General introduction

Abstract

The present study treats two main research questions: what acoustic-prosodic means do speakers use to mark the information structure of spoken discourse and which cues are most important for the listener to detect this structure? More specifically, how does this process work in spontaneous speech? In this first chapter, the importance of such research is described, from a theoretical point of view as well as from an acoustic-phonetic point of view. The research goals are discussed in more detail, and the outline of the thesis is given.
1.1. Introduction

Spoken discourse basically involves three different aspects: the speaker, the listener, and the message itself. * Speakers may use various acoustic-phonetic means to signal the structure of the message they are producing. They will mark certain words as more important than others, and they will also divide the whole message into smaller parts, such as paragraphs and sentences. When listening to such a spoken discourse, *listeners* have certain ideas about the structure of the incoming text. They perceive certain words or word groups as more important than others, while they are also able to detect different types of boundaries, such as sentence boundaries and paragraph boundaries (Lehiste, 1979). The *message* itself, i.e., the text as produced by the speaker, also has a structure. Assuming that it is more or less coherent (a requirement for the listener to understand it), it can be divided into paragraphs, sentences, clauses, phrases, and so forth. Apart from acoustic means, the speaker also has a variety of linguistic means available to indicate the structure of the message.

The variety of acoustic-phonetic means used to convey the structure of a spoken message will be referred to as the *prosodic realization* of the message. Prosodic characteristics are supra-segmental in that they are present in the speech signal, independent of the sequence of segments (vowels and consonants) that build up the signal (cf. Rietveld & Van Heuven, 1997). In the present study we will focus on two prosodic aspects, namely, intonation and pausing. It is generally assumed for Dutch that speakers use mainly intonation, i.e. pitch accents and boundary marking pitch movements, and pausing to signal prominence and phrasing. A combination of both intonational and pausal features may also be used (e.g. Blaauw, 1995; Swerts, 1994). A more elaborate discussion of prosodic features in relation to prominence and boundary marking can be found in chapters 4 and 5.

The present study has two main questions. First of all, from a production point of view: what acoustic means are used by the speaker to signal the structure of spoken discourse? Secondly, from a perception point of view: how are the acoustic cues, as provided by the speaker, used by the listener to detect the structure of what has been said?

Spoken discourse can be produced in a number of ways, and the prosodic realization of spoken discourse may be influenced by a number of factors. First of all, a distinction can be made between speech read aloud, and spontaneously uttered speech. So far, most research into the relation between discourse structure and its prosodic realization was done using speech read aloud. Furthermore, during the last ten years, researchers have also been interested in the prosodic difference itself between speaking styles (Koopmans-van Beinum, 1992a, 1992b; Blaauw, 1995; Laan, 1997; Bruce & Touati, 1992; Eskénazi, 1993; Granström, 1992; Harmegnies & Poch-Olivé, 1992; Howell & Kadi-
Hanifi, 1991). Researchers have become increasingly aware of the importance of investigating spontaneous speech (see, for instance, books as the one edited by Sagisaka, Campbell & Higuchi, 1996). However, spontaneous speech itself can also be produced in different ways, as e.g. unstructured narratives, extended descriptive narratives (the type we use in the present study), instruction monologues or dialogues, and database querying dialogues (cf. Beckman, 1996). All these types can be viewed as spontaneous speech, but they represent, of course, different subtypes.

Emotion is another factor that may influence the prosodic realization of spoken discourse. So far, research into the prosodic characteristics of emotions in speech has been limited. This is due among other things to the difficulties of objective measuring procedures (see, for instance, Scherer, 1989 for a review). However, some attempts - for the most part successful - have been made (e.g. Carlson et al., 1992; Mozziconacci, 1998).

We will concentrate on spontaneously spoken discourse rather than speech read aloud. We are interested in finding out the relation between discourse structure and prosodic realization, as can be observed in everyday communicative situations. In this first chapter, we will first give a short description of the relation between discourse structure and prosodic aspects. (A detailed discussion of the literature on this issue is given in chapter 2.) Secondly, we will present our own research goals, as they will be dealt with in this thesis. Finally, we will give an outline of the thesis.

