

Streefkerk, B. M., Pols, L. C. W. & Ten Bosch, L. F. M. (1998) Automatic detection of prominence (as defined by listeners' judgements) in read aloud Dutch sentences. In *ICSLP'98 Proceedings* (R. H. Mannell & J. Robert-Ribes, eds.), pp. 683-686. Sydney: ASSTA.

t' Hart, J. (1981) Differential sensitivity to pitch distance, particularly in speech, *Journal of the Acoustical Society of America,* **69**(3), 811-821.

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

Taylor, P. A. (2000) Analysis and synthesis of intonation using the Tilt model, *Journal of the Acoustical Society of America,* **107**(3), 1697-1714.

van Katwijk, A. (1974) *Accentuation in Dutch.* Amsterdam/Assen: Van Gorcum.

Wightman, C. W. & Ostendorf, M. (1994) Automatic labeling of prosodic patterns, *IEEE Transactions on Speech and Audio Processing,* **2**(4), 469-481.

Öhman, S. E. G., Zetterlund, S., Nordstrand, L. & Engstrand, O. (1979) Predicting segment durations in terms of a gesture theory of speech production. In *Proceedings of the Ninth International Congress of Phonetic Sciences*, pp. 305-311. Copenhagen: Institute of Phonetics, University of Copenhagen.

Paper I In *Focal accent – f_0 movements and beyond*
pp. 29–54

Labeling of boundaries and prominences by phonetically experienced and non-experienced transcribers[1]

Eva Strangert and Mattias Heldner

The purpose of this study is to analyze transcriptions of read and spontaneous spoken material using an agreed-upon IPA-based transcription system for Swedish. The transcriptions included labels for boundaries as well as prominences. Though our main concern is labeling made by expert phoneticians, we also included a group of transcribers having no previous experience of phonetics in the study. In doing so, we hoped to be able to estimate the degree of professionalism demanded for transcribing prosody. Our main objective is to evaluate the system by estimating the extent of reliability between transcribers. We also consider other criteria that a transcription system should meet, e.g. coverage and learnability, within the framework of current approaches to speech and language processing (see Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg, 1992a). In conclusion, high inter-transcriber reliability was achieved. We report this and other evaluations demonstrating the capacity of the system.

1. Introduction

1.1. The research area

A great number of systems have been used over the years for transcribing prosody. Most of them appear to have been developed for dealing with prominence distinctions; see Beckman (1986) for an overview. Others have aimed at a more extensive coverage of prosodic dimensions, e.g. Bruce & Touati (1990) and Bagshaw & Williams (1992). Bruce & Touati (1990) describe an IPA-based system for transcribing prosody in Swedish and French spontaneous dialogue with symbols for (a) stress and accentual prominence, (b) prosodic grouping and phrasing, (c) voice and pitch range, (d) boundary tones and (e) pausing.

The Bruce & Touati (1990) system aimed at a categorization of the prosodic structure to be used for further acoustic-phonetic analysis. Thus, this system is based on an auditory analysis of the speech to be transcribed. This characteristic is shared by most transcription systems developed in the past. In more recent systems, however, the transcriber, in addition to the auditory impression, may use graphically displayed information of the speech waveform. For example, the procedure described for labeling German (accents and

[1] This material has been published as Strangert, E. & Heldner, M. (1995) Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In *PHONUM 3*, pp. 85-109. Umeå: Department of Phonetics, Umeå University.

boundaries) in the VERBMOBIL framework allows the transcriber to use information from the speech waveform as presented on a computer screen (Reyelt, 1993). Even more sophisticated acoustic information is supplied to transcribers using the ToBI system. ToBI (short for Tones and Break Indices), developed for transcribing different varieties of English, relies heavily on graphic displays of acoustic information supplied by the ESPS/Waves+ speech analysis software. ToBI transcriptions are made via interactive listening and graphic displays of the speech waveform, total signal energy, and extracted fundamental frequency (Silverman *et al.*, 1992a; Silverman, Blaauw, Spitz & Pitrelli, 1992b).

However, both the ToBI and the VERBMOBIL systems are basically phonologically or perceptually determined; e.g. the tonal component of ToBI is founded on the phonological theory of Pierrehumbert (1980) and the VERBMOBIL procedure was developed on the basis of a number of tests following which the labeling categories were modified in accordance with the auditory capabilities of the transcribers (Reyelt, 1993). The auditory categorization is similarly the basis of systems for automatic and semi-automatic prosodic labeling. Thus, the IPO automatic labeling of pitch movements for Dutch ('ten Bosch, 1993) uses an algorithm posed in the theory relating acoustic realization and perceptual labels developed by t' Hart, Collier & Cohen (1990). And the perceptually based labeling described by Bagshaw & Williams (1992) is seen as the first step towards automatic labeling of speech.

While in the past, prosodic transcriptions were developed for the sole purpose of classifying prosodic events for phonetic/linguistic reasons, quantitative computational modeling of speech is thus an additional motivation for labeling prosody today. Speech technology research requires corpora of prosodically transcribed speech, which should not only be large but also meet needs such as reliability and learnability. Thus, before they can be used extensively, they have to be evaluated. (The evaluation of transcription systems, however, is not a new phenomenon; e.g. Lieberman (1965) reported on the considerable variability between phonetically experienced transcribers in prominence labeling tasks.) Such evaluations are presently undertaken for a number of systems. Reyelt (1993) reported the results of an evaluation of the labeling of German phrase and secondary accents and phrase boundaries. Further, ToBI has been evaluated in a number of studies (Silverman *et al.*, 1992a; Silverman *et al.*, 1992b; Pitrelli, Beckman & Hirschberg, 1994). This system is the most widespread today, as far as we know, and it has been evaluated and elaborated on by a large group of researchers with expertise in prosody. The conclusion by Pitrelli *et al.* (1994) is that "the ToBI standard and its training materials have been refined to the point that they can be used fruitfully for large-scale annotation of prosodic phenomena in speech databases." Thus, ToBI as it has been developed furnishes us with important background information for the study we are undertaking.

ToBI has been developed to be used for a number of variants of English. The tonal component, based on the intonational phonology framework of Pierrehumbert (1980), include a pitch accent, a phrase accent and a boundary tone part. The break component with five categories is inspired by recent work by Price, Ostendorf, Shattuck-Hufnagel & Fong (1991). In order to use the system, guidelines have been worked out and training materials developed (Beckman & Ayers, 1994; Beckman & Hirschberg, 1994).

According to Silverman *et al.* (1992a) ToBI meets the following criteria: "(1) reliability: agreement between different transcribers must be at least 80%; (2) coverage: sufficiently comprehensive to capture the most important prosodic phenomena in spontaneous speech; (3) learnability in a relatively short time, in order to be used in multi-site data collections, and (4) capability of being related to current approaches to speech recognition, to parser outputs, and to formal representations of semantics and pragmatics."

1.2. Purpose of the study

Summarized, the purpose of the study is to collect data on labeling based on a recently agreed-upon transcription system for Swedish (see Bruce, 1994) and to evaluate the system in a number of respects. The labelings have been made of boundaries and prominences in read as well as spontaneous speech materials. Though our main concern is expert labeling, we also included a group of transcribers with no previous experience of phonetics in the study. In doing so, we hoped to be able to estimate the degree of professionalism demanded for carrying out the prosodic-transcription tasks.

Our main objective in this first large-scale evaluation is to test the extent of reliability of the system as measured by the agreement between transcribers. We will also consider and discuss the other criteria for a usable system enumerated in the study by Silverman *et al.*, (1992a), that is coverage, learnability and capability of being related to current approaches to speech and language processing.

1.3. Structure of the report

Section 2 gives a detailed description of the transcription system evaluated. Section 3 reports on and discusses the labeling made by the experts and contains, in addition to procedural details, labeling raw data, reliability and agreement computations, data showing the distribution of label categories and a summary of comments on the transcription procedure made by the transcribers. The next section, 4, presents the non- expert labeling following a similar scheme as for the expert labeling. The non-expert data are further compared to the expert labeling data. In Section 5, finally, we summarize and discuss the results and make suggestions for improvements.

2. The transcription system

The transcription system used in the study is the result of discussions among Swedish phoneticians specializing in prosody. The system is proposed as a common basic module for transcribing Swedish prosody within the framework of a national prosody database. The notation, using IPA-symbols, is restricted to the symbolization of prominence and boundary phenomena. The description below is a slightly modified version of a description previously given by Bruce (1994):

Boundary categories:

cv ‖ cv	extra strongly marked boundary	(= corresponding to e.g. speech paragraph)
cv || cv	strongly marked boundary	(= corresponding to e.g. prosodic utterance)
cv | cv	weakly marked boundary	(= corresponding to e.g. prosodic phrase)
cv cv	no boundary	

Prominence levels:

ˈˈcv	focused, accent I	(= extra strong prominence)
ˈˈcv̀	focused, accent II	(as above)
ˈcv	primary stressed, or accented, accent I	(= strong prominence)
ˈcv̀	accented, accent II	(as above)
ˌcv	secondary stressed	(= weak prominence)
cv	unstressed	(no marking)

'cv' refers to any syllable, with 'c' and 'v' representing the consonant and the vowel, respectively.

By being restricted to the most basic prosodic distinctions, the transcription system is assumed to be useful over a wide area of research on Swedish prosody. Additional modules required for specific purposes could easily be added. For example, Bruce (1994) studying the prosodic structuring of dialogue has added a specific module for tonal analysis with some similarities to the ToBI tonal transcription. Thus, the IPA-based transcription gives an auditory-phonologic categorization which can then serve as a basis for acoustic-phonetic analysis and be used in the search for regularities in prosodic-acoustic patterning (Bruce & Touati, 1990).

The base prosody system was used by the experts participating in the study. The analysis and evaluation, however, did not include all the categories, see 3.1.2. The non-experts used a modified version of the system; for details, see 4.1.2.

3. Expert labeling

3.1. Method

3.1.1. Subjects, material and labeling procedure

Nine subjects from different universities in Sweden participated in the expert-labeling part of the study. They are all experienced phoneticians or speech researchers specializing in prosody. All speak Swedish as their mother tongue. The expert transcribers are henceforth referred to as E1–E9. Of these E9 is the one with the most experience in prosodic transcription.

Two kinds of recorded speech material were transcribed. One was an excerpt, 233 words long from an authentic news cable read aloud. The other was a 252-word-long excerpt of spontaneous speech, a retelling of the story in read-aloud speech. Both recordings were made in a soundproof room and rendered by the same male Swedish speaker.

All transcription tests were run individually. The subjects were given a tape-recording with the read and the spontaneous speech together with the written versions of both, as well as explanations and instructions for labeling according to the base prosody system (see 2). The written versions contained the orthographic symbols corresponding to the spoken material but without punctuation. The subjects were asked to listen to the recordings and to write the prosodic symbols into the written text. Thus, the labelings are based exclusively on auditory impression; no graphically displayed acoustic information was supplied. As the base system aims at transcribing boundaries in a strict sense, pauses due to hesitation, slips of the tongue etc. should be left unmarked. Attention was directed to this distinction between

boundaries and different kinds of hesitation phenomena and the transcribers were instructed not to pay attention to hesitations etc. The texts corresponding to the read and the spontaneous material appear in the Appendix with the labelings made by one of the experts, E9.

3.1.2. Analysis and evaluation procedures

Although the labeling task involved all the categories of the base prosody (see 2), some were excluded from the following analyses. Thus, the distinction between words with accent I and II, respectively, was given no attention. Furthermore, secondary stress was noted only in those cases where it was the only prominence marking in a word. (In Swedish, secondary stress also co-occurs with primary stress in compounds, which, accordingly, have two prominences. Thus, compounds, as well as all other types of words in this study, were analyzed as having just one prominence.) For the analysis and evaluation the labelings were coded accordingly:

Boundary categories:

cv ‖‖ cv	extra strongly marked boundary	= 3
cv ‖ cv	strongly marked boundary	= 2
cv | cv	weakly marked boundary	= 1
cv cv	no boundary	= 0

Prominence levels:

˝cv	focused, accent I	= 3
˝cv̀	focused, accent II	= 3
ˈcv	primary stressed, accented, accent I	= 2
ˈcv̀	accented, accent II	= 2
ˌcv	secondary stressed	= 1
cv	unstressed	= 0

To evaluate the extent to which the nine experts vary in their labeling of the two kinds of material, reliability and agreement indices were computed. These indices will be complemented with more detailed analyses of the behavior of the subjects as well as a summary of the comments on the transcription procedure by the transcribers. As a basis for this evaluation excerpts containing raw data on boundary and prominence markings and some calculations made will be presented.

3.2. Raw data

Tables 1a–d show the labeling of boundaries and prominences in samples of read and spontaneous speech, respectively, by the nine experts (E1–E9). The tables contain the words in the samples ordered vertically in the first column. For each word the individual labeling by each expert is given in the following nine columns. The columns furthest to the right contain means, standard deviations and ranges based on all expert subjects for each specific word.

3.3. Evaluations

A first rough estimation of inter-transcriber variability may be achieved by checking the data presented in Tables 1a–d. The calculated standard deviations give some indications of the reliability of the labelings.

3.3.1. The concept of reliability

"Reliability concerns the extent to which measurements are *repeatable* – when different persons make the measurements, on different occasions, with supposedly alternative instruments for measuring the same thing and when there are small variations in circumstances for making measurements that are not intended to influence results. In other words, measurements are intended to be *stable* over a variety of conditions in which essentially the same results should be obtained." (Nunnaly, 1978, p. 191). This is a definition on the basis of which different aspects of reliability may be dealt with.

One aspect concerns the extent to which measuring instruments (e.g. transcribers) covary, i.e. give relative values which are correlated. Another aspect concerns the extent to which instruments return identical values. That is, having a set of instruments that covary, does not necessarily imply that we have instruments that perfectly 'agree' with regard to the absolute values they indicate. Following Rietveld & van Hout (1993) on which the preceding is based, we will refer to the first aspect as *reliability* and to the second as *agreement*, although both are in a general sense 'reliability'.

Table 1a. Labeling of boundaries in read speech by nine experts (E1–E9) and mean, standard deviation and range of the labelings.

Word	E1	E2	E3	E4	E5	E6	E7	E8	E9	Mean	SD	Range
enligt	0	0	0	0	0	0	0	0	0	0	0	0
libyska	0	0	0	0	0	0	0	0	0	0	0	0
uppgifter	2	1	1	1	1	0	1	1	1	1	0.5	2
föll	0	0	0	0	0	1	0	0	0	0.11	0.33	1
åtta	0	1	0	0	0	0	0	0	0	0.11	0.33	1
450-kilosbomber	2	1	1	1	1	1	1	1	1	1.11	0.33	1
över	0	0	0	0	0	0	0	0	0	0	0	0
Tripoli	0	0	0	0	0	0	0	0	0	0	0	0
och	0	0	0	0	0	0	0	0	0	0	0	0
Bengazi	2	1	1	1	1	1	1	2	1	1.22	0.44	1
när	0	0	0	0	0	0	0	0	0	0	0	0
de	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	0	0	0	0	0	0	0	0	0	0	0	0
bombplanen	0	0	0	0	0	0	0	0	0	0	0	0
slog	0	0	0	0	0	0	0	0	0	0	0	0
till	0	1	0	1	0	0	1	0	0	0.33	0.5	1
natten	0	0	0	0	0	0	0	0	0	0	0	0
till	0	0	0	0	0	0	0	0	0	0	0	0
tisdagen	3	2	2	2	3	2	2	3	2	2.33	0.5	1
en	0	0	0	0	0	0	0	0	0	0	0	0
av	0	0	0	0	0	0	0	0	0	0	0	0
bomberna	0	1	0	1	0	0	1	0	0	0.33	0.5	1

Table 1b. Labeling of boundaries in spontaneous speech by nine experts (E1–E9) and mean, standard deviation and range of the labelings.

Word	E1	E2	E3	E4	E5	E6	E7	E8	E9	Mean	SD	Range
det	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	0	0	0	0	0	0	0	0	0	0	0	0
luftangreppet	0	1	1	1	0	1	1	1	1	0.78	0.44	1
sattes	0	0	0	0	0	0	0	0	0	0	0	0
in	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0	0
tisdags	3	2	2	2	3	2	1	2	2	2.11	0.6	2
och	0	0	0	0	0	0	0	0	0	0	0	0
det	0	0	0	0	0	0	0	0	0	0	0	0
genomfördes	0	1	0	0	0	1	0	1	0	0.33	0.5	1
främst	0	0	0	1	0	0	0	0	0	0.11	0.33	1
utav	0	0	0	0	0	1	0	0	0	0.11	0.33	1
bombare	0	1	0	1	1	0	0	1	0	0.44	0.53	1
som	0	0	0	0	0	0	0	0	0	0	0	0
startade	0	0	0	0	0	0	0	0	0	0	0	0
från	0	0	0	0	0	0	0	0	0	0	0	0
England	3	2	2	2	3	2	2	3	2	2.33	0.5	1
under	0	0	0	0	0	1	0	0	0	0.11	0.33	1
själva	0	0	0	0	0	1	0	0	0	0.11	0.33	1
anfallet	2	1	1	1	0	1	1	1	1	1	0.5	2
så	0	0	0	0	0	0	0	0	0	0	0	0
stördes	0	0	0	1	0	1	0	0	0	0.22	0.44	1

Table 1c. Labeling of prominences in read speech by nine experts (E1–E9) and mean, standard deviation and range of the labelings.

Word	E1	E2	E3	E4	E5	E6	E7	E8	E9	Mean	SD	Range
enligt	0	0	1	0	0	0	0	0	0	0.11	0.33	1
libyska	3	2	3	3	3	3	2	3	3	2.78	0.44	1
uppgifter	2	2	2	2	2	2	2	2	2	2	0	0
föll	0	0	2	0	0	1	0	0	0	0.33	0.71	2
åtta	2	3	3	2	3	3	2	3	2	2.56	0.53	1
450-kilosbomber	2	2	3	3	3	2	2	2	2	2.33	0.5	1
över	0	0	1	0	0	0	0	0	0	0.11	0.33	1
Tripoli	2	2	3	3	3	3	2	3	3	2.67	0.5	1
och	0	0	0	0	0	0	0	0	0	0	0	0
Bengazi	2	3	3	2	3	2	2	3	2	2.44	0.53	1
när	0	0	0	0	0	0	0	0	0	0	0	0
de	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	2	2	3	2	2	2	2	2	2	2.11	0.33	1
bombplanen	2	3	3	3	3	3	2	3	3	2.78	0.44	1
slog	0	0	0	0	0	0	0	0	0	0	0	0
till	2	3	3	2	2	2	2	3	2	2.33	0.5	1
natten	2	2	2	2	2	2	2	2	2	2	0	0
till	0	0	1	0	0	0	0	0	0	0.11	0.33	1
tisdagen	2	3	2	3	3	2	2	2	2	2.33	0.5	1
en	2	0	2	0	0	0	1	0	0	0.56	0.88	2
av	0	0	0	0	0	0	0	0	0	0	0	0
bomberna	2	2	3	3	3	3	2	3	2	2.56	0.53	1

Table 1d. Labeling of prominences in spontaneous speech by nine experts (E1–E9) and mean, standard deviation and range of the labelings.

Word	E1	E2	E3	E4	E5	E6	E7	E8	E9	Mean	SD	Range
det	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	2	2	2	2	2	1	2	0	2	1.67	0.71	2
luftangreppet	2	2	3	3	2	3	3	2	3	2.56	0.53	1
sattes	0	0	0	0	2	1	0	0	0	0.33	0.71	2
in	2	0	2	2	0	2	2	0	2	1.33	1	2
i	0	0	0	0	0	0	0	0	0	0	0	0
tisdags	3	3	3	3	3	3	3	2	3	2.89	0.33	1
och	0	0	0	0	0	0	0	0	0	0	0	0
det	0	0	0	0	0	0	0	0	0	0	0	0
genomfördes	2	2	2	2	2	2	2	0	2	1.78	0.67	2
främst	3	2	3	3	2	2	2	2	0	2.11	0.93	3
utav	0	2	1	2	0	0	0	0	0	0.56	0.88	2
bombare	2	2	3	3	2	2	2	2	2	2.22	0.44	1
som	0	0	0	0	0	0	0	0	0	0	0	0
startade	2	2	2	2	2	2	0	2	2	1.78	0.67	2
från	0	0	0	0	0	0	0	0	0	0	0	0
England	3	3	3	3	3	3	3	3	3	3	0	0
under	2	0	1	2	2	0	0	1	1	1	0.87	2
själva	2	2	1	0	2	2	2	2	1	1.56	0.73	2
anfallet	3	2	3	3	3	3	3	3	3	2.89	0.33	1
så	0	0	1	0	0	0	2	0	0	0.33	0.71	2
störde	2	2	3	2	2	2	2	2	2	2.11	0.33	1

3.3.2. Reliability indices

The basic formula for determining reliability based on internal consistency, i.e. covariation, is Cronbach's alpha. Table 2a shows the inter-rater reliability coefficients of the composite rating for boundaries and prominences in the read and spontaneous speech computed by using the procedures outlined by e.g. Winer, Brown & Michels (1991, pp. 1011-1015). The coefficient $R_{k(f)}$, equivalent to Cronbach's alpha, gives an estimation of the correlation between the labelings made by the group of transcribers and an estimated repeated labeling of the same material by the same transcribers. (The subscripts k(f) indicate the group, or composite, k, of transcribers constituting a fixed factor, f, that is, a non-random choice of transcribers (experts specializing in prosody), respectively.) As may be seen, the reliability indices for the expert group are very high, being close to 1 in all four conditions. The differences, moreover, are not significant (cf. Hakstian & Whalen, 1976).

In addition the average reliability of a single transcriber has been computed. This is high, too and does not vary much over the four conditions. (The differences are insignificant.) This may be seen in Table 2b, containing single transcriber reliability coefficients. The coefficient $R_{1(f)}$ estimate the correlation expected between one transcriber and another judging the same material, that is, the typical, or average, reliability of a single transcriber. (The subscripts 1(f) indicate correlations between single, 1, transcribers constituting a fixed factor, f, that is, chosen non-randomly.)

Table 2a. Inter-rater reliability of composite rating $R_{k(f)}$ based on nine experts.

	Boundaries	Prominences
Read	0.98	0.98
Spontaneous	0.97	0.97

Table 2b. Inter-rater reliability of single rater $R_{1(f)}$ based on nine experts.

	Boundaries	Prominences
Read	0.87	0.85
Spontaneous	0.76	0.77

However, high reliability in the sense of internal consistency is no indication that the transcribers agree in an absolute sense on their labelings. To estimate the agreement between transcribers other tests are necessary.

3.3.3. Agreement proportions: comparisons with ToBI evaluations

As mentioned above, tests of agreement estimate, in one way or another, the degree to which measuring instruments, e.g. transcribers, agree in the values assigned to objects (e.g. words). First we will use the same approach as in the ToBI evaluation, described by Silverman *et al.* (1992a, p. 868): "Agreement was calculated across all possible pairs of transcribers for each word of each utterance. For example, four labelers (a, b, c, d) would produce six possible transcriber pairs (ab, ac, ad, bc, bd, cd). Our agreement criterion is stringent: if 3 of 4 transcribers (a, b and c) agree, only 3 of 6 pairs will match (ab, ac and bc but not ad, bd and cd) and we would report 50% agreement." To this can be added that the reported indices are averages of the proportions calculated for each word. For further details, see also Pitrelli *et al.*, (1994).

As may be recalled (see 1), one of the evaluation criteria used for the ToBI system (Silverman *et al.*, 1992a) is an inter-transcriber agreement of at least 80%. The agreement proportions reported in the study by Silverman *et al.*, (1992a) with 20 transcribers participating come close to, or exceed the 80% criterion for the tonal component as a whole, while for the boundary component they do not reach 80% agreement. A slightly relaxed criterion (to +/- one boundary category) raised the agreement on boundaries to above 90%.

Looking at the details of agreement proportions for boundaries reported by Silverman *et al.* (1992a) for the 4 most distinguished of their 20 transcribers they reached 69% exact agreement and an agreement of 94% with a criterion relaxed to +/- one boundary category on their combined read and spontaneous speech material. (The figures for the whole group are only slightly lower, 67% and 93%, respectively.) For our data the corresponding *exact* agreement proportions are 91% for the read and 81% for the spontaneous speech (see Table 3). However, the boundary component of ToBI distinguishes five categories to be compared to the four categories in our system, a fact to be taken into account upon comparison of the two systems.

Table 3. Agreement proportions (%) based on nine experts.

	Boundaries	Prominences
Read	91	78
Spontaneous	81	71

Thus, the boundary categories in ToBI and the system we are evaluating only partially overlap. The differences are even more striking upon comparison of the tonal component of ToBI and the prominence labeling in our system. The tonal transcriptions in ToBI are based on the intonational phonology of Pierrehumbert (see Pierrehumbert & Hirschberg, 1990) and span a wider area of tonal phenomena than the prominence features of our system; (see Silverman *et al.*, 1992a). The prominence component of our system has its closest counterpart in the pitch accent part of the ToBI tonal component. Concerning pitch accents, the ToBI transcribers made their decision in two steps. They decided first whether a word had a pitch accent or not, and if so, they decided on what kind of pitch accent. The agreement proportions reported for these tasks were 86% and 64% respectively for their four most experienced transcribers. Relaxing the criterion on the second task, the one most closely corresponding to our prominence-labeling task, raised the proportion to 79%. Our corresponding figures are 78% and 71% for the read and spontaneous speech, respectively, applying the exact match criterion (see Table 3). Thus, our exact match proportions fall in between the proportions with an exact match and relaxed criterion on the pitch accent part of ToBI.

The ToBI material evaluated by Silverman *et al.* (1992a) as well as Silverman *et al.* (1992b) and Pitrelli *et al.* (1994) consists of utterances (= isolated sentences) representing different kinds of read and spontaneous speech, while we are using longer stretches of connected speech, either read or spontaneous. In the ToBI evaluations, read and spontaneous speech data are pooled. Thus, it is not possible to compare read and spontaneous speech separately. However, our results, reported separately for read and spontaneous speech, point to greater difficulties with the spontaneous speech; the agreement proportions are higher for read as compared to spontaneous speech for both boundaries and prominences.

Another difference between ToBI and the Swedish IPA-based system is that ToBI transcriptions are done using the ESPS/Waves+ speech analysis software, that is, via interactive listening and graphic displays of the speech waveform, total signal energy and extracted fundamental frequency contours (Silverman *et al.*, 1992b). The present system, on the other hand, is exclusively auditory-based. This difference, we think, favors the ToBI transcribers over ours.

In summary, then, our agreement proportions, compared as above, are very similar to those reported by Silverman *et al.* (1992a). Moreover, of the differences between ToBI and our IPA-based system some seem to favor ToBI and some our system. So, if the evaluation reported by Silverman *et al.* (1992a) can be said to have produced reliable results, as claimed by the authors, then our evaluation data seem to be encouraging, too.

3.3.4. Exact agreement between all transcribers

We will now turn to a test with an even stricter criterion for agreement. It is actually a measure of the extent to which *all* the nine transcribers make *exactly* the same judgement. Thus, if 8 out of 9 of the transcribers agree on a word, we would report no agreement using this procedure, while using the ToBI-inspired procedure (3.3.3) the agreement would be very

high (actually 78%). Like similar tests of agreement, this one has an associated test statistic that enables the researcher to determine whether the observed extent of agreement can be attributed to chance (see Rietveld & van Hout, 1993 for a general discussion about agreement indices). In the present study coefficients of inter-rater agreement were calculated on the basis of the formula developed by Tinsley & Weiss (1975):

$$T = \frac{N_1 - pN}{N - pN}$$

where

N = the number of judged objects (number of rows, or words; see Table 1)
N_1 = the number of agreements, i.e. number of rows containing identical labelings
p = the probability that the agreement of k judges is solely due to chance

The formula for calculating p depends on the criterion of agreement (c). If by agreement is meant identical scores, then c = 0. However, if a more relaxed criterion is permitted, with the inclusion of scores differing by one level, then c = 1. The formulas for calculating p when c = 0 and c = 1, respectively, are:

if c = 0:

$$p = (1/v)^{k-1}$$

if c = 1

$$p = \frac{(v-1)\Sigma_{i=1}^{k} 2^{i-1} + 1}{v^k}$$

where
v = number of scale points
k = number of judges

Both kinds of criteria have been used in the calculations in the present study, although our primary concern here is the extent of complete inter-transcriber agreement (c = 0). We present the coefficients (T) in Table 4a for c = 0 and 4b for c = 1, respectively. All coefficients are significantly greater than chance agreement, when applying the test developed by Lawlis & Lu (1972); see also Tinsley & Weiss (1975).

Table 4a. Inter-rater agreement T (Identical scores; c=0) based on nine experts.

	Boundaries	Prominences
Read	0.79	0.45
Spontaneous	0.59	0.32

Table 4b. Inter-rater agreement T (One level differences; c=1) based on nine experts.

	Boundaries	Prominences
Read	0.97	0.80
Spontaneous	0.89	0.66

The proportions of agreement between all the transcribers present the same *pattern* as observed when applying the ToBI-inspired agreement calculations. That is, comparing boundaries and prominences, the figures indicate that agreement is higher on boundaries, and similarly, comparing read speech and spontaneous, agreement is higher on read speech. Thus, the greatest agreement is reached on the transcription of boundaries in read speech and the least on the transcription of prominences in the spontaneous speech. This holds for both of the criteria applied as may be inferred from the two parts of the table. Concerning the comparison of boundaries and prominences our results are corroborated by another study using Swedish material; Bannert (1994) has reported greater agreement on boundaries as compared to prominences in a study of spontaneous Swedish by both expert and non-expert transcribers.

The new information given above is then the extent of *exact* inter-transcriber agreement. It is apparent from Table 4a that it varies over the different conditions. For boundaries it is high, and for read speech boundaries it is almost .80, thus being very close to the 80% agreement criterion introduced by Silverman *et al.* (1992a). Relaxing the criterion for agreement results, as expected, in higher coefficients, that is higher agreement. Allowing for +/- one level differences, all but the prominences-in-spontaneous-speech condition meet an 80% agreement criterion.

Although we do not a priori know what agreement level should be required in order to tell a good transcription system from a bad one, we find the operationally defined 80 % agreement criterion a reasonable one. From the data presented above, then, it is obvious that the low agreement on prominences is a particular weakness that should be remedied, if possible. To find out if there are grounds for making suggestions to improve the system we are evaluating, we will present other aspects of the labeling data.

3.3.5. Distribution data

In Tables 5a–d the distribution of labels over the different boundary and prominence categories (0, 1, 2, 3, respectively) is shown for each transcriber for both the read and the spontaneous speech. Looking at the figures for boundary assignment (Table 5a-b) the dominance of the 0-boundary is very obvious. In effect, agreement on the absence of a boundary may be the main reason for the high inter-rater agreement on boundaries (see 3.3.3; 3.3.4). In contrast, category 3 boundaries (extra strongly marked and corresponding to boundaries between paragraphs) occur very infrequently and with some of the transcribers this category does not occur at all. One could therefore argue that special attention should be devoted to training transcribers to use this category. Alternatively, category 2 and 3 boundaries could be collapsed into one category (all defined as strongly marked and corresponding to boundaries between prosodic utterances). Either of these actions, we think, will increase the agreement to some extent.

The distribution of prominence categories is less clear-cut (Tables 5c–d). The transcribers apparently vary widely in their assignment of the categories 0, 1, 2 and 3. The most

conspicuous diversity concerns the category 1 prominences (words judged as having a single, secondary stress). For all but one of the transcribers, it is an infrequently used category and some do not use it at all. On the basis of the data presented here we suggest that this category be eliminated except for the marking of a second prominence of compound words (see 3.1.2).

We will return in Section 5 to this discussion of possible improvements to the transcription system based on the presented data and the comments of the transcribers summarized in the following section (3.3.6).

Table 5a. Number of words labeled in each boundary category. Read speech (233 words); nine experts.

	E1	E2	E3	E4	E5	E6	E7	E8	E9
3	12	4	1		12	1	3	12	1
2	6	9	10	11	6	11	9	12	11
1	12	28	15	39	15	18	21	9	21
0	203	192	207	183	200	203	200	200	200

Table 5b. Number of words labeled in each boundary category. Spontaneous speech (252 words); nine experts.

	E1	E2	E3	E4	E5	E6	E7	E8	E9
3	18	3	1		16	4		15	3
2	16	17	13	20	9	14	17	12	14
1	13	37	12	45	19	55	27	40	44
0	205	195	226	187	208	179	208	185	191

Table 5c. Number of words labeled in each prominence category. Read speech (233 words); nine experts.

	E1	E2	E3	E4	E5	E6	E7	E8	E9
3	8	22	39	44	26	49	11	33	31
2	122	98	83	93	107	82	101	90	97
1			21			5	6		1
0	103	113	90	96	100	97	115	110	104

Table 5d. Number of words labeled in each prominence category. Spontaneous speech (252 words); nine experts.

	E1	E2	E3	E4	E5	E6	E7	E8	E9
3	26	15	56	53	41	49	17	30	50
2	116	109	53	82	101	89	108	88	87
1		1	55			9	4	13	8
0	110	127	88	117	110	105	123	121	107

3.3.6. Comments on the transcription procedure

Six of the nine transcribers gave written comments on the transcription procedure. Most of them looked upon the task as both time-consuming and difficult. Generally, the transcribers felt unsure about how to draw the limits between the different categories, or levels, of boundaries and prominences. Some had difficulties with the definitions.

General comments on the labeling of boundaries concerned problems experienced with hesitation phenomena and pauses reflecting speech planning. The transcribers pointing to these problems found it extremely difficult to make a distinction between these kind of phenomena and boundaries as defined in the instructions for labeling, according to which hesitations and similar phenomena should not be taken notice of (3.1.1). A number of transcribers suggested that disfluencies of any kind, associated with boundaries as well as hesitations, should be treated together, without considering what caused the disfluency. Separating different kinds of disfluency could then be made in a later analysis.

The difficulties were, as might be expected, most apparent in the transcription of the spontaneous speech. However, with the exception of problems encountered with distinguishing boundaries proper from other types of disfluency, boundaries seem to have been felt easier to label than prominences.

The general comments on prominences concerned problems with distinguishing between the two highest levels (focused versus primary stressed/accented). The transcribers also felt insecure about level 1 (secondary stress) in accent I words, that is, they were insecure about secondary stress representing the highest prominence level in a word. This insecurity seems to have been reflected in the very infrequent use of level 1 prominences (see Tables 5c–d). Some found the definitions of the prominence levels unsatisfactory. The terms chosen were also felt to be confusing by some transcribers.

4. Non-expert labeling

In this section we will present the study of non-experts. The procedure was made in a similar fashion as the one described featuring expert transcribers. The non-expert part of the investigation was based on the same transcription system as used by the experts, though it was modified in some respects to make it appropriate for non-expert use. These modifications will be explained below (4.1.2). The presentation of the labeling procedure and the results of the labeling follow basically the same scheme as for the experts. Moreover, upon presentation of the results, comparisons will be made with the expert data as presented in Section 3.

4.1. Method

4.1.1. Subjects, material and labeling procedure

Ten university students without any prior experience in phonetics acted as transcribers. They were all paid for their participation. The subjects, referred to as NE1–NE10, all speak Swedish as their mother tongue.

The same recorded material was transcribed as in the expert study (see 3.1.1). The transcription tests were run individually or in groups of two or three subjects in the phonetics laboratory in Umeå. The subjects listened to the material on headphones and made markings of the boundaries and prominences they perceived in written versions of the material in which all punctuation had been removed.

All the tests were supervised by a test leader, who handled the equipment, distributed the written instructions for labeling and could answer questions about the labeling. There was a short pre-test to get the subjects acquainted with the procedure and to make sure that they had understood what to do.

4.1.2. Modifications of the base prosody system

The transcriptions by the non-experts were made according to a modified base system. Instead of the four levels of boundaries and prominences in the base system, only three levels were distinguished. In addition, the non-experts were not given exactly the same definitions of boundaries and prominences as the experts. The definitions were adjusted to be suitable for subjects without prior knowledge of phonetics. The modified system is described below with the symbols used by the transcribers to the left, the definitions given in the middle and the numerical coding used to handle the data in the subsequent analyses to the right:

Boundary categories:

cv || cv	stronger marked boundary	= 2
cv | cv	weaker marked boundary	= 1
cv cv	no boundary	= 0

Prominence levels:

<u>cv</u>	having strong prominence	= 2
<u>cv</u>	having weaker prominence	= 1
cv	not prominent	= 0

4.1.3. Analysis and evaluation procedures

As mentioned, we will follow a procedure for presenting the data similar to that used for the experts, that is, we will start by presenting some excerpts from the raw data of boundary and prominence markings together with some simple calculations. For evaluating the non-expert transcriptions we have used the same kind of tests as for the experts. The non-experts, however, were not specifically asked to give any comments on the transcription procedure. Nevertheless, it seemed to us that the subjects had no obvious difficulties with understanding the instructions and definitions and could manage quite well with the procedure.

4.2. Raw data

Tables 6a–d show the labeling of boundaries and prominences in samples of read and spontaneous speech, respectively, by the ten non-experts (NE1–NE10). The tables contain the words in the samples ordered vertically in the first column. For each word the individual labeling by each expert is given in the following ten columns, and the columns furthest to the right contain means, standard deviations and ranges based on all non-expert subjects for each specific word.

Table 6a. Labeling of boundaries in read speech by ten non-experts (NE1–NE10) and mean, standard deviation and range of the labelings.

