Speech Variability and Emotion:

Production and Perception

Sylvie Mozziconacci

SPEECH VARIABILITY AND EMOTION

Production and Perception

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

MOZZICONACCI, SYLVIE J. L.
Speech variability and emotion : production and perception / by Sylvie J. L. Mozziconacci.-
Eindhoven : Technische Universiteit Eindhoven, 1998.
Proefschrift. - ISBN 90-386-1191-9
NUGI : 719 / 949
Trefw. : Prosodie, Emoties , Intonatie
Subject headings : Prosody, Emotions, Intonation

Cover design: Ben Mobach
Printed in the Netherlands, by the Printers of the University of Eindhoven

SPEECH VARIABILITY AND EMOTION

Production and Perception

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof. dr. M. Rem,
voor een commissie aangewezen door het College voor Promoties
in het openbaar te verdedigen op
vrijdag 20 november 1998 om 14.00 uur

door

Sylvie Jeannette Laure Mozziconacci

geboren te Parijs

Dit proefschrift is goedgekeurd door de promotoren:

prof. dr. A. J. M. Houtsma
en
prof. dr. ir. L. C. W. Pols

Copromotor:
dr. D. J. Hermes

Table of contents

Acknowledgements

This research has originally been granted by the SOBU (SamenwerkingsOrgaan Brabantse Universiteiten), and was later supported by the graduate school J. F. Schouten Institute for User-System Interaction Research.

I thank IPO (Intitute for Perception Research, now Center for Research on User-System Interaction, Eindhoven), KTH (Royal Institute of Technology, Dept. of Speech, Music and Hearing, Stockholm), and IFA (Institute of Phonetic Sciences, Amsterdam) for providing me with all necessary facilities.

In these three institutes, where I spent part of my academic life, there were always helpful people around. I highly appreciate their support.

I want to thank my supervisor Dik Hermes, my first promotor Aad Houtsma, my second promotor Louis Pols, and the other members of my committee, Gunnar Fant and Sieb Nooteboom, who all provided me with many useful comments. In addition to the people above, I also thank Don Bouwhuis and Vincent van Heuven, who were always available for all kinds of advices.

I feel undebted and very thankful to the friends and the colleagues who were helping me when I needed them most.

I certainly want to thank my friends and relatives for understanding my lack of availability, putting up with my frustrations, and sharing in the good laughs of every-day life.

A very special thank to Arlette, and to Janny en Gerrit!

Amsterdam, October 1998

Sylvie

Chapter I

General introduction

1. Problem statement and framework of the study

Spoken communication is a part of everyday life. Among the different means of communication that people have developed, speech is a convenient and powerful one. For most people, speaking is a potentially efficient and agreeable way of communicating. Therefore, the notions of efficiency, pleasure, and quality, have to be taken into consideration when developing speech technologies. Technological developments such as the telephone and the radio are successful because they transmit spoken communication. Even though telephone and radio restrict the communication channel merely to what sounds convey, they offer an efficient and agreeable way to communicate. The aim of speech technologies, implying speech synthesis and speech recognition, is to make it possible to rely on speech communication in an increasing number of situations. An example is the communication with machines in dialog systems giving access to information by phone around the clock. However, people expect this spoken communication to remain efficient and agreeable.

Unfortunately, today's synthetic speech still sounds rather unnatural and uninvolved. Who wants to communicate with an interlocutor whose speech sounds as if the communication at hand is boring? Traditionally, in terms of synthesis performance, the main concern used to be with the intelligibility aspects. Nowadays, despite the reasonably good intelligibility of speech-synthesis-systems, there is still a general lack of naturalness. Synthetic speech seems to lack the variability that makes speech so lively and maintains the interlocutors' attention on the communication. The fact that this need for 'additional variability' is the result of an incomplete specification of the signal has been discussed by

Liénard (1995). He qualifies the traditional view of speech processing, in which the speech signal is implicitly considered as if it was merely an oral realization of the written message, as a 'reductive' view.