1.2. Discourse structure and prosodic aspects

The structure of information in written texts usually becomes clear by the use of typographic means, such as indents, punctuation and capitals, highlighting, and the use of different lay-outs and fonts. In spoken texts, it is generally assumed that the speaker may use various acoustic means to assign structure, for instance by accenting important words or by pausing at boundaries. In written texts, words can also be perceived as being more or less important on the basis of grammatical and pragmatic characteristics. This structure of spoken or written texts in terms of important information and boundaries can be referred to as the overall information structure of a text.

In the often used elicitation method of question/answer pairs, the information structure (focus distribution) can be described using labels such as ‘new’ vs. ‘old’ information, where ‘new’ usually refers to ‘accented’ and ‘old’ to ‘not accented’ (see, for instance, Nooteboom & Kruyt, 1987; Cooper et al., 1985; Eady et al., 1986). Focus is thus defined through intonation. However, this kind of definition may lead to circularity in that the possible acoustic features are already included in the definition itself. This is not necessarily a problem, unless
the subject of the study is the acoustic realization of focus, as it is in the present case. This circularity has also been noted by other researchers (e.g. Brown, Currie & Kenworthy, 1980). How the focal structure of a whole discourse in terms of 'new' vs. 'old' information should be traced is (still) less clear.

Much work has been done recently to investigate the relation between discourse structures and their prosodic features, using independent frameworks of discourse structure, to avoid the above-mentioned circularity (e.g. Geluykens & Swerts, 1994; Grosz & Hirschberg, 1992; Hirschberg & Grosz, 1992, 1994; Nakatani, 1995; Nakatani, Hirschberg & Grosz, 1995; Swerts & Geluykens, 1994; Terken & Hirschberg, 1994). These studies are primarily concerned with the overall global structure of whole discourses. However, Hirschberg et al. (see references mentioned above), analyze their discourses at both a global level (the structure of the discourse constituents which form the whole discourse) and a local level (parentheticals, quotations, tags, and indirect reported speech).

In the present study we want to concentrate not only on the structure of the discourse in terms of global ‘chunking’, but also on the internal focal structure of a discourse in terms of ‘new’ and ‘given’ information. That is, we want to apply the given/new distinction to complete discourse rather than to isolated and/or pre-cooked utterances, to analyze the structure of different types of information. The ‘global’ and ‘local’ levels as defined by Hirschberg et al. are not sufficient for this purpose. In the first place because the location of discourse boundaries is determined beforehand, and secondly, because the information status of the various words and word groups in the discourse is not available. Therefore, an independent framework will be developed for discourse analysis, one in which the internal focal structure of a discourse is based on pragmatic features of discourse structure rather than on acoustic features such as intonation and accentuation. The work by Mann & Thompson (1988), Chafe (1987, 1994), and Prince (1981, 1992) will play a central role in the development of this framework. By using this framework we can avoid the circularity of acoustic features being already included in the definition of focal structure. This method applies to both written texts and verbatim transcriptions of spontaneously produced spoken texts. The analysis thus obtained reflects the structure of a text, based on the written text alone.

1.3. Research goals

The general aim of the present study, as indicated in the introduction, was first of all to investigate the various acoustic-prosodic means available to the speaker to mark discourse structure in spontaneous discourse. Secondly, we wanted to investigate which of these cues are most important for the listener to detect this structure, and how the listener selects and uses them to attain his/her goal (i.e. to
understand the structure of the message). In the present study we will use speech material that is representative of the speech used in everyday communicative situations. From a phonetic point of view, more insight into the human process of speaking and hearing is essential for a better comprehension of the interaction between speaker and hearer, and is also beneficial for the further development of technological applications. From a linguistic point of view, it is valuable to acquire more knowledge about the internal structure of discourse, i.e., above the level of the clause (e.g. Dik, 1997, part II).

The discourse serves as a reference point in the present research. By discourse, we mean a coherent collection of utterances, either spoken or written, produced by one speaker, but with the intention of communicating the content to a listener. Spoken discourse, for the purpose of this study, thus is a monologue, as opposed to a dialogue, where at least two speakers are involved. Every discourse has an internal structure in terms of a hierarchical division into smaller parts, such as paragraphs, sentences, clauses, phrases, concepts, and words. This hierarchy can, of course, be extended below the level of the word, but in the present study we will be dealing mainly with words and higher levels. This discourse structure should ideally be determined on the basis of the text alone, i.e., without recourse to any acoustic features. In written discourse, this structure is usually indicated by typographic means (indents, capitals, underlining, italics, etc.). Spoken discourse, whether carefully prepared or spontaneously uttered, can, of course, be structured along the same lines: the same global build-up into words, concepts, phrases, clauses, sentences, and paragraphs applies.