Word	1	2	3	4	5	6	7	8	9	10	Mean	SD	Range
enligt	0	0	0	0	0	0	0	0	0	0	0	0	0
libyska	0	0	0	0	0	0	0	0	0	0	0	0	0
uppgifter	1	1	0	0	1	0	0	1	1	0	0.5	0.52	1
föll	0	0	0	0	0	0	0	0	0	0	0	0	0
åtta	0	0	0	0	0	0	0	0	0	0	0	0	0
450-kilosb...	1	2	1	1	1	1	1	1	1	0	1	0.47	2
över	0	0	0	0	0	0	0	0	0	0	0	0	0
Tripoli	0	0	0	0	0	0	0	0	0	0	0	0	0
och	0	0	0	0	0	0	0	0	0	0	0	0	0
Bengazi	1	1	1	1	1	1	1	1	1	1	1	0	0
när	0	0	0	0	0	0	0	0	0	0	0	0	0
de	0	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	0	0	0	0	0	0	0	0	0	0	0	0	0
bombplanen	0	0	0	0	0	0	0	0	0	0	0	0	0
slog	0	0	0	0	0	0	0	0	0	0	0	0	0
till	0	0	0	0	0	0	0	0	0	0	0	0	0
natten	0	0	0	0	0	0	0	0	0	0	0	0	0
till	0	0	0	0	0	0	0	0	0	0	0	0	0
tisdagen	2	2	2	2	2	2	2	2	2	2	2	0	0
en	0	0	0	0	0	0	0	0	0	0	0	0	0
av	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6b. Labeling of boundaries in spontaneous speech by ten non-experts (NE1–NE10) and mean, standard deviation and range of the labelings.

Word	1	2	3	4	5	6	7	8	9	10	Mean	SD	Range
det	0	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	0	0	0	0	0	0	0	0	0	0	0	0	0
luftangreppet	1	1	0	1	1	0	1	1	1	0	0.7	0.48	1
sattes	0	0	0	0	0	0	0	0	0	0	0	0	0
in	0	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	1	0	0	0	0	1	0	0.2	0.42	1
tisdags	2	2	2	2	2	1	2	2	2	0	1.7	0.67	2
och	0	0	0	0	0	0	0	0	0	0	0	0	0
det	0	0	0	0	0	0	0	0	0	0	0	0	0
genomfördes	1	1	0	1	1	0	1	1	1	0	0.7	0.48	1
främst	0	0	0	0	0	0	0	0	0	0	0	0	0
utav	1	1	0	1	0	0	1	1	1	0	0.6	0.52	1
bombare	0	0	0	0	0	0	0	1	0	0	0.1	0.32	1
som	0	0	0	0	0	0	0	0	0	0	0	0	0
startade	0	0	0	0	0	0	0	0	0	0	0	0	0
från	0	0	0	0	0	0	0	0	0	0	0	0	0
England	2	2	2	2	2	1	2	2	2	2	1.9	0.32	1
under	1	1	1	1	1	1	1	1	2	0	1	0.47	2
själva	1	1	1	1	0	1	1	1	1	0	0.8	0.42	1
anfallet	1	2	1	1	1	1	0	2	1	0	1	0.67	2
så	0	0	0	0	0	0	0	1	1	0	0.2	0.42	1
störtes	1	0	0	0	0	0	1	1	1	0	0.4	0.52	1

Table 6c. Labeling of prominences in read speech by ten non-experts (NE1–NE10) and mean, standard deviation and range of the labelings.

Word	1	2	3	4	5	6	7	8	9	10	Mean	SD	Range
enligt	0	0	0	0	0	0	0	0	0	0	0	0	0
libyska	2	0	0	0	2	2	2	2	1	1	1.2	0.92	2
uppgifter	1	0	0	0	0	1	1	1	0	0	0.4	0.52	1
föll	0	0	0	0	0	0	0	0	0	0	0	0	0
åtta	2	1	2	1	2	0	2	2	2	2	1.6	0.7	2
450-kilosb...	2	0	2	0	2	2	1	1	1	1	1.2	0.79	2
över	0	0	0	0	0	0	0	0	0	0	0	0	0
Tripoli	1	2	1	0	2	2	1	2	2	2	1.5	0.71	2
och	0	0	0	0	0	0	0	0	0	0	0	0	0
Bengazi	1	2	1	0	2	2	1	2	2	2	1.5	0.71	2
när	0	0	0	0	0	0	0	0	0	0	0	0	0
de	0	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	0	0	0	0	0	0	0	0	0	1	0.1	0.32	1
bombplanen	1	0	1	0	1	1	1	1	1	2	0.9	0.57	2
slog	0	0	0	0	0	0	0	0	0	0	0	0	0
till	1	0	0	0	0	0	2	1	1	2	0.7	0.82	2
natten	0	0	0	0	0	0	1	0	0	0	0.1	0.32	1
till	0	0	0	0	0	0	0	0	0	0	0	0	0
tisdagen	1	0	0	0	0	0	1	1	0	1	0.4	0.52	1
en	0	0	1	0	2	0	0	0	0	1	0.4	0.7	2
av	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6d. Labeling of prominences in spontaneous speech by ten non-experts (NE1–NE10) and mean, standard deviation and range of the labelings.

Word	1	2	3	4	5	6	7	8	9	10	Mean	SD	Range
det	0	0	0	0	0	0	0	0	0	0	0	0	0
amerikanska	1	0	0	0	0	0	0	0	1	0	0.2	0.42	1
luftangreppet	2	1	0	0	0	0	1	1	1	1	0.7	0.67	2
sattes	0	0	0	0	0	0	0	0	0	0	0	0	0
in	0	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0	0	0
tisdags	1	2	2	1	2	1	2	1	2	2	1.6	0.52	1
och	0	0	0	0	0	0	0	0	0	0	0	0	0
det	0	0	0	0	0	0	0	0	0	0	0	0	0
genomfördes	0	0	0	0	0	0	0	0	1	1	0.2	0.42	1
främst	2	0	1	0	0	0	1	1	1	0	0.6	0.7	2
utav	0	0	0	0	0	0	0	0	0	0	0	0	0
bombare	1	1	0	0	0	0	1	1	1	2	0.7	0.67	2
som	0	0	0	0	0	0	0	0	0	0	0	0	0
startade	1	0	0	0	0	0	1	0	0	1	0.3	0.48	1
från	0	0	0	0	0	0	0	0	0	0	0	0	0
England	2	2	2	2	2	1	2	2	2	2	1.9	0.32	1
under	0	0	0	0	0	0	1	0	1	1	0.3	0.48	1
själva	0	0	0	0	0	0	1	0	1	1	0.3	0.48	1
anfallet	2	2	1	1	0	0	2	2	2	2	1.4	0.84	2
så	0	0	0	0	0	0	1	0	1	0	0.2	0.42	1
störtdes	1	1	0	0	0	0	1	1	1	1	0.6	0.52	1

4.3. Evaluations

4.3.1. Reliability indices

Table 7a shows the inter-rater reliability coefficients of the composite rating, $R_{k(f)}$, and Table 7b the average reliability of a single transcriber, $R_{1(f)}$, for boundaries and prominences in the read and spontaneous speech using the same procedures as described for the expert transcribers (see 3.3.2).

Table 7a. Inter-rater reliability of composite rating $R_{k(f)}$ based on 10 non-experts.

	Boundaries	Prominences
Read	0.99	0.86
Spontaneous	0.97	0.89

Table 7b. Inter-rater reliability of single rater $R_{1(f)}$ based on 10 non-experts.

	Boundaries	Prominences
Read	0.94	0.38
Spontaneous	0.74	0.45

The reliability of the composite rating is high for both boundaries and prominences in the read and spontaneous speech. In read and spontaneous speech alike, the coefficients for boundaries are as high as for the experts, while for prominences, the coefficients are somewhat lower than the experts'. There are significant ($p = < .01$) differences between the conditions.

The average reliability of the single transcribers varies considerably, and there are significant ($p = < .01$) differences between the conditions. The reliability is highest for boundaries in read and next highest for boundaries in spontaneous speech. The coefficients for prominences are only about half of those for boundaries. Accordingly, the group/single rater difference is considerable in judgements of prominence.

Thus, comparing the *group* of non-experts with the expert group, we find only minor differences irrespective of the condition. Comparing the *average* rater reliability, on the other hand, we find quite low non-expert values for prominences. We interpret this result as meaning that non-experts are a less homogeneous group than experts, at least as far as prominence labeling is concerned. Based on these data we could argue that, to guarantee reliable results on prominence labelings, we should not use a single non-expert transcriber, or a small group. For experts, on the other hand, a small group of transcribers would not introduce similar unreliability.

4.3.2. Agreement proportions; comparisons with ToBI evaluations

Table 8 shows agreement proportions calculated as for the experts (see 3.3.3).

Table 8. Agreement proportions (%) based on 10 non-experts.

	Boundaries	Prominences
Read	98	69
Spontaneous	84	70

Similarly high proportions as for the experts are obtained with the non-experts. In effect, the proportions are even somewhat higher for the two boundary conditions for the non-experts. Thus, as a group, the non-experts agree on their labelings to about the same extent as the experts.

4.3.3. Exact agreement between all transcribers

The non-expert agreement indices (T) computed in the same way as for the experts (see 3.3.4 for explanations) are shown in Tables 9a–b.

Table 9a. Inter-rater Agreement T (Identical judgements) based on 10 non-experts.

	Boundaries	Prominences
Read	0.92	0.35
Spontaneous	0.60	0.35

Table 9b. Inter-rater Agreement T (One level differences) based on 10 non-experts.

	Boundaries	Prominences
Read	0.98	0.68
Spontaneous	0.91	0.72

First, all agreement coefficients in Tables 9a–b, expressing the extent to which all the 10 transcribers make exactly the same judgement, are significantly greater than could be expected on the basis of chance (for the significance testing, see Lawlis & Lu, 1972 and Tinsley & Weiss, 1975). Second, whatever criteria we use, the agreement is considerably greater on boundaries than on prominences. Third, comparing the figures for exact match and one-level differences, it is apparent that relaxing the criterion has a strong effect on all conditions except for boundaries in read speech in which even the exact match coefficient is very high (> .90).

Comparing the non-expert with the corresponding expert data (3.3.4), the patterning of the results are very similar on all points considered. Also, the coefficients are quite similar comparing the two data sets. In making these comparisons, however, we shall keep in mind the differences between the tasks of the two groups, differences which we assume made the task less complicated for the non-experts. The non-experts only had to distinguish between three categories for boundaries and prominences, while the experts were instructed to use four categories.

4.3.4. Distribution data

Tables 10a–d show the distribution of labels over the different boundary and prominence categories (0, 1, 2 respectively) for each transcriber for both the read and the spontaneous speech.

Table 10a. Number of words labeled in each boundary category. Read speech (233 words); 10 non-experts.

	NE1	NE2	NE3	NE4	NE5	NE6	NE7	NE8	NE9	NE10
2	11	15	11	11	11	11	11	12	11	12
1	20	18	14	17	18	16	17	19	19	15
0	202	200	208	205	204	206	205	202	203	206

Table 10b. Number of words labeled in each boundary category. Spontaneous speech (252 words); 10 non-experts.

	NE1	NE2	NE3	NE4	NE5	NE6	NE7	NE8	NE9	NE10
2	12	20	14	10	14	9	16	22	21	32
1	65	39	17	56	31	29	45	57	67	25
0	175	193	221	186	207	214	191	173	164	195

Table 10c. Number of words labeled in each prominence category. Read speech (233 words); 10 non-experts.

	NE1	NE2	NE3	NE4	NE5	NE6	NE7	NE8	NE9	NE10
2	31	12	13	2	24	9	39	16	7	25
1	77	19	32	8	30	32	69	47	37	88
0	125	202	188	223	179	192	125	170	189	120

Table 10d. Number of words labeled in each prominence category. Spontaneous speech (252 words); 10 non-experts.

	NE1	NE2	NE3	NE4	NE5	NE6	NE7	NE8	NE9	NE10
2	18	14	18	4	32	5	49	16	23	46
1	69	27	36	14	21	21	84	37	66	79
0	165	211	198	234	199	226	119	199	163	127

Except for the labeling of boundaries in read speech, the transcribers vary considerably in the number of words classified in each category. This is roughly the same pattern as found for the experts (3.3.5), although the variability is greater for the non-experts, especially concerning prominences. For example, comparing 0-boundaries and 0-prominences for the two groups, the number of words classified in these categories varies over a wider range for the non-experts, except for boundaries in read speech, the easiest of the four conditions. The

ranges for both boundaries and prominences scored 0 are shown in Table 11 for both the experts and non-experts.

Table 11. Ranges (minimum and maximum) of the number of words referred to the 0-boundary and 0-prominence categories over the 9 experts and 10 non-experts, respectively.

	Experts	Non-experts
Boundaries: read	183–207	200–208
Boundaries: spontaneous	179–226	164–221
Prominences: read	90–115	120–202
Prominences: spontaneous	88–127	119–234

5. Summary discussion and conclusions

Summarizing the results comparing the expert and non-expert labeling, we generally do not find any very striking differences. In most respects the experts and non-experts are equally reliable and agree to roughly the same extent on the tasks carried out. That is, the non-experts distinguish between boundary and prominence categories in a similarly consistent manner as the experts do, although the agreement is somewhat lower on prominences for the non-experts. However, the similarities may seem more apparent than they really are, considering the somewhat different tasks given the two groups. Thus, the more simplified procedure *might* have raised the scores for the non-experts. Although we do not know for sure, we have to consider this possibility in evaluating the results. The sole apparent differences, the considerably lower average reliability indices for the non-experts on the two prominence conditions (cf. Tables 2b and 7b) seem to indicate that individual non-experts vary more between themselves than experts. That is, choosing one expert subject instead of another might affect the results less than choosing one rather than another non-expert. In conclusion, the experts constitute a more homogeneous group with individuals behaving in a very similar manner.

Even if the non-expert subjects appear to distinguish between boundary and prominence categories in a similarly consistent manner as the experts do, from the data we have presented we cannot determine the extent of agreement *between* the experts and the non-experts. Theoretically, the agreement could be low or even non-existent. However, it could be noted in passing that some preliminary data on the expert/non-expert agreement have been presented before (Strangert & Heldner, 1994). The subjects in that study were the same non-experts as in the present study and the expert E9. We found that the labelings of the group of non-experts clearly correlated with the labelings of the expert. The regression coefficients were close to 1 for boundaries, while for prominences they were about .50, indicating that the non-experts generally underestimated the strength of prominences in relation to the expert.

Furthermore, comparing the labeling of boundaries and prominences, a consistent pattern observed is the higher reliability and agreement reached on boundaries. This is the case for the experts and non-experts alike and in all calculations in this study. Similarly, there is a pattern with somewhat higher values for the read than for the spontaneous speech. This pattern is however only found for the expert transcribers.

The indices for agreement compare well with other similar evaluations (e.g. Silverman *et al.*, 1992a; Reyelt, 1993), though differences between the transcription systems complicate the interpretation of agreement values. The 80% agreement criterion favored in the ToBI evaluations is met (without being relaxed, see Silverman *et al.*, 1992a) by the experts in all conditions in our study, except in the case of prominences in spontaneous speech. As for the non-experts, the agreement on boundaries exceeds the absolute match criterion and the agreement on prominences comes close, even if it does not meet it. Thus, we must conclude that the transcription system we have used seems to work about equally well as other systems as far as reliability is concerned. This is encouraging, especially considering that the evaluation presented here is the first large-scale test of the system.

The other criteria, discussed by Silverman *et al.* (1992a) in connection with ToBI (see 1.2) are, at least to some extent, also met by the presented IPA-based system. Concerning coverage, the system has some limitations. Intended as a base prosody system, it is restricted to the transcription of very primary prosodic categories (though it may easily be enlarged by including additional prosodic aspects, cf. the system used by Bruce & Touati, 1990). According to the learnability criterion, the system should be possible to learn in a relatively short time. About this we do not have much to say, as the system we used has not been taught, except for the short written instruction given before the labeling. There was no training whatsoever. However, it could be argued that the experts had been trained, as the IPA-symbols used are well known for all the experts. The non-experts, on the other hand, had not been trained at all and still performed almost as well on their somewhat simplified task. This seems to indicate that the prosodic categories labeled are in some sense 'natural' and easily available irrespective of formal training.

However, this does not mean that the system cannot be improved. On the contrary, we believe that agreement may be increased by certain modifications of the existing procedure. Our proposals for improvement are based on the analysis of the labeling (primarily the expert data, see especially 3.3.5) but also on the comments given by the expert transcribers (3.3.6). Thus, we think that category 2 and 3 boundaries could be collapsed into one. Further, the relevance of distinguishing prominences associated with secondary stress in accent I words should be questioned, primarily because of the infrequent use of this label, but also because the transcribers witnessed to the insecurity they experienced with the use of this category. Another proposal is to refine the definitions of the different prosodic categories, which in the existing system were felt to be to some extent unclear. Also, the problem with the distinction between boundaries and other types of disfluencies must be solved. Finally, the very obvious feeling of insecurity about how to distinguish between contiguous categories expressed by the majority of transcribers promotes the development of a special training program (cf. Beckman & Ayers, 1994; Beckman & Hirschberg, 1994).

Yet, even without such a training program, the base prosody, when used both by experts and non-experts, produce very reliable data, although the highest-quality labeling in terms of inter-transcriber agreement is clearly obtained with the experts. Therefore, based on this evaluation study, we may conclude that the system has the potential for being a usable means for transcribing Swedish prosody.

Acknowledgements

We gratefully acknowledge the following contributors, without whose help this study would not have been possible: Gösta Bruce, Rolf Carlson, Claes-Christian Elert, Anders Eriksson, Gunnar Fant, Eva Gårding, Olle Kjellin, Anita Kruckenberg, Per Lindblad, Ulla Sundberg and the group of non-expert students.

The research was supported by grants from the Swedish HSFR/NUTEK Language Technology Program.

References

Bagshaw, P. C. & Williams, B. J. (1992) Criteria for labelling prosodic aspects of English speech. In *Proceedings ICSLP 92*, pp. 859-862. Alberta: Department of Linguistics, University of Alberta.

Bannert, R. (1994) Listeners' identification of prominence and chunking in spoken Swedish: Variation and consistency. In *Papers from the Eighth Swedish Phonetics Conference*, pp. 10-13. Lund: Department of Linguistics, Lund University.

Beckman, M. E. (1986) *Stress and non-stress accent.* Dordrecht: Foris Publications.

Beckman, M. E. & Ayers, G. (1994) *Guidelines for ToBI labelling, version 2.0.* Ohio State University, Obtain by writing to tobi@ling.ohio-state.edu.

Beckman, M. E. & Hirschberg, J. (1994) *The ToBI annotation conventions.* Ohio State University, Obtain by writing to tobi@ling.ohio-state.edu.

Bruce, G. (1994) Prosodisk strukturering i dialog. In *Svenskans beskrivning 20*, Lund, pp. 9-23.

Bruce, G. & Touati, P. (1990) On the analysis of prosody in spontaneous dialogue. In *Working Papers 36*, pp. 37-55. Lund: Department of Linguistics and Phonetics, Lund University.

Hakstian, R. A. & Whalen, T. E. (1976) A k-sample significance test for independent alpha coefficients, *Psychometrica,* **41**(2), 219-231.

Lawlis, G. F. & Lu, E. (1972) Judgment of counseling process: Reliability, agreement, and error, *Psychological Bulletin,* **78**(1), 17-20.

Lieberman, P. (1965) On the acoustic basis of the perception of intonation by linguists, *Word,* **21**, 40-54.

Nunnaly, J. C. (1978) *Psychometric Theory.* New York: McGraw-Hill.

Pierrehumbert, J. B. & Hirschberg, J. (1990) The meaning of intonation contours in the interpretation of discourse. In *Intentions in communication* (P. R. Cohen, J. L. Morgan & M. E. Pollack, eds.). Cambridge, Mass: MIT Press.

Pierrehumbert, J. B. B. (1980) *The Phonology and Phonetics of English Intonation*. Ph D Dissertation. Bloomington, Indiana: Distributed by Indiana University Linguistics Club 1987.

Pitrelli, J. F., Beckman, M. E. & Hirschberg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings ICSLP 94*, pp. 123-126. Yokohama, Japan.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. (1991) The use of prosody in syntactic disambiguation, *Journal of the Acoustical Society of America,* **90**(6), 2956-2970.

Reyelt, M. (1993) Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German. In *Working Papers 41: Proceedings of the ESCA Workshop on Prosody* (D. House & P. Touati, eds.), pp. 238-241. Lund: Department of Linguistics and Phonetics, Lund University.

Rietveld, T. & van Hout, R. (1993) *Statistical techniques for the study of language and language behaviour.* Berlin: Mouton de Gruyter.

Silverman, K. E. A., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B. & Hirschberg, J. (1992a) ToBI: A standard for labeling English prosody. In *Proceedings ICSLP 92*, pp. 867-870. Alberta: Department of Linguistics, University of Alberta.

Silverman, K. E. A., Blaauw, E., Spitz, J. & Pitrelli, J. F. (1992b) A prosodic comparison of spontaneous speech and read speech. In *Proceedings ICSLP 92*, pp. 1299-1302. Alberta: Department of Linguistics, University of Alberta.

Strangert, E. & Heldner, M. (1994) Prosodic labelling and acoustic data. In *Working Papers 43: Papers from the EIGHTH SWEDISH PHONETICS CONFERENCE*, pp. 120-123. Lund: Department of Linguistics and Phonetics, Lund University.

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

'ten Bosch, L. F. M. (1993) Algorithmic classification of pitch movements. In *Working Papers 41: Proceedings of the ESCA Workshop on Prosody* (D. House & P. Touati, eds.), pp. 242-245. Lund: Department of Linguistics and Phonetics, Lund University.

Tinsley, H. E. A. & Weiss, D. J. (1975) Interrater reliability and agreement of subjective judgements, *Journal of Counseling Psychology,* **22**(4), 358-376.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design.* New York: McGraw-Hill.

Appendix

Read speech

enligt "libyska 'ùppgifter | föll 'å̀tta fyrahundra'fèmtiokilos,bomber | över "Tripoli och Ben'gazi | när de ameri'kanska "bòmb,planen slog 'till 'natten till 'tisdagen || en av bomberna ˌföll bara ett "tì̀o,tal 'meter | från den 'bỳggnad där 'lèdaren Mu'ammar Gad"dafis fa'milj låg och 'sov | i en mili'tä̀rför,läggning 'nä̀ra 'Tripoli || en av 'dem som 'dȍdades var Gad'dafis 'sèxton 'må̀nader 'gàmla adop'tì̀v,dotter "Hànna | som 'bòdde med fa'miljen i "Tripoli || 'en av offi'cèrarna i för'läggningen be'rättar | att Gad'dafi "själv be'fann sig i sitt 'tält | ett par 'hundra 'meter längre "bort | då ameri'kanerna 'slä̀ppte 'bomberna || i 'nä̀rheten av 'tältet finns en 'bòmb,krater | och "sànd,säckarna kring 'tä̀lt,väggarna | har pressats 'samman av 'trỳck,vågen || enligt "libyska 'ùpp,gifter har ameri"kanerna 'också 'prì̀ckat en "tènnis,plan med 'en av 'bomberna || i "nä̀r,heten av mili'tä̀rför,läggningen 'dràbbades "flera ci'vila || en "libysk af'fä̀rsman och hans fa'milj | låg och 'sov i sitt 'hus | i ett av de "fì̀nare kvar'teren i "Tripoli | när de ameri'kanska 'bòmb,planen 'natten till 'tisdagen kom ˈìn,svepande över den 'libyska 'hùvud,staden || det var sannolikt inte "mèningen att de ameri'kanska 'bomberna | skulle "träffa af'fä̀rsmannens 'hus | utan i 'stä̀llet den 'libyska "sä̀kerhetspo,lisens 'bỳggnad | omkring 'femhundra 'meter "bort || men det "blev i 'stä̀llet en "fùll,träff i pri'vàt,villan | och af"fä̀rs,mannen och hans fa'milj på 'sju per'soner "ùt,plånades | enligt det offi'ciella be'skedet i 'Tripoli || enligt "libyska 'ùpp,gifter | har 'ùppröjningsar,betena så 'smått kommit i"gång i de ci'vila kvar'teren i 'Tripoli | men "må̀nga frùktar ett "nytt ameri'kanskt 'lùftan,fall || en "vild 'rỳktes,flora gras'serar 'också om 'lèdaren Mu'ammar Gad'dafi |||

Spontaneous speech

det am(e)ri'kanska "lùftanˌgreppet | sattes 'in i "tisda(g)s || och de(t) 'gènomˌfördes "främst utav 'bòmbare som 'stàrtade från "England || ˌunder ˌsjälva "ànˌfallet | så 'stȍrdes de(t) 'libyska 'lùftförˌsvaret "ut | elek"troniskt | utav am(e)ri'kanska 'sȃndare på 'fȃrˌtyg i "Medelˌhavet | "ùtanför 'Tripoli || så att ˌsjälva "lùftanˌgreppet blev ju 'då en 'fràmˌgång | 'libyerna kunde '(i)nte för'svara sig || mot "Tripoli 'fȃllde man | "ȃtta 'stycken fyrahundra"fèmtikilosˌbomber || dom ˌverkar va ha ˌvarit koncen"trerade | framför 'allt mot en mili"tärförˌläggning 'där | Gad"dafi 'tìllbringade "natten || "en av 'bomberna 'träffade de(t) "tält där hans fa'milj 'bòdde || o(ch) en adop"tìvˌdotter ti(ll) honom | 'sèxton 'månader 'gàmmal | 'dȍdades || Gad'dafi "själv | han "klàra(de) sig | genom att han 'låg i ett 'tält | ett 'hùndrantal "meter 'däri'från || det fanns en 'krater 'ùtanför 'där | som 'vìsade emell(er)'tid att man | va(r) "nära a(tt) "träffa hoˌnom || "sändˌsäckarna som 'òmgav 'tältet hade | 'tryckts i'hop || men som 'sagt | 'klàra(de) sej Gad'dafi 'òskadd ||| på ett "annat 'ställe i 'stan | i lite 'fìnare kvar'ter så | 'ùtˌplånades en af"fȃrsˌman | o(ch) 'hèla hans fa'milj på 'sju per'soner || de(t) va(r) natu(r)li(g)tvis inte "heller 'mèningen utan | "mȃlˌtavlan 'där | ˌverkar ha 'vàrit | "sȃkerhetsˌtjänsten | som hade en 'bỳggnad | 'bàra | "fem hundra 'meter däri'från ||| "ùppˌgifterna | 'från ''Libyen | har 'vàrit 'vȃldigt "mòtsägelseˌfulla | men "i och 'med att | de(t) 'fȍrsta "planet | från 'Libyen | 'làndade i "Rom | 'nu i "torsda(g)s | "fȍrmidˌda(g) | så 'fick man en del "ȍgonvittnesˌskildringar || bla(nd a)nnat be'rättade en 'ung span"jor om | den "vìllerˌvalla o(ch) de(t) "kaos som 'rȃdde på 'gàtorna | be"väpnade 'yngre "män sprang om'kring och 'sköt "vilt | under peri'oden strax "èfter de 'här 'ànˌfallet || och informa"tionen om 'va(d) som "pȃˌgick | var ˌpraktis(kt) taget "òbeˌfintlig | 'rỳkten gras'serade 'vilt |||

Paper II In *Focal accent – f_0 movements and beyond*
pp. 55–60

The labeling of prominence in Swedish by phonetically experienced transcribers[1]

Eva Strangert and Mattias Heldner

An IPA-based system has been agreed upon for labeling Swedish prosody. In the present study this system is evaluated by assessing the intertranscriber reliability in prominence labeling of nine expert subjects. The study also explores the acoustic (f_0) basis for observed variability in the assignment of focus accent, the highest prominence label.

1. Introduction

Recently, as large corpora of prosodically labeled speech are needed for quantitative computational modeling of speech, great efforts are being taken to develop transcription systems meeting high standards on reliability. Thus, before extensive use of a system is initiated, it must be evaluated. The ToBI (Tones and Break Indices) system developed for transcribing English prosody has been evaluated in a number of studies (e.g. Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg, 1992; Pitrelli, Beckman & Hirschberg, 1994). Reyelt (1993) evaluated a number of variants of prosodic transcription for German within the VERBMOBIL project. For Swedish, an IPA-based system has been agreed upon for labeling prosody (prominence and boundary phenomena), the details of which have been described in Bruce (1994). We have used this system in two studies Strangert & Heldner (1994) and Strangert & Heldner (1995) comparing the labeling of boundaries and prominences in spoken Swedish made by phonetically experienced and non-experienced transcribers.

In the present study, the scope has been widened. One purpose, which it shares with the former studies (Strangert & Heldner, 1994; Strangert & Heldner, 1995), is to evaluate the transcription system used for labeling. In particular we want to estimate the extent to which experienced phoneticians and speech researchers vary in their labeling of prominences when presented with samples of read and spontaneous Swedish. In addition, the study aims at exploring the acoustic basis, specifically f_0-characteristics, for the variability in labeling that we predict will occur. In particular, we want to establish the extent to which the variability associated with the assignment of focus accent is explainable in terms of f_0-cues.

Beckman (1986) reviews the research on acoustic correlates to perceived stress in English. Referring to Nakatani & Aston (1978), Beckman (1986, p. 60-62) makes clear that the dependence of perceived stress on f_0-cues is complex, and varies with the position of the

[1] This material has been published as Strangert, E. & Heldner, M. (1995) The labelling of prominence in Swedish by phonetically experienced transcribers. In *Proceedings ICPhS 95*, pp. 204-207. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

word in the sentence. Further, Wells (1986) concludes that f_0 cues play an important role for perceived prominence in English, although various other cues contribute, too. Although f_0 is not assumed to be the only cue to prominence in Swedish – Bruce (1983) also mentions temporal correlates, and there are also data reported in Sundberg (1994) indicating temporal correlates – it is believed to be an important determiner of focus accent. Thus, relating perceived focus accent to f_0-events seems reasonable in the light of previous research (Bruce, 1977) according to which focus accent is intimately tied to a f_0-rise following a word accent f_0-fall timed differently for words with acute and grave accent, respectively.

2. Evaluation of the transcription system

2.1. Method

The nine subjects participating in the study are all phoneticians or speech researchers with wide experience in prosody from different sites in Sweden. All are native-born Swedes.

The subjects transcribed two kinds of recorded speech material. One was an excerpt, 233 words long, from an authentic news cable read aloud. The other was a 252-word-long excerpt of spontaneous speech, a retelling of the story read aloud. Both recordings were made in a soundproof room and rendered by the same male Swedish speaker.

Each expert was sent the recorded material and instructions for labeling prominence according to the IPA-based Swedish system. Following this, four levels of prominence were distinguished and labeled accordingly for each word in the material: no stress (unmarked), secondary stress ($_{\iota}$), primary stress/accented ($'$) and focus accent ($''$).

Subsequent analyses included coding the data (no stress=0; secondary stress=1; primary stress=2; focus accent=3) and statistical analyses to estimate reliability.

2.2. Labeling data

Table 1 shows the labeling of prominences by the nine experts in a sample of the read material. The words in the text are ordered vertically in the first column. The following nine columns contain the individual labelings of the transcribers and the tenth column the means of these labelings for each word. The data presented give a rough indication of the reliability of labeling.

2.3. Inter-transcriber reliability

Generally, reliability concerns the extent to which measurements are repeatable in a variety of conditions. Within this framework we will consider two aspects, the one concerning the extent to which the transcribers covary, that is, give relative labeling values that are correlated, and the other concerning the extent to which the transcribers give identical labels. We will henceforth refer to the first as ‘reliability’ and the second as ‘agreement’. All computations are made with acute and grave accent words pooled.

The inter-transcriber reliability (Cronbach’s alpha) for prominence is .98 for read and .97 for spontaneous speech (difference not significant). That is, the transcriptions are highly reliable in the sense of relative labeling consistency irrespective of the material.

Table 1. Labeling by nine transcribers. 0=no stress, 1=secondary stress, 2=primary stress, 3=focus accent.

Word	Transcribers 1 – 9									Mean
enligt	0	0	1	0	0	0	0	0	0	0.1
libyska	3	2	3	3	3	3	2	3	3	2.8
uppgifter	2	2	2	2	2	2	2	2	2	2.0
föll	0	0	2	0	0	1	0	0	0	0.3
åtta	2	3	3	2	3	3	2	3	2	2.6
450-kilosbomber	2	2	3	3	3	2	2	2	2	2.3
över	0	0	1	0	0	0	0	0	0	0.1
Tripoli	2	2	3	3	3	3	2	3	3	2.7
och	0	0	0	0	0	0	0	0	0	0
Bengazi	2	3	3	2	3	2	2	3	2	2.4
när	0	0	0	0	0	0	0	0	0	0
de	0	0	0	0	0	0	0	0	0	0

To determine the reliability in the more strict sense of agreement, that is identical matching, we used the same test as Silverman *et al.* (1992) and Pitrelli *et al.* (1994). They calculated the agreement across all possible pairs of transcribers for each word of each utterance labeled. The index was calculated as the average percentage of agreeing pairs and, according to the criterion set in (Silverman *et al.*, 1992), the agreement should be at least 80%. Calculated on our data, this index is 78% and 71%, for the read and spontaneous speech respectively, thus indicating a somewhat higher agreement on the read speech. There are several differences between ToBI and our system, which make comparisons complicated. For the ToBI transcribers, the task was to decide whether a word had a pitch accent or not, and if so, what kind of pitch accent. The indices reported for these tasks were 86% and 64% respectively for the four most experienced of their 20 transcribers.

We also calculated an index estimating the extent to which *all* the transcribers made *exactly* the same judgements on each word. A detailed account of these calculations and other evaluation data presented here are given in (Strangert & Heldner, 1995).

3. f_0 in relation to prominence labels

3.1. Method

The subsequent analysis was made on 60 acute and 55 grave accent words judged to be focused (that is, having a prominence degree of 3, according to our coding) by two or more of the nine transcribers. For each of these words a prominence mean score based on the labeling of all nine transcribers was calculated. The words were digitized at 44.1 kHz. Measurements were made in both the read and spontaneous speech of the size of the word accent fall and the focus accent rise.

To calculate the falls and rises four measuring points were defined, primarily on the basis of the f_0 tracings, see the illustrations in Figure 1: (1) The beginning of the word accent fall; the highest point in the word accent fall. (2) The end of the word accent fall; the lowest point in the word accent fall. (3) The beginning of the focus accent rise; the lowest point in the focus accent rise. For acute accent words this point coincides with (2). For grave accent

words it either coincides with (2) or, in the case of longer words, may be located at some distance from (2). (4) The end of the focus accent rise; the highest point in the focus accent rise. In a few cases in which the critical f_0-events were not easily located, additional criteria were used, determined on the basis of the patterns observed in the unequivocal cases. We also used Engstrand (1989) as a reference when deciding on these additional criteria.

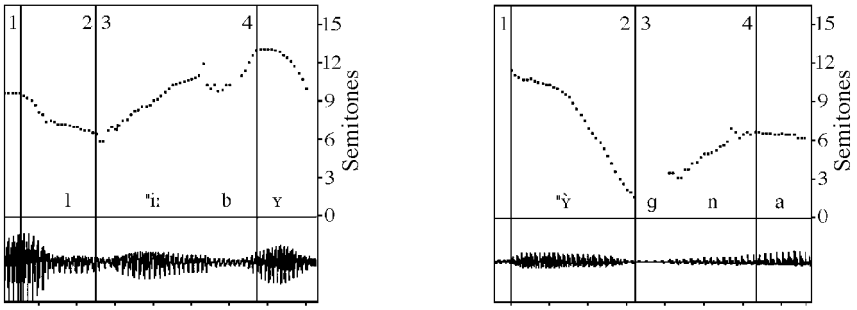

Figure 1. Measurement points. Left: underlined portion of libyska (acute accent); right: underlined portion of byggnad (grave accent).

The word accent fall is defined as the difference between points (1) and (2) and the focus accent rise is defined as the difference between points (3) and (4) measured in semitones. In addition we tested two other f_0-parameters, differences (focus accent rise–word accent fall) and ratios (focus accent rise/word accent fall).

3.2. Results

The majority of the prominence mean scores for all acute and grave accent words included in the analysis fell in the range between 2 and 3. (It should be recalled that a word judged to be focused is coded as 3 in our analysis. Therefore, mean scores close to 3 indicate a general agreement on the word as being focused.) The prominence mean scores were then used in multiple regression analyses to determine if, and to what extent, the measured f_0 movements (with word accent fall and focus accent rise as the independent variables) could explain the variability in the prominence scores.

The results demonstrate insignificant effects of the word accent fall in the read as well as the spontaneous speech and for words with acute and grave accent alike. The focal accent rise, on the other hand, is significantly correlated with perceived scores ($p<.05$) both in the read and spontaneous speech and for acute and grave accent words (Figure 2 a-d). That is, the greater the size of the rise, the stronger the agreement on focus accent. Both kinds of data therefore corroborate previous results demonstrating greater effects on perceived prominence of the rise than the fall (Bruce, 1977; Sundberg, 1994). However, the R-square values, correlations in terms of explained variance, are quite low for all four regression models, .14 and .26 for the acute accent and .41 and .38 for the grave, indicating other influences than f_0 on perceived prominence cf. (Sundberg, 1994).

We also did regression tests with differences as well as ratios between the focus accent rise and the word accent fall as independent variables, but neither of them reached significance.

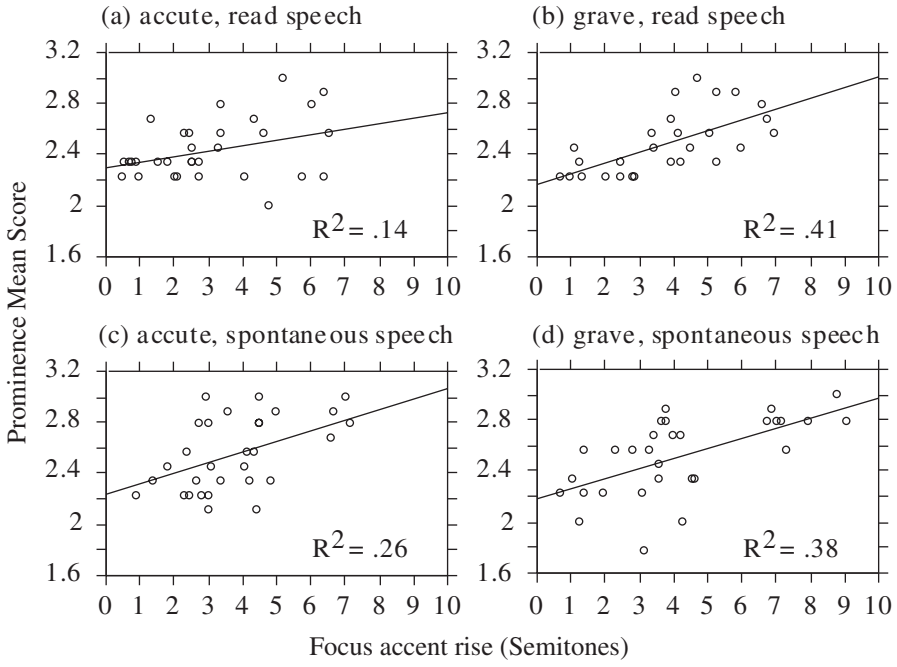

Figure 2. Regression analyses of size of focus accent rise and prominence mean score for 60 acute and 55 grave accented words in read and spontaneous speech.