A spoken message carries more information than its written counterpart, even if, out of all communication channels, we restrict ourselves to the auditory one, and if out of all sounds humans can produce vocally, we restrict ourselves to speech sounds, as we do in this study. Speech does not only convey the content of the message formulated in a language, it also provides information concerning, among others, the identity of the speaker, his or her gender, age, regional and social background. It also contains information about the speaker's state of health and emotional state, and his or her attitude towards the speaker, the situation and the topic of conversation. Naturally, the speech signal must primarily carry speech sounds that allow the listener to identify the message. It also provides cues on how to structure this message by signaling, for instance, the important parts of the message, and most frequently also cues for structuring the communication, by indicating, for instance, that the speaker has completed the message. This additional information to the listener, that is not explicitly included in the lexical or syntactic contents of the sentences, is conveyed by means of prosody. The function of prosody in oral communication is to convey accentuation/prominence, to indicate the phrasing of sentences and the structure of the discourse/dialogue (also turn-taking), to contribute to speaker identification and verification, and to convey the expression of emotion and attitude.

Paying attention to modeling variability is a current trend in speech synthesis (e.g., Carlson, 1991; Eskénazi, 1993; Carlson, 1994; Cole et al., 1995; Pisoni, 1997; Pols, 1998). Speaker characteristics are being increasingly taken into account, and concepts such as focus, emphasis and emotion are nowadays relevant to speech technology. Modeling variability would, therefore, have an impact on the quality of synthesis and, consequently, on applications in speech technology. However, improving the naturalness of synthetic speech eventually involves understanding how speech variations are performed and perceived. This brings us to a study of speech displaying a large scale of variations. Such variations in speech occur when the speaker becomes emotional. The study of emotion in speech has the advantage that it provides two means of enriching the synthesis. First, allowing the synthesis of emotional speech could be useful for interactive dialog systems or multimedia systems, and also for a reading machine. Secondly,

improving our knowledge concerning the way variability is realized in speech would provide the possibility to include a range of variations in non-emotional speech synthesis, thereby improving speech synthesis by making it sound more natural. Additionally, understanding variability in speech would make it possible to improve the robustness of speech recognition.

- ***Emotion***

A methodological difficulty is that there is, as yet, no widely accepted definition and taxonomy of emotion. Different authors have proposed different lists of terms indicating emotions (e.g., Izard, 1977; Plutchik, 1980; Ekman, 1982; Frijda, 1986). A review of definitions was given by Plutchik (1980, p. 81-83). It should also be appreciated that a single emotion can be uttered in different ways. Scherer (1986) distinguishes different categories in a single emotion, for instance, the category 'cold anger/irritation' and the category 'hot anger/rage'.

In literature, the term emotion is used with different meanings and sometimes includes notions such as attitude or intention. Two main tendencies in theories have developed. One tendency is to consider emotions as discrete categories (Ekman, 1973; Izard, 1977; Plutchik, 1980). A distinction is made between basic emotions and combinations of these basic ones. Another tendency is to view emotions as characterized by progressive, smooth transitions (Schlosberg, 1954). Similarities and dissimilarities between emotions are characterized in terms of gradual distances on dimensions such as pleasant/unpleasant, novel/old, consistent/discrepant, control/no control. In the present study, the first approach has been chosen, for methodological reasons; considering discrete categories allows the use of identification paradigms, which means that subjects in an experiment can be asked to identify the intended emotion in a natural or synthetic utterance.

The notion of emotion is used here to refer to emotional states or dispositions related to physical states that influence the speech produced. The vocal expression of emotion is perceived by listeners as indicators of the speaker's mood. Emotions are considered to be complex phenomena with biological, social, and personal components (Izard et al., 1988). Attitudes are considered as a tendency to evaluate a particular entity with some degree of favor or disfavor (Eagly et al., 1993). This notion of attitude is used to refer to the disposition of the speaker towards the topic, the situation and/or the listener; it seems to involve cognition more than emotion does. For the sake of conciseness, and because there

is no compelling theoretical base for a distinction between attitudes such as indignation and