The main questions underlying the present research can be formulated as follows:

- What acoustic-prosodic means are used by speakers to mark the structure of a spontaneously spoken discourse?
- Which acoustic-prosodic cues, as given by the speaker, do listeners use to detect the structure in spoken discourse, and how do they do this?

Spontaneously spoken discourses were collected from speakers of standard Dutch. The discourse text, i.e., the reference point mentioned above, was obtained by making verbatim transcriptions of the discourses. This text was then analyzed for discourse structure using a purely text-based method. The material was then viewed from the perspective of the three different components of the communication chain: the message, the speaker, and the listener. As mentioned before, the discourse text served as a reference point. The results of both the acoustic-prosodic analysis and the perceptual analysis will be interpreted in relation to the textual discourse analysis.

Our procedure in investigating the prosodic aspects of information structure in discourse will be the following. On the basis of the textual information, we are
able to determine which words in the text are 'new' and 'given', and where discourse boundaries are located in the text. Determining the information structure in terms of new/given information and discourse boundaries is not always unambiguous. This is, however, a consequence of using natural speech. This is illustrated by the fact that everyone perceives and interprets spoken or written text in relation to his/her own personal context (Prince, p.c.). By analyzing the acoustic realization of the speaker, we are able to verify whether he/she has actually realized the discourse structure acoustically as predicted on the basis of the textual analysis. By having naive listeners (i.e. not phonetically trained) evaluate the spoken versions of the discourse, and by asking them to indicate which words are spoken with emphasis by the speaker, and where they perceive discourse boundaries of some type, we are able to verify whether this perceived structure is the same as that predicted by the text-based textual analysis. Ultimately, the results from the acoustic analyses and the perceptual analyses are related, to see what acoustic cues, as provided by the speaker, are used by the listener to detect the structure of the discourse. An evaluation by means of rule-based synthesis lies outside the scope of this thesis.

Summarizing, the following analyses are performed:

- **Textual analyses** of the discourse structure in terms of important words and discourse boundaries.
- **Acoustic-prosodic analyses** of the means used by the speaker to mark important information and discourse boundaries, e.g. pitch accents, boundary tones, and pausing strategies.
- **Perceptual analyses** of the discourse structure as perceived by naive listeners, in terms of prominence (perceived important words) and phrasing (perceived discourse boundaries).

In the following section, the outline of the thesis is given.

**1.4. Outline of the thesis**

Chapter 2 gives an overview of three different methods of discourse analysis. These methods served as a starting point for the development of our own framework for analyzing information structure in spoken discourse. This framework is presented there, as well as the pilot experiment in which its usefulness for the present research was tested. Finally, the adapted framework to be used is given.

In chapter 3, the selection of speech material is presented, as well as the procedure for collecting and recording the material. Furthermore, an overall characterization of the speakers is given in terms of temporal and intonational
features. Perceptual evaluations of the (prosodic) performance of the speakers in their production of the discourse complete this characterization. Speaker profiles are presented, indicating 'good' and 'bad' speakers.

In chapter 4, the prosodic characterization of discourse boundaries is explored. The location of boundaries in the text-based discourse structure is taken as a reference point, and the actual realization by the speaker and the perception of boundaries by naive listeners is investigated, in terms of both temporal (pausing) and intonational (boundary tones) aspects. The relation between discourse, speaker, and listeners regarding discourse boundaries concludes this chapter.

Chapter 5 presents a study on the focal structure in discourse, i.e., the relation between information structure (cf. the given/new distinction) on the one hand and the realization and perception of that structure on the other hand. This relation is investigated by studying the location and realization of pauses and pitch accents relative to the information structure. The perceptual evaluation is obtained from prominence judgements by naive listeners.

Chapter 6, finally, presents a general discussion. The conclusions of this thesis are presented, as well as recommendations for future research.
...