4. Conclusions

In this prosodic transcription evaluation we have demonstrated the capacity of the system as used by expert transcribers. The reliability is high as well as the intertranscriber agreement. Exploring the acoustic basis for observed variability associated with the assignment of focus accent, we found that the greater the f_0-rise, the stronger the agreement on focus accent. That is, the size of the focus accent cues the degree of prominence. Yet it explains only part of the variation. In conclusion then, there are other important cues to perceived prominence (focus accent) than those investigated here. We are in the process of conducting a study including temporal as well as other cues to perceived focus accent.

Acknowledgements

We gratefully acknowledge the contributions of Gösta Bruce, Rolf Carlson, Claes-Christian Elert, Anders Eriksson, Gunnar Fant, Eva Gårding, Olle Kjellin, Anita Kruckenberg, Per Lindblad, Ulla Sundberg, without whose help this study would not have been possible.

This research was supported by grants from the Swedish HSFR/NUTEK Language Technology Program.

References

Beckman, M. E. (1986) *Stress and non-stress accent.* Dordrecht: Foris Publications.

Bruce, G. (1977) *Swedish word accents in sentence perspective.* Lund: CWK Gleerup.

Bruce, G. (1983) Accentuation and timing in Swedish, *Folia Linguistica,* **17**, 221-238.

Bruce, G. (1994) Prosodisk strukturering i dialog. In *Svenskans beskrivning 20*, Lund, pp. 9-23.

Engstrand, O. (1989) Phonetic features of the acute and grave word accents: data from spontaneous speech, *PERILUS: Phonetic Experimental Research at the Institute of Linguistics University of Stockholm,* **X**, 13-37.

Nakatani, L. H. & Aston, C. H. (1978) *Acoustic and linguistic factors in stress perception*. Unpublished manuscript.

Pitrelli, J. F., Beckman, M. E. & Hirschberg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings ICSLP 94*, pp. 123-126. Yokohama, Japan.

Reyelt, M. (1993) Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German. In *Working Papers 41: Proceedings of the ESCA Workshop on Prosody* (D. House & P. Touati, eds.), pp. 238-241. Lund: Department of Linguistics and Phonetics, Lund University.

Silverman, K. E. A., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B. & Hirschberg, J. (1992) TOBI: A standard for labeling English prosody. In *Proceedings ICSLP 92*, pp. 867-870. Alberta: Department of Linguistics, University of Alberta.

Strangert, E. & Heldner, M. (1994) Prosodic labelling and acoustic data. In *Working Papers 43: Papers from the EIGHTH SWEDISH PHONETICS CONFERENCE*, pp. 120-123. Lund: Department of Linguistics and Phonetics, Lund University.

Strangert, E. & Heldner, M. (1995) Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In *PHONUM 3* (R. Bannert & K. Sullivan, eds.), pp. 85-109. Umeå: Department of Phonetics, Umeå University.

Sundberg, U. (1994) Tonal and temporal aspects of child directed speech. In *Working Papers 43: Papers from the EIGHTH SWEDISH PHONETICS CONFERENCE* (G. Bruce, D. House & P. Touati, eds.), pp. 128-131. Lund: Department of Linguistics and Phonetics, Lund University.

Wells, W. H. G. (1986) An experimental approach to the interpretation of focus in spoken English. In *Intonation in Discourse* (C. Johns-Lewis, ed.), pp. 53-75. London: Croom Helm.

Paper III In *Focal accent – f_0 movements and beyond*
pp. 61–66

To what extent is perceived focus determined by f_0-cues?[1]

Mattias Heldner and Eva Strangert

Two experiments were designed to investigate the perceptual strength of a f0-rise relative to other possible local and global cues to focus in Swedish. The contribution of f0 relative to other possible local cues was investigated by manipulating the f0-contour in naturally produced Swedish sentences. Manipulations involved a gradual reduction of the f0-rise in focused words and a gradual addition of a f0-rise in non-focused words. The possible influence of global cues was explored by varying the amount of global information. In the first experiment, subjects were presented with complete sentences. In the second experiment, they heard sentences with the last part excluded. The results indicate that the f0-rise is neither necessary nor sufficient to perceive focus; other cues, both local and global, appear also to play a role.

1. Introduction

It is widely held that focus is signaled mainly by prosodic means such as accents. Pitch movements, moreover, are generally seen as the most reliable acoustic correlates and also the most reliable perceptual cues to accents.

The experiments described below are part of a series dealing with the perception of focus in Swedish (Heldner, 1998). The main question raised here concerns pitch as a cue to focus, and more specifically, whether a f_0-rise is a necessary or sufficient cue to focus in Swedish. The reason for evaluating the f_0-rise is, on the one hand, the assumption reflected in the literature that an f_0-rise is the crucial cue to perceived focus in Swedish (Bruce, 1977) and to accents in general (Beckman, 1986). On the other hand, there are studies indicating that the f_0-rise is not a necessary cue. Strangert & Heldner (1994) reported on cases of perceived focus in Swedish without the crucial f_0-rise, and Strangert & Heldner (1995) found only weak correlations between the size of the f_0-rise and perceived prominence. Another issue touched on is the extent to which focus is globally signaled. It is generally claimed that the domain of focus is a unit larger than the constituent focused, e.g. a phrase or a sentence (Bruce, 1977). In the light of this, it seems reasonable to assume focus to be perceived on the basis of the prominence relations within this larger unit.

These questions were investigated in two perception experiments with manipulations of naturally produced sentences. The contribution of f_0 relative to other possible cues, e.g. duration and intensity, was investigated by removing the f_0-rise in focused words and

[1] This material has been published as Heldner, M. & Strangert, E. (1997) To what extent is perceived focus determined by F0-cues? In *Eurospeech '97 Proceedings,* pp. 875-877. Rhodes, Greece: ESCA.

inserting a f_0-rise in non-focused words. The possible influence of global signaling was explored by restricting the amount of global information.

2. Method

The speech material (See Table 1) was composed of pairs of questions (Q) and answers (A) similar to those used by Bruce (1977). Four target words, two with acute accent (be'nämna and 'nummer) and two with grave accent ('ànmäla and 'nùnnor) were embedded in two positions (medial and final) in the answer sentences. Depending on the question, either the medial or the final target word was in narrow focus. The material was recorded by a male Central Standard Swedish speaker.

Table 1. Speech material. Target words are underlined, capitals mark focus placement.

Q:	Vilka linjära tecken vill man benämna? 'Which linear signs does one want to name?'
A:	Man vill benämna några linjära NUMMER. 'One wants to name a few linear numbers.'
Q:	Vad vill man göra med några linjära nummer? 'What does one want to do with a few linear numbers?'
A:	Man vill BENÄMNA några linjära nummer. 'One wants to name a few linear numbers.'
Q:	Vilka lama damer vill man anmäla? 'Which paralyzed ladies does one want to report?'
A:	Man vill anmäla några lama NUNNOR. 'One wants to report a few paralyzed nuns.'
Q:	Vad vill man göra med några lama nunnor? 'What does one want to do with a few paralyzed nuns?'
A:	Man vill ANMÄLA några lama nunnor. 'One wants to report a few paralyzed nuns.'

2.1. First and second experiments

The answer sentences shown in Table 1 were analyzed and resynthesized with five artificial f_0-contours on the medial target words. Two series of stimuli were constructed; the first involving a gradual addition (4 equal steps; semitones) of a f_0-rise in the non-focused words in medial position and the second a gradual reduction (4 equal steps; semitones) in the f_0-rise in the focused words in medial position. The manipulations were modeled on findings from a production experiment (Heldner, 1998). Thus, the steps were about one semitone for the target word 'benämna' and about two semitones for 'anmäla'. A stylized image of the manipulations in the sentences containing the target word 'anmäla' is given in Figure 1.

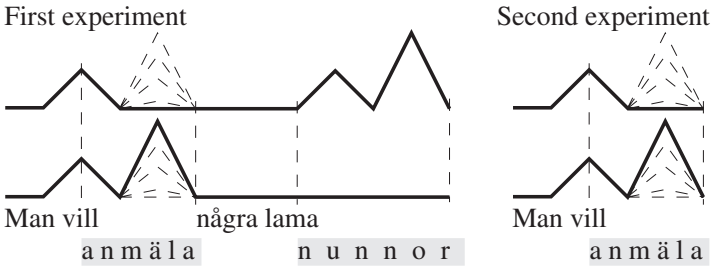

Figure 1. Stylized images of the f_0-manipulations in the target word 'anmäla' in the first (left-hand figure) and in the second experiment (right-hand figure). Manipulations included a gradual addition of a f_0-rise in the non-focused words (top panels) and a gradual reduction in the f_0-rise in the focused words (bottom panels). Thick lines mark original contours.

In the first experiment, subjects were presented with the complete sentences as shown in the left-hand part of Figure 1. They were instructed to mark the most prominent word and also to indicate on a 5-point rating scale how confident they were in their judgements.

As can be seen in Figure 1 (left part), the manipulations of the originally non-focused words resulted in stimuli with two f_0-markings of focus. Correspondingly, the manipulations of originally focused words resulted in sentences without any f_0-marking of focus. As f_0 was the only correlate manipulated, there were possibly both local and global cues conflicting with the manipulated f_0-contours. Therefore, in the second experiment the amount of global information was reduced. The same stimuli and setup as in the first experiment were used with the exception that only the part of the utterances up to and including the target words was presented as shown Figure 1 (right part).

Six subjects participated in each experiment. The 20 different stimuli in each experiment were presented 5 times to each subject resulting in a total of 30 judgements of each stimulus.

There seems to be at least two possible outcomes of the two experiments: If f_0 outweighs all other cues and the signaling is entirely local, the stimuli in the second experiment should be less ambiguous and thus easier to classify than those in the first because the amount of conflicting f_0-information is reduced. If, on the other hand, there are other important cues than f_0 and there is a non-local part of the signaling, the second experiment should be more difficult than the first.

3. Results

The results of the first and second experiments are given in Tables 2 and 3. The columns furthest to the left show the target word, whether it is originally focused or not, and the magnitude in semitones (ST) of the manipulation of the respective stimuli. The next two columns show the number of judgements for focused and non-focused respectively. The last two columns show means and standard deviations of the confidence ratings.

Table 2. First experiment: complete sentences. Judgements by six subjects of the target words in medial position. Number of focused and non-focused judgements and mean confidence ratings (1 low, 5 high rating of confidence) for each stimulus.

Stimuli	+F	–F	Confidence	
			Mean	Std. Dev.
'benämna' –F original	0	30	4.8	.6
'benämna' –F + 1 ST	0	30	4.6	.6
'benämna' –F + 2 ST	0	30	4.1	.9
'benämna' –F + 3 ST	2	28	3.9	1.3
'benämna' –F + 4 ST	11	19	3.6	1.1
'benämna' +F original	30	0	5.0	.0
'benämna' +F - 1 ST	30	0	4.8	.4
'benämna' +F - 2 ST	30	0	4.8	.6
'benämna' +F - 3 ST	30	0	4.9	.3
'benämna' +F - 4 ST	30	0	4.8	.5
'anmäla' –F original	2	28	4.7	.8
'anmäla' –F + 2 ST	0	30	4.5	.9
'anmäla' –F + 4 ST	2	28	4.2	.8
'anmäla' –F + 6 ST	5	25	3.7	1.3
'anmäla' –F + 8 ST	9	21	3.8	1.1
'anmäla' +F original	30	0	4.9	.3
'anmäla' +F - 2 ST	30	0	4.8	.4
'anmäla' +F - 4 ST	30	0	4.8	.8
'anmäla' +F - 6 ST	30	0	4.8	.4
'anmäla' +F - 8 ST	30	0	4.5	.8

Apparently, the subjects were not very sensitive to the f_0-manipulations occurring in the first experiment. The gradual removal of the f_0-rise in the stimuli based on focused target words produced no effects at all. The addition of a f_0-rise in the stimuli based on non-focused target words resulted in only a few ratings for focused. Examination of the confidence ratings supports these results.

Table 3. Second experiment: lasts part of sentences excluded. Judgements by six subjects of the target words in medial position. Number of focused and non-focused judgements and mean confidence ratings (1 low, 5 high rating of confidence) for each stimulus.

Stimuli	+F	–F	Confidence	
			Mean	Std. Dev.
'benämna' –F original	0	30	4.3	.7
'benämna' –F + 1 ST	1	29	3.7	1.0
'benämna' –F + 2 ST	7	23	3.4	.8
'benämna' –F + 3 ST	5	25	3.2	.8
'benämna' –F + 4 ST	16	14	3.3	.7
'benämna' +F original	30	0	4.4	.9
'benämna' +F - 1 ST	27	3	4.1	1.0
'benämna' +F - 2 ST	24	6	3.3	1.0
'benämna' +F - 3 ST	26	4	3.6	.9
'benämna' +F - 4 ST	15	15	3.5	.9
'anmäla' –F original	0	30	4.3	.9
'anmäla' –F + 2 ST	2	28	3.9	1.0
'anmäla' –F + 4 ST	7	23	3.7	.9
'anmäla' –F + 6 ST	5	25	3.8	1.0
'anmäla' –F + 8 ST	7	23	3.8	1.0
'anmäla' +F original	30	0	4.5	.8
'anmäla' +F - 2 ST	29	1	4.6	.7
'anmäla' +F - 4 ST	30	0	4.4	.8
'anmäla' +F - 6 ST	28	2	3.9	.8
'anmäla' +F - 8 ST	29	1	4.2	.8

The pattern found in the second experiment is more complex. Obviously, it is possible to judge whether a word is focused or not without having access to the entire utterance since the subjects perceived the original utterances 100% correctly. Moreover, the manipulations produced stronger effects in 'benämna' as compared to 'anmäla' and, also in non-focused words as compared to focused words. When there was a f_0-rise of 4 ST added to an originally non-focused 'benämna', the stimuli could just as well be perceived either as focused or as non-focused. Correspondingly, when the f_0-rise was reduced by 4 ST in an originally focused 'benämna', it might just as well be perceived as non-focused. Thus, in neither case was there a majority for one category; the f_0-manipulations only gave rise to ambiguity.

When the overall results of the two experiments are compared, stimuli seem to have been perceived as more ambiguous in the second experiment. Also, the confidence ratings were somewhat lower in the second experiment.

4. Conclusions

There were two main findings in the first experiment. First, focus can be perceived in the absence of a f_0-rise. Secondly, words with a f_0-rise can be perceived as non-focused.

Although these findings indicate that the f_0-rise is neither necessary nor sufficient to perceive a word as focused, it was believed that the results may have been influenced by the fact that some of the manipulated utterances contained f_0-signalling of focus in two positions, while others contained none. The second experiment was conducted to reduce the influence of these unnatural stimuli and thus to investigate the possible effects of contextual f_0-information. Although the manipulations produced stronger effects in this case, the same overall tendencies as in the first experiment were found, that is, focus can be perceived in the absence of a f_0-rise and words with a f_0-rise can be perceived as non-focused.

Both the first and the second experiments thus seem to indicate that the f_0-rise is neither necessary nor sufficient for the perception of focus. Accordingly, the findings by Strangert & Heldner (1994) and Strangert & Heldner (1995) are supported.

What is held to be the most important cue to perceived focus in Swedish (Bruce, 1977) and to accents in general (Beckman, 1986), that is f_0-movements, thus seem to be optional from the listener's point of view. Moreover, the different results in the two experiments suggest that focus may be perceived on the basis of prominence relations within a unit larger than the word. That is, it seems that there are non-local (global) components at play.

If this is the case, the signaling of focus cannot possibly be restricted to the local f_0-rise only. There must be other cues, probably both local and global ones, which affect the perception of focus.

Acknowledgements

This work was supported by a grant from the Swedish Council for Research in the Humanities and Social Sciences.

References

Beckman, M. E. (1986) *Stress and non-stress accent.* Dordrecht: Foris Publications.

Bruce, G. (1977) *Swedish word accents in sentence perspective.* Lund: CWK Gleerup.

Heldner, M. (1998) Is an f_0-rise a necessary or a sufficient cue to perceived focus in Swedish? In *Nordic Prosody: Proceedings of the VIIth Conference, Joensuu 1996* (S. Werner, ed.), pp. 109-125. Frankfurt am Main: Peter Lang.

Strangert, E. & Heldner, M. (1994) Prosodic labelling and acoustic data. In *Working Papers 43: Papers from the EIGHTH SWEDISH PHONETICS CONFERENCE*, pp. 120-123. Lund: Department of Linguistics and Phonetics, Lund University.

Strangert, E. & Heldner, M. (1995) The labelling of prominence in Swedish by phonetically experienced transcribers. In *Proceedings ICPhS 95* (K. Elenius & P. Branderud, eds.), pp. 204-207. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

Paper IV In *Focal accent – f_0 movements and beyond*
pp. 67–99

Temporal effects of focus in Swedish [1]

Mattias Heldner and Eva Strangert

The four experiments reported concern the amount and domain of lengthening associated with focal accents in Swedish. Word, syllable and segment durations were measured in read sentences with focus in different positions. As expected, words with focal accents were longer than nonfocal words in general, but the amount of lengthening varied greatly, primarily due to speaker differences but also to position in the phrase and the word accent distinction. Most of the lengthening occurred within the stressed syllable. An analysis of the internal structure of stressed syllables showed that the phonologically long segments –whether vowels or consonants– were lengthened most, while the phonologically short vowels were hardly affected at all. Through this nonlinear lengthening, the contrast between long and short vowels in stressed syllables was sharpened in focus. Thus, the domain of focal accent lengthening includes at least the stressed syllable. Also an unstressed syllable immediately to the right of the stressed one was lengthened in focus, while initial unstressed syllables, as well as unstressed syllables to the right of the first unstressed one, were not lengthened. Thus, we assume the domain of focal accent lengthening in Swedish to be restricted to the stressed syllable and the immediately following unstressed one.

1. Introduction

This article reports on experiments investigating aspects of the lengthening associated with focal accents in Swedish. More precisely, these experiments concern the amount and variability of focal accent lengthening, that is, the degree to which words and constituents within words such as segments and syllables are lengthened under the influence of focal accents. In addition, factors that influence the amount of lengthening such as position in the phrase and the Swedish word accent distinction are investigated. Finally, the domain of focal accent lengthening, that is, *what* is lengthened within the words carrying focal accents, is studied.

The notion of 'focus' is central in the work to be reported. It is used as in e.g., Ladd (1980), Gussenhoven (1984), Nooteboom & Kruyt (1987), and Ladd (1996). That is, single words or larger constituents in utterances (possibly also constituents smaller than the word, cf. van Heuven (1994)) can be put in focus by the speaker to indicate that they are new or otherwise informative to the listener. In many languages, focus is reflected in the prosodic signaling and considerable interest has been paid to acoustic correlates of focus over the years. In languages such as English, Dutch and Swedish, the primary correlate of focus is a

[1] This material has been published as Heldner, M. & Strangert, E. (2001) Temporal effects of focus in Swedish, *Journal of Phonetics,* **29**(3), pp. 329-361. Available online at http://www.idealibrary.com

pitch accent associated with the prosodic head of the focus domain. This pitch accent is usually also accompanied by other correlates such as longer durations within the word carrying the accent. Such temporal effects of focus are the main concern in the present article.

As far as languages such as English and Dutch are concerned, we shall refer to such effects as 'accentual lengthening', whereas for Swedish, 'focal accent lengthening' is a more appropriate term. The reason is that prominence models for Dutch and English usually distinguish two levels of prominence, stressed and accented, while in the prominence model developed by Bruce for Swedish (e.g., Bruce, 1977; Bruce, 1999) and used here, there are three phonologically distinct prominence levels, 'stressed', 'accented' and 'focused'. The extra level is needed to account for the Swedish word accents.

The word accent distinction is reflected in the tonal patterning. The accents are characterized by a tonal fall (an HL gesture), with a distinctively different timing of the fall relative to the stressed syllable, the H of accent I being timed earlier (HL*) than accent II (H*L). The tonal manifestation of the focal accent is an extra pitch rise to an H tonal point after the word accent HL (e.g., Bruce, 1999). However, as for the temporal patterning, Bruce (1981) found approximately the same durations for accent I and accent II words with focal accents and concluded that the same temporal program is used in focus for accent I and accent II.

Thus, 'accented' is used for the word accents only (accent I, or acute, and accent II, or grave), while 'focused' is used for words with a word accent combined with a focal accent. As a consequence, what is called 'accented' in the aforementioned models for English and Dutch is to be compared with what is called 'focused' in the Swedish model. When referring to Swedish data we will thus henceforth use the term 'focal accent lengthening' for what is called 'accentual lengthening' in other studies.

Although many studies have reported on lengthening as a correlate of accent, the observed increase in duration varies greatly between them. Factors such as word length and position in the phrase seem to play a role, but still the results are far from clear-cut. When different languages are involved – the most extensively studied languages are English and Dutch – the results often diverge. The influence of word length is but one example. Thus, in a study on Dutch, Eefting (1991) reports an average accentual lengthening of about 25% irrespective of word length, while Cooper, Eady, & Mueller (1985) find an inverse relationship between word length and relative increase in duration for English words in focus. For one- to three-syllable words they reported a lengthening of between 37% and 43%.

The figures given in these studies were obtained with focused words in phrase-medial position. One of them (Cooper *et al.*, 1985) also included phrase-final words in focus, and for this position considerably less lengthening was obtained – 14% to 17% – than for the words in medial position (37% to 43%, see above). In a study of tonal-temporal interplay in Swedish, Bruce (1981) presented segment durations for disyllabic words, occurring in medial and final position in the phrase, and with and without focal accents. Calculations (by us) using these data showed that one of his two speakers, similar to the American English speakers, lengthened focally accented words in medial position in the phrase more than those in final position. Phrase-medial words were lengthened by about 30% and phrase-final words by about 13%. However, the other speaker lengthened the words in phrase-medial position less than those in final position. Phrase-medial words were lengthened by 10% and the words in phrase-final position by 17%.

Thus, apart from the variation in the magnitude of lengthening, the studies by Bruce (1981) and by Cooper *et al.* (1985) also indicate that the amount of lengthening may be dependent on the word's position in the phrase. The opposite results for the two speakers in the Swedish study add to the complexity of the phenomenon of lengthening. Generally,

speaker variability seems to be far from negligible in the area of accentual/focal accent lengthening. However, neither the study of Swedish (Bruce, 1981) nor that of American English (Cooper *et al.*, 1985) were optimally designed for investigating effects of position in the phrase, as there were different words in the different positions. Therefore, what may have appeared to be effects of position in the phrase could also have been effects of the specific words used.

In an experiment explicitly designed to test for effects of phrase-position (similarly constructed material, same design and methodology) on the amount of accentual lengthening, Cambier-Langeveld (2000) found that accented Dutch words were lengthened significantly less in phrase-final position as compared to other positions. In RP-English words, on the other hand, there were no significant effects of phrase-position. Thus, she concluded that there was an interaction between final lengthening and accentual lengthening in Dutch, but not in RP-English. This difference is related to differences in "durational expandability" between English and Dutch, allowing longer durations in English due to a phonological rule of vowel lengthening before voiced obstruents in English, a rule lacking in Dutch. Thus, it is assumed that there is an interplay between lengthening due to focus and other language-specific phonological constraints.

Summarizing the results of the above-mentioned studies, the only consistent pattern is the existence of an increase in duration of words in focus. The variation concerns the amount of lengthening and also whether there is a dependence on the amount of lengthening of word-length as well as of position in the phrase. Also, variation may at a deeper level be related to language-specific phonologic characteristics.

There is also the issue of the domain of accentual lengthening, that is, *what* is lengthened in the focused word. Despite the prosodic similarities between Dutch and English, different domains have been proposed for the two languages. For Dutch, Eefting (1991) suggested the word. Subsequently, van Heuven (1993) tried to determine exactly what this word domain is. He concluded that the domain of accentual lengthening extends beyond the morpheme or the monomorphemic word and includes all segments in a compound word. Moreover, Sluijter & van Heuven (1995) and Sluijter & van Heuven (1996) claimed that accenting a particular word results in an almost linear time expansion of the entire word.

The word has also been proposed as the domain of accentual lengthening for English. Sluijter (1995) reported similar results for American English as for Dutch, that is, an accentual lengthening domain extending over the word as a whole and also including initial unstressed syllables. However, Turk & Sawusch (1997) also basing their study on American English, proposed a smaller linguistic unit. They found a relatively large amount of lengthening within the pitch accented syllable where all segments were lengthened, although with less lengthening on final consonants than on initial ones. In addition, they found that word-final, but not word-initial, unstressed syllables were lengthened in disyllabic accented words. Thus, they suggested that the domain begins with the pitch accented syllable and includes at least one following unstressed syllable within the word, while initial unstressed syllables are excluded from the domain.

Turk & White (1999) replicated and extended the findings of Turk & Sawusch (1997) using Scottish English, and found that the lengthening extended throughout all syllables in a trisyllabic word with primary stress on the initial syllable. Thus, they assumed a domain beginning with the pitch accented syllable and extending rightward until a word boundary. They moreover showed that some lengthening occurred in initial unstressed syllables, at least for some speakers. Nevertheless, initial unstressed syllables were not included in the domain. Furthermore, they found that the spread of accentual lengthening was merely attenuated by constituent boundaries such as the left edge of a pitch accented syllable and the left and right

edges of a word. Therefore, spread of accentual lengthening across word boundaries was possible.

As this short overview has shown, there has been some disagreement regarding the domain of accentual lengthening. However, a recent comparative study based on Dutch and English by Cambier-Langeveld & Turk (1999) suggests that the discrepancy between the studies for Dutch and English, and indeed for the different studies on English, may in fact be due to experimental artifacts. The discrepancies disappear when the effects are measured using the same methodology. Both languages have less lengthening on initial unstressed syllables than on the stressed syllable or a following unstressed syllable. Consequently, it is shown that the claim by Sluijter & van Heuven (1995) and Sluijter & van Heuven (1996) that accenting a particular word results in an almost linear time expansion of the entire word is untenable for Dutch as well as for English.

To our knowledge, there have been no explicit attempts to establish the domain of focal accent lengthening in Swedish. However, it has been shown that increased prominence is reflected in lengthening of at least the stressed syllable (Bruce, 1981; Fant, Kruckenberg & Nord, 1991) and also, in some cases, in lengthening of unstressed syllables following the stressed syllable (Bruce, 1981). Thus, these previous studies suggest that the domain of focal accent lengthening in Swedish is either the stressed syllable and the following unstressed syllable or the entire word.

A detail of the domain of lengthening is what happens within the stressed syllable in words with focal accents. Swedish is a quantity language and there is a distinction between long and short vowels within the stressed syllable. Also, there is a complementarity between the vowel and the following consonant. If the syllable is closed, the post-vocalic consonant is short, when the vowel is long (V:C). If the vowel is short, the consonant is long (VC:) or part of a consonant cluster (e.g., VCC) (Elert, 1964). The effect of variations of prominence on the temporal structure within the stressed syllable has previously been investigated by Bannert (1979) and by Fant *et al.* (1991). In both studies the long segment, that is the long vowel in a V:C-sequence as well as the long consonant in a VC:-sequence, was lengthened more than the short segment. In Bannert's (1979) interpretation, the specific lengthening pattern contributed to an enhancement of the quantity distinction in the stressed syllable in words in focus. However, Bannert (1979) examined only Accent II words, and the results were based on quite a small corpus. (Two speakers read 12 sentences each, where two test words occurred three times in narrow focus and three times out of focus.) Moreover, in the study by Fant *et al.* (1991), only two levels of prominence (unstressed and stressed in their terminology) were distinguished. As a consequence, the results for accented (i.e., words with accents I or II only) and focused words (i.e., words with word accents *and* focal accents) were mixed. This also had the consequence that no distinction was made between words with accents I and II. Also in this case, the corpus was relatively small (133 words with 72 stressed words read by one speaker).

Even though the domain of focal accent lengthening in Swedish maybe restricted to the focused word or part of it, the possibility that there may be temporal effects spanning longer stretches of speech, that is, more global effects, cannot be excluded. As mentioned above, small temporal effects to the right of the accented word have been observed in some studies on Dutch (Cambier-Langeveld & Turk, 1999) and Scottish English (Turk & White, 1999). Furthermore, more global effects of focal accents have been observed for other acoustic dimensions in Swedish, for example, compression of the pitch range following a focal accent (Bruce, 1977). From this perspective, it is reasonable to search for similar phenomena – effects on nonfocused portions of the utterance – also in the temporal manifestation of focus.

The present study relates to a project on acoustic correlates of focus in Swedish and their perceptual relevance. Within this larger project, aiming at a more thorough description of the complexity of focus signaling in Swedish, the purpose of the present study is to add to the incomplete picture of the temporal aspects of focusing. Compared to the research efforts spent on the tonal signaling of focus (e.g., Bruce, 1977; Bruce, 1999), the work devoted to temporal and other correlates of focus accent is very restricted, as we have attempted to show above. The results so far are restricted in scope, based on a limited number of speakers and a restricted material. Yet, temporal as well as other focus correlates deserve to be more thoroughly investigated. In addition to contributing to the description of the acoustic aspects of focus, a more detailed description of duration patterns is motivated for reasons having to do with perception. Though tonal cues without doubt play an important role for the perception of focus, they are not necessary cues. Perceptual experiments by Heldner & Strangert (1997) and by Heldner (1998) have shown that focus may be perceived even in the absence of the focal accent rise held to be the primary focus cue. In other words, other cues than tonal have the potential for signaling focus to the listener. Thus, the study of the temporal patterning should contribute to the further understanding of the perception of focus.

The experiments described in this paper have been designed to extend previous studies of the temporal aspects of focal accenting in Swedish (the studies referred to above and preliminary work by the present authors, e.g., Strangert & Heldner, 1998). We aim to quantify the amount and variability of focal accent lengthening, and also attempt to investigate whether, and to what extent, focal accent lengthening is influenced by factors such as the Swedish word accent distinction and position in the phrase. In addition, we want to specify the domain of accentual lengthening and examine the lengthening patterns within the stressed syllable.

To this end, we have measured temporal effects of focus at the word, syllable and segment levels, in a series of experiments. All experiments were based on meaningful (or near meaningful) read-aloud Swedish sentences in which the location of focus was systematically varied. However, though the main purpose in all experiments was to explore the amount and domain of focal accent lengthening, each one had additional, more specific goals. Thus, in Experiment 1, the lengthening of the focally accented words and the distribution of the lengthening between stressed and unstressed syllables were investigated. In addition, we wanted to verify the observation that the Swedish word accent distinction does not affect the amount of focal accent lengthening (Bruce, 1981). Experiment 2 was designed to study focal accent lengthening in different positions in the phrase. In addition, possible global effects of focal accents (that is, the effects on surrounding nonfocused portions of speech) were explored. Experiment 3 was designed to replicate and extend previous results obtained by Bannert (1979) and by Fant *et al.* (1991) by examining the part of the focal accent lengthening that occurs within the stressed syllable. The intra-syllable lengthening was examined with respect to the Swedish quantity and word accent distinctions. Finally, Experiment 4 was designed to study whether the domain of focal accent lengthening extends beyond the two-syllable words investigated in the previous experiments. To that end, unstressed syllables before and after the stressed syllable were examined in three-syllable words.

2. Experiment 1: Lengthening of words and syllables

2.1. Introduction

Experiment 1 was designed to study temporal effects of focal accents at the word- and syllable-level in disyllabic Swedish words. We wanted to investigate the amount and variability of focal accent lengthening, the distribution of the lengthening between stressed and unstressed syllables, and whether and to what extent focal accent lengthening is influenced by the Swedish word accent distinction. In addition, we aimed at a preliminary analysis of more global effects of focal accents, that is, adjustments of the temporal structure of surrounding nonfocal words as determined by position and distance relative to the focused word.

2.2. Material

The experiment was based on the two sentences shown in Table I. One had accent I words in all sentence positions (*Mannen tömmer dammen*) and the other accent II words (*Kvinnan dammar kannan*). All the words were two-syllable noncompounds, the vowels in the stressed syllables were short and followed by long consonants, and stress was always on the first syllable. For syllabification, the Swedish quantity distinction had to be considered. Thus, because of the compensatory relationship between the vowel and the following consonant in stressed syllables, the post-vocalic consonant was included in the stressed syllable in the syllabification of the words.

The sentences occurred as answers in a question-answer context. The questions were designed to elicit focal accents on each of the three words in turn. Thus, each sentence occurred in three versions, one focal and two nonfocal, as can be seen in Table II. The words in initial position in the sentences occurred in focal and in two pre-focal positions (i.e., one and two words before the focally accented word), the medial words occurred in post-focal, focal and pre-focal position, and the final words in focal and in two post-focal positions (i.e., one and two words after the focally accented word).

TABLE I. The sentences in Experiment 1 with translations and broad phonemic transcriptions including prosodic and syllabic structure.

Mannen ‘The man’	tömmer ‘is draining’	dammen ‘the pond’
/ ˈman.ən.	ˈtøm.ər.	ˈdam.ən. /
Kvinnan ‘The woman’	dammar ‘is dusting’	kannan ‘the jug’
/ ˈkvɪ̀n.an.	ˈdàm.ar.	ˈkàn.an. /

TABLE II. Questions and answers used in Experiment 1 together with translations. Capitals indicate focally accented words.

Vem tömmer dammen?	'Who is draining the pond?'
MANNEN tömmer dammen.	'The MAN is draining the pond.'
Vad gör mannen med dammen?	'What is the man doing with the pond?'
Mannen TÖMMER dammen.	'The man is DRAINING the pond.'
Vad tömmer mannen?	'What is the man draining?'
Mannen tömmer DAMMEN.	'The man is draining the POND.'
Vem dammar kannan?	'Who is dusting the jug?'
KVINNAN dammar kannan.	'The WOMAN is dusting the jug.'
Vad gör kvinnan med kannan?	'What is the woman doing with the jug?'
Kvinnan DAMMAR kannan.	'The woman is DUSTING the jug.'
Vad dammar kvinnan?	'What is the woman dusting?'
Kvinnan dammar KANNAN.	'The woman is dusting the JUG.'

2.3. Speakers

The speakers were five males and four females. They were all native speakers of Swedish, without any strong dialectal influence and without any known hearing or speaking disorders. They were not paid for their services. For reasons described below (Section 2.5), the recordings from two of the males and one of the females were discarded. Thus, only the results from six speakers will be presented.

2.4. Recording

The recordings took place in a sound-treated room using a high quality condenser microphone at a fixed distance from the speaker's mouth. The speakers were presented with one question-answer pair at a time on a computer screen and were instructed to read the answers appropriately according to the questions. A computer program handled the presentation of text on the screen as well as the recording. The answers were recorded onto hard disk (48 kHz, 16 bit).

At the recording session, the speakers initially read all test sentences once for practice. Then they repeated the material 11 times, yielding a total of 66 sentences per speaker (2 sentences x 3 focus conditions x 11 repetitions). The question-answer pairs were presented in random order. If, during the recording, the experimenter noticed that an answer did not have a focal accent at the prompted location, the question-answer pair was presented again on the screen.

2.5. Missing data

As the speakers were not explicitly instructed to read the answers as one phrase, some of them inserted a pause after the first word, when there was a focal accent on that word. Although this is a perfectly normal way of pronouncing these sentences, it still introduced unwanted variation in the material. Therefore, after the recording sessions, all sentences were listened to a second time by both authors and all the productions from the speakers inserting pauses after the first word were discarded, as well as other unsuccessful productions not

noticed during the recording. Thus, the recordings from two of the males and one of the females were discarded. Out of the remaining sentences, ten repetitions of each sentence by each speaker were selected randomly, and accordingly, a total of 360 sentences were submitted to further analyses.

2.6. Measurements and data analysis

All words in the finally selected sentences were included in the subsequent analysis. Segment boundaries were determined and labeled using ESPS/waves+^{TM}. Segmentation was guided by visual criteria, for example, abrupt changes in the amplitude and shape of successive glottal periods, and verified by auditory feedback. Word-, stressed syllable- and unstressed syllable-durations were calculated using the labeled segment boundaries. For syllabification criteria, see Section 2.2.

The data were analyzed in three separate "by item" MANOVAs: one for word durations, one for stressed syllable durations and one for unstressed syllable durations. There were three dependent variables in each design: durations in the first second and third words in the sentences. Apart from the different dependent variables, the three designs were identical. There were three Between-Subjects Factors: Focal accent with three levels (focal accented vs. two nonfocal conditions), Word accent with two levels (word accent I vs. accent II) and Speaker with six levels (Sp1-Sp6). All factors were treated as fixed as random effects are not permitted in MANOVAs in SPSS (1999). The main effect of Focal accent was analyzed using Bonferroni *post hoc* tests.

2.7. Results

Word duration data will be presented first. Fig. 1 shows the mean word durations (and the variability expressed as standard errors) for focally accented words and for the two nonfocal conditions for each word and each speaker. Several effects may readily be observed. First, regarding the differences between words with and without focal accents, it is clear from the diagrams that, in general, words with focal accents are considerably longer than those without. Second, comparing the nonfocal versions of each word, the effects of position (pre- or post-focal) and distance relative to the focally accented word are generally small. Third, the speaker differences are striking. Most speakers apparently lengthened the words with focal accents considerably, while others (speaker 5) hardly had any lengthening at all. Fourth, Fig. 1 suggests that there may be a minor effect of the Swedish word accents. It seems that the accent II words were lengthened somewhat more than the accent I words. This effect is most pronounced in the final words. Fifth, it is obvious that the words in phrase-final position, whether focused or not, are considerably longer than those in initial and medial position in the phrase. Nonfocal words in final position are often longer than focally accented words in other positions, and in the medial position in particular. All of these observations are supported by the outcome of the "by item" MANOVA for word duration as presented below.

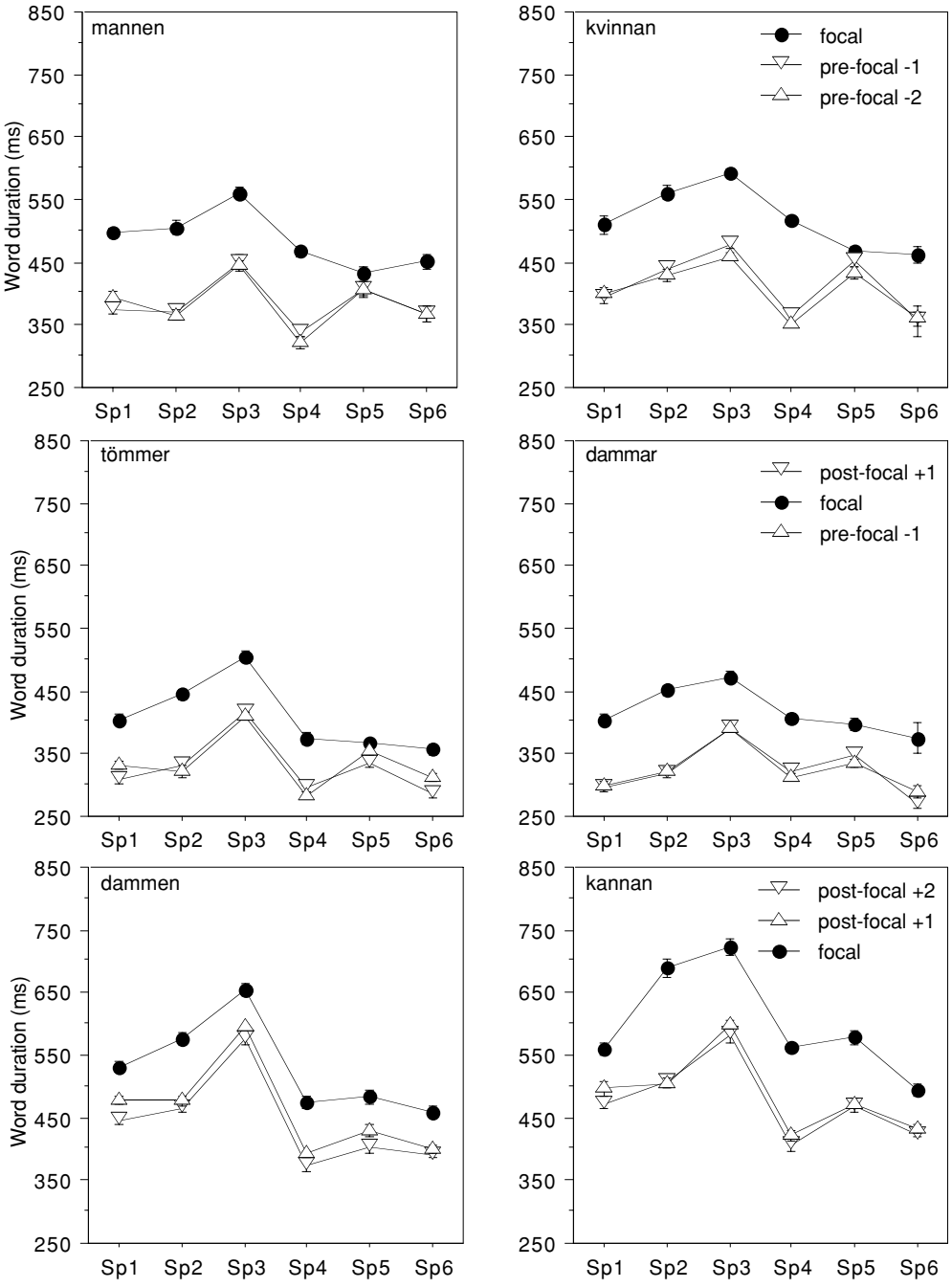

Figure 1. Word duration means (ms) and standard errors for the different focus conditions and speakers in Experiment 1. The top panels show the words in phrase-initial position, the middle panels the words in phrase-medial and the bottom panels the words in phrase-final position. The accent I words are shown to the left and the accent II words to the right.

The Multivariate tests (the test statistics reported in the multivariate tests is Pillai's Trace) in the MANOVA for word duration showed significant main effects of Focal accent ($F(6, 646)=611.98$; $p<0.01$), of Word accent ($F(3, 322)=99.33$; $p<0.01$) and of Speaker ($F(15, 972)=36.44$; $p<0.01$). Moreover, the two-way interactions of Focal accent and Word accent ($F(6, 646)=10.57$; $p<0.01$), of Focal accent and Speaker ($F(30, 972)=11.28$; $p<0.01$), and of Word accent and Speaker ($F(15, 972)=6.95$; $p<0.01$), were also significant. However, the three-way interaction of Focal accent, Word accent and Speaker was not significant ($F(30, 972)=1.36$; $p=0.09$). Also, the observed power for the three-way interaction was high (0.97).

Furthermore, the Tests of Between-Subjects Effects showed that the main effect of Focal accent was significant in all positions in the phrase ($F_{w1}(2, 324)=436.36$; $p<0.01$; $F_{w2}(2, 324)=441.69$; $p<0.01$; $F_{w3}(2, 324)=539.14$; $p<0.01$). The main effect of Word accent was significant in the initial and final but not in the medial words ($F_{w1}(1, 324)=68.17$; $p<0.01$; $F_{w2}(1, 324)=0.47$; $p=0.50$; $F_{w3}(1, 324)=231.25$; $p<0.01$), although the observed power for the medial word was low (0.11). The main effect of Speaker was also significant in all positions in the phrase ($F_{w1}(5, 324)=86.76$; $p<0.01$; $F_{w2}(5, 324)=148.50$; $p<0.01$; $F_{w3}(5, 324)=411.59$; $p<0.01$). Moreover, the interaction of Focal accent and Word accent was significant in the medial and final words only ($F_{w1}(2, 324)=0.87$; $p=0.42$; $F_{w2}(2, 324)=4.70$; $p=0.01$; $F_{w3}(2, 324)=24.64$; $p<0.01$). Also the two-way interactions involving speaker were significant in all the positions in the phrase in the tests of between-subjects effects: Focal accent and Speaker ($F_{w1}(10, 324)=12.53$; $p<0.01$; $F_{w2}(10, 324)=9.38$; $p<0.01$; $F_{w3}(10, 324)=8.75$; $p<0.01$); Word accent and Speaker ($F_{w1}(5, 324)=6.99$; $p<0.01$; $F_{w2}(5, 324)=8.87$; $p<0.01$; $F_{w3}(5, 324)=7.18$; $p<0.01$). As the three-way interaction was not significant in the multivariate tests, no results from the Tests of Between-Subjects Effects will be reported.

Bonferroni *post hoc* tests on the main effect of Focal accent were used to examine differences between the two nonfocal versions of each word. These tests showed that the two nonfocal levels were significantly different in the phrase-final words. The phrase-final words 'dammen' and 'kannan' immediately following the focally accented word were about 16 ms longer than when they occurred two words after the focally accented word. No significant differences between the nonfocal versions of each word were found for the words in initial and medial position and the mean differences were only about 4 ms for both words.

To be able to quantify the amount of lengthening of words with focal accents, a nonfocal reference had to be established. As the differences between the two nonfocused versions of each word were usually small (cf. Fig. 1) and not significant in two out of three positions in the sentence, the average of the nonfocused durations will be used in the following for assessing the magnitude of lengthening. Table III presents such averaged nonfocal reference durations (-F) for words as well as durations for focused (+F) words. The table also includes nonfocused reference durations for the stressed and unstressed syllables within the words, as well as durations for the same syllables when in focus. Table IV shows the amount of lengthening of the words, stressed syllables and unstressed syllables calculated on the basis of Table III. Table IV in addition includes ratios between the lengthening in the stressed syllable and the lengthening in the word.

The grand means across all words in Table IV show that the words with focal accents were on average 96 ms or 25% longer than nonfocal words. The different words vary in the amount of lengthening between 17% and 29 %. Though not explicitly shown in the table, there are also considerable speaker differences, as can be inferred from Fig. 1 (and confirmed by the statistical analysis). The average lengthening of the words as produced by the two most extreme speakers of the six participating in this experiment was 5% and 45%, respectively.

TABLE III. Means and standard deviations across all speakers for word-, stressed syllable-, and unstressed syllable-durations (in ms) for words with (+F) and without (–F) focal accents. –F here refers to averaged nonfocal durations.

	Word duration		Stressed syllable duration		Unstressed syllable duration	
	+F Mean (S.D.)	*–F* Mean (S.D.)	*+F* Mean (S.D.)	*–F* Mean (S.D.)	*+F* Mean (S.D.)	*–F* Mean (S.D.)
Mannen	486 (50)	384 (47)	315 (43)	243 (36)	171 (19)	141 (21)
tömmer	408 (58)	332 (46)	319 (41)	251 (32)	89 (22)	81 (21)
dammen	530 (74)	453 (72)	317 (41)	266 (43)	214 (43)	186 (40)
Kvinnan	519 (57)	409 (54)	324 (46)	239 (33)	195 (20)	170 (27)
dammar	418 (51)	324 (40)	288 (38)	209 (29)	130 (18)	116 (15)
kannan	601 (87)	482 (63)	344 (50)	269 (32)	257 (42)	213 (38)

TABLE IV. The lengthening of words in ms (ΔW) and as percentages (%W), stressed and unstressed syllables, also in ms (ΔS, ΔU) and as percentages (%S, %U) in words with focal accents compared to nonfocal words, and the stressed-syllable-to-word-lengthening ratio (%ΔS/ΔW) as a percentage.

	Δ*W* Mean	*%W* Mean	Δ*S* Mean	*%S* Mean	Δ*U* Mean	*%U* Mean	*%ΔS/ΔW* Mean
Mannen	102	27%	72	30%	30	21%	71%
tömmer	76	23%	68	27%	8	10%	89%
dammen	77	17%	51	19%	28	15%	66%
Kvinnan	110	27%	85	36%	25	15%	77%
dammar	94	29%	79	38%	14	12%	84%
kannan	119	25%	75	28%	44	21%	63%
Grand mean	96	25%	72	30%	25	16%	75%

The ratio – 75% on average – between the lengthening in the stressed syllable and the lengthening in the word, that is, the stressed-syllable-to-word-lengthening ratio, shows that both the stressed and the unstressed syllables contributed to the lengthening of the word. This was also verified by the outcome of the MANOVA models for stressed and unstressed syllable durations.

The Multivariate tests for stressed syllables showed that there were significant Focal accent differences ($F (6, 646)=708.13$; $p<0.01$), Word accent differences ($F (3, 322)=161.45$; $p<0.01$), and Speaker differences ($F (15, 972)=58.96$; $p<0.01$). Moreover, the interactions of Focal accent and Word accent ($F (6, 646)=9.00$; $p<0.01$), Focal accent and Speaker ($F (30, 972)=10.36$; $p<0.01$), Word accent and Speaker ($F (15, 972)=27.50$; $p<0.01$), and Focal accent and Word accent and Speaker ($F (30, 972)=4.58$; $p<0,01$) were all significant.

The Tests of Between-Subject Effects showed that the interactions of Focal accent and Word accent ($F_{w1} (2, 324)=2.38$; $p=0.09$; $F_{w2} (2, 324)=10.38$; $p<0.01$; $F_{w3} (2, 324)=9.71$; $p<0.01$), as well as of Focal accent and Word accent and Speaker ($F_{w1} (10, 324)=1.52$; $p=0.13$; $F_{w2} (10, 324)=2.56$; $p<0.01$; $F_{w3} (10, 324)=9.47$; $p<0.01$), were significant in the medial and final words only. All of the other effects that reached significance in the Multivariate tests were also significant ($p<0.01$) in all positions in the phrase in the Tests of

Between-Subject Effects. More detailed results from this test will not be reported for these effects.

A Bonferroni *post hoc* test showed that there were significant differences between the nonfocal productions in initial position (i.e., one and two words before the focal accented) and in final position (i.e., one and two words after the focal accented). However, there were no significant differences between pre-focal and post-focal words in medial position in the sentence. The stressed syllable in the initial word was on average 9 ms shorter, when there was a focal accent on the medial as compared to on the final word. The stressed syllable in the final word was on average 8 ms longer when there was a focal accent on the medial as compared to the initial word.

Significant speaker differences and a significant contribution to the lengthening of the word were also found in the MANOVA for unstressed syllable duration. Again, the Multivariate tests showed significant effects of Focal accent ($F(6, 646)=103.50$; $p<0.01$), Word accent ($F(3, 322)=111.96$; $p<0.01$), and Speaker ($F(15, 972)=46.16$; $p<0.01$). In addition, the interactions between Focal accent and Word accent ($F(6, 646)=12.62$; $p<0.01$), Focal accent and Speaker ($F(30, 972)=5.18$; $p<0.01$), Word accent and Speaker ($F(15, 972)=35.37$; $p<0.01$), and Focal accent Word accent and Speaker ($F(30, 972)=2.77$; $p<0.01$) were significant.

Again, the Bonferroni *post hoc* test showed that there were significant differences between the nonfocal productions in initial and final position in the sentence but not in medial position.

Summarizing, as shown by the grand mean of the stressed-syllable-to-word-lengthening ratio in Table IV, most of the lengthening of the word (75% on average) occurred in the stressed syllable. Moreover, the stressed syllables were lengthened relatively more than the unstressed syllables and therefore more than the words as a whole.

2.8. Discussion

As expected, Experiment 1 shows that words with focal accents are longer than nonfocal words. The grand mean of the amount of focal accent lengthening of words in Experiment 1 is 25%. This figure is close to the results obtained for Dutch (Eefting, 1991). However, as the amount of lengthening seems to be dependent on several factors, such as speaker, word accent and position in the sentence (see below), we do not believe that a general figure expressed as a percentage is an appropriate description of accentual lengthening, at least not for Swedish.

One factor that determines the amount of lengthening is the speaker. Some speakers apparently lengthen words in focus more than others, and it appears that lengthening may even be an optional correlate of focal accent, as one speaker hardly lengthened words with focal accents at all. Another factor is whether the word is an accent I or an accent II word, as the accent II words are lengthened somewhat more than the accent I words. Thus, this experiment indicates that the Swedish word accent distinction may affect the amount of focal accent lengthening, a result which is contrary to that obtained by Bruce (1981), who observed the same lengthening pattern for focally accented words irrespective of the type of word accent.

Another factor that might affect the lengthening is position in the sentence, as the initial and final words were usually longer than the medial ones. However, even though the words in Experiment 1 were similar with respect to characteristics such as number of syllables and consonants in the rhyme of the stressed syllable, what appears to be an effect of position in the sentence could also be an effect of the actual words used, as there were different words in the different positions in the sentence. Therefore, another experiment with the same lexical

item in different positions in the sentence would be needed to test positional effects thoroughly.

Furthermore, the usually small differences between the two nonfocal versions of each word indicate that the effects of focal accents on the durations of surrounding nonfocal words are weak. There were no tendencies to leftward effects of focal accent lengthening across word boundaries. Rightward effects across word boundaries were found in nonfocused final words only, and the experiment failed to show any rightward effects in medial words. Thus, it seems that position and distance relative to the focally accented word are relatively unimportant in determining the duration of nonfocal words. It should be noticed, however, that when such effects occurred, they appeared *after* the focused words.

Moreover, this experiment has shown that most of the lengthening of words with focal accents occurs in the stressed syllable, 75% on average. Also, since the stressed syllables are lengthened more than the unstressed, and more than the word as a whole, focal accent lengthening in Swedish cannot be described as a linear time expansion of the entire word. Thus, the results for Swedish syllable durations run counter to the earlier descriptions of Dutch and English by Sluijter & van Heuven (1995) and Sluijter & van Heuven (1996) while they agree with the more recent study by Cambier-Langeveld & Turk (1999).

3. Experiment 2: Lengthening of words in different positions in the phrase

3.1. Introduction

Though data were obtained for words in different positions in the phrase in Experiment 1, positional effects of lengthening could not be substantiated due to shortcomings in experimental control; there were different words in the different positions. Nevertheless, the results of Experiment 1 indicate that there *may* be positional effects.

First, whether focused or not, the words in phrase-final position were observed to be longer than the others. These long durations may be ascribed at least partly to the phenomenon of final lengthening (e.g., Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992; Horne, Strangert, & Heldner, 1995).

Second and of main concern here, also the focal accent lengthening appeared to be dependent on position. These positional effects were observed in the focused word *and* the immediately following word. Thus, we concluded that the results of Experiment 1 did not exclude the possibility of more global effects of focus and that position and distance relative to the focused word might play a role in that.

The lengthening affecting the words *in focus* revealed differences between the positions in the phrase with the initial words being lengthened more than the medial words. In final position the situation was rather complex, with the accent I word being lengthened less and the accent II word being lengthened more than in other positions (at least when measured in absolute figures). Similar results of less lengthening in final position were obtained for Dutch in a study by Cambier-Langeveld (2000) and for American English by Cooper *et al.* (1985), while other experiments, as referred to in the Introduction, show either insignificant effects of position (Cambier-Langeveld, 2000, for RP-English), or unstable patterns (Bruce (1981, for Swedish).

Thus, the results so far point in very different directions; there seems to be no general trend as far as the previous studies are concerned. This, in combination with possible positional dependencies in Experiment 1, led us to an experiment especially designed to study such effects. To that end we constructed material where the same words occurred in different positions in the phrase.

3.2. Material

Six phrases were selected where the same three content words (*mannen* 'the man', *kvinnan* 'the woman', *barnen* 'the children') separated by *och* ('and') occurred in initial, medial and final position in the phrase (e.g., *Mannen och kvinnan och barnen* 'The man and the woman and the children'). The content word *barnen* was included only to allow *mannen* and *kvinnan* to occur in three phrase positions. Each of the phrases occurred in three focus conditions, one with focal accent on the initial content word in the phrase, one with focal accent on the medial, and one with focal accent on the final content word. This yielded a total of 18 test phrases (6 phrases x 3 focus conditions). The test material is listed in Table V.

Since all the content words occurred with focal accents (as in Experiment 1), there was one focal accented and two nonfocal versions of each word in each phrase. The initial words occurred in focal and in two pre-focal positions (one or two words before the focally accented word), the medial words occurred in post-focal, focal and pre-focal position, and the final words in focal and in two post-focal positions (one or two words after the focally accented word).

TABLE V. Test material used in Experiment 2. Capitals indicate focally accented words.

MANNEN och kvinnan och barnen	KVINNAN och barnen och mannen
Mannen och KVINNAN och barnen	Kvinnan och BARNEN och mannen
Mannen och kvinnan och BARNEN	Kvinnan och barnen och MANNEN
MANNEN och barnen och kvinnan	BARNEN och mannen och kvinnan
Mannen och BARNEN och kvinnan	Barnen och MANNEN och kvinnan
Mannen och barnen och KVINNAN	Barnen och mannen och KVINNAN
KVINNAN och mannen och barnen	BARNEN och kvinnan och mannen
Kvinnan och MANNEN och barnen	Barnen och KVINNAN och mannen
Kvinnan och mannen och BARNEN	Barnen och kvinnan och MANNEN

3.3. Speakers

The speakers were three males and one female. They were all native speakers of Swedish without any strong dialectal influence and without any known hearing or speaking disorders. They were not paid for their services.

3.4. Recording

In general, the procedures were the same as in Experiment 1. However, focally accented words were not elicited by questions. Instead they were indicated with capital letters (as shown in Table V above) and the speakers were instructed to emphasize the words with capitalization. In addition, to avoid phrase internal pauses, the speakers were explicitly asked to read the phrases without any pauses.

The speakers initially read all 18 phrases once for practice. Then they repeated the material 5 times, yielding a total of 360 phrases (18 phrases x 5 repetitions x 4 speakers). The phrases were presented in random order.

The speaker's productions were monitored by the speakers themselves and by the experimenter. Either the speakers or the experimenter could decide whether a phrase should be reread. In addition, both authors listened to all phrases after the recording sessions to

eliminate erroneous readings, but since no unsuccessful productions were found, all the 360 productions were used in the following analysis.

3.5. Measurements and data analysis

The analysis was restricted to the two words *mannen* and *kvinnan* occurring in initial, medial and final position, respectively, and in and out of focus. Word durations were calculated as in Experiment 1.

First, since there was one focal accented and two nonfocal versions of each word in each phrase, we wanted to investigate whether there were differences between these nonfocal versions. Significant differences should force us to keep them apart, while insignificant differences would allow us to collapse them into one nonfocused category. Therefore, in initial position, the nonfocal words one or two words before the focally accented word were compared. In medial position, post-focal and pre-focal words were compared and in final position, post-focal words one and two words after the focally accented word were compared. The differences between these respective conditions were tested in two ANOVAs, one for each word (*mannen* and *kvinnan*). The dependent variables were the word durations of *mannen* and *kvinnan*. The independent variables were Focal condition nested under Position in the phrase with three levels in each position and Speaker with four levels. Focal condition nested under Position in the phrase was treated as a fixed and Speaker as a random factor. Planned comparisons were then performed to examine differences between the nonfocal conditions in each position.

Second, in order to determine whether position in the phrase affected word duration and the amount of focal accent lengthening, another two ANOVAs were run. The dependent variables were the word durations of *mannen* and *kvinnan*, respectively. Speaker was treated as a random factor (with four levels). Position in the phrase (initial vs. medial vs. final) and Focal accent were treated as fixed factors. We made the number of levels of the factor Focal accent dependent on the outcome of the analysis of the nonfocal conditions as described above. Thus, with the two nonfocal conditions collapsed, a distinction between 'focal' and 'nonfocal' would suffice with 'nonfocal' including positions one as well as two before the focused word (initial position), post-focal as well as pre-focal words (in medial position), and (in final position) word one as well as word two after the focused word. With the alternative outcome of the analysis, we would have to distinguish between the two nonfocused conditions resulting in three levels of the Focal accent factor.

3.6. Results

First, the results of the comparisons of the two nonfocal versions of each word in each position will be reported. Table VI shows that the mean differences between the nonfocal versions of each word were generally small. However, there was a trend revealing that the words immediately to the right of the focused word were longer than those before, or two words after, the focused word.

The planned comparisons did not reveal any significant differences between the two nonfocal versions of each word in any of the positions ($F_{mannen,\ initial}(1)=0.1$; $p=0.76$; $F_{mannen,\ medial}(1)=0.4$; $p=0.56$; $F_{mannen,\ final}(1)=0.2$; $p=0.69$; $F_{kvinnan,\ initial}(1)<0.1$; $p=0.98$; $F_{kvinnan,\ medial}(1)=0.1$; $p=0.77$; $F_{kvinnan,\ final}(1)<0.1$; $p=0.88$). The small and nonsignificant differences justified collapsing the nonfocal words into one category in the following analyses. Thus, 'nonfocused' henceforth refers to this collapsed category and 'nonfocused duration' to an average of the durations of the two nonfocused conditions for each position.

The mean word durations of focused and nonfocused *mannen* and *kvinnan* in different positions in the phrase are presented in Fig. 2. The amount of focal accent lengthening (calculated as a percentage on the basis of the difference between the focused and average nonfocused word duration, as in Experiment 1) in the different positions is listed in Table VII.

As in Experiment 1, words with focal accents were longer than the nonfocal versions of the same words. In this experiment, the focally accented words were on average 12% longer than the nonfocal words.

TABLE VI. The mean differences (in ms) between the durations of the nonfocal versions of each word in each position (ΔW) and the lengthening/shortening as percentages (%W).

	Mannen		Kvinnan	
	ΔW	*%W*	*ΔW*	*%W*
Initial position (one word before the focused vs. two words before)	-7	-2%	-1	<-1%
Medial position (one word after the focused vs. one word before)	14	4%	8	2%
Final position (one word after the focused vs. two words after)	9	2%	4	<1%

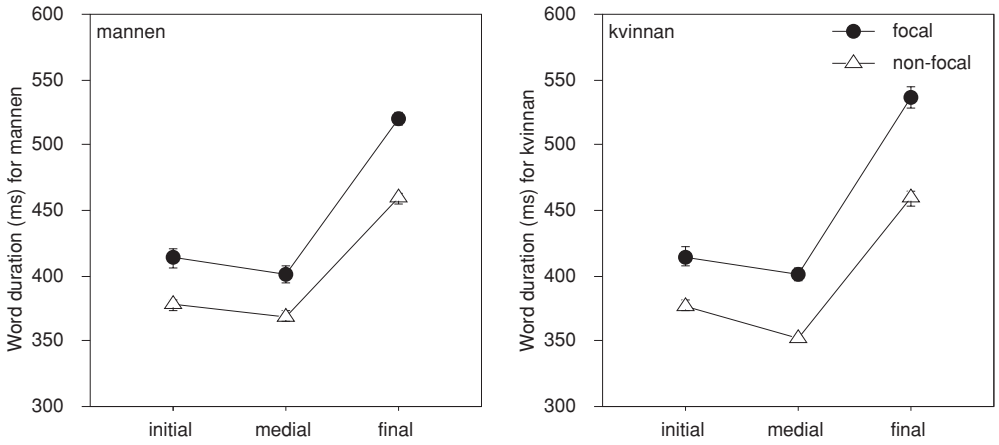

Figure 2. Word duration means (ms) and standard errors for focal and nonfocal words in three different positions in the phrase in Experiment 2. The test word 'mannen' is shown to the left, 'kvinnan' to the right.

TABLE VII. The amount of focal accent lengthening (in ms and as percentages) for the test words 'Mannen' and 'Kvinnan' in different positions in the phrase.

	ΔW	*%W*
Mannen initial position	36	10%
Mannen medial position	32	9%
Mannen final position	61	13%
Kvinnan initial position	38	10%
Kvinnan medial position	48	14%
Kvinnan final position	77	17%

In addition, there were effects of position in the phrase. First - whether focused or not - each respective word was considerably longer in final position than in either the initial or medial position. For example, the nonfocal words in final position were 90 ms *(mannen)* and 106 ms *(kvinnan)* longer than the same nonfocal words in medial position. Second, the focally accented words in final position were lengthened more than the words in initial and medial position. For example, *kvinnan* was lengthened by 77 ms or 17% in final position, by 38 ms or 10% in initial position, and by 48 ms or 14% in medial position. The details of the ANOVAs are presented below.

The main effects of Focal accent (F_{mannen} (1, 3)= 18.65; p=0.02; $F_{kvinnan}$ (1, 3)=20.16; p=0.02) were significant in both analyses showing that the focally accented words were significantly longer than the nonfocal words. In addition, both analyses showed significant main effects of Position in the phrase (F_{mannen} (2, 6)= 11.05; p=0.01; $F_{kvinnan}$ (2, 6)=11.00; p=0.01). Bonferroni *post hoc* tests on the main effect of Position in the phrase showed that the observed means were significantly different in all three positions in the phrase. In other words, the final words were significantly longer than the words in initial and medial position and the initial words were significantly longer than the medial words. However, the differences between initial and medial position were only a fraction (10 ms for *mannen* and 21 ms for *kvinnan*) of those found between, for example, medial and final position (100 ms for *mannen* and 116 ms for *kvinnan*). The main effect of Speaker was not significant (F_{mannen} (3, 7.48)= 0.45; p=0.73; $F_{kvinnan}$ (3, 8.28)=0.47; p=0.71).

The two-way interaction of Focal accent and Position in the phrase failed to show significant differences in the amount of lengthening in the different positions in the phrase for *mannen,* whereas there were such differences for *kvinnan* (F_{mannen} (2, 6)=3.15; p=0.12; $F_{kvinnan}$ (2, 6)=14.98; p<0.01). However, it should be noted that the power of the test was low for *mannen* (0.40) and that there were trends in the same direction for both words.

Moreover, there were apparent speaker differences. Several interactions involving the factor Speaker were significant. The interaction of Focal accent and Speaker (F_{mannen} (3, 6)=3.99; p=0.07; $F_{kvinnan}$ (3, 6)=15.79; p<0.01) showed significant differences in the analysis of *kvinnan* but not in the analysis of *mannen*. The interaction of Position in the phrase and Speaker was significant in both analyses (F_{mannen} (6, 6)=16.28; p<0.01; $F_{kvinnan}$ (6, 6)=55.64; p<0.01). Finally, the three-way interaction of Position in the phrase, Focal accent and Speaker was significant in the analysis of *mannen* but not in *kvinnan*. (F_{mannen} (6, 336)=3.08; p<0.01; $F_{kvinnan}$ (6, 336)=1.72; p=0.12).

3.7. Discussion

The first part of this experiment showed that the distance and the position relative to the focally accented word has little influence on the duration of nonfocal words. The differences between the nonfocal versions of each word in each position were generally small and insignificant. However, there was a trend that the words immediately to the right of the focused word were longer than those before or two words after the focused word. The same tendencies were also observed in Experiment 1. Thus, there might be a small rightward effect of focal accents across word boundaries in Swedish similar to that reported for English by Turk & White (1999). They showed monosyllabic post-focal words in medial position in the phrase to be significantly different (5% longer) from pre-focal words in the same position. Possibly, given a larger material, mean differences in the magnitudes observed in our experiment could also be shown to be statistically significant.

Still, it is questionable whether such small differences, though significant, have any relevance as a cue to focus for the listener. Rather, given these results, we do not believe that there are perceptually relevant temporal effects of focal accents in Swedish spanning longer stretches of speech than the focused word itself. That is, there seems to be no reason to take more global temporal effects of focus into account, at least not as far as Swedish is concerned. We will come back to this issue in 6.6 in the general discussion.

The second part of the experiment investigated the effects of position in the phrase on the duration of focal and nonfocal words as well as on the amount of focal accent lengthening. The longer word durations in final position in the phrase compared to initial and medial position obtained in this experiment show that there is final lengthening in nonfocal as well as in focally accented words in Swedish (as has previously been shown by Horne *et al.*, 1995). In addition it was found that the initial words were slightly longer than the medial. Thus, word duration, whether in or out of focus, was clearly affected by position.

Turning now to our main concern, the focal accent lengthening, the average found in this experiment was 12%. This figure is only about half of the average lengthening in Experiment 1. We ascribe this difference between the two experiments to speaker variability as observed in Experiment 1, where the spread around the average of 25% was great with extremes between 5% and 45% lengthening for the different speakers. Speaker differences are also reflected in the significant interactions of Speaker and Focal accent, as well as Speaker and Position.

The influence of position in the phrase on the amount of focal accent lengthening in this experiment is far from clear. There is a trend indicating that words in phrase-final position are lengthened more than in other positions, but this trend reached statistical significance only in one of the test words. Altogether, however, our data indicate that focal accented Swedish words in phrase final position are longer than the same words in nonfinal and nonfocal position for several reasons. First, they are longer because they are focal accented. Second, they are longer because they are in phrase final position, that is, a position where there is final lengthening. Third, they are longer because they are focal accented *and* in phrase final position.

These results are in conflict with those from other studies; the observed tendency to an interaction of focal accent lengthening and final lengthening in Swedish is opposite to that reported for American English by Cooper *et al.* (1985) and for Dutch by Cambier-Langeveld (2000). In both these studies, there was less lengthening in final position than in others. Our results also disagree with those for RP-English reported in Cambier-Langeveld (2000), where no interaction at all was found. Rather, in the RP-English study, accentuation was additive, contributing about the same amount of durational increase in all positions of the phrase. (Relatively, however, there was also less lengthening in final position in RP-English,

as nonfocused duration was longer finally than in the other positions.) Although we lack an explanation for these divergent results, it should be noted that the various studies are not easily comparable in all aspects using different methodologies and reporting results in different ways (for example, lengthening measured in absolute vs. relative terms). The study by Cambier-Langeveld (2000) using the same methodology and the same kind of material for RP-English and Dutch is an exception. Also, when studying different languages, the material used to study lengthening cannot be the same and, sometimes, not even similar. Thus, the observed variation between the different studies, though based on different languages, need not necessarily reflect language differences, but could just as well be ascribed to other factors. Also the great variability between speakers, at least for Swedish, is illustrated by the various results of position dependencies of our speakers as well as the Swedish speakers in the study by Bruce (1981).

Finally, this experiment indicates that the phrase-medial position may be the most suitable position for target words in order to study temporal effects of focal accents, as it is least affected by factors other than the focal accents, for example final lengthening. This is also the procedure we will adopt in the following experiments.

4. Experiment 3: Lengthening within the stressed syllable

4.1. Introduction

From Experiment 1 it was clear that most of the focal accent lengthening occurred in the stressed syllable. However, the lengthening patterns *within* the stressed syllable were not examined. In order to understand the distribution of lengthening, syllable-internal conditions also have to be taken into account. This is especially true for stressed syllables in Swedish where, due to the quantity distinction, one should not a priori expect a linear time expansion over the entire syllable.

Vowels in stressed syllables in Swedish are either long or short. Also, the length of the consonants following the vowels is complementary to the length of the vowels, that is, if the vowel is short, the consonant is long and vice-versa. Given this, we might not exclude specific adjustments in order to maintain, or even enhance, the long-short distinction of vowels and consonants in focus position (cf. Bannert, 1979). Thus, the material in Experiment 1 was too restricted as the stressed syllables only contained short vowels followed by long consonants. Therefore, Experiment 3 was designed to examine the lengthening patterns within stressed syllables with long vowels followed by short consonants as well as short vowels followed by long consonants.

4.2. Material

The material consisted of 40 sentences, where 20 accent I and 20 accent II transitive verbs with stress on the first syllable occurred in medial position in the sentence. Half the verbs contained a short vowel and half of them a long one. Owing to the complementary relationship between the vowel and the following consonant in Swedish, the syllable internal structure was either a short vowel followed by a long consonant (VC:) or a long vowel followed by a short consonant (V:C) within the stressed syllable. Also, the stressed vowel was preceded by either one or two consonants. Except for these restrictions, the verbs were chosen so as to include a great variety of vowels and consonants in order to avoid unwanted systematic influences on the results. The accent I verbs were embedded in the sentence frame *Mannen 'The man' {Verb} kvinnan 'the woman'* and the accent II verbs were placed in the

similar frame: *Kvinnan 'The woman' {Verb} mannen 'the man*'. Two examples are: *Mannen biter kvinnan* 'The man bites the woman' and *Kvinnan kammar mannen* 'The woman combs the man'. Focal accents were elicited on each of the three words in each of the 40 sentences through the same kind of questions as in Experiment 1, giving a total of 120 different question-answer pairs. The complete set of question-answer pairs used can be found in Appendix A.

4.3. Speakers

Four speakers, two males and two females, read the test material. They were all native speakers of Swedish without any strong dialectal influence and without any known hearing or speaking disorders. They were not paid for their services.

4.4. Recording

In the recording sessions, the speakers initially read a few of the test sentences for practice. Then each speaker read each sentence once, giving a total of 480 productions (120 test sentences x 1 repetition x 4 speakers). Due to technical problems and unsuccessful productions not discovered during the recording, three tokens were excluded from the analyses.

4.5. Measurements and data analysis

Only the verbs were included in the subsequent analyses and instead of calculating word durations, the durations of the segments within the stressed syllable were measured. As in the previous experiments, there were two nonfocal versions of each word, a pre-focal and a post-focal version. As the differences between the two nonfocal versions were usually small and statistically insignificant in the previous experiments, the nonfocal versions were collapsed into one category (cf. 3.6). Apart from this, the procedures were identical to those of Experiment 1.

In order to determine how the lengthening of words with focal accents was distributed within the stressed syllables, a MANOVA was used. The dependent variables were (i) the duration of the consonant (or consonants) before the vowel (C(C)), (ii) the duration of the vowel (V), and (iii) the duration of the consonant following the vowel (C), that is, all parts of the stressed syllable. The independent variables were Focal accent (focal vs. nonfocal), Word accent (accent I vs. accent II), Vowel length (long vs. short) and Speaker (four levels). All independent variables were fixed, as random variables are not permitted in MANOVAs in SPSS (1999).

4.6. Results

The mean segment durations within the stressed syllables for the focal and nonfocal conditions across all words and speakers are shown in Fig. 3. The influence of focal accent is apparent (and indeed, statistically significant in most segments, see below). Going into detail, the lengthening in the words with long vowels followed by short consonants (left panels in Fig. 3) affected all segments, but especially the consonants preceding the vowel and the long vowels. C(C) was lengthened by 28 ms or 31%, V by 35 ms or 24%, and C by 20 ms or 21%, on average. However, in the words with short vowels followed by long

consonants (right panels in Fig. 3), almost all of the lengthening occurred in the consonants preceding and following the vowel. C(C) was lengthened by 29 ms or 30% and C by 39 ms or 35%, on average. The lengthening in the short vowels was only 5 ms, or about 5%. Thus, all segments in stressed syllables were affected by a focal accent. In addition, there were differences between words with accent I and II, though much smaller than the differences induced by focal accent. Nevertheless, word accent showed up beside focal accent as a significant main effect in the statistical analysis, as did speaker and vowel length. The details of the analysis are presented below.

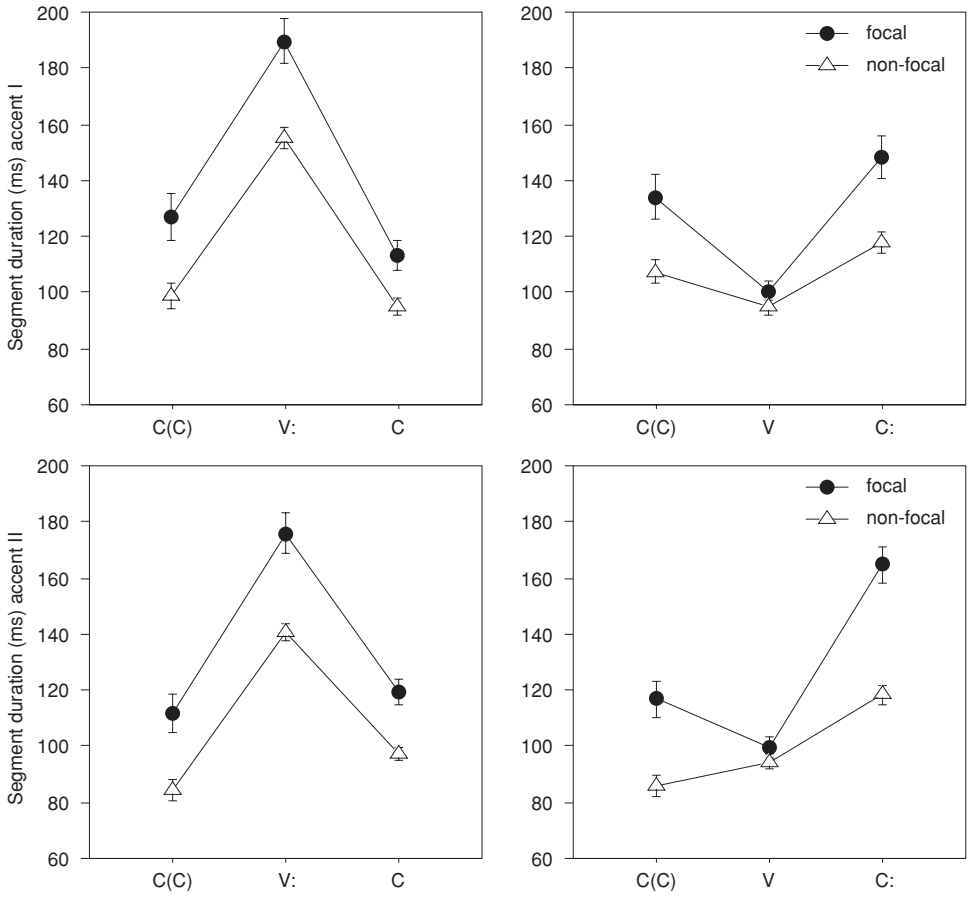

Figure 3. Mean durations (ms) and standard errors for vowels and consonants within the stressed syllable for focal and nonfocal words. Syllables with long vowels followed by short consonants are shown to the left and syllables with short vowels and long consonants to the right. Accent I words are shown in the top panels and accent II words in the bottom panels.

The Multivariate tests in the MANOVA (the test statistics used in the multivariate tests is Pillai's Trace) showed significant main effects of Focal accent ($F(3, 443)=136.01$; $p<0.01$), Word accent ($F(3, 443)=12.63$; $p<0.01$), Vowel length ($F(3, 443)=315.19$; $p<0.01$), and Speaker ($F(9, 1335)=48.24$; $p<0.01$). Furthermore, the two-way interactions of Focal accent and Vowel length ($F(3, 443)=15.93$; $p<0.01$), Word accent and Vowel length ($F(3, 443)=3.61$; $p=0.01$), Focal accent and Speaker ($F(9, 1335)=8.43$; $p<0.01$), and, Vowel length and Speaker ($F(9, 1335)=4.10$; $p<0.01$) were all significant. Neither of the other two-way

interactions (Focal accent and Word accent; Word accent and Speaker), three-way interactions (Focal accent, Word accent and Vowel length; Focal accent, Word accent and Speaker; Focal accent, Vowel length and Speaker; Word accent, Vowel length and Speaker), nor the four-way interaction (Focal accent, Word accent,Vowel length and Speaker), were significant. Hence, we will not report any results for these effects in the Tests of Between-Subjects Effects, that is, the tests for each segment separately.

The Tests of Between-Subjects Effects showed that the effect of Focal accent was significant in all three segments ($F_{C(C)}$ (1, 445)=56.52; p<0.01; F_V (1, 445)=78.31; p<0.01; F_C (1, 445)=116.42; p<0.01). In other words, all the segments in the stressed syllable were affected by a focal accent. Also Word accent ($F_{C(C)}$ (1, 445)=20.01; p<0.01; F_V (1, 445)=10.20; p<0.01; F_C (1, 445)=5.85; p=0.02) and Speaker ($F_{C(C)}$ (3, 445)=15.01; p<0.01; F_V (3, 445)=119.33; p<0.01; F_C (3, 445)=49.04; p<0.01) had significant effects in all three segments in the stressed syllable. However, the effect of Vowel length was significant in the vowel and in the following consonant only ($F_{C(C)}$ (1, 445)=2.01; p=0.15; F_V (1, 445)=905.23; p<0.01; F_C (1, 445)=131.37; p<0.01).

Furthermore, the effect of Focal accent was significantly different in the long vowels as compared to in the short ones as well as in the following short and long consonants, while there were no significant differences on the consonant or consonants preceding the vowel. This was shown by the fact that the interaction of Focal accent and Vowel length was significant in V and C but not in C(C) ($F_{C(C)}$ (1, 445)<0.01; p=0.93; F_V (1, 445)=43.01; p<0.01; F_C (1, 445)=11.18; p=0.01). Furthermore, planned comparisons showed that Focal accent had a significant effect on all segments in the syllables with long vowels ($F_{C(C)}$ (1)=27.5; p<0.01; F_V (1)=118.5; p<0.01; F_C (1)=27.9; p<0.01) but only on the consonants in the syllables with short vowels ($F_{C(C)}$ (1)=29.0; p<0.01; F_V (1)=2.6; p=0.11; F_C (1)=100.1; p<0.01). In other words, the short vowel was not significantly affected by focal accents. The effect of Word accent was significantly different in the long and short vowels as the interaction of Word accent and Vowel length was significant in V only ($F_{C(C)}$ (1, 445)=0.38; p=0.54; F_V (1, 445)=8.64; p<0.01; F_C (1, 445)=0.72; p=0.40). The interaction of Focal accent and Speaker was significant in all segments ($F_{C(C)}$ (3, 445)=5.72; p<0.01; F_V (3, 445)=7.31; p<0.01; F_C (3, 445)=4.68; p<0.01). Finally, the interaction of Vowel length and Speaker was significant in V and C but not in C(C) ($F_{C(C)}$ (3, 445)=0.23; p=0.87; F_V (3, 445)=8.54; p<0.01; F_C (3, 445)=6.32; p<0.01). As mentioned above, none of the other two-, three- or four-way interactions were significant in the Multivariate test. Nor were any of them significant in the Tests of Between-Subjects effects.

4.7. Discussion

These results show that a long vowel in a stressed syllable is lengthened when there is a focal accent on that word. However, a short vowel in a stressed syllable is only marginally (and not significantly) affected by the presence of a focal accent. Instead, the long consonant following the short vowel is lengthened considerably. Thus, the contrast between the short and long segments within the stressed syllable is enhanced when there is a focal accent on a word.

Furthermore, not only the phonologically long segments are lengthened in a focally accented word. Consonants preceding the vowel – whether it is long or short – are also lengthened, as well as the phonologically short consonants following long vowels. Thus, short consonants may be lengthened, while short vowels may not. So, it seems that the vowel is the most important segment for preserving the Swedish quantity distinction, that is, the distinction between VC: and V:C patterns.

In an experiment based on a small corpus of accent II words, Bannert (1979) concluded that the temporal contrast between long and short segments is enhanced in focus. The current experiment confirms these findings and, in addition, extends them to words with accent I. The results of this experiment are also consistent with the results of Fant *et al.* (1991). Although Fant *et al.* (1991) found lengthening in the entire CV(:)C(C)-sequence, irrespective of vowel length, it is obvious from their data that the lengthening in the short vowel was marginal.

Lengthening of stressed syllables with long vowels in Swedish appears to be fairly similar to the lengthening found in Dutch and English, as all segments in the stressed syllable are lengthened (cf. Turk & Sawusch, 1997; Cambier-Langeveld & Turk, 1999; Turk & White, 1999). However, what does appear to be different in Swedish compared to Dutch and English is the lengthening pattern on stressed syllables with short vowels.

Finally, in the same way as the lengthening patterns at the syllabic level were shown to be incompatible with the idea of a linear time expansion of the entire word (cf. Experiment 1) also the lengthening patterns within the stressed syllables are incompatible with linear expansion. If there was a linear expansion we would have found lengthening of the short vowels in the stressed syllables as well. We will return to this issue in 6.3 in the general discussion.

5. Experiment 4: Lengthening of unstressed syllables before and after the stressed syllable

5.1. Introduction

In the previous experiments, we have touched upon the issue of the domain of focal accent lengthening. In Experiment 1 we saw that most of the lengthening occurred in the stressed syllable (and in Experiment 3 we made observations of the syllable-internal distribution of lengthening). We also observed, again in Experiment 1, that an unstressed syllable immediately to the right of the stressed syllable in a disyllabic word was lengthened to some extent.

The material in the previous experiments contained only disyllabic words with stress on the first syllable. Preliminary work of our own on Swedish (Strangert & Heldner, 1998), as well as work on English (Turk & White, 1999) indicates that the lengthening may extend to final unstressed syllables in trisyllabic words, as well as to initial unstressed syllables, although the unstressed syllable lengthening was much less than that on stressed syllables. Therefore, in the present experiment, we use trisyllabic words, where stress is either on the initial or on the medial syllable. The purpose is to examine whether the domain of focal accent lengthening extends beyond the stressed syllable and the immediately following unstressed one by examining unstressed syllables before and after the stressed syllable in longer words.

5.2. Material

Ten trisyllabic words with stress on the first syllable (S.U.U-words) and ten with stress on the second syllable (U.S.U-words) were chosen. All the words were verbs, the S.U.U-words in past and the U.S.U-words in present tense. As all S.U.U-words in Swedish have accent II and all U.S.U-words accent I, word accent variation is inevitable. The S.U.U- and the U.S.U-words are listed in Table VIII with a phonemic transcription and syllabification as well as a translation into English.

TABLE VIII. The test words used in Experiment 4 with phonemic transcriptions, syllabification and translations

S.U.U-words			U.S.U-words		
agade	ˈɑːg.a.dɛ	beat	anammar	a.ˈnam.ar	accepts
hotade	ˈhuː.ta.dɛ	threatened	besöker	bɛ.ˈsøːk.ɛr	visits
häktade	ˈhɛk.ta.dɛ	detained	bevakar	bɛ.ˈvɑːk.ar	guards
väntade	ˈvɛn.ta.dɛ	expected	betraktar	bɛ.ˈtrak.tar	watches
skadade	ˈskɑː.da.dɛ	injured	förlåter	fœr.ˈloːt.ɛr	forgives
fängslade	ˈfɛŋ.sla.dɛ	imprisoned	föraktar	fœr.ˈak.tar	despises
frestade	ˈfrɛs.ta.dɛ	tempted	förbannar	fœr.ˈban.ar	curses
smickrade	ˈsmik.ra.dɛ	flattered	gestaltar	jɛ.ˈstal.tar	shapes
granskade	ˈgran.ska.dɛ	reviewed	kontaktar	kɔn.ˈtak.tar	contacts
strandade	ˈstran.da.dɛ	stranded	välsignar	vɛl.ˈsiŋ.nar	blesses

The syllabification was based on the outcome of an optimality theory parser (Eriksson & Andersson, 1997) using a sonority- and a maximum onset-rule, in that order. The sonority-rule was based on a list of permitted consonant combinations for Swedish compiled by Elert (1970, p. 90). The maximum onset rule simply joined all consonants to the nearest vowel to the right. In addition to these rules, one violation of the onset rule was permitted in the unstressed syllable following the stressed. This was done due to the complementary relationship between the vowel and the following consonant in stressed syllables in Swedish. Thus, all the stressed syllables were closed and moreover varied in complexity from two to five segments. The unstressed syllables contained between one and three segments.

The target words were placed in medial position in the sentence frame *Kvinnan {Target word} mannen* (‘The woman’ {Target word} ‘the man’). Each of these sentences occurred with three different focal accent structures, that is, with a focal accent on the initial word, on the medial, and on the final word, respectively.

5.3. Speakers

Three speakers, two males and one female read the test material. They were all native speakers of Swedish without any strong dialectal influence and without any known hearing or speaking disorders. They were not paid for their services.

5.4. Recording

The procedures were the same as in Experiment 2. Focally accented words were indicated with capital letters and the speakers were instructed to emphasize the capitalized words. They were also explicitly asked to read the sentences as one phrase, that is, without any pauses.

The speakers initially read a few of the test sentences as a practice. Then, they read each sentence four times giving a total of 720 sentences (20 test sentences x 3 focal accent conditions x 4 repetition x 3 speakers).

The recordings were monitored during and after the recording. No sentences had to be excluded from the analyses due to technical problems or unsuccessful productions not discovered during the recording.

5.5. Measurements and data analysis

Only the target words, the verbs, were included in the subsequent analyses. Segment boundaries were determined and labeled as in Experiment 1, and the durations of each of the three syllables in each verb were calculated. In accordance with the syllabification principles adopted (see 5.2), the consonant following the vowel was included in the stressed syllable because of the compensatory relationship between the vowel and the following consonant in stressed syllables in Swedish.

As in the previous experiments, there were two nonfocal versions of each word, a pre-focal and a post-focal version. The two nonfocal versions were collapsed into one nonfocal category (with reference to the results in Experiment 2 showing only insignificant differences between the nonfocal conditions, cf. 3.6).

As the purpose was to examine whether the domain of focal accent lengthening extends beyond the stressed syllable and the immediately following unstressed syllable, only the duration of the final unstressed syllable in the S.U.U-words and the initial unstressed syllable in the U.S.U-words were submitted to inferential testing. To determine whether these syllables were affected by focal accent lengthening, two separate univariate ANOVAs were run, one for each of the two syllables. The dependent variables were the durations of the final unstressed syllable for the S.U.U-words and the initial unstressed syllable for the U.S.U-words. The independent variables were Focal accent (focal vs. nonfocal), Word (ten words per analysis), and Speaker (three levels). Focal accent was treated as a fixed factor, Word and Speaker were treated as random factors.

5.6. Results

The mean durations of all three syllables in the focal and nonfocal S.U.U- and U.S.U-words across all speakers are presented in Fig. 4. As shown previously (Experiment 1), most of the lengthening of words with focal accents is found in the stressed syllable. Thus, the stressed syllable is lengthened by 63 ms or 22% in the S.U.U-words and by 54 ms or 20% in the U.S.U-words. There is also lengthening in the unstressed syllable immediately to the right of the stressed. These unstressed syllables are lengthened by 22 ms or 18% in the S.U.U-words and by 27 ms or 21% in the U.S.U-words. The final unstressed syllable in the S.U.U-words and the initial unstressed syllable in the U.S.U-words are lengthened considerably less, 5 ms or 5% in the S.U.U-words and 9 ms or 8% in the U.S.U-words.

The ANOVA for the final unstressed syllable in the S.U.U-words showed a significant main effect of the Word ($F(9, 10.24)=3.61$; $p=0.03$) and a significant interaction of Word and Speaker ($F(18, 18)=2.35$; $p=0.04$). These effects indicate that there were significant differences in the duration of the final unstressed syllable between the different words and speakers. None of the other effects was significant: Focal accent ($F(1, 1.43)=11.35$; $p=0.12$); Speaker ($F(2, 4.67)=1.03$; $p=0.43$); Focal accent and Speaker ($F(2, 18)=2.07$; $p=0.12$); Focal accent and Word ($F(9, 18)=0.72$; $p=0.68$); Focal accent, Speaker and Word ($F(18, 300)=1.26$; $p=0.21$). However, the observed power for these effects was generally low. For example, the observed power for the Focal accent effect (that is, the difference between focal and nonfocal words) given the observed mean difference of 5 ms was only 0.32.

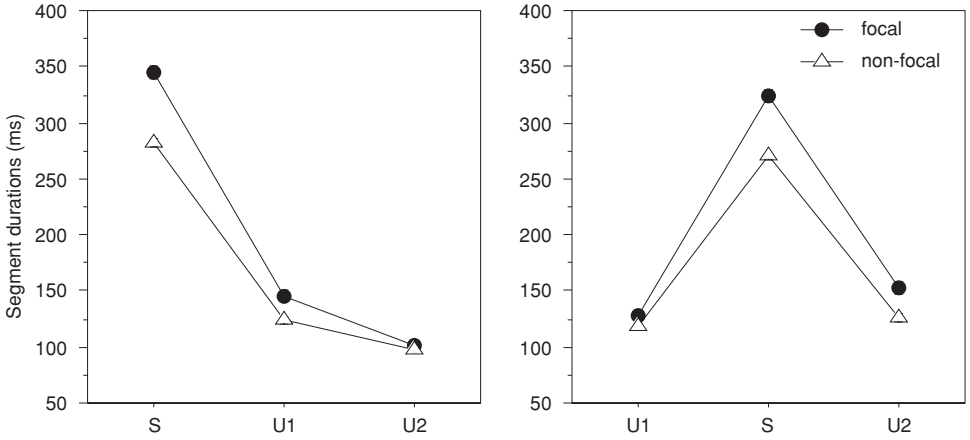

Figure 4. Mean syllable durations (ms) and standard errors in focal and nonfocal conditions for the S.U.U-words to the left, and the U.S.U-words to the right. S = stressed syllable, U1, U2 = unstressed syllables.

The ANOVA for the initial unstressed syllables in the U.S.U-words showed a significant main effect of Word ($F(9, 19.06)=17.95$; $p<0.01$). There were also significant interactions of Word and Speaker ($F(18, 18)=8.40$; $p<0.01$) and of Focal accent, Speaker and Word ($F(18, 300)=1.77$; $p=0.03$). These effects show that the duration of the word-initial unstressed syllable differed significantly between the words and speakers and also that the amount of lengthening differed among the words and speakers. None of the other effects were significant: Focal accent ($F(1, 2.79)=9.06$; $p=0.06$); Speaker ($F(2, 13.73)=0.96$; $p=0.41$); Focal accent and Speaker ($F(2, 18)=2.53$; $p=0.11$); Focal accent and Word ($F(9, 18)=1.62$; $p=0.18$). Again, the observed power for the nonsignificant effects was quite low.

Thus, similar to the final unstressed syllable in the S.U.U-words, the mean differences between focal and nonfocal conditions did not reach significance in the initial unstressed syllable in the U.S.U-words.

5.7. Discussion

This experiment supports the findings from Experiment 1 that most of the focal accent lengthening occurs in the stressed syllable and that there is some lengthening in the immediately following unstressed syllable, too. All the words in Experiment 1 were disyllabic with stress on the first syllable. However, in the trisyllabic S.U.U-words in the present experiment, there was no significant lengthening of the final unstressed syllable and the mean focused vs. nonfocused difference was only 5 ms. Neither was there any significant lengthening of the initial unstressed syllable in the trisyllabic U.S.U-words with a mean focused vs. nonfocused difference of just 9 ms. Thus, the present experiment failed to show any lengthening to the right of the first unstressed syllable and to the left of the stressed syllable in Swedish.

Consequently, our results concerning the final unstressed syllable in Swedish S.U.U-words are not in agreement with those of Turk & White (1999) for English. They found accentual lengthening extending throughout all syllables in a trisyllabic word with primary stress on the initial syllable. A final syllable in an accented trisyllabic word was 13.7% longer than in an unaccented word. Regarding initial unstressed syllables in U.S.U-words our results for Swedish are generally in agreement those for English, although Turk & White

(1999) found some lengthening in initial unstressed syllables, at least for some speakers. These results, together with those from Experiment 1 investigating the lengthening in disyllabic words, have consequences for our interpretation of the domain of focal accent lengthening in Swedish. We will return to this issue in 6.2 in the general discussion.

6. General discussion and conclusions

6.1. Amount of focal accent lengthening

This study has shown that focal accent lengthening occurs in Swedish to about the same extent as in languages such as English and Dutch (cf. Cooper *et al.*, 1985; Eefting, 1991; Sluijter & van Heuven, 1995). Focally accented words in Experiment 1 were on average 25% longer than nonfocal ones. In addition, this study has shown that the amount of lengthening may be influenced by several factors, speaker being the single most important one. This is illustrated by the fact that in Experiment 2, using fewer and other speakers than in Experiment 1, the average focal accent lengthening was only 12%. Thus, as there is substantial speaker variation in the amount of lengthening, a general figure of, say 25%, should be interpreted with some caution.

Experiments 1 and 2 also revealed the influence of another factor on the amount of focal accent lengthening, the position in the phrase of the focused word. In particular, there was a trend that phrase-final words were lengthened more than words in other positions. However, this trend found in Swedish contrasts with that reported for American English by Cooper *et al.* (1985) and for Dutch by Cambier-Langeveld (2000). In both these studies, there was less lengthening in final position than in others. Our results also disagree with those for RP-English in Cambier-Langeveld (2000), where no influence whatsoever of position in the phrase could be substantiated. Our observations are even partly in conflict with an earlier Swedish study (Bruce, 1981), in which only one of the two speakers lengthened the words in phrase-final position more than those in phrase-medial position.

Cambier-Langeveld (2000) relates the differences between RP-English and Dutch to differences in the phonological rules in the two languages. However, we lack an explanation for the observed differences in lengthening between Swedish on the one hand, and English and Dutch on the other. These differences *might be* real ones, but could just as well be a reflection of different methodologies, as well as other factors not controlled for. In addition, we cannot exclude the possibility that speaker variability might have played a role here (cf. the different patterns of the two Swedish speakers in the study by Bruce (1981).

The present study further indicates that yet another factor may influence the amount of focal accent lengthening. Experiment 1 showed that accent II words were lengthened more than words with accent I, at least in some positions in the phrase. Similar tendencies were also observed in Experiment 2, although only one word of each accent type was included in that material and no statistical tests of word accent differences were made. However, there were no indications of lengthening effects in the stressed syllables due to word accents in Experiment 3. Thus, the Swedish word accent distinction has an unstable influence on the amount of focal accent lengthening, and moreover the effect when present is not very strong. Therefore, we tend to agree with the conclusion reached in Bruce (1981), that the same temporal program is used in focus for accent I and accent II.

6.2. Domain and distribution of focal accent lengthening

As to the distribution of lengthening, the findings of Experiments 1 and 4, taken together, show (i) that most of the focal accent lengthening occurs in the stressed syllable, (ii) that there is also lengthening of an unstressed syllable immediately to the right of the stressed syllable, though to a lesser extent, and (iii) that neither unstressed syllables to the left of the stressed syllable, nor the second unstressed syllable to the right of the stressed syllable are affected by lengthening.

Thus, the present study indicates that the domain of focal accent lengthening in Swedish (at least for noncompound words) begins with the stressed syllable and includes one unstressed syllable to its right. Neither unstressed syllables to the left of the stressed syllable nor the second unstressed syllable after the stressed one should be included in the domain.
These results differ from those reported for English by Turk & White (1999). Their data point to a somewhat larger domain, as they found that the lengthening extended throughout all syllables in a three-syllable word with stress on the first syllable. Thus, their domain begins with the pitch accented syllable and extends rightward until a word boundary. However, it should be noted that even though the lengthening of the final unstressed syllable was statistically significant, it was not great. Final unstressed syllables were 13.7% longer when focused.

Turk & White (1999) also found small but statistically significant lengthening effects on initial unstressed syllables. These syllables were 4.1% longer when focused. While it is true that no significant lengthening on initial unstressed syllables was found for Swedish, English is probably not different from Swedish in this regard – there is no perceptually significant lengthening of initial unstressed syllables in either language. We will discuss differences of magnitudes like these in relation to perceptual aspects of lengthening in 6.6.

6.3. Lengthening within the stressed syllable and the quantity distinction

As concerns the distribution of lengthening *within* the stressed syllable, it is clearly influenced by the quantity distinction in Swedish as shown in Experiment 3. First of all, the phonologically long segments within the stressed syllable – vowels as well as consonants – are lengthened considerably. Thus, in syllables with a long vowel followed by a short consonant, the long vowel is heavily affected. In addition, the consonants preceding as well as those following the vowel are lengthened to almost the same degree as the vowel. In syllables with a short vowel followed by a long consonant, the long consonant is lengthened as well as consonants preceding the vowel, while the short vowel is practically unaffected. Thus, we see a pattern of selective lengthening with all segments affected except the phonologically short vowel.

It is not altogether obvious why the lengthening associated with focal accents in Swedish should be selective, or nonlinear, in the way described above. A linear time expansion of all segments would still preserve the relations between long and short segments within the stressed syllable as well as the quantity distinction between VC:- and V:C-syllables. Nevertheless we see a nonlinear lengthening pattern, a lengthening of the phonologically long segments that – in combination with the almost unaffected short vowel – even gives rise to a strengthening of the quantity distinction, confirming previous conclusions by Bannert (1979) and Fant *et al.* (1991). Thus, strengthening of temporal contrast is one possible reason for the observed lengthening pattern. Also contrasts in other acoustic dimensions appear to be strengthened in focus. For example, Fant & Kruckenberg (1994) have shown that the difference in intensity between vowels and consonants increases in focus. Furthermore, in a

recent study (Heldner, forthcoming) it is shown that this nonlinear lengthening of CVC:-syllables is important for the perceived naturalness of focally accented Swedish words.

Finally, the nonlinear lengthening of CVC:-syllables taken together with the findings from Experiments 1 that the stressed syllable is lengthened relatively more than the word as a whole, provides ample evidence that it is inappropriate to describe focal accent lengthening in Swedish as a linear time expansion of the entire word. Cambier-Langeveld & Turk (1999) reached the same conclusion for English and Dutch.

6.4. Interplay with other aspects of phonology

The deviation from linearity, that is, the selective lengthening demonstrated in the stressed syllable, indicates that focal accent lengthening does not occur "blindly", but rather, as we have shown, seems to be adjusted to the demands of the phonological quantity distinction in Swedish. Similarly, in Experiment 1, we saw different amounts of lengthening of the stressed and unstressed syllable, respectively. This adjustment to the stressed-unstressed distinction is another indication of a complex interaction between focal accent lengthening and other aspects of Swedish phonology.

Cambier-Langeveld (2000) similarly points to the possibility of adjustments to language-specific phonologic constraints in her comparative study of Dutch and English. She assumed the more extensive lengthening in final syllables in English (as compared to Dutch) could be explained with reference to a phonologic lengthening rule for English without correspondence in Dutch. In addition to providing further evidence for an interplay with other aspects of phonology, the Dutch-English study points to the influence of structural differences between languages as a source of variability. In other words, language-specific characteristics might be responsible for some of the variability reported between studies based on different languages.

6.5. Global effects of lengthening

A specific issue in the exploration of the distribution and domain of focal accent lengthening in Swedish has been the investigation of whether the lengthening extends across a word boundary; that is, whether there are more global temporal effects. Such effects have been demonstrated for the tonal domain, for example the compression of the pitch range following a focal accent in Swedish (Bruce, 1977).

The present study, however, gives no clear indications of similar adjustments in the temporal domain. The effects of focal accents on surrounding nonfocal words were weak in Experiment 1, and absent in Experiment 2. Significant rightward effects across word boundaries were found in nonfocused final words in Experiment 1 only. And, there were no tendencies to leftward effects of focal accent lengthening across word boundaries in either of the two experiments.

However, in both Experiments 1 and 2, there was a trend that the words immediately to the right of the focally accented word were longer than those before, or two words after, the focused word. Thus, there might be small rightward effects of focal accents across word boundaries in Swedish similar to those reported for English (Turk & White, 1999). Possibly, given a larger material, mean differences in the magnitudes observed in our experiment could also be shown to be statistically significant. Furthermore, it should be noted that Turk & White (1999) compared monosyllabic words before and after monosyllabic focused words whereas all words in our experiment were disyllabic. It could well be the case that more

lengthening would be found on unstressed syllables immediately following a stressed syllable across a word boundary.

Still, it is questionable whether such small differences, though significant, have any relevance as cues to focus for the listener (see 6.6 below). Given these results, we do not believe that there are temporal effects of focal accents in Swedish spanning longer stretches of speech than the focused word itself, at least not any effects playing a significant role. In other words, there seems to be no reason to take more global temporal effects of focus into account, at least not as far as Swedish is concerned. As a consequence, we do not believe that the domain of focal accent lengthening in Swedish extends beyond the word.

6.6. Perceptual relevance

When reporting durational differences, as done in this paper, the question arises as to how large a durational difference has to be in order to be perceived. This is important because temporal effects that reach statistical significance may still be so small that they have no perceptual relevance – the differences may just not be noticeable. It is beyond the scope of this article to cover the literature on just noticeable differences (JND); for a review, see Eriksson (1991). However, to cite just one example, Klatt & Cooper (1975) reported average JNDs across different contexts for stressed vowels to be 41 ms corresponding to an average lengthening of 18%, and 48 ms corresponding to an average lengthening of 35% for a following consonant. By no means all of the lengthening effects we have reported in this study exceed these values. In particular, and in light of what is known about difference limens, the very marginal (and indeed nonsignificant) effect of 5% lengthening in the short stressed vowel in Experiment 3 is unlikely to be perceivable. Taking limitations of this kind into account is necessary when evaluating the acoustic data reported in this study in a perceptual perspective. We know from previous studies (Heldner & Strangert, 1997; Heldner, 1998) that other cues than those held to be primary, that is, tonal cues, may signal focus to the listener. If, and to what extent, temporal effects such as those obtained in the present study could function as cues to the listener depends on the extent to which they are detectable by the human ear.

Acknowledgements

The authors would like to thank Vincent van Heuven, Sieb Nooteboom and Alice Turk for helpful comments and discussion. We are also grateful to Thierry Deschamps for technical assistance and to Åke Olofsson and Hans Nyquist for statistical advice. This research was funded by a grant from the Swedish Council for Research in the Humanities and Social Sciences (HSFR).

References

Bannert, R. (1979) The effect of sentence accent on quantity. In Proceedings of the Ninth International Congress of Phonetic Sciences, pp. 253-259. Copenhagen: Institute of Phonetics, University of Copenhagen.

Bruce, G. (1977) Swedish word accents in sentence perspective. Lund: CWK Gleerup.

Bruce, G. (1981) Tonal and temporal interplay. In Nordic prosody II (T. Fretheim, ed.), pp. 63-74. Trondheim: Tapir.

Bruce, G. (1999) Word tone in Scandinavian languages. In Word prosodic systems in the languages of Europe (H. van der Hulst, ed.), pp. 605-633. Berlin, New York: Mouton de Gruyter.

Cambier-Langeveld, T. (2000) The interaction between final and accentual lengthening: Dutch vs. English. In Temporal marking of accents and boundaries, pp. 99-124. Amsterdam: Holland Institute of Generative Linguistics.

Cambier-Langeveld, T. & Turk, A. E. (1999) A cross-linguistic study of accentual lengthening: Dutch vs. English, Journal of Phonetics, 27, 255-280.

Cooper, W. E., Eady, S. J. & Mueller, P. R. (1985) Acoustical aspects of contrastive stress in question-answer contexts, Journal of the Acoustical Society of America, 77(6), 2142-2156.

Eefting, W. (1991) The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments, Journal of the Acoustical Society of America, 89(1), 412-424.

Elert, C.-C. (1964) Phonologic studies of quantity in Swedish. Uppsala: Almqvist & Wiksell.

Elert, C.-C. (1970) Ljud och ord i svenskan. Stockholm: Almqvist & Wiksell.

Eriksson, A. (1991) Aspects of Swedish speech rhythm. Gothenburg: Department of Linguistics, University of Göteborg.

Eriksson, A. & Andersson, O. (1997) Three phonological parsers written in JAVA. Available online: http://www.ling.umu.se/~anderse/education/Parsers.html. Last accessed on the 30 November 1999.

Fant, G. & Kruckenberg, A. (1994) Notes on stress and word accent in Swedish, STL-QPSR, 2-3/1994, 125-144.

Fant, G., Kruckenberg, A. & Nord, L. (1991) Durational correlates of stress in Swedish, French and English, Journal of Phonetics, 19, 351-365.

Gussenhoven, C. (1984) On the grammar and semantics of sentence accents. Dordrecht: Foris.

Heldner, M. (1998) Is an F0-rise a necessary or a sufficient cue to perceived focus in Swedish? In Nordic Prosody: Proceedings of the VIIth Conference, Joensuu 1996 (S. Werner, ed.), pp. 109-125. Frankfurt am Main: Peter Lang.

Heldner, M. (forthcoming) On the non-linear lengthening of focally accented Swedish words. In To appear in Nordic Prosody: Proceedings of the VIIIth Conference, Trondheim 2000 (W. van Dommelen & T. Fretheim, eds.). Frankfurt am Main: Peter Lang.

Heldner, M. & Strangert, E. (1997) To what extent is perceived focus determined by F0-cues? In Eurospeech '97 Proceedings, pp. 875-877. Rhodes, Greece: ESCA.

Horne, M., Strangert, E. & Heldner, M. (1995) Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In Proceedings of the XIIIth International Congress of Phonetic Sciences, pp. 170-173. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

Klatt, D. H. & Cooper, W. E. (1975) Perception of segment duration in sentence contexts. In Structure and Process in Speech Perception (A. Cohen & S. G. Nooteboom, eds.), pp. 69-86. Berlin, Heidelberg, New York: Springer-Verlag.

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

Ladd, D. R. (1996) Intonational phonology. Cambridge: Cambridge University Press.

Nooteboom, S. G. & Kruyt, J. G. (1987) Accents, focus distribution, and the perceived distribution of given and new information: An experiment, Journal of the Acoustical Society of America, 82(5), 1512-1524.

Sluijter, A. M. C. (1995) Phonetic correlates of stress and accent. The Hague: Holland Academic Graphics.

Sluijter, A. M. C. & van Heuven, V. J. (1995) Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch, Phonetica, 52, 71-89.

Sluijter, A. M. C. & van Heuven, V. J. (1996) Spectral balance as an acoustic correlate of linguistic stress, Journal of the Acoustical Society of America, 100(4, Pt 1), 2471-2485.

SPSS. (1999) SPSS Advanced Models 9.0. Chicago: SPSS Inc./Prentice Hall.

Strangert, E. & Heldner, M. (1998) On the amount and domain of focal lengthening in Swedish. In ICSLP'98 Proceedings (R. H. Mannell & J. Robert-Ribes, eds.), pp. 3305-3308. Sydney: ASSTA.

Turk, A. E. & Sawusch, J. R. (1997) The domain of accentual lengthening in American English, Journal of Phonetics, 25(1), 25-41.

Turk, A. E. & White, L. (1999) Structural influences on accentual lengthening in English, Journal of Phonetics, 27(2), 171-206.

van Heuven, V. J. (1993) On the temporal domain of focal accent. In Proceedings of an ESCA Workshop on Prosody (D. House & P. Touati, eds.), pp. 132-135. Lund: Lund University, Department of Linguistics.

van Heuven, V. J. (1994) What is the smallest prosodic domain? In Phonological structure and phonetic form: Papers in laboratory phonology III (P. A. Keating, ed.), pp. 76. Cambridge: Cambridge University Press.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P. J. (1992) Segmental durations in the vicinity of prosodic phrase boundaries, Journal of the Acoustical Society of America, 91(3), 1707-1717.

Appendix A: Test sentences used in Experiment 3

A.1. Questions

Vem är det som {[$_{VP}$ V} {NP]]}? e.g., Vem är det som biter kvinnan? *'Who is that biting the woman?'*
Vad gör {[$_{S}$ NP} med {NP]]}? e.g., Vad gör mannen med kvinnan? *'What is the man doing to the woman?'*
Vem är det som {[S NP} {[VP V}? e.g., Vem är det som mannen biter? *'Who is it that the man is biting?'*

A.2. Answers

Verbs with short vowels and long consonants	*Verbs with long vowels and short consonants*
Accent I words	*Accent I words*
Mannen kniper ('pinches') kvinnan.	Mannen klipper ('cuts') kvinnan.
Mannen dräper ('slays') kvinnan.	Mannen släpper ('releases') kvinnan.
Mannen biter ('bites') kvinnan.	Mannen gitter (nonsense) kvinnan.
Mannen mäter ('measures') kvinnan.	Mannen sätter ('puts') kvinnan.
Mannen sviker ('jilts') kvinnan.	Mannen sticker ('pricks') kvinnan.
Mannen läker ('heals') kvinnan.	Mannen väcker ('wakes') kvinnan.
Mannen grämer ('grieves') kvinnan.	Mannen stämmer ('summons') kvinnan.
Mannen bryner ('browns') kvinnan.	Mannen finner ('finds') kvinnan.
Mannen kyler ('chills') kvinnan.	Mannen fyller ('stuffs') kvinnan.
Mannen mäler (nonsense) kvinnan.	Mannen fäller ('convicts') kvinnan.
Accent II words	*Accent II words*
Kvinnan slipar ('grinds') mannen.	Kvinnan tippar ('dumps') mannen.
Kvinnan kapar ('cuts') mannen.	Kvinnan tappar ('drops') mannen.
Kvinnan ritar ('draws') mannen.	Kvinnan hittar ('finds') mannen.
Kvinnan matar ('feeds') mannen.	Kvinnan fattar ('grasps') mannen.
Kvinnan pikar ('taunts') mannen.	Kvinnan kickar ('kicks') mannen.
Kvinnan hakar ('hooks') mannen.	Kvinnan hackar ('minces') mannen.
Kvinnan mimar ('mimes') mannen.	Kvinnan trimmar ('trims') mannen.
Kvinnan kramar ('hugs') mannen.	Kvinnan kammar ('combs') mannen.
Kvinnan tinar ('defrosts') mannen.	Kvinnan skinnar ('skins') mannen.
Kvinnan manar ('bids') mannen.	Kvinnan stannar ('stops') mannen.

Paper V In *Focal accent – f_0 movements and beyond*
pp. 101–109

On the non-linear lengthening of focally accented Swedish words[1]

Mattias Heldner

1. Introduction

This study deals with the perceptual relevance of a specific non-linear lengthening pattern in focally accented Swedish words, namely that found in stressed syllables where a phonologically short vowel is followed by a long consonant.

Swedish is a quantity language and there is a distinction between phonologically long and short vowels in the stressed syllable. In addition, there is a complementary relation between the vowel and the immediately following consonant – the consonant being short whenever the vowel is long (V:C) and long (VC:) or part of a consonant cluster (e.g. VCC) when the vowel is short (Elert, 1964). In other words, a quantity contrast may also be said to appear within the stressed syllable.

Lengthening is a well-established acoustic correlate of focal accents in Swedish, as well as in many other languages. The amount of lengthening reported in the literature varies, but, typically, words with a focal accent are about 25% longer than the same words when non-focal. However, previous studies of Swedish (e.g. Bannert, 1979; Fant, Kruckenberg & Nord, 1991; Heldner & Strangert, 2001) have shown that the segments within the focally accented words are sometimes lengthened in a non-linear fashion. That is, some segments are lengthened relatively more than others.

Figure 1 shows typical focal accent lengthening patterns for stressed syllables with a short vowel followed by a long consonant (CVC:) as well as for syllables with a long vowel and a short postvocalic consonant (CV:C) (data from Heldner & Strangert, 2001). The non-linear lengthening patterns occur within the CVC:-syllable. Here, the short vowel remains practically unaffected by the presence of a focal accent, while the postvocalic long consonant is lengthened considerably. This pattern is qualitatively different from that observed in the stressed CV:C-syllable, where all segments are lengthened, including the short postvocalic consonant.

Figure 1 also shows hypothetical linear lengthening patterns for both syllable types. The stressed syllable durations are the same for the focally accented (and non-linearly lengthened) syllables and the linearly lengthened ones, but the lengthening is differently distributed. A comparison between the linear lengthening patterns and the non-linear ones shows that the short vowel in the CVC:-syllable is lengthened relatively less and the long consonant relatively more than the syllable as a whole. As a result, the temporal contrast between the short vowel and the long consonant is strengthened when the word carries a focal accent. In the CV:C-syllable, however, all segments, including the short postvocalic

[1] This material has been published as Heldner, M. (2001) On the non-linear lengthening of focally accented Swedish words. In *Nordic Prosody: Proceedings of the VIIIth Conference, Trondheim 2000* (W. van Dommelen & T. Fretheim, eds.), pp. 103-112. Frankfurt am Main: Peter Lang.

consonant, are lengthened approximately linearly, which means that the temporal contrast is left unaffected.

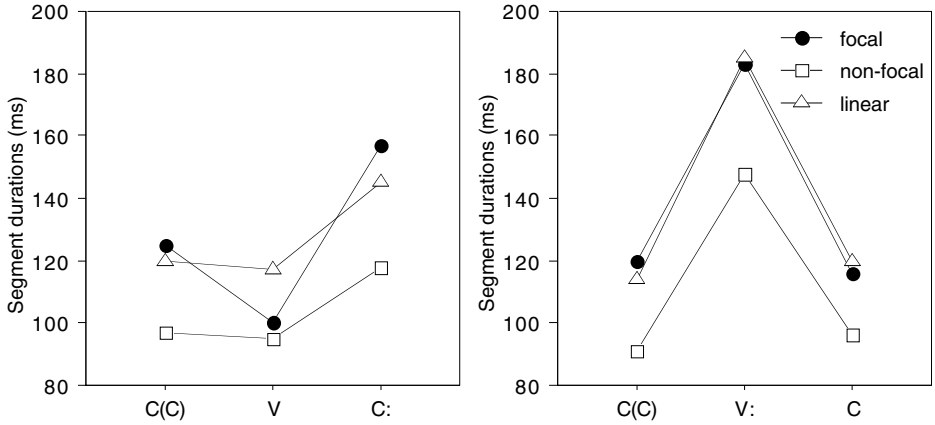

Figure 1. Examples of lengthening patterns in CVC:- (left panel) and CV:C-syllables (right panel). Segment durations (in ms) within the stressed syllable in a focal accented (filled circles) and non-focal word (empty squares). The empty triangles show a linear lengthening pattern where the amount of lengthening across the syllable is equal to that in the focally accented syllable, but where the lengthening is evenly distributed.

Thus, the non-linear lengthening of CVC:-syllables has two distinguishing features as compared to the approximately linear lengthening of CV:C-syllables. Firstly, the phonologically short segment in the CVC:-syllable (i.e. the vowel) resists lengthening, while everything else is lengthened. Secondly, there is a strengthening within the CVC:-syllable of the temporal contrast between the short segments and the long ones (no strengthening occurs in CV:C-syllables). It is not altogether obvious why the stressed CVC:-syllable should be lengthened in this non-linear fashion but there seems to be something special about CVC:-syllables and phonologically short vowels.

The first experiment in the present study has been designed to investigate experimentally whether this non-linear lengthening pattern in CVC:-syllables is important for the perceived naturalness of focal accented words. What it attempts to establish, in other words, is whether it is important to model focal accent lengthening in CVC:-syllables in this non-linear fashion, or whether one might just as well lengthen all the segments linearly, as is done in CV:C-syllables. An obvious motivation for this question, of course, is that if non-linear lengthening has an impact on perceived naturalness, CVC:- and CV:C-syllables have to be treated differently in speech synthesis.

In addition, the relative importance of the two features distinguishing the non-linear from the linear lengthening pattern, that is, (i) the resistance to lengthening in the short vowel and (ii) the strengthening of temporal contrast will be examined in a second experiment.

2. Method

2.1. Collection of production data

As the main question in this paper pertains to the perceived naturalness of non-linear lengthening patterns, the principal part of the investigation will be the perceptual

experiments. However, prior to this, some production data were collected to serve as a model for the stimuli to be used in the perceptual experiments. To this end, six three-word sentences were constructed, in which the middle words (the test words) were disyllabic verbs with stress on the first syllable. These sentences are shown in Table 1. All the test words contained a stressed syllable with a phonologically short vowel followed by a long consonant.

Table 1. Sentences used in the study with transcriptions of the test words and translations.

Mannen dricker vinet.	/drikːər/	'The man is drinking wine.'
Mannen klipper persilja.	/klipːər/	'The man is cutting parsley.'
Mannen knäcker äggen.	/knɛkːər/	'The man is breaking eggs.'
Mannen grillar köttet.	/grilːar/	'The man is grilling meat.'
Mannen gräddar brödet.	/grɛdːar/	'The man is baking bread.'
Mannen pressar citronen.	/prɛsːar/	'The man is squeezing lemon.'

There were two versions of each sentence, one of which had a focal accent on the middle word and the other on the final word. The test words thus occurred in focal as well as in non-focal (i.e. pre-focal) position. One speaker (the author) read each of these sentences six times. Subsequently, the durations of all segments in all sentences were measured, and the amount of focal accent lengthening for each segment of the test word, as well as for the whole test word, was determined. In addition, a set of f_0 turning points associated with word accents, focal accents and boundary tones were measured. The measured segment durations are shown in Figure 2, while the corresponding word durations and amounts of lengthening are shown in Table 2. Figure 2 also presents a linear lengthening relative to the segment durations in non-focal position using the amounts of focal accent lengthening recorded in Table 2.

Table 2. Mean word durations (in ms) for the test words in focal and non-focal position and the amount of focal accent lengthening.

	dricker	klipper	knäcker	grillar	gräddar	pressar
focal	516	514	571	524	523	506
non-focal	340	325	400	383	360	360
% lengthening	52%	58%	43%	37%	45%	41%

2.2. Preparation of stimuli

As mentioned above, the production data was collected only to serve as a model for stimuli in perceptual experiments. Thus, the stimuli in the two experiments in this study were different synthetic versions of the sentences in Table 1. All of these sentences were spoken by a male mbrola voice and prepared using the software WaveSurfer (Sjölander & Beskow, 2000) with a text-to-speech plug-in.

The stimuli in the two experiments were similar in many respects. The general aspects of the stimuli will be described here, whereas the particularities will be dealt with in the following sections. In general, the segmental durations and f_0-contours in all sentences were modeled after sentences with focal accents on words in medial position in the sentence. Moreover, the experimental variation occurred in the sentence-medial words only. Everything was kept constant in the initial and final words. Furthermore, all experimental variations concerned segmental durations. However, as a consequence of these manipulations the slopes of the f_0-movements within the test words were also affected. This was because the values of the f_0-turning points were kept constant while these points were anchored relative to the segments (e.g. to the boundary between the stressed vowel and the postvocalic consonant). Therefore, lengthening of a segment also resulted in a less steep slope of f_0-movements in that segment.

2.3. The first perceptual experiment

The first perceptual experiment was designed to investigate whether a non-linear lengthening pattern in CVC:-syllables is important for the perceived naturalness of focal accented words. Accordingly, listeners were asked in this experiment to compare pairs of synthesized sentences, where the amount of lengthening of the focally accented word was the same, while the distribution of this lengthening differed. One test word in each pair reflected the non-linear lengthening patterns in focally accented words (as in the filled circles of Figure 2). The other test word in the pair contained a linear lengthening of all segments relative to the segment durations for the non-focal words, that is, all segments within the word were lengthened by the same percentage (as in the empty triangles of Figure 2). Six pairs of sentences were prepared: one pair for each sentence in Table 1.

Eleven native speakers of Swedish participated in the first experiment. They were given the forced choice task to judge whether the first or second sentence in each pair sounded more natural. The order of presentation within each pair, that is whether the linear or non-linear version occurred first, was random. The presentation order of the sentence pairs was also random. Each listener judged each pair ten times. Thus, 110 judgements were obtained of each pair.

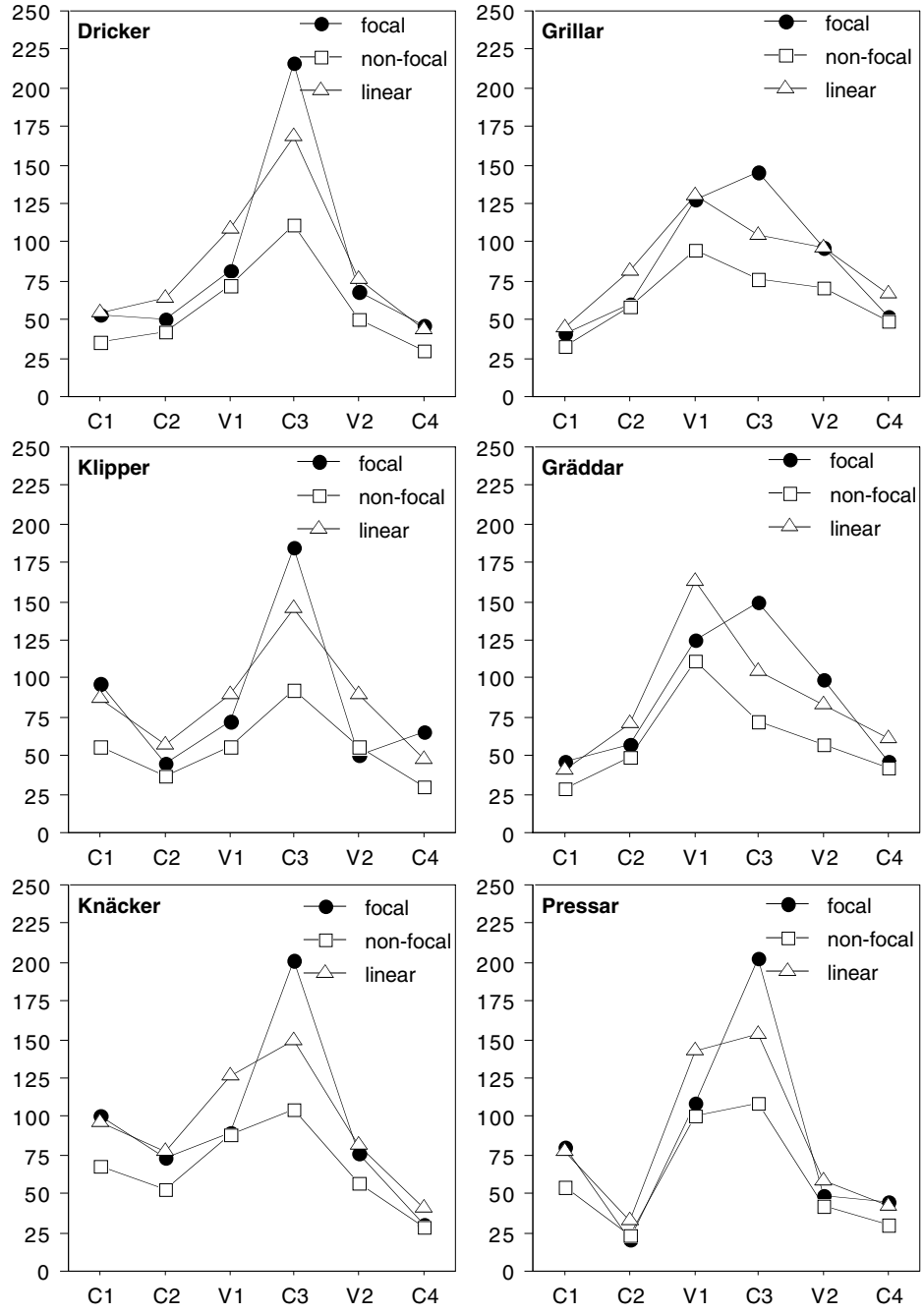

Figure 2. Mean segment durations (in ms) for each of the test words in focal and non-focal position and a linear lengthening relative to the segment durations in non-focal position. C1=first consonant, C2=second consonant, V1=first vowel etc.

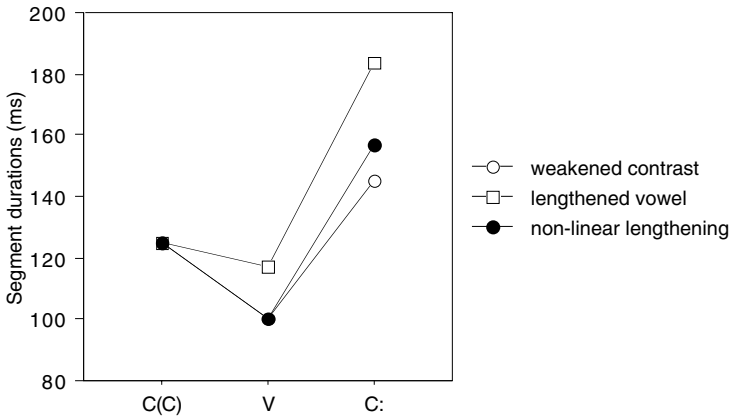

Figure 3. An illustration of the lengthening patterns weakened contrast (empty circles) and lengthened vowel (empty squares) together with the non-linear lengthening reference (filled circles) in the second perceptual experiment.

2.4. The second perceptual experiment

The second perceptual experiment was conducted to compare the perceptual importance of the two features distinguishing the non-linear from the linear lengthening patterns. Thus, the purpose was to find out whether one of the features (i) strengthening of temporal contrast and (ii) short vowel remaining short was more important than the other.

Again, listeners were asked to compare the naturalness of pairs of synthesized sentences where the segmental durations in the sentence-medial word differed. However, in this experiment both sentences in each pair deviated from the patterns observed in production. Thus, the listeners had to judge which deviation affected the naturalness least.

The experimental variations were restricted to the stressed vowel and the immediately following consonant and started out from the non-linear lengthening patterns (as shown in Figure 2). In one of the sentences in each pair, the vowel remained short while the consonant was shortened to the same duration as in the linear pattern (see Figure 2). This means that the temporal contrast was weakened compared to what was the case in the non-linear lengthening pattern, while the vowel stayed short. In the other sentence in the pair, both the vowel and the consonant were lengthened in such a way as not to affect the temporal contrast compared to the contrast displayed by the non-linear version. Here, the vowel had the same duration as in the linear pattern (see Figure 2) while the consonant was longer than in the non-linear reference. These lengthening patterns are illustrated in Figure 3.

Contrary to the manipulations carried out in the first experiment, those in the second experiment affected the total duration of the stressed syllable in the test words and consequently also the rhythm of the sentence.

Ten native speakers of Swedish participated in the second experiment. Their task and the number of presentations of each sentence pair was the same as in the first experiment. Every sentence pair was judged 100 times (10 times per speaker).

3. Results of the first perceptual experiment

Turning now to the results of the first perception experiment, Table 3 shows the observed frequencies for linearly and non-linearly lengthened stimuli judged to be more natural than the other sentence in a pair. It is quite obvious that stimuli with non-linear lengthening were judged to be more natural. The Chi-square tests across all speakers are significant for all test words except *grillar:* [*dricker:* $\chi^2(1, N=110)=44.54$, $p<.01$; *klipper*: $\chi^2(1, N=110)=44.54$, $p<.01$; *knäcker*: $\chi^2(1, N=110)=37.24$, $p<.01$; *grillar:* $\chi^2(1, N=110)=2.33$, $p=.13$; *gräddar:* $\chi^2(1, N=110)=47.13$, $p<.001$; *pressar*: $\chi^2(1, N=110)=26.51$, $p<.001$]. Moreover, the non-significant Chi-square test for *grillar* is probably due to the small durational differences between the linear and non-linear versions (cf. Figure 2).

Table 3. Frequencies across all listeners of linearly and non-linearly lengthened stimuli judged to be more natural than the other sentence in a pair.

	dricker	klipper	knäcker	grillar	gräddar	pressar	Totals
Linear	20	20	23	47	19	28	157
Non-linear	90	90	87	63	91	82	503
Totals	110	110	110	110	110	110	660

Although a majority of the listeners were sensitive to the durational differences they were exposed to during the experiment, two of the listeners reported that they did not notice any difference between the two versions in each sentence pair. An examination of their judgments showed no preference for either non-linear or linear lengthening. In other words, not all listeners were sensitive to the durational variations.

4. Results of the second perceptual experiment

The results of the second perception experiment are shown in Table 4. The table shows observed frequencies for stimuli judged to be more natural than the other sentence in a pair. Clearly, words with a lengthened vowel in combination with the maintaining of temporal contrast between the short vowel and the long consonant were judged to be less natural than words where the vowel remained short while the temporal contrast was weakened.

Table 4. Frequencies across all listeners of stimuli with lengthened vowels and weakened contrasts, respectively, judged to be more natural than the other sentence in a pair.

	dricker	klipper	knäcker	grillar	gräddar	pressar	Totals
Lengthened vowels	6	16	5	33	9	12	81
Weakened contrasts	94	84	95	67	91	88	519
Totals	100	100	100	100	100	100	600

The preference for the weakened contrast (or short vowel remaining short) stimuli was statistically significant in all words as the Chi-square tests across all speakers were significant for all test words: [*dricker:* $\chi^2(1, N=100)=77.44$, $p<.01$; *klipper*: $\chi^2(1, N=100)=46.24$, $p<.01$; *knäcker*: $\chi^2(1, N=100)=81.00$, $p<.01$; *grillar:* $\chi^2(1, N=100)=11.56$, $p<.01$; *gräddar:* $\chi^2(1, N=100)=67.24$, $p<.01$; *pressar*: $\chi^2(1, N=100)=57.76$, $p<.01$].

Moreover, a comparison of the results of the two experiments shows that the listeners were more certain of their judgments in the second experiment. The listeners preferred the non-linear lengthening pattern in 76% of the judgments in the first experiment while the stimuli where the short vowel remained short (weakened contrast) in the second experiment were preferred in 87% of the judgements (cf. Tables 3 and 4).

5. Discussion

The first perceptual experiment in this study showed that a majority of the listeners were sensitive to the durational differences they were exposed to during the experiment. These listeners preferred a non-linear lengthening of focally accented CVC:-syllables to a linear expansion. In other words, the listeners preferred the kind of lengthening patterns previously observed in production studies (e.g. Bannert, 1979; Fant *et al.*, 1991; Heldner & Strangert, 2001). Taken together, these findings show that a linear time expansion is neither an appropriate description of focal accent lengthening in Swedish from a production perspective, nor from a perception one. Moreover, Swedish seems to be different in this respect from other languages. For example, as shown by Sluijter (1995), the temporal contribution of an accent in Dutch and English is an almost linear time expansion of the entire word.

Previous production studies have shown that the non-linear lengthening of CVC:-syllables has two distinguishing features as compared to the lengthening of CV:C-syllables: firstly, the phonologically short segment (i.e. the vowel) restricts lengthening, and, secondly, there is a strengthening of the temporal contrast between short and long segments within the syllable (e.g. Bannert, 1979; Heldner & Strangert, 2001).

The second perceptual experiment has shown that it is more important for the perceived naturalness that the short vowel remains short than that the temporal contrast be strengthened. This is an indication that the linearly lengthened stimuli in the first experiment were judged to be less natural than non-linearly primarily because the short vowel was lengthened. Perhaps more importantly, it indicates the existence of an expandability constraint at the phonemic level analogous to the compressibility constraint proposed by Klatt (e.g. Klatt, 1976). The precise explanation of an expandability constraint in short vowels in Swedish remains to be discovered, but it is certainly not physiologically determined, as compressibility constraints will often be. A phonological explanation in terms of maintaining sufficient dissimilarity between long and short vowels (or CVC:- and CV:C-syllables) when lengthened (Bannert, 1979) would seem to be more plausible.

Moreover, the second perceptual experiment indicates that the strengthening of temporal contrast observed in production data is a consequence of the expandability constraint in the short vowel. If the vowel remains short while the syllable is lengthened, the following consonant has to be lengthened more than the syllable as a whole.

To summarize, perceived naturalness increases as a result of modeling focal accent lengthening of a CVC:-syllable with a non-linear pattern as compared to using a linear lengthening pattern. Moreover, the most important feature of this non-linear pattern is for the

vowel to be maintained short. For text-to-speech applications, this clearly implies that the lengthening of CVC:-syllables should be modeled differently from that of CV:C-syllables.

Acknowledgements

The author would like to thank Jonas Beskow for technical assistance.

References

Bannert, R. (1979) The effect of sentence accent on quantity. In *Proceedings of the Ninth International Congress of Phonetic Sciences*, pp. 253-259. Copenhagen: Institute of Phonetics, University of Copenhagen.

Elert, C.-C. (1964) *Phonologic studies of quantity in Swedish.* Uppsala: Almqvist & Wiksell.

Fant, G., Kruckenberg, A. & Nord, L. (1991) Durational correlates of stress in Swedish, French and English, *Journal of Phonetics,* **19**, 351-365.

Heldner, M. & Strangert, E. (2001) Temporal effects of focus in Swedish, *Journal of Phonetics,* **29**(3), 329-361.

Klatt, D. H. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, *Journal of the Acoustical Society of America,* **59**(5), 1208-1221.

Sjölander, K. & Beskow, J. (2000) WaveSurfer. Stockholm: Centre for Speech Technology (CTT) at KTH. Available for download at http://www.speech.kth.se/wavesurfer/.

Sluijter, A. M. C. (1995) *Phonetic correlates of stress and accent.* The Hague: Holland Academic Graphics.

Paper VI In *Focal accent – f_0 movements and beyond*
pp. 111–133

On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish[1]

Mattias Heldner

This study shows that increases in overall intensity and spectral emphasis are reliable acoustic correlates of focal accents in Swedish. They are both reliable in the sense that there are statistically significant differences between focally accented words and non-focal ones for a variety of words, in any position of the phrase and for all speakers in the analyzed materials, and in the sense of their being useful for automatic detection of focal accents. Moreover, spectral emphasis turns out to be the more reliable correlate, as the influence on it of position in the phrase, word accent and vowel height was less pronounced and as it proved a better predictor of focal accents in general and for a majority of the speakers. Finally, the study has resulted in data for overall intensity and spectral emphasis that might prove important in modeling for speech synthesis.

1. Introduction

This study deals with the acoustic signaling of focal accent in Swedish, and in particular with the reliability of two acoustic features – overall intensity and spectral emphasis – that have been mentioned among the acoustic correlates of focal accents. 'Focal accent' is a term used in the Swedish intonation model about an accent signaling that a word (or some other constituent within a phrase which may be smaller or larger) is 'focused' or 'in focus' (Bruce, 1977; Bruce & Gårding, 1978; Gårding & Bruce, 1981; Bruce, Granström, Gustafson, Horne, House & Touati, 1997; Bruce, 1999). Overall intensity and spectral emphasis, furthermore, represent two different operationalizations of loudness. Overall intensity, as the name suggests, is the intensity (or SPL) of the whole spectrum, as opposed to spectral emphasis, which may be described as the relative intensity in the higher frequency bands. Two aspects of the reliability of these acoustic correlates will be considered. The first is investigating whether there are statistically significant differences between focally accented and non-focal words in paradigmatic – or between-phrase – comparisons. The second approach is exploring the usefulness of these correlates for the detection of focally accented words within phrases, i.e. in syntagmatic comparisons.

It is generally agreed that the most important and reliable acoustic correlates of accents marking focus in languages such as English, Dutch and Swedish are fundamental frequency (f_0) movements (e.g. Bolinger, 1958; Fry, 1958; van Katwijk, 1974; Bruce, 1977; Beckman, 1986; t' Hart, Collier & Cohen, 1990) and prolonged segmental durations (e.g. Cooper, Eady

[1] This material has been submitted for publication 2001.

& Mueller, 1985; Eefting, 1991; Fant, Kruckenberg & Nord, 1991; Sluijter & van Heuven, 1995; Cambier-Langeveld & Turk, 1999; Turk & White, 1999; Heldner & Strangert, 2001). At the same time, some kind of loudness variation is also intuitively felt to be part of the signaling of prominence distinctions (cf. Lehiste & Peterson, 1959). Indeed, increases in loudness, as measured using several different operationalizations such as overall intensity (e.g. Fry, 1955), intensity summed over time (Beckman, 1986), spectral tilt (Sluijter, Shattuck-Hufnagel, Stevens & van Heuven, 1995), and spectral balance (Sluijter & van Heuven, 1996) have also been shown to be reliable acoustic correlates of accents.

Thus, f_0 and duration, as well as the different operationalizations of loudness are all potentially useful for automatic detection of accented words. In fact, systems for automatic classification of prosodic categories, including detection of accented words, typically use some combination of duration, f_0 and overall intensity (or energy) features (e.g. House & Bruce, 1990; Campbell, 1992; Campbell, 1994; Wightman & Ostendorf, 1994; Sautermeister & Lyberg, 1996; Ostendorf & Ross, 1997; Nöth, Batliner, Kießling, Kompe & Niemann, 2000; Shriberg, Stolcke, Hakkani-Tür & Tür, 2000). Although less frequent, various features related to the slope of the spectrum (e.g. spectral balance, spectral emphasis or spectral tilt) have also been exploited for automatic detection of prominence distinctions (e.g. Campbell, 1995; Sluijter, Shattuck-Hufnagel, Stevens & van Heuven, 1995; Sluijter & van Heuven, 1996; van Kuijk & Boves, 1999).

Just as there are several terms to denote the phenomena related to the slope of the spectrum (i.e. spectral balance, spectral emphasis, and spectral tilt), there are several methods for measuring them. Furthermore, there seems to be no consensus as to which term is to be associated with which method. Therefore, it is tentatively proposed that there are two classes of measures, which will be referred to as 'spectral tilt' and 'spectral emphasis'. 'Spectral tilt' will be used for measures explicitly representing the slope of the spectrum, while 'spectral emphasis' will be used for measures of the relative energy in the higher frequency bands, or, put differently, the relative contribution of the high frequency parts of the spectrum to the overall intensity. Although the two classes are related to each other, spectral emphasis is – as will be shown below – distinct from spectral tilt in several respects, a salient one being that an increase in spectral emphasis results in a decrease in spectral tilt.

A commonly used measure of spectral tilt is the difference (in dB) between the first harmonic (H1) and the strongest harmonic in the third formant peak (A3) with corrections (marked by asterisks) for the influence of the first formant on H1 and of the first and second formants on A3. This spectral tilt measure is thus defined as H1*-A3* (e.g. Stevens & Hanson, 1994; Sluijter *et al.*, 1995). A related estimate of spectral tilt is the difference between the first and second harmonics (H1-H2) (Jackson, Ladefoged, Huffman & Antoñanzas-Barroso, 1985; Titze & Sundberg, 1992; Campbell, 1995; Campbell & Beckman, 1997).

There exist several measures that would fall into the spectral emphasis category. In the influential work by Sluijter & van Heuven (1996) a measure called 'spectral balance' was defined as the intensity in four contiguous frequency bands: 0-0.5, 0.5-1, 1-2, 2-4 kHz. Moreover, an estimate referred to as 'spectral tilt' and used in recent studies by Fant and colleagues (Fant, 1997; Fant, Kruckenberg & Liljencrants, 2000a; Fant, Kruckenberg, Liljencrants & Hertegård, 2000c) is the difference (in dB) between signals with a high frequency pre-emphasis and a flat frequency weighting (defined as SPHL-SPL). Several authors have also measured spectral emphasis as the difference between the overall intensity and the intensity in a low-pass-filtered signal (e.g. Childers & Lee, 1991; Campbell, 1995; Traunmüller, 1997; Traunmüller & Eriksson, 2000). The latter methods differ mainly in the low-pass filter cut-off frequency.

Several spectral emphasis measures of the last mentioned type were also used in a previous study of our own (Heldner, Strangert & Deschamps, 1999). These measures included one calculating the difference (in dB) between the overall intensity and the intensity in a signal that was low-pass filtered at 1.5 times the f_0 mean for each utterance (as was also done in Traunmüller, 1997; Traunmüller & Eriksson, 2000). The other measures were inspired by the work of Sluijter & van Heuven (1996). In these measures, too, the difference between the overall intensity and the intensity in a low-pass filtered signal was calculated, but fixed low-pass filters with cut-off frequencies at 0.5 kHz, 1 kHz and 2 kHz were used. The rationale behind a filter cut-off frequency at 1.5 times f_0 is to 'separate' the fundamental from the rest of the harmonics (the second harmonic being at 2 times f_0) and to obtain a normalized measure of the energy in the higher frequency bands. (Strictly speaking, however, the filter has a slope of 12 dB/octave and is only attenuating the rest of the harmonics and especially the second harmonic will be included to some extent.) However, determining the low-pass filter from the f_0 mean of a whole utterance does not seem altogether satisfactory. In the case where f_0 is below the f_0 mean of the whole utterance, more energy will pass through the filter than just the fundamental thereby resulting in a lower spectral emphasis value. Similarly, when f_0 is above the mean, the result will be a higher value. To overcome this problem, we have developed a new and fully automatic technique for measuring spectral emphasis applying a dynamic low-pass filter with a cut-off frequency following the course of the fundamental frequency. This technique will be described in more detail below (section 2.2.).

Although several acoustic features have been shown to be reliable correlates of accentuation, and thus also potentially useful for automatic detection, this investigation has been restricted to the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. One approach to this subject is paradigmatic (or between-phrase) comparisons of focally accented and non-focal words. If the correlates are to be considered reliable, these comparisons should establish statistically significant differences between focal and non-focal words. Previous work in this area includes a series of studies by Fant and his associates. Fant *et al.* (2000a) recently summarized their own work on acoustic correlates of prominence in Swedish in general and of focal accents in particular. Regarding the correlates of interest in the present study, they reported the gain in overall intensity (or SPL) in focally accented words compared to non-focal to be in the order of 4–6 dB. The corresponding gain in their measure of 'spectral tilt' (SPLH-SPL) was in the order of 2–3 dB. These results were based on five speakers' readings of a five-word sentence occurring in six versions, one of which had a neutral reading and the rest a systematically varied focal accent distribution. Fant *et al.* (2000a) concluded that overall intensity and spectral tilt (i.e. SPLH-SPL) are fairly reliable correlates of focal accents in Swedish. In the present study, additional data for non-focal and focally accented words were collected using a larger and more varied material.

It is well known that the overall intensity of the human voice increases with fundamental frequency, at least up to a mid-frequency of the speaker's f_0-range (e.g. Fant *et al.*, 2000a; Fant, Kruckenberg & Liljencrants, 2000b). For example, an increase in fundamental frequency of six semitones is typically accompanied by an increase in overall intensity of about 6 dB, mainly due to increased voice source amplitude and a larger number of excitations per second. Inversely, a decrease in fundamental frequency is typically accompanied by decreased overall intensity. Pierrehumbert (1979) observed that the general downdrift of the fundamental frequency over the course of an intonation group (a tendency that has been observed in many languages) was accompanied by a downdrift in overall intensity of 3–4 dB. Thus, there may be an influence (at least an indirect one) of position on overall intensity and possibly also on spectral emphasis. Moreover, given the covariation of

overall intensity and fundamental frequency, it also seems warranted to examine if the differences in f_0 patterns between pre- and post-focal words in Swedish, that is, a compressed pitch range after the focal accent (Bruce, 1982), are reflected in the overall intensity and spectral emphasis patterns.

For this reason, besides treating the effects of focal accents, we will also touch upon the possible influence of position on overall intensity and spectral emphasis; that is, position of the focally accented word in the phrase and position and distance of non-focal words relative to the focally accented word. If the correlates are to be considered reliable, there should be significant differences between focal and non-focal words in all positions in the phrase. Moreover, if positional influences do exist, they might prove important in modeling for synthesis. Therefore, the results from different positions will be presented separately both in the paradigmatic comparisons and in the detection experiment.

Another approach to studying the reliability of overall intensity and spectral emphasis as acoustic correlates is investigating to what extent focally accented words may be detected automatically on the sole basis of these correlates. Given such an approach, a high degree of correct detections will obviously have to be taken to indicate high reliability. The work on automatic detection of focal accents in Swedish using overall intensity and spectral emphasis was initiated by (Heldner *et al.*, 1999) in a study where several measures of overall intensity and spectral emphasis were evaluated. As mentioned earlier, these spectral emphasis measures were all calculated as the difference (in dB) between the overall intensity and the intensity in a low-pass filtered signal, and differed only in the choice of low-pass filter cut-off frequency. One of the measures used a low-pass filter at 1.5 times the f_0 mean for each utterance, and the others used fixed low-pass filters with cut-off frequencies at 0.5 kHz, 1 kHz and 2 kHz, respectively. These experiments showed that overall intensity generally scored better than the different spectral emphasis measures. Moreover, the spectral emphasis measures using low-pass filters adjusted to the f_0 mean of the utterance resulted in more correct detections than those using fixed cut-off frequencies. However, as noted above, none of these spectral emphasis measures seemed satisfactory, as they might have been dependent on f_0 and might have favored words with higher f_0 than the mean and disfavored those with lower f_0 than the mean. Although this probably meant favoring focally accented words, it might also have favored words in phrase initial position and disfavored final words given a general declining trend in f_0 over the course of the utterance.

A solution to this problem would be a dynamic low-pass filter with a cut-off frequency following the course of the fundamental frequency. Using a dynamic low-pass filter had not been feasible in the previous study (Heldner *et al.*, 1999), for lack of adequate tools. Since then, however, tools using this kind of filters have been developed. In the present study, this improved technique for measuring spectral emphasis was used for revisiting automatic detection of focal accents in Swedish. In addition, we wished to test whether this new technique yields higher recognition scores than the previous method and, moreover, whether overall intensity is a better predictor than the improved spectral emphasis measure.

To summarize, then, the primary aim of this study is to assess the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. This problem is approached from two angles. The first consists in paradigmatic comparisons of non-focal and focally accented words, using statistical methods to assess the reliability of the correlates. Here, for the correlates to be considered reliable, the experiment must establish statistically significant differences between focal and non-focal versions of words for all speakers, for all words and in all positions in the phrase. The second approach to the reliability of overall intensity and spectral emphasis is investigating to what extent focally accented words may be detected automatically using these correlates. More exactly, what was being evaluated here was the usefulness of overall intensity and an improved spectral

emphasis measure as predictors in an automatic focal accent detector for Swedish. If the correlates are to be considered reliable, automatic detection using these correlates should yield a fairly high degree of correct detections. A secondary aim of this research is to collect data for overall intensity and spectral emphasis to be used in modeling for speech synthesis.

2. Method

Recordings taken from three different sets of phrases were used for both the paradigmatic comparisons and for the detection experiment. However, the material was primarily designed for paradigmatic comparisons. Two of the phrase sets were recorded for a study on temporal effects of focal accents in Swedish (Heldner & Strangert, 2001). A short description of the material and the recording procedures will be provided below. Although the composition of the three sets was different, they all contained short, mainly meaningful Swedish phrases or sentences each corresponding to one prosodic phrase. All the words were disyllabic and stress was always on the first syllable. The material was chosen to cover effects of position in the phrase, vowel quality and quantity, and word accents. Taken together it provides a relatively large basis for generalizations. The entire material was manually segmented into phonemic units and overall intensity and spectral emphasis were measured for each unit. The measurements were subsequently used for the paradigmatic comparisons as well as in the detection experiment.

2.1. Analyzed material

The first recording was based on 40 phrases, where 40 verbs occurred in medial position in the carrier phrase *Mannen VERB kvinnan*. The verbs were chosen so as to balance the number of phonologically long and short vowels in the stressed syllables, of open and closed vowels in the stressed syllables, of accent I and II words, and also to include a variety of consonantal contexts. Two examples are *Mannen biter kvinnan* ‘The man bites the woman’ and *Mannen kammar kvinnan* ‘The man combs the woman’. A list of all the verbs together with transcriptions is found in the Appendix. Focal accents were elicited on each word in each phrase using questions. Two male and two female speakers read each question-answer pair once yielding a total of 480 recorded phrases.

While the first recording was recorded specifically for the present study, the second and third ones were also used in a previous study (Heldner & Strangert, 2001). The second recording was based on six phrases. The content words *mannen* ‘the man’, *kvinnan* ‘the woman’, and *barnen* ‘the children’, separated by *och* ‘and’, were combined in order for all the three words to occur successively in initial, medial and final position (e.g. *Mannen och kvinnan och barnen* ‘The man and the woman and the children’). As in the other recordings, each phrase occurred in three versions, with a focal accent on either the initial, medial or the final content word in the phrase. However, instead of being elicited using questions, the focal accents were indicated by means of capital letters and the speakers were instructed to emphasize the words with capitalization. Each version of each phrase was repeated five times by each speaker. There were three male speakers and one female. The total number of phrases was 360.

The third recording was based on two phrases: *Mannen tömmer dammen* ‘The man is draining the pond’ and *Kvinnan dammar kannan* ‘The woman is dusting the jug’. These phrases occurred as answers in a question-answer context. The questions were designed to elicit focal accents on each of the three words in turn. Thus, each phrase occurred in three

versions and each word occurred in one focal and two non-focal conditions. Three female and three male speakers read ten repetitions of each version yielding a total of 360 phrases.

All the speakers participating in the recordings were native speakers of Central Swedish without any strong dialectal influence and without any known hearing or speaking disorders. They were not paid for their services.

All the recordings were made in a sound-treated room with a high quality condenser microphone mounted on a headset, so that a constant distance from the mouth to the microphone was maintained. The gain control on the microphone preamplifier was adjusted to obtain approximately the same sound pressure level for all speakers. The productions were monitored by the speakers themselves and by the experimenter. Either the speakers or the experimenter could decide whether a phrase should be reread. In addition, the author together with a colleague listened to all the phrases after the recording sessions to eliminate erroneous readings. However, no phrases had to be discarded in this process.

2.2. Acoustic measurements

A number of different measures of overall intensity and spectral emphasis were derived from the speech signal. Among these, there were four overall intensity measures differing in the time over which intensity was averaged (i.e. in the amount of smoothing) including a "standard method" using a 25 ms Hamming window with 12.5 ms frame advance as well as the means of the "standard method" across each segment, syllable and word (as in Heldner *et al.*, 1999). The overall intensity values were measured in dB relative to the arbitrary reference level of the maximum amplitude of a 16 bit AD converter.

There were eight spectral emphasis measures differing in how they were measured in the frequency domain as well as in the time over which they were averaged. To enable comparisons with the 'best' spectral emphasis measure in our previous study (Heldner *et al.*, 1999) the procedure used there was replicated. Thus, the difference (in dB) between the overall intensity and the intensity in a signal that was low-pass filtered at 1.5 times the f_0 mean for each utterance was calculated.

However, as was noted already in the introduction, using a low-pass filter determined by the f_0 mean of each utterance is not altogether satisfactory, as it is sensitive to f_0 movements above and below the f_0 mean. Therefore, a new implementation of spectral emphasis applying a dynamic low-pass filter with a cut-off frequency following the course of the fundamental frequency was developed. The low-pass filter is a two-pole filter with a slope of 12 dB per octave. As for overall intensity, four measures differing in the time over which they were averaged were calculated for both these spectral emphasis measures. As spectral emphasis is meant to reflect features present in the voice source, it was only calculated for segments classified as voiced by the f_0 analysis, and the values in the unvoiced segments were set to zero. The spectral emphasis measures were implemented in an ESPS/Waves+™ environment. They were adopted with a full understanding of the fact that they will be influenced not only by the source, but also by the filter function (Fant *et al.*, 2000a). However, no corrections for the influence of supralaryngeal settings were made in this study.

Thus, four overall intensity measures, four spectral emphasis measures using a filter determined by the f_0 mean, and four using a dynamic filter were calculated. However, to reduce the amount of figures, the presentation of data for the full set of measures will be restricted to the detection experiment. In the paradigmatic comparisons, only the means per segment of overall intensity and the new spectral emphasis measure will be presented.

2.3. The focal accent detector

The focal accent detector was based on the assumption that the focally accented word is the most prominent word in the phrase. It was assumed, in other words, that there can be only one focally accented word in each phrase. It was also assumed that these prominence relations would show up in the measures of overall intensity and spectral emphasis. The detector was thus expected to select the word containing the highest value in the phrase for a given measure and classify it as focally accented. Separate detection experiments were run for each measure of overall intensity and spectral emphasis and for each recording. The performance of the detector was evaluated based on comparisons with the intended positions for focal accents in the test materials as produced by the speakers and as verified through listening during and after the recordings.

3. Results: Paradigmatic comparisons

In this section, the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents is assessed by comparing focally accented and non-focal words from different phrases (i.e. paradigmatic comparisons). Here, to be considered reliable, statistically significant differences between focal and non-focal words should be established for all words, in all positions in the phrase and for all speakers. In addition, it is examined whether the choice of non-focal reference matters when comparing focal and non-focal words. The presentation of data in the paradigmatic comparisons will be restricted to the means per segment. Furthermore, only data for the new spectral emphasis measure will be presented (see section 2.2.).

The experiment is divided into four subsections. The first three deal with the influence of focal accents on overall intensity and spectral emphasis. The fourth section examines non-focal words, and specifically the effects of position and distance relative to the focally accented word on the non-focal words.

3.1. Is an increase in overall intensity and spectral emphasis a reliable correlate of focal accents? Generalizations across words

In the first subsection, we will address the question whether overall intensity and spectral emphasis are reliable correlates of focal accents in the sense that significant differences show up between focal and non-focal versions of all words. Thus, the basis for generalizations across words is made as broad as possible. Data for the 40 different verbs in medial position in the phrase from the first recording (see section 2.1.) were examined. Thus, overall intensity and spectral emphasis measures from 160 focally accented words were compared with those of 320 non-focal words, in all 480 words. Within each of these words, the overall intensity and spectral emphasis within the consonant or consonants preceding the stressed vowel C(C), the stressed vowel V, the consonant following the vowel C, and the vowel and consonant in the unstressed syllable VC, were compared in focal and non-focal conditions. The reason why the segments in VC were not separated was the difficulties in determining the boundary between the vowel and the consonant which was always an /r/.

The data was analyzed in eight ANOVAs. However, prior to these ANOVAs, two MANOVAs, one for overall intensity and another for spectral emphasis, with four dependent variables each were run to control for correlations between the dependent variables. As the qualitative results of the MANOVAs did not differ from those of the ANOVAs, these results have been omitted in the following.

There was one ANOVA model for overall intensity and another for spectral emphasis for each segment, that is C(C), V, C and VC. The between-subjects factors in each model were Focal accent (focal vs. non-focal; i.e. the two non-focal versions of each word were collapsed into one non-focal category) and Word with 40 different levels. Focal accent was included as a fixed and Word as a random factor.

Table I shows descriptive statistics for overall intensity and spectral emphasis for each measured segment in the focal and non-focal conditions. Apparently, the differences in overall intensity and spectral emphasis between focal and non-focal words ranged from no difference to an increase of 3 dB in the different segments. Clearly, the overall intensity and spectral emphasis increased at least in the vowel in the stressed syllable and in the unstressed syllable.

Table I. The means and standard deviations (in dB) across all words of overall intensity and spectral emphasis in focal accented and non-focal words. N focal=160; N non-focal=320.

		Overall intensity		Spectral emphasis	
		Mean	Std. Dev.	Mean	Std. Dev.
C(C)	focal	-13.8	5.6	3.1	2.4
	non-focal	-14.0	5.9	3.3	2.4
V	focal	-5.2	2.1	9.0	2.1
	non-focal	-8.1	3.1	7.0	1.9
C	focal	-15.3	7.2	3.2	3.4
	non-focal	-16.1	6.4	2.7	2.6
VC	focal	-6.7	2.3	9.0	1.9
	non-focal	-10.2	3.3	6.0	1.7

As for the outcome of the ANOVAs, the results for overall intensity will be presented first, followed by those for spectral emphasis. First, there was a significant increase in overall intensity in focally accented words, at least in the V and VC segments. The increase did not differ significantly among the different words. This is shown by the fact that the main effect of Focal accent was significant in V and VC [$F_{C(C)}$ (1,39) = 0.1; p = 0.73; F_{V} (1,39) = 234.4; p <0.01; F_{C} (1,39) = 3.6; p = 0.06; F_{VC} (1,39) = 610.3; p <0.01] while the interaction of Focal accent and Word was not significant for any of the segments [$F_{C(C)}$ (39,400) = 0.6; p = 0.97; F_{V} (39,400) = 0.5; p = 0.99; F_{C} (39,400) = 1.1; p = 0.29; F_{VC} (39,400) = 0.2; p = 1].

Second, as expected, there were significant differences in overall intensity among the different vowels and consonants. The main effect of word was significant in all segments [$F_{C(C)}$ (39,39) = 13.1; p <0.01; F_{V} (39,39) = 3.9; p <0.01; F_{C} (39,39) = 21.2; p <0.01; F_{VC} (39,39) = 4.8; p <0.01].

Turning now to the results for spectral emphasis, the first observation was that also the spectral emphasis increased significantly in focally accented words compared to non-focal words, at least in the V, C and VC segments. The main effect of Focal accent was significant for all segments but C(C) [$F_{C(C)}$ (1,39) = 4.1; p = 0.05; F_{V} (1,39) = 181.3; p <0.01; F_{C} (1,39) = 6.2; p = 0.02; F_{VC} (1,39) = 419.4; p <0.01]. There were hardly any significant differences in the amount of increase for the different words as the interaction of Focal accent and Word was significant for C only [$F_{C(C)}$ (39,400) = 0.3; p = 1; F_{V} (39,400) = 0.6; p = 0.96; F_{C} (39,400) = 2.5; p <0.01; F_{VC} (39,400) = 0.8; p = 0.76].

Finally, there were significant differences in spectral emphasis among the different vowels and consonants as the main effect of word was significant in all segments [$F_{C(C)}$ (39,39) = 30.2; p <0.01; F_{V} (39,39) = 5.8; p <0.01; F_{C} (39,39) = 18.3; p <0.01; F_{VC} (39,39) = 3.3; p <0.01].

3.2. Is an increase in overall intensity and spectral emphasis a reliable correlate of focal accents? Generalizations across different positions in the phrase

In the second section, we will continue investigating whether overall intensity and spectral emphasis are reliable correlates of focal accents also in the sense that significant differences can be found between focal and non-focal words in all positions in the phrase. In addition, we will investigate *how* position in the phrase influences overall intensity and spectral emphasis. For example, do focally accented and non-focal words in phrase final position have a lower overall intensity and spectral emphasis than in other positions?

When dealing with positional effects, it is crucial that the same words are compared in all positions. Otherwise, other factors such as vowel intrinsic intensities might rule out the positional differences. Thus, only data from the second recording (see section 2.1.) will be examined here. This recording was specifically designed to explore possible influences of position in the phrase (and of position and distance relative to the focally accented word, see also 3.4.) with *mannen*, *kvinnan* and *barnen* occurring in initial, medial and final position in the phrase. To reduce the amount of figures, the analyses in this section will be restricted to the vowels in the stressed and unstressed syllables in the word *mannen* only.

Four ANOVA models were designed and overall intensity and spectral emphasis were examined in separate models. The dependent variables were values taken from the vowels in the stressed and unstressed syllables in *mannen*. One random and two fixed factors were included in each design: Focal accent (focal vs. non-focal) and Position (initial vs. medial vs. final) were fixed, while Speaker (4 levels) was included as a random factor. Table II shows descriptive statistics for the vowels in the stressed and unstressed syllables in *mannen* in focally accented and non-focal words in different positions in the phrase.

The analyses showed significant positional effects on overall intensity as well as on spectral emphasis. The interaction between Focal accent and Position was significant for both vowels in the analyses of overall intensity [stressed /a/: F (2,6) = 8.9; p = 0.02; unstressed /e/: F (2,6) = 7.4; p = 0.02] but only for the vowel in the stressed syllable for spectral emphasis [/a/: F (2,6) = 12.3; p <0.01; /e/: F (2,6) = 2.7; p = 0.15].

The difference between focal and non-focal words increased the further to the right in the phrase the word was located. For example, the gains in overall intensity in the stressed vowel /a/ were about 3, 5 and 6 dB in initial, medial and final position, respectively, and the corresponding gains in spectral emphasis were 2, 3 and 4 dB (cf. Table II). These effects were mainly due to the fact that the values in the non-focal words decreased faster with position than the focal words (cf. Table II). A marginal downdrift in overall intensity over the course of the utterance was also observed in the focally accented words, while the non-focal words decreased considerably more. Although a similar pattern was found for spectral emphasis in the stressed vowels, there was no downdrift in the focally accented words while the downdrift was substantial only in the non-focal words. However, in the unstressed vowels, the focally accented and non-focal words drifted downwards to about equal amounts.

Table II. The means and standard deviations (in dB) of overall intensity and spectral emphasis for the vowels in the stressed and unstressed syllables in 'mannen' in focal accented and non-focal words in different positions in the phrase. N focal in each position=40; N non-focal in each position=80.

		Overall intensity		Spectral emphasis	
/a/		Mean	Std. Dev.	Mean	Std. Dev.
initial	focal	-3.1	1.3	10.3	2.3
initial	non-focal	-5.7	2.8	8.0	2.2
medial	focal	-2.6	1.2	10.3	2.2
medial	non-focal	-7.4	2.9	7.4	1.6
final	focal	-4.3	2.1	10.3	2.2
final	non-focal	-10.6	3.6	6.0	1.2
/e/		Mean	Std. Dev.	Mean	Std. Dev.
initial	focal	-7.9	2.1	8.7	3.0
initial	non-focal	-9.1	3.3	6.9	2.8
medial	focal	-7.5	2.1	8.2	2.2
medial	non-focal	-10.7	3.2	5.0	1.5
final	focal	-8.5	2.2	6.4	1.3
final	non-focal	-14.4	3.7	4.5	1.2

Finally, the analyses showed that there were speaker differences in the amount of increase in different positions, as the interaction between Focal accent, Position and Speaker was significant in three out of four ANOVAs. There were significant differences in overall intensity for both vowels [/a/: F (6,336) = 2.6; p = 0.02; /e/: F (3,336) = 4.5; p <0.01]. For spectral emphasis, only the unstressed vowel showed significant results for this effect [/a/: F (6,336) = 1.9; p = 0.08; /e/: F (3,336) = 9.4; p <0.01]. Still, the focally accented words had a higher overall intensity and spectral emphasis than non-focal ones for all speakers and in all positions.

3.3. Is an increase in overall intensity and spectral emphasis a reliable correlate of focal accents? Generalizations across speakers

Also the third section of the paradigmatic comparisons deals with the reliability of overall intensity and spectral emphasis as correlates of focal accent, but here in the sense that there should be significant differences between focal and non-focal words for all speakers. The third recording (see section 2.1.) provides the best basis for generalizations across speakers. As in the previous section, overall intensity and spectral emphasis were examined in separate models and the dependent variables were values taken from the vowels in the stressed (V1) and unstressed (V2) syllables in all words in all positions in the third recording. The total number of words in each model was 1080 (360 phrases x 3 words). Thus, there were four ANOVA models. Speaker (6 levels) was included as a random factor and Focal accent (focal vs. non-focal) as a fixed factor.

These analyses showed that all speakers increased overall intensity and spectral emphasis significantly in focally accented words compared to non-focal. The main effect of Focal accent was significant in the models for overall intensity [F_{V1}(1,5) = 19.9; p <0.01; F_{V2}(1,5)

= 16.2; $p = 0.01$] as well as in those for spectral emphasis [$F_{V1}(1,5) = 45.8$; $p < 0.01$; $F_{V2}(1,5) = 85.1$; $p < 0.01$]. Across the six speakers there was an average increase in overall intensity of about 3 dB in the stressed and in the unstressed vowels. The corresponding value for spectral emphasis was about 2 dB.

In addition, there were significant speaker differences. The interaction of Focal accent and Speaker was significant in the models for overall intensity [$F_{V1}(5,1068) = 12.1$; $p < 0.01$; $F_{V2}(5,1068) = 6.5$; $p < 0.01$] as well as in those for spectral emphasis [$F_{V1}(5,1068) = 6.3$; $p < 0.01$; $F_{V2}(5,1068) = 4.6$; $p < 0.01$]. Moreover, the main effect of Speaker was significant in three out of four models: Overall intensity [$F_{V1}(5,5) = 2.3$; $p = 0.19$; $F_{V2}(5,5) = 8.3$; $p = 0.02$], Spectral emphasis [$F_{V1}(5,5) = 34.5$; $p < 0.01$; $F_{V2}(5,5) = 55.4$; $p < 0.01$]. However, the speaker differences were only due to different amounts of increase. All speakers increased the overall intensity and spectral emphasis in focally accented words.

3.4. Does position and distance relative to the focally accented word influence the overall intensity and spectral emphasis of non-focal words?

The final section of the paradigmatic comparisons deals with two issues relating specifically to the overall intensity and spectral emphasis in the non-focal words. The first question concerns whether there are influences of position relative to the focally accented word; that is, whether post-focal words are different from pre-focal ones. Then there is the question whether there are effects of distance relative to the focally accented word. In other words, are post-focal words located two words after the focally accented word different from those one word after? Similarly, are pre-focal words whose location is two words before the focally accented word different from those located one word before? This experiment has been done primarily to find out whether the choice of non-focal reference matters when comparing focal and non-focal words.

First, to deal with the influences of position relative to the focally accented words, the data from the 40 different verbs in medial position in the phrase in the first recording (see section 2.1.) was reanalyzed. Again, in order to reduce the amount of figures the ANOVAs were restricted to the vowel in the stressed syllable and the unstressed final VC. As before, separate analyses were made for overall intensity and spectral emphasis. Thus, there were four ANOVA models with two independent variables in each. Position relative to the focally accented word (focal vs. pre-focal vs. post-focal) was included as a fixed factor and Word (40 levels) was included as a random factor. The variable Speaker was not included, as that would have eliminated all variance. Planned comparisons were performed to examine differences between pre- and post-focal words.

The analyses showed that post-focal words had significantly lower overall intensity than pre-focal in the stressed vowel as well as in the unstressed syllable [V: $F(1) = 257.2$; $p < 0.01$; VC: $F(1) = 938.7$; $p < 0.01$]. In the stressed vowel, the overall intensity was on average 3 dB lower in post-focal words as compared to pre-focal. The corresponding figure for the unstressed syllable was 5 dB. Furthermore, the analyses showed that post-focal words also had significantly lower spectral emphasis than pre-focal ones in the unstressed syllable, while there was no significant difference in the stressed vowel [V: $F(1) = 2.6$; $p = 0.11$; VC: $F(1) = 14.8$; $p < 0.01$]. However, the difference in the unstressed syllable was only of the order of 0.6 dB.

Next, to deal with the influences of distance relative to the focally accented word, data from the second recording (see section 2.1.) were reanalyzed. The dependent variables were the overall intensity and spectral emphasis in the stressed and unstressed vowels in *mannen.* The independent variables were Speaker (4 levels) and Position relative to the focally accented word nested under Position in the phrase. There were one focal and two non-focal

conditions in each position. In initial position, the non-focal words were either one or two words before the focally accented word (pre-focal -1 and pre-focal -2). In medial position the non-focal words were either one word before (pre-focal -1) or one word after (post-focal +1) the focally accented word. In final position, the non-focal words were either one or two words after the focally accented word (post-focal +1 and post-focal +2). Planned comparisons were then performed to examine differences between the non-focal conditions in each position. Figure 1 shows the means of overall intensity and spectral emphasis in the different positions and focal and non-focal conditions. Table III shows the results of these planned comparisons for overall intensity and spectral emphasis.

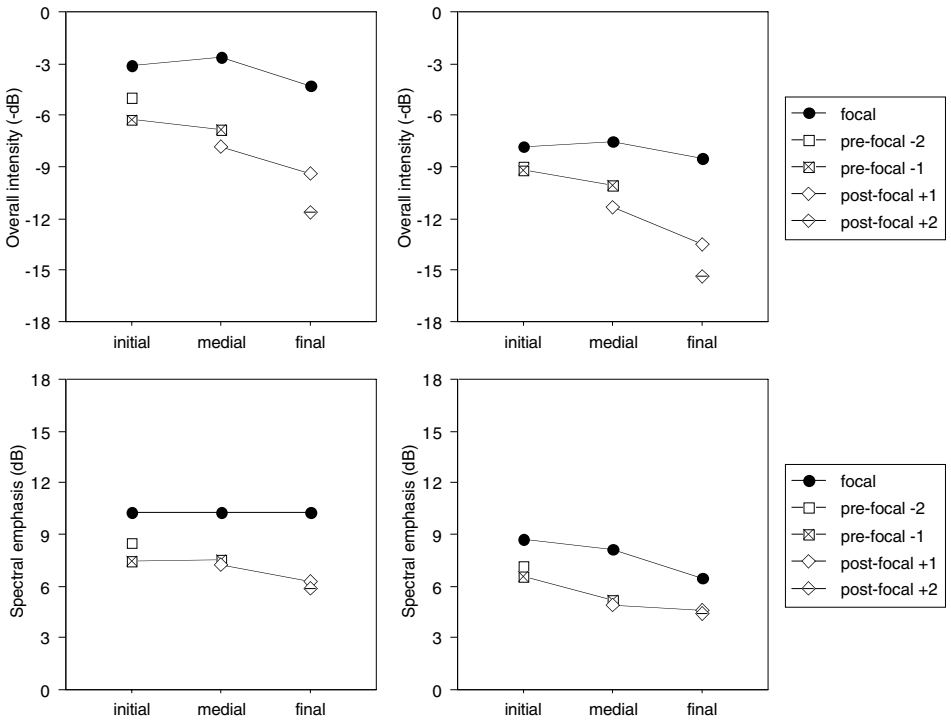

Figure 1. Means of overall intensity and of spectral emphasis (in dB) from the vowels in the stressed /a/ (left panels) and unstressed /e/ syllables (right panels) in 'mannen'. Overall intensity is shown in the top and spectral emphasis in the bottom panels. N in each phrase position=120.

A few observations can be made from Figure 1 and Table III. First, it seems that position and distance relative to the focally accented word influence overall intensity more than spectral emphasis.

In phrase-initial position, distance relative to the focally accented word (i.e. distance to the left of this word) affected both the overall intensity and spectral emphasis. The stressed vowel had about 1.3 dB higher overall intensity and 1 dB higher spectral emphasis when occurring two words before the focally accented word compared to one word before it. However, in the unstressed vowel, position relative to the focally accented word had a significant effect on spectral emphasis only, where the difference was 0.7 dB.

In medial position in the phrase – and just as in the re-analysis of the first recording above – position relative to the focally accented word affected overall intensity, whereas no effects

were recorded on spectral emphasis. The overall intensity in the stressed vowel was 1.3 dB lower in post-focal compared to pre-focal words. The difference in the unstressed vowel was 1 dB.

In phrase-final position, distance relative to the focally accented word (here: distance to the right of the focally accented word) had a significant effect on overall intensity but not on spectral emphasis. The stressed vowel had 2.3 dB and the unstressed 1.8 dB lower overall intensity when the word was two words after compared to one word after the focally accented word.

Table III. Planned comparisons of overall intensity and spectral emphasis values in the non-focal conditions for /a/ and /e/ in 'mannen'. N in each comparison = 40 vs. 40.

		Initial position (pre-focal-1 vs. pre-focal-2)	Medial position (pre-focal-1 vs. post-focal+1)	Final position (post-focal+1 vs. post-focal+2)
Overall intensity	/a/	$F(1) = 8.8$; $p < 0.01$	$F(1) = 5.6$; $p = 0.02$	$F(1) = 28.6$; $p < 0.01$
	/e/	$F(1) = 0.1$; $p = 0.70$	$F(1) = 8.8$; $p < 0.01$	$F(1) = 16.9$; $p < 0.01$
Spectral emphasis	/a/	$F(1) = 22.4$; $p < 0.01$	$F(1) = 1.2$; $p = 0.28$	$F(1) = 2.4$; $p = 0.12$
	/e/	$F(1) = 14.1$; $p < 0.01$	$F(1) = 2.8$; $p = 0.10$	$F(1) = 0.8$; $p = 0.37$

4. Discussion: Paradigmatic comparisons

The first part of the experiment with paradigmatic comparisons has shown that although there were differences among the words, focally accented words were characterized by statistically significant increases in overall intensity and spectral emphasis compared to non-focal words. Moreover, these effects were found primarily in the vowels in the stressed and unstressed syllables.

Across all 40 words, the increase in overall intensity was about 3 dB both in the stressed vowels and in the unstressed syllable. The corresponding values for spectral emphasis were 2 and 3 dB. Thus, our values were in the same range as those reported in (Fant *et al.*, 2000a) where an increase in overall intensity in the order of 4 to 6 dB and in spectral emphasis in the order of 2 to 3 dB was reported.

However, although the distributions of the values of focally accented and non-focal segments were statistically significant, they overlapped to a considerable extent (cf. the standard deviations in Table I). Clearly, focal accents do not always result in increased overall intensity and spectral emphasis in all segments in paradigmatic comparisons. Furthermore, the gains in both measures were dependent on the words. We would still argue, however, that both overall intensity and spectral emphasis are reliable correlates of focally accented words in the sense that there are statistically significant differences between focally accented and non-focal words. Furthermore, as the analyzed material was fairly varied (40 different disyllabic words differing in vowel quality and quantity, word accents and consonantal context), it seems reasonable to generalize at least across disyllabic words in medial position in the phrase.

The second part of the experiment has shown that there was a significant increase in overall intensity as well as in spectral emphasis in focally accented words in all positions in the phrase. Therefore, such increases may be considered as reliable correlates of focal accents also in the sense that they occur in all positions in the phrase. However, the

differences between focal and non-focal words increased significantly for both measures the further to the right in the phrase the words were located. Moreover, there was a downdrift over the utterance in both measures, similar to that previously observed for English by Pierrehumbert (1979). This downdrift was visible in the focally accented words as well as in the non-focal ones. The reason why the differences grew larger the further to the right in the phrase the words were located was that the non-focal words had a steeper downdrift than the focal ones. Thus, in addition to the effect of focal accents, there were clear positional effects on the amount of increase in overall intensity or spectral emphasis. Further research is needed to investigate whether these positional influences have any perceptual relevance.

The third section has shown that increases in overall intensity and spectral emphasis are reliable correlates of focal accents also in the sense that all speakers in this study employ them. It is true that there were only six speakers in the third recording, but the results generalize to all of them. This finding taken together with the results from the second section with another four speakers leads us to believe that it is also reasonable to generalize across Swedish speakers.

The fourth section has shown that, apart from the differences between focal and non-focal words established in the previous sections, there were also differences among the non-focal words depending on position and distance relative to the focally accented word. Thus, the choice of the non-focal reference also matters in comparisons of focal and non-focal words and especially for comparisons of overall intensity.

Position relative to the focally accented word affected overall intensity, as post-focal had lower overall intensity than pre-focal ones. These results bear strong resemblance to the previously observed differences in fundamental frequency between pre- and post-focal words in Swedish (Bruce, 1982). The compressed pitch range after the focally accented word is accompanied by lower overall intensity.

Finally, there were indications of effects on non-focal words of the distance relative to the focally accented word. Post-focal words two words after the focally accented word had lower overall intensity compared to those located one word after. Thus, the overall intensity was lower the further to the right of the focally accented word the non-focal word was situated. Inversely, pre-focal words two words before the focally accented word were observed to have both higher values of overall intensity and spectral emphasis than those occurring one word before. Thus, the values were higher the further to the left of the focally accented word the non-focal word was situated.

As we can see, the data reflect the general downdrifting trends across the utterance previously observed for fundamental frequency and overall intensity (Pierrehumbert, 1979). In addition, the effect of distance relative to the focally accented word indicates that the downdrifting trends were steeper *after* that word. This is also in line with the work of Gårding (e.g. Gårding, 1993) who observed that declination usually changes direction in connection with focal accents.

5. Results: The detection experiment

In the detection experiment, a different approach is taken to assess the reliability of overall intensity and spectral emphasis as acoustic correlates to focal accents in Swedish. Instead of making paradigmatic comparisons of focally accented and non-focal words the reliability is assessed by investigating to what extent it is possible to tell focally accented and non-focal words apart automatically using these correlates. As noted before, they should yield a high degree of correct detections in order to be considered reliable. In addition, the detection

scores for the improved spectral emphasis measure will be compared here with the method that gave the best results in the study by Heldner *et al.* (1999).

Table IV allows a comparison of the performance of the different measures of overall intensity and spectral emphasis as predictors of focal accents across all three recordings. This table shows the percentage of phrases across all three recordings where the highest value of overall intensity or spectral emphasis in the phrase was found in the focally accented word. Apparently, the best measure of overall intensity allowed detection of 69% of the focally accented words. The best measure of spectral emphasis using a low-pass filter determined by the f_0 mean in each utterance detected 63% of the focally accented words. However, the best measure of the improved method using a dynamic low-pass filter following the course of f_0 detected as much as 75%. Clearly, the new spectral emphasis measure improved the results as compared to those obtained with the previous method. Furthermore, detection using the new measure actually resulted in more correct detections than that based on overall intensity.

Table IV. Percentages correct detections for overall intensity and the different measures of spectral emphasis as well as for the different integration times. Data from all three recordings, N=1200.

	Overall intensity	Mean f_0 LP-filter	Dynamic LP-filter
25 ms Hamming	68%	57%	69%
Mean/Segment	69%	63%	75%
Mean/Syllable	66%	61%	74%
Mean/Word	65%	57%	65%

Regarding the different times over which the measures were averaged, it seems that some smoothing or averaging over a certain stretch of speech is favorable. The means across each segment was the best measure for overall intensity as well as for the different spectral emphasis measures. To reduce the amount of figures, only the results of the means per segment will be presented in the following. For the same reason, we will restrict ourselves to presenting data for the new spectral emphasis measure besides data for overall intensity. Turning now to a more detailed analysis of the detection scores, Tables V, VI and VII show the percentages correct detections for each word and each position in the phrase for each of the three recordings.

Table V. Percentages correct detections for overall intensity and spectral emphasis in the first recording. N=480.

	mannen	VERB	kvinnan	Totals
Overall intensity	98%	77%	26%	67%
Spectral emphasis	95%	87%	33%	71%

Table VI. Percentages correct detections for overall intensity and spectral emphasis in the second recording. N=360.

Initial position	mannen	kvinnan	barnen	Totals
Overall intensity	100%	98%	85%	94%
Spectral emphasis	98%	75%	95%	89%
Medial position	mannen	kvinnan	barnen	Totals
Overall intensity	78%	73%	63%	71%
Spectral emphasis	100%	85%	88%	91%
Final position	mannen	kvinnan	barnen	Totals
Overall intensity	73%	45%	45%	54%
Spectral emphasis	90%	68%	45%	68%

Table VII. Percentages correct detections for overall intensity and spectral emphasis in the third recording. N=360.

	mannen	tömmer	dammen	Totals
Overall intensity	92%	78%	65%	78%
Spectral emphasis	98%	90%	92%	93%
	kvinnan	dammar	kannan	Totals
Overall intensity	65%	88%	7%	53%
Spectral emphasis	37%	98%	15%	50%

Tables V, VI and VII show that the best detection results for overall intensity were achieved in phrase-initial position, where 91% correct detections (counts correct divided by the total counts) were obtained across all three recordings. The scores were lower (77%) in medial position, and approaching what you would expect to find by chance (38%) in final position in the phrase. For spectral emphasis, the best results were obtained in medial position in the phrase with 90% correct detections across all three recordings. The scores in initial position were 85% and in final position 49%. In general, the same relations were also present in the second recording analyzed by itself (c.f. Table VI). By comparing the same words in all positions in the phrase, the possibility that the differences were *only* due to the different words in the different positions can be ruled out. Thus, there is a genuine effect of position in the phrase on the detection scores. Still, there were differences that might be attributed to the specific sounds occurring in the words or to the particular speakers involved in the different recordings.

As the verbs in medial position in the phrase in the first recording provide the broadest basis for generalizations across words, we will examine these results in more detail. Thus, Table VIII presents the detection scores for accent I and II words as well as for words with

open and closed vowels in the stressed syllables separately. As there were only minor differences between words with long or short vowels (words with short vowels had 1–2% higher scores than those with long vowels), this detail was omitted from Table VIII. A number of observations can be made. As in the results across all three recordings, spectral emphasis was a better predictor of focal accents than overall intensity. 87% of the verbs were correctly detected using spectral emphasis and 77% using overall intensity. However, the more detailed analysis also revealed that word accent as well as vowel height in the stressed syllable affected the detection scores. Accent II words were detected correctly more often than accent I words and words with open vowels more often than those with closed vowels. Moreover, especially vowel height, but also word accent to some extent, had a greater influence on the overall intensity scores than on those for spectral emphasis.

Table VIII. Percentages correct detections for overall intensity and spectral emphasis for accent I and II words with open and closed vowels in the stressed syllables from the first recording. N=160.

Overall intensity	Open	Closed	Totals
Accent I	72%	53%	62%
Accent II	100%	82%	91%
Totals	86%	68%	77%
Spectral emphasis	**Open**	**Closed**	**Totals**
Accent I	80%	70%	75%
Accent II	100%	97%	99%
Totals	90%	84%	87%

Finally, a closer examination of the individual scores for the 13 speakers involved in the recordings (one of the speakers participated in two of the recordings) revealed that also the speaker affected the usefulness of the correlates as predictors of focal accents. Spectral emphasis was the best predictor for eight of the speakers and overall intensity was the best for three of them, while both predictors were equally good for the remaining two speakers. Furthermore, the percentage correct detections ranged between 48% and 93% for overall intensity and between 60% and 92% for spectral emphasis for the different speakers. A comparison with the results from the paradigmatic comparisons shows that speakers with larger paradigmatic differences between focally accented and non-focal words also tended to have larger syntagmatic (or within-phrase) differences and higher detection scores.

6. Discussion: The detection experiment

First of all, this experiment has shown that the new method of measuring spectral emphasis improved the detection scores by 12% compared to that used in our previous study (Heldner *et al.*, 1999). Moreover, this new spectral emphasis measure turned out to be a better predictor of focal accents than overall intensity, a result which is not in conformity with that of our previous study. The experiment also showed that the usefulness of overall intensity and spectral emphasis as predictors of focal accents was influenced by the position of the focally accented word in the phrase. The scores were well above chance level in initial and medial position in the phrase, while in final position the detection scores, and especially

those for overall intensity, approached what you would expect to find by chance. Furthermore, the Swedish word accents seem to have influenced detection scores. Focally accented words with word accent II were correctly detected more often than those carrying word accent I. However, although the words had been balanced with respect to open and closed vowels and included variation in consonantal context, the possibility that the differences between accent I and II were due to the segmental composition of the words cannot be ruled out completely. For this, recordings of minimal pairs differing only in word accent are needed. Finally, vowel height also affected the detection scores. Words with open vowels in the stressed syllable were detected correctly more often than those with closed vowels. There were also clear indications of speaker dependencies in the detection scores.

As to the question of the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents, this detection experiment has shown that the reliability was fairly high when making syntagmatic comparisons. The overall scores for both measures were certainly better than what could have been expected by chance. However, the experiment has also shown that spectral emphasis is the more reliable correlate in the sense that factors such as position in the phrase, word accent, and vowel height influenced the scores for spectral emphasis to a lesser extent than those for overall intensity. Spectral emphasis was also the best predictor for a majority of the speakers.

Although the reliability in this sense was fairly high, the highest value in the phrase of overall intensity (or of spectral emphasis) was not always found in the focally accented word. There may have been several reasons for this, but intrinsic intensity was certainly one of them. The overall intensity of the vowel is dependent on factors such as the degree of openness and consonantal context (Lehiste & Peterson, 1959; Fant, 1960). For example, the intrinsic intensity of an /a/ was almost 6 dB higher than that of an /i/ (Lehiste & Peterson, 1959). Thus, if the vowels in the focally accented word are closed and a non-focal word in the same phrase contains an open vowel, the peak in the phrase may be found in the non-focal word. Similarly, the spectral emphasis is dependent on the specific formant pattern and increases with the degree of articulatory opening (Fant, 1960).

Another reason why focally accented words do not always have the highest values in the phrase might be the general declining trend in intensity accompanying the f_0 downdrift on an intonation group. This intensity downdrift typically totals 3–4 dB towards the end of the phrase (Pierrehumbert, 1979). Such a decrease may well explain why the peak in the phrase is seldom found in phrase-final words.

A conclusion to be drawn from this is that there might be room for further improvement of the detection, if corrections are included for the prosodic and segmental factors that influenced the detection.

7. General discussion and conclusions

Overall intensity is generally considered a weak prominence cue. Perceptual experiments, including the classic experiments by Fry (1955, 1958), have shown that overall intensity is relatively unimportant as a cue in the perception of stress. More recent work, however, has shown that spectral emphasis is a relevant cue for the perception of lexical stress; it is more reliable than overall intensity and close in strength to duration as a cue of lexical stress (Sluijter, van Heuven & Pacilly, 1997).

The present study has been concerned with the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. However, the reliability of acoustic correlates is not the same as the reliability of perceptual cues. A cue without perceptual relevance may still be a reliable acoustic correlate. Nor is focal accent equivalent

to lexical stress in the aforementioned studies. In fact, lexically stressed words with and without focal accents have been studied here. This study has shown that both overall intensity and spectral emphasis as measured by the improved technique described in section 2.2. are reliable acoustic correlates of focal accents. They are reliable in the sense that there are statistically significant differences between focal and non-focal words for all words, in all positions and for all speakers in the analyzed material as well as in the sense that they are useful for automatic detection. Furthermore, spectral emphasis turned out to be the more reliable correlate in several respects.

On a more detailed level, the paradigmatic comparisons have shown that in general, focally accented words were characterized by both higher overall intensity and spectral emphasis. The average increase in overall intensity in the stressed vowel across 40 different words in medial position in the phrase was about 3 dB; the corresponding value for spectral emphasis was about 2 dB. However, the amount of increase in the correlates was also shown to be dependent on factors such as the segmental composition of the words, to some extent on the speaker, on the position in the phrase and on the non-focal reference chosen (e.g. pre- vs. post-focal). Thus, there were clear positional effects. These results confirm the observations on position dependencies in (Fant *et al.*, 2000b). The paradigmatic comparisons moreover indicated that spectral emphasis was the more reliable correlate, since within-speaker factors such as position in the phrase and the segmental composition of the test words had lesser influence on spectral emphasis than on overall intensity. Apart from being less susceptible to within-speaker influences, spectral emphasis has also been shown to be less affected by the between-speaker factors age and sex (Traunmüller & Eriksson, 2000). Moreover, spectral tilt (defined as SPLH-SPL) has also been shown to be a better predictor (in terms of explained variance or R^2) of perceived prominence than overall intensity (Fant *et al.*, 2000c).

Furthermore, spectral emphasis turned out to be the more reliable correlate in the sense that it gave more correct detections than overall intensity in an experiment with automatic detection of focal accents. The detector was based on the assumptions that the focally accented word would be the most prominent in the phrase and that the prominence would be reflected in overall intensity and spectral emphasis. Both correlates yielded fairly high degrees of correct detections – about 69% of the focally accented words were detected correctly using overall intensity and about 75% using spectral emphasis. Thus, it seems possible to use overall intensity and spectral emphasis to detect focally accented Swedish words to an extent comparable to that reported for accented words in English, Dutch and German (c.f. Nöth *et al.*, 1991; Campbell, 1992; Campbell, 1995; van Kuijk & Boves, 1999; Nöth *et al.*, 2000).

Still, it might be possible to improve the detection by utilizing corrections for vowel height or formant positions. Not surprisingly, the vowel height of the stressed vowels affected the detection scores for both correlates. Focally accented words with open vowels were correctly detected more often than those with closed vowels. The scores were also affected by position in the phrase and by word accent. However, spectral emphasis was also the more reliable correlate in the sense that its detection scores were affected to a lesser extent by these factors than those of overall intensity. Spectral emphasis was also the best predictor for a majority of the speakers.

However, it deserves to be stressed at this point that the primary aim of the detection experiment has been to assess the reliability of overall intensity and spectral emphasis as correlates of focal accents, and that the approach taken to do this has been exploring the usefulness of these acoustic features for automatic detection of focal accents. The intention has by no means been to present a fully-fledged system for accent detection. The detector presented here must obviously be regarded as fairly rudimentary compared to the elaborate systems which, using a wealth of acoustic information, are capable of detecting all kinds of

prosodic categories (Nöth *et al.*, 2000; Shriberg *et al.*, 2000). Nevertheless, this study indicates that even these more ambitious systems for prosodic classification could benefit from the inclusion of information about spectral emphasis as it has been confirmed here and elsewhere that spectral emphasis is a more reliable acoustic correlate than overall intensity as far as detection of accents is concerned (cf. Campbell, 1995; Sluijter *et al.*, 1995; Sluijter & van Heuven, 1996; van Kuijk & Boves, 1999). Therefore, using spectral emphasis as an information source in those systems rather than overall intensity ought to be an advantage.

In addition to assessing the reliability of the correlates, the investigations have resulted in a solid ground of data for overall intensity and spectral emphasis in focally accented and non-focal words that might prove important in modeling for speech synthesis. Future work will most certainly include experiments where the perceptual relevance of increases in overall intensity and spectral emphasis for focally accented words is tested. In addition, effects due to position and distance relative to the focally accented word might prove important in modeling for synthesis, as well.

As the material used in this study was restricted to short phrases read in an artificial situation, it would be premature to generalize the results to hold for the spontaneous speech of all Swedish speakers in all situations. Still, the material included 13 different speakers, both men and women. There were some 40 different words with variation in word accent, vowel quality and vowel quantity as well as in consonantal context. Moreover, the material included words in initial, medial and final position in the phrase. Thus, we feel fairly confident in generalizing the results to controlled productions by Central Swedish speakers without any strong dialectal influence.

Finally, and in addition to these findings, this study has presented a new implementation of spectral emphasis, which yields a continuous estimate of the relative energy in the higher frequency band in the voiced segments. Compared to previous implementations of spectral emphasis (Childers & Lee, 1991; Campbell, 1995; Sluijter & van Heuven, 1996; Traunmüller, 1997; Traunmüller & Eriksson, 2000), it has the advantage of being insensitive to f_0 movements in the vicinity of the low-pass filter cut-off frequency.

Acknowledgments

The research reported here was carried out while I was a guest at the Centre for Speech Technology (CTT) at KTH in Stockholm, an opportunity for which I am extremely grateful. I would also like to thank Eva Strangert, Rolf Carlson, Hartmut Traunmüller, Anders Eriksson, Gunnar Fant, Nick Campbell and two anonymous reviewers for helpful comments and discussion, and Thierry Deschamps for technical assistance. Finally, I would like to thank Hartmut Traunmüller and Anders Eriksson again for providing the dynamic low-pass filter used to improve the spectral emphasis measure.

References

Beckman, M. E. (1986) *Stress and non-stress accent.* Dordrecht: Foris Publications.

Bolinger, D. L. (1958) A theory of pitch accent in English, *Word,* **14**(2-3), 109-149.

Bruce, G. (1977) *Swedish word accents in sentence perspective.* Lund: CWK Gleerup.

Bruce, G. (1982) Developing the Swedish intonation model. In *Working papers*, pp. 51-116. Lund: Lund University, Department of Linguistics.

Bruce, G. (1999) Word tone in Scandinavian languages. In *Word prosodic systems in the languages of Europe* (H. van der Hulst, ed.), pp. 605-633. Berlin, New York: Mouton de Gruyter.

Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D. & Touati, P. (1997) On the analysis of prosody in interaction. In *Computing Prosody* (Y. Sagisaka, N. Campbell & N. Higuchi, eds.), pp. 43-59. New York: Springer-Verlag.

Bruce, G. & Gårding, E. (1978) A prosodic typology for Swedish dialects. In *Nordic Prosody*, Lund, pp. 219-228.

Cambier-Langeveld, T. & Turk, A. E. (1999) A cross-linguistic study of accentual lengthening: Dutch vs. English, *Journal of Phonetics,* **27**, 255-280.

Campbell, N. (1992) Prosodic encoding of English speech. In *Proceedings ICSLP 92*, pp. 663-666. Alberta: Department of Linguistics, University of Alberta.

Campbell, N. (1994) Combining the use of duration and F0 in an automatic analysis of dialogue prosody. In *Proceedings ICSLP 94*, pp. 1111-1114. Yokohama: The Acoustical Society of Japan.

Campbell, N. (1995) Loudness, spectral tilt, and perceived prominence in dialogues. In *Proceedings ICPhS 95*, pp. 676-679. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

Campbell, N. & Beckman, M. E. (1997) Stress, prominence, and spectral tilt. In *Intonation: Theory, models and applications* (A. Botinis, G. Kouroupetroglou & G. Carayiannis, eds.), pp. 67-70. Athens: ESCA.

Childers, D. G. & Lee, C. K. (1991) Vocal quality factors: Analysis, synthesis, and perception, *Journal of the Acoustical Society of America,* **90**(5), 2394-2410.

Cooper, W. E., Eady, S. J. & Mueller, P. R. (1985) Acoustical aspects of contrastive stress in question-answer contexts, *Journal of the Acoustical Society of America,* **77**(6), 2142-2156.

Eefting, W. (1991) The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments, *Journal of the Acoustical Society of America,* **89**(1), 412-424.

Fant, G. (1960) *Acoustic theory of speech production.* The Hague: Mouton.

Fant, G. (1997) The voice source in connected speech, *Speech Communication,* **22**(2-3), 125-139.

Fant, G., Kruckenberg, A. & Liljencrants, J. (2000a) Acoustic-phonetic analysis of prominence in Swedish. In *Intonation: Analysis, modelling and technology* (A. Botinis, ed.), pp. 55-86. Dordrecht: Kluwer Academic Publishers.

Fant, G., Kruckenberg, A. & Liljencrants, J. (2000b) The source-filter frame of prominence, *Phonetica,* **57**(2-4), 113-127.

Fant, G., Kruckenberg, A., Liljencrants, J. & Hertegård, S. (2000c) Acoustic-phonetic studies of prominence in Swedish, *TMH-QPSR*(2-3), 1-51.

Fant, G., Kruckenberg, A. & Nord, L. (1991) Durational correlates of stress in Swedish, French and English, *Journal of Phonetics,* **19**, 351-365.

Fry, D. B. (1955) Duration and intensity as physical correlates of linguistic stress, *Journal of the Acoustical Society of America,* **27**(4), 765-768.

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

Gårding, E. (1993) On parameters and principles in intonation analysis. In *Working papers 40*, pp. 25-47. Lund: Lund University, Dept. of Linguistics.

Gårding, E. & Bruce, G. (1981) A presentation of the Lund model for Swedish intonation. In *Nordic Prosody II*, Trondheim, pp. 33-39.

Heldner, M. & Strangert, E. (2001) Temporal effects of focus in Swedish, *Journal of Phonetics,* **29**(3), 329-361.

Heldner, M., Strangert, E. & Deschamps, T. (1999) A focus detector using overall intensity and high frequency emphasis. In *Proceedings ICPhS'99*, pp. 1491-1493. San Francisco: Linguistics Department, University of California, Berkeley.

House, D. & Bruce, G. (1990) Word and focal accents in Swedish from a recognition perspective. In *Nordic Prosody V* (K. Wiik & I. Raimo, eds.), pp. 156-173: Turku University.

Jackson, M., Ladefoged, P., Huffman, M. K. & Antoñanzas-Barroso, N. (1985) Measures of spectral tilt, *ULCA Working Papers in Phonetics,* **61**, 72-78.

Lehiste, I. & Peterson, G. E. (1959) Vowel amplitude and phonemic stress in American English, *Journal of the Acoustical Society of America,* **31**(4), 428-435.

Nöth, E., Batliner, A., Kießling, A., Kompe, R. & Niemann, H. (2000) Verbmobil: The use of prosody in the linguistic components of a speech understanding system, *IEEE Transactions on Speech and Audio Processing,* **8**(5), 519-532.

Nöth, E., Batliner, A., Kuhn, T. & Stallwitz, G. (1991) Intensity as a predictor of focal accent. In

Proceedings ICPhS 91, pp. 230-233. Aix-en-Provence: Université de Provence.
Ostendorf, M. & Ross, K. (1997) A multilevel model for recognition of intonation labels. In *Computing Prosody* (Y. Sagisaka, N. Campbell & N. Higuchi, eds.), pp. 291-308. New York: Springer-Verlag.
Pierrehumbert, J. (1979) The perception of fundamental frequency declination, *Journal of the Acoustical Society of America,* **66**(2), 363-369.
Sautermeister, P. & Lyberg, B. (1996) Detection of sentence accents in a speech recognition system, *Journal of the Acoustical Society of America,* **99**(4, pt 2), 2493.
Shriberg, E., Stolcke, A., Hakkani-Tür, D. & Tür, G. (2000) Prosody-based automatic segmentation of speech into sentences and topics, *Speech Communication,* **32**, 127-154.
Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K. N. & van Heuven, V. J. (1995) Supralaryngeal resonance and glottal pulse shape as correlate of stress and accent in English. In *Proceedings ICPhS 95*, pp. 630-633. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.
Sluijter, A. M. C. & van Heuven, V. J. (1995) Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch, *Phonetica,* **52**, 71-89.
Sluijter, A. M. C. & van Heuven, V. J. (1996) Spectral balance as an acoustic correlate of linguistic stress, *Journal of the Acoustical Society of America,* **100**(4, Pt 1), 2471-2485.
Sluijter, A. M. C., van Heuven, V. J. & Pacilly, J. J. A. (1997) Spectral balance as a cue in the perception of linguistic stress, *Journal of the Acoustical Society of America,* **101**(1), 503-513.
Stevens, K. N. & Hanson, H. M. (1994) Classification of glottal vibration from acoustic measurements. In *Vocal fold physiology: Vocal quality control* (O. Fujimura & M. Hirano, eds.), pp. 147-170. San Diego: Singular Publishing Group.
t' Hart, J., Collier, R. & Cohen, A. (1990) *A perceptual study of intonation.* Cambridge: Cambridge University Press.
Titze, I. R. & Sundberg, J. (1992) Vocal intensity in speakers and singers, *Journal of the Acoustical Society of America,* **91**(5), 2936-2946.
Traunmüller, H. (1997) Perception of speaker sex, age, and vocal effort. In *PHONUM 4* (R. Bannert, M. Heldner, K. Sullivan & P. Wretling, eds.), pp. 183-186. Umeå: Department of Phonetics.
Traunmüller, H. & Eriksson, A. (2000) Acoustic effects of variation in vocal effort by men, women, and children, *Journal of the Acoustical Society of America,* **107**(6), 3438-3451.
Turk, A. E. & White, L. (1999) Structural influences on accentual lengthening in English, *Journal of Phonetics,* **27**(2), 171-206.
van Katwijk, A. (1974) *Accentuation in Dutch.* Amsterdam/Assen: Van Gorcum.
van Kuijk, D. & Boves, L. (1999) Acoustic characteristics of lexical stress in continuous telephone speech, *Speech Communication,* **27**(2), 95-111.
Wightman, C. W. & Ostendorf, M. (1994) Automatic labeling of prosodic patterns, *IEEE Transactions on Speech and Audio Processing,* **2**(4), 469-481.

Appendix: Phrases and words used in the first recording

Questions

Vem är det som {VERB} kvinnan? *'Who is that {VERB} the woman?'*
Vad gör mannen med kvinnan? *'What is the man doing to the woman?'*
Vem är det som mannen {VERB}? *'Who is it that the man is {VERB}?'*

Answers: Mannen {VERB} kvinnan.

Accent I words

kniper ('pinches')	/ˈkniːpər/	klipper ('cuts')	/ˈklipər/
dräper ('slays')	/ˈdrɛːpər/	släpper ('releases')	/ˈslɛpər/
biter ('bites')	/ˈbiːtər/	gitter (nonsense)	/ˈjitər/
mäter ('measures')	/ˈmɛːtər/	sätter ('puts')	/ˈsɛtər/
sviker ('jilts')	/ˈsviːkər/	sticker ('pricks')	/ˈstikər/
läker ('heals')	/ˈlɛːkər/	väcker ('wakes')	/ˈvɛkər/
grämer ('grieves')	/ˈgrɛːmər/	stämmer ('summons')	/ˈstɛmər/
bryner ('browns')	/ˈbryːnər/	finner ('finds')	/ˈfinər/
kyler ('chills')	/ˈçyːlər/	fyller ('stuffs')	/ˈfylər/
mäler (nonsense)	/ˈmɛːlər/	fäller ('convicts')	/ˈfɛlər/

Accent II words

slipar ('grinds')	/ˈslɪ̀ːpar/	tippar ('dumps')	/ˈtɪ̀par/
kapar ('cuts')	/ˈkɑ̀ːpar/	tappar ('drops')	/ˈtàpar/
ritar ('draws')	/ˈrɪ̀ːtar/	hittar ('finds')	/ˈhɪ̀tar/
matar ('feeds')	/ˈmɑ̀ːtar/	fattar ('grasps')	/ˈfàtar/
pikar ('taunts')	/ˈpɪ̀ːkar/	kickar ('kicks')	/ˈkɪ̀kar/
hakar ('hooks')	/ˈhɑ̀ːkar/	hackar ('minces')	/ˈhàkar/
mimar ('mimes')	/ˈmɪ̀ːmar/	trimmar ('trims')	/ˈtrɪ̀mar/
kramar ('hugs')	/ˈkrɑ̀ːmar/	kammar ('combs')	/ˈkàmar/
tinar ('defrosts')	/ˈtɪ̀ːnar/	skinnar ('skins')	/ˈʃɪ̀nar/
manar ('bids')	/ˈmɑ̀ːnar/	stannar ('stops')	/ˈstànar/

Paper VII In *Focal accent – f_0 movements and beyond*
pp. 135–144

Spectral emphasis as a perceptual cue to prominence[1]

Mattias Heldner

This paper is a first attempt at investigating whether spectral emphasis has any relevance for the perception of accented words. In particular, this study focuses on (i) whether an increase in spectral emphasis will cause accented words to be perceived as more prominent, and (ii) whether modeling of spectral emphasis in connection with accents is liable to improve the quality and naturalness of speech synthesis. A method for experimentally manipulating spectral emphasis in natural and synthesized speech without adverse effects on speech quality is proposed. This method is subsequently used to create stimuli for two perceptual experiments. The first experiment involved increasing spectral emphasis in accented words in natural speech and asking listeners to compare the prominence of the manipulated and original words. Similarly, the second experiment included comparisons of the naturalness of manipulated and original words, this time generated by an mbrola synthesis. The results of these two experiments were on the whole quite negative. Increased spectral emphasis, at least as it was implemented here, did not cause words to be perceived as more prominent. Neither did it improve the naturalness of speech synthesis. A few possible explanations of these results are discussed. However, it remains an open question whether other and perhaps more realistic implementations of increased spectral emphasis may produce more salient effects.

1. Introduction

It has long been recognized that fundamental frequency (f_0) and duration are the most important cues for the perception of accents and of prominence distinctions in general. Recent work seems to indicate, however, that these two features may, after all, be insufficient to model the characteristics of different levels of prominence. Among additional features to take into account the slope of spectrum has emerged as an interesting candidate.

Part of the evidence to support such a hypothesis comes from production studies. Various measures of spectral slope (e.g. spectral balance, spectral tilt or spectral emphasis) have thus been shown to be reliable acoustic correlates in several languages for distinguishing between stressed and unstressed syllables (Sluijter & van Heuven, 1996), or (focally) accented and non-accented words (e.g. Sluijter & van Heuven, 1996; Campbell & Beckman, 1997; Fant, Kruckenberg & Liljencrants, 2000; Heldner, forthcoming). What this amounts to in practice is that stressed syllables and accented words may be expected to have relatively more energy in the higher frequency bands than comparable unstressed syllables and non-accented words do – in addition to the differences in duration and f_0.

[1] This material has been published as Heldner, M. (2001) Spectral emphasis as a perceptual cue to prominence, *TMH-QPSR,* **2/2001**, 51-57.

The crucial evidence for the hypothesis that spectral slope features ought to be included in the modeling of prominence, however, should be sought in perception studies. It must be shown that the observed acoustic differences in spectral slope also have a perceptual relevance. However, perceptual experiments with manipulations of spectral slope are rare, perhaps due to the fact that such features have been more difficult to modify compared to f_0, duration or overall intensity without noticeable degradation of speech quality. One important exception, though, is the elaborate experiments performed by Sluijter, van Heuven and Pacilly (1997). They showed that spectral balance – implemented in terms of increasing the levels of the frequency components above 500 Hz – provided a relatively strong perceptual cue for lexical stress, that is, for distinguishing stressed syllables from unstressed. They moreover showed that spectral balance was almost as important in this respect as duration. It deserves to be stressed here that they studied a prominence distinction at the lower end of the scale. A Dutch reiterant nonsense word *nana*, concatenated from duplicates of an unaccented and unstressed syllable *na,* was used as the starting point for the acoustic manipulations. Furthermore, the subjects were instructed to determine the position of the stressed syllable in the target word, that is, whether the word was *NAna* or *naNA*.

Taken together, the observations from production and perception studies suggest that differences in spectral slope may contribute to the perception of prominence. The energy in the higher frequency bands increases with prominence in natural speech and the observed acoustic differences seem to have a perceptual relevance, at least for distinguishing stressed and unstressed syllables. However, it remains to be shown that the acoustic differences have perceptual relevance also for distinctions at the upper end of the prominence scale, for example for distinguishing (focally) accented and unaccented words. Furthermore, the stronger the evidence for perceptual relevance, the more meaningful it will be to test the importance of spectral emphasis for speech synthesis. Given that spectral emphasis contribute to the perception of prominence an explicit modeling of spectral slope in connection with prominence could possibly improve the quality and the naturalness of speech synthesis (cf. Campbell & Beckman, 1997; Sluijter *et al.*, 1997).

The current paper and the experiments reported in it have a double goal. A first aim has been to investigate whether spectral emphasis – a specific implementation of changes in spectral slope – contributes to the perception of prominence and especially to the perception of focal accents. In particular, it has been tested whether an increase in spectral emphasis causes words to be perceived as more prominent given an increase of the same magnitude as that found when comparing focally accented and non-accented words in natural speech. A second aim has been to explore whether modeling of spectral emphasis in connection with focal accents is liable to improve speech synthesis as far as quality and naturalness are concerned.

2. Perception experiment I

2.1. Method

2.1.1. Material

The speech material used in the first experiment was a read-aloud passage from a short story about a Robinson Crusoe not wanting to be rescued from his desert island[2]. One male

[2] The title of the short story originally written by Jean Ferry is *Robinson* and it first appeared in 1953 in *Le mécanicien et autres contes* (Swedish translation by Claes Hylinger).

Swedish speaker read this passage and rendered it as six shorter phrases (henceforth Phrases 1 – 6) and accented the words indicated by capitals in Table 1.

The recording was made in a sound-treated room with a high quality condenser microphone. The signal was digitized and stored on hard disk (16 bit, 44.1 kHz).

For each phrase an alternative version was subsequently created involving increased spectral emphasis in each one of the accented words. The fact that spectral emphasis was increased in words that were already accented deserves to be stressed here. The original phrases together with the manipulated version of each phrase yielded a total of twelve stimuli.

Table 1. Speech material used in the first experiment. Phrase boundaries and accents, as produced by the speaker, are indicated by a new line and capitals, respectively.

Phr 1	Jag kastade min KIKARE i havet, 'I threw my BINOCULARS into the sea,'
Phr 2	och jag satte INTE upp något STÄNGSEL runt min mark. 'and I did NOT put a FENCE around my territory.'
Phr 3	Tidvattnet hade fört med sig ÅTSKILLIGT vrakgods, 'The tide had brought PLENTY of wreckage,'
Phr 4	som kunde vara till STOR nytta för en skeppsbruten, 'that could be of GREAT use for a shipwrecked,'
Phr 5	och för att slippa SE det 'and to spare me from SEEING it'
Phr 6	gick jag och slog mig ned på andra sidan ÖN. 'I went and sat down on the other side of the ISLAND.'

Increased spectral emphasis was implemented in terms of amplifying the frequency components above the fundamental frequency by 4 dB, while attenuating the fundamental approximately 2 dB. As a result overall intensity increased by roughly 3 dB. The spectral slices seen in Figure 1 can be used to illustrate the effect of the variations. The manipulations were brought about by means of digital filtering with a time varying high shelf filter whose corner frequency followed the course of the fundamental frequency at 1.5 times f_0. A high shelf is a filter that either amplifies or attenuates everything above its corner frequency by equal amounts. The digital filtering was carried out by means of software that was primarily intended for music production (Renaissance Equalizer from Waves Ltd. used as a plug-in module in ProTools Free from Digidesign Inc.). It should be noted that this method of increasing spectral emphasis does not degrade speech quality and that implementations with other filter characteristics are easily accomplished using the same tools.

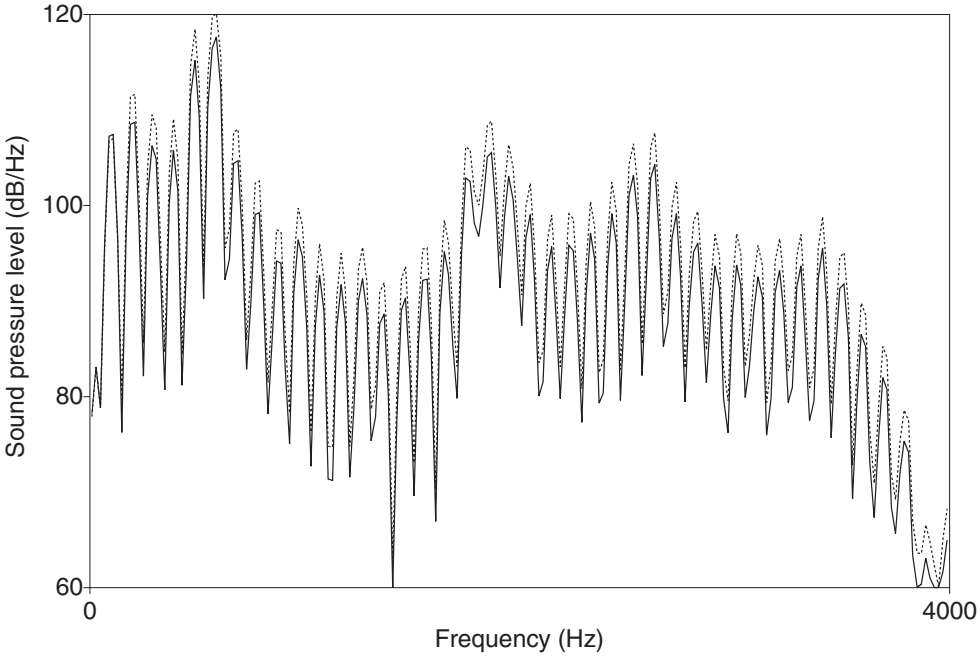

Figure 1. Spectral slices to illustrate the effect of the filtering in an open-mid vowel. The slice drawn with a solid line shows the original version and that drawn with a dashed line is the version provided with increased spectral emphasis.

2.1.2. Subjects and procedure

The strength of spectral emphasis as a perceptual cue to prominence was assessed by means of a paired comparisons procedure. The listeners were exposed to pairs of stimuli with the original version of each phrase constituting one member of the pair and the other member a version of the same phrase where spectral emphasis had been increased in the accented words. To neutralize a possible bias of order of presentation within each pair (concerning the order of the original version relative to the manipulated one), as well as any contextual effects between pairs, these two presentation orders were balanced, randomized and different for all subjects. Stimuli were presented over headphones and the listeners could repeat each pair as many times as they wished. Simultaneously with the presentation of each pair the corresponding text – where accented words had been capitalized (as in Table 1) – was shown on a computer screen. The listeners were instructed to concentrate on the accented words and to determine whether the accentuation was stronger in the first phrase of the pair or in the second (binary forced choice).

Ten native speakers of Swedish (five males, five females) participated in the experiment. All were employees, students or visitors at the Department of Philosophy and Linguistics, Umeå University, or at the Department of Speech, Music, and Hearing at KTH. Each listener had to judge each pair four times. In all, 40 judgements were obtained of each pair. Each session lasted approximately 15 minutes.

2.2. Results

The responses from the first listening test were coded according to the number of times a given version of a phrase was judged to have stronger accentuation than the other version of the same phrase. These figures are shown in Table 2. If an increase in spectral emphasis does make accented words appear more prominent, the listeners can be expected to pick out the higher spectral emphasis version as having stronger accentuation more often than the original version. If they fail to do so, this means that increased spectral emphasis does not in fact increase word prominence and possibly also that the listeners have not been able to discriminate between the two versions of each phrase.

Table 2. Number of judgements of stronger accentuation for each phrase across all speakers.

	Phr 1	Phr 2	Phr 3	Phr 4	Phr 5	Phr 6	Totals
Increased emphasis	19	24	28	23	18	35	147
Original	21	16	12	17	22	5	93
Totals	40	40	40	40	40	40	240

As is evident from Table 2, increased spectral emphasis did not prove to be an unambiguous cue to prominence in this experiment. Although the accented words of the phrases with increased emphasis were, on the whole, perceived as more strongly accentuated than those of the original versions (cf. Totals column in Table 2), there were also two phrases (Phr 1 and 5) for which the original version was perceived as stronger in the indicated sense. Furthermore, the Chi-square tests across all speakers for the individual phrases were significant for the third and sixth phrases only: [Phr 1: $\chi^2(1)=0.1$, $p=0.75$; Phr 2: $\chi^2(1)=1.6$, $p=0.21$; Phr 3: $\chi^2(1)=6.4$, $p=0.01$; Phr 4: $\chi^2(1)=0.9$, $p=0.34$; Phr 5: $\chi^2(1)=0.4$, $p=0.53$; Phr 6: $\chi^2(1)=22.5$, $p<0.01$].

2.3. Discussion

The results presented above did not reveal any dramatic effect on perceived prominence of an increase in spectral emphasis. As a matter of fact, spectral emphasis, as implemented here, did not cause already accented words to be perceived as more prominent in any straightforward manner, since perceived strength was significantly affected only in two out of six comparisons. This result may seem somewhat unexpected in the light of those obtained in previous listening experiments involving manipulations of spectral slope (Sluijter *et al.*, 1997).

It is true that several factors concerning the speech material were not controlled for in this experiment. Nevertheless, there seems to be no simple ad hoc explanation why an increase in spectral emphasis would not yield a significant effect on perceived prominence in the remaining four comparisons. First of all, the observed results are unlikely to be accounted for by properties inherent in the accented words. The reason is obviously that the phrases showing a significant difference between the original version and the increased emphasis version on the one hand and the ones where such a difference is lacking on the other are quite comparable both as for the distribution of word length in accented words and of vowel height in the stressed vowels (cf. Table 1). Nor does it seem very probable that the differences could be accounted for by phrase characteristics, such as length (in terms of number of words) or complexity (in terms of embeddedness). In fact, both groups of phrases

– whether the differences were significant or not – contained phrases of about the same length and complexity. Furthermore, all the phrases were longer than the one used by Sluijter *et al.* (1997).

The explanation for the weak effect of spectral emphasis on perceived prominence, however, could of course be the way spectral emphasis was increased. Firstly, the method chosen in the present study might just not have been realistic enough to affect perceived prominence. The uniform amplification of components above f_0 was perhaps to crude an approximation of increased spectral emphasis. Secondly, the spectral change caused by this particular implementation may not have been large enough. Indeed, the implementation was slightly different from the one used by Sluijter *et al.* (1997). They increased the levels of the frequency components above 500 Hz by up to 9 dB, whereas in the present experiment, the lower limit of the amplified frequency range was variable (i.e. determined by 1.5 times f_0 in each instant) and the components above this limit were amplified by 4 dB. Thus, the lower limit of the amplified frequency range as well as the amount of amplification was generally lower. As a consequence was the spectral change smaller and therefore possibly more difficult to perceive in the present experiment. It should be emphasized, however, that the increases in spectral emphasis operated on the test material were of the same magnitude as, or even slightly larger than, those typically found when comparing focally accented and non-accented words in production studies (Heldner, forthcoming).

Furthermore, the stimuli used in Sluijter *et al.* (1997) may have varied both in formant frequency and in spectral balance, the reason being that an amplification of components above 500 Hz may also affect the formant frequencies in the vicinity of 500 Hz. This effect is non-negligible. In a preliminary test of our own, F1 in the beginning of an unstressed /a/ uttered by a male speaker increased from about 590 Hz to 650 Hz, that is about 0.4 barks, as a result of the amplification of components above 500 Hz by 9 dB. Thus, the vowel quality changed towards a more open vowel. Such an effect is likely to be perceivable in itself under normal listening conditions (e.g. Kewley-Port & Zheng, 1999). The positive results, therefore, *may* have been due to additional formant frequency differences among the stimuli rather than the variations in spectral slope. Spectral emphasis, as implemented in the present experiment, may also affect F1. However, the actual changes in F1 were minor and generally within a range of ±10 Hz.

Finally, there may possibly be limitations to the increase in prominence to be expected when raising spectral emphasis in already accented words where, presumably, spectral emphasis has already been elevated as compared to some non-accented baseline. If this is so, it might simply be unfair to make comparisons with the previously mentioned listening experiments, where positive results were reported. In fact, Sluijter *et al.* (1997) started from a much lower baseline in their experiments by increasing the spectral balance in non-accented and unstressed syllables.

We must not forget, however, that there were two comparisons where the increase in spectral emphasis had a significant effect on perceived prominence. If nothing else, this shows that an increase of the magnitude used in the present experiment *may* be above a perceptual threshold.

In the following experiment, mbrola synthesis utterances were manipulated in the same way as in the first experiment. In these utterances there was no previous modeling of spectral emphasis in connection with the accented words in the original utterances.

3. Perception experiment II

The aim of the second experiment was to examine whether modeling of spectral emphasis in connection with accents will improve the quality and the naturalness of synthetic speech. To this end, utterances produced by a mbrola speech synthesis were manipulated in the same way as in the first experiment and listeners were asked to compare the naturalness of original and manipulated utterances.

3.1. Method

3.1.1. Material

The phrases from the first experiment were also used in the second experiment (cf. Table 1). However, this time they were produced by a male mbrola synthesis voice. The current version of the synthesizer contains no modeling of spectral emphasis in connection with prominence, so accented words in the second experiment were marked by durational and tonal means only. The synthetic utterances were prepared using the software WaveSurfer (Sjölander & Beskow, 2000) with a text-to-speech plug-in.

A version with increased spectral emphasis in the accented words was subsequently created of each phrase using the same digital filtering technique as in the first experiment. Thus, the frequency components above the fundamental frequency were amplified by approximately 4 dB while the fundamental was attenuated by 2 dB, and the overall intensity level in the vowels was raised about 3 dB. All in all, the six original phrases in combination with the manipulated version of each phrase yielded a total of twelve stimuli.

3.1.2. Subjects and procedure

In a listening test it was then explored whether any gain in naturalness had been obtained from the inclusion of a realistic increase in spectral emphasis in the accented words. The details of the procedure were practically identical to those of the first experiment. The only difference concerned the actual stimuli – as mbrola utterances were used here – and the task given to the subjects. As in the first experiment, the listeners were instructed to compare the accented words in two versions of a phrase. This time, however, they were asked to judge which version sounded more natural (binary forced choice).

Another ten native speakers of Swedish (five males, five females) participated in the second experiment. None of them had participated in the first experiment. All of them were employees, students or visitors at the Department of Speech, Music, and Hearing at KTH. Each listener had to judge each pair four times. Thus, 40 judgements were obtained for each pair. In this experiment, too, each session lasted approximately 15 minutes.

3.2. Results

The responses from the second experiment were coded according to the number of times a given version of a phrase was judged to have a more natural accentuation than an alternative version of the same phrase. The figures obtained can be studied in Table 3. In case an increase in spectral emphasis causes accented words in the mbrola synthesis to be perceived as more natural, the listeners could be expected to judge the version with elevated spectral emphasis to be more natural more often than the original version.

Table 3. Number of judgements of more natural accentuation for each phrase across all speakers.

	Phr 1	Phr 2	Phr 3	Phr 4	Phr 5	Phr 6	Totals
Increased emphasis	20	19	22	17	19	16	113
Original	20	21	18	23	21	24	127
Totals	40	40	40	40	40	40	240

As is shown by Table 3, manipulation of spectral emphasis did not affect the perceived naturalness of the accented words to any great extent. Rather, the original versions seem to have been perceived as slightly more natural (cf. Totals column in Table 3). Moreover, none of the Chi-square tests for the individual phrases were significant: [Phr 1: $\chi^2(1)=0.0, p=1;$ Phr 2: $\chi^2(1)=0.1, p=0.75;$ Phr 3: $\chi^2(1)=0.4, p=0.53;$ Phr 4: $\chi^2(1)=0.9, p=0.34;$ Phr 5: $\chi^2(1)=0.1, p=0.75;$ Phr 6: $\chi^2(1)=1.6, p=0.21$].

3.3. Discussion

All in all, the second experiment yielded clearly negative results. As we have just seen, there was no significant improvement in the perceived naturalness of accented words in mbrola utterances provided with an addition of a realistic increase in spectral emphasis. In other words, although an increase in spectral emphasis may in principle be a reliable acoustic correlate for distinguishing various levels of prominence (e.g. Sluijter & van Heuven, 1996; Campbell & Beckman, 1997; Fant *et al.*, 2000; Heldner, forthcoming), the inclusion of this particular feature did not manage to make synthesized utterances appear any more natural to the listeners in this experiment.

Now, this result is perhaps not very surprising. After all, the perceptual influence of spectral emphasis found in the first experiment was already extremely feeble, in spite of the fact that the manipulated natural speech it contained had a high sound quality as compared to that of the mbrola speech in the second experiment. Since subtle acoustic differences are likely to be more difficult to perceive in speech of degraded quality, one would expect those used in the second experiment to have an even weaker effect on perception. Supposing, furthermore, the actual implementation of spectral emphasis in the first experiment to be somehow problematic, results of the second experiment would have been equally affected, since the stimuli of both experiments were created using the same technique.

On the other hand, the second experiment had the merit of making the manipulated versions more similar to natural speech, as realistic variations of spectral emphasis in connection with accents were introduced. In contrast, the manipulations carried out in the first experiment diverged from naturalness, as the increases in spectral emphasis were performed on already accented words. Therefore, the second experiment ought to have been the better test of whether realistic increases in spectral emphasis also increase the naturalness of accented words. Apparently, however, spectral emphasis as implemented here did not improve the naturalness of accented words in mbrola utterances. A possible explanation for this could be that the listeners were not able to perceive the difference.

4. Conclusions

This study has presented a flexible method for increasing spectral emphasis without degrading the quality of the output. It works equally well for natural and for synthesized speech. The technique also allows for local increases in spectral emphasis, as for example in accented words or in stressed syllables. Potentially this method ought to be useful for exploring spectral emphasis in connection with the perception of prosodic phenomena.

Furthermore, this study has reported on two listening experiments in which this technique was used. Interestingly, both experiments yielded negative results, which was somewhat unexpected given the results of previous production and perception experiments. Manipulations of spectral emphasis of the magnitude found when comparing the production of accented and non-accented words – i.e. what could be considered a realistic level – did not make accented words more prominent in any straightforward fashion. Nor did they improve the naturalness of accented words in synthetic utterances.

So, a tentative conclusion must be that spectral emphasis, at least as implemented here, seems to be fairly weak as a cue to prominence at the upper end of the prominence scale, and moreover, of little value for improving the quality of speech synthesis. Future research will have to show if other ways of implementing spectral emphasis may produce more salient effects on perceived prominence and on perceived naturalness of accented words in speech synthesis.

Acknowledgments

This research was carried out when the author was a guest at the Centre for Speech Technology, a competence center at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The author would like to thank Eva Strangert and Inger Karlsson for invaluable comments on earlier drafts.

References

Campbell, N. & Beckman, M. E. (1997) Stress, prominence, and spectral tilt. In *Intonation: Theory, models and applications* (A. Botinis, G. Kouroupetroglou & G. Carayiannis, eds.), pp. 67-70. Athens: ESCA.

Fant, G., Kruckenberg, A. & Liljencrants, J. (2000) Acoustic-phonetic analysis of prominence in Swedish. In *Intonation: Analysis, modelling and technology* (A. Botinis, ed.), pp. 55-86. Dordrecht: Kluwer Academic Publishers.

Heldner, M. (forthcoming) On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish, *Submitted to Journal of Phonetics*.

Kewley-Port, D. & Zheng, Y. (1999) Vowel formant discrimination: Towards more ordinary listening conditions, *Journal of the Acoustical Society of America,* **106**(5), 2945-2957.

Sjölander, K. & Beskow, J. (2000) WaveSurfer. Stockholm: Centre for Speech Technology (CTT) at KTH. Available for download at http://www.speech.kth.se/wavesurfer/.

Sluijter, A. M. C. & van Heuven, V. J. (1996) Spectral balance as an acoustic correlate of linguistic stress, *Journal of the Acoustical Society of America,* **100**(4, Pt 1), 2471-2485.

Sluijter, A. M. C., van Heuven, V. J. & Pacilly, J. J. A. (1997) Spectral balance as a cue in the perception of linguistic stress, *Journal of the Acoustical Society of America,* **101**(1), 503-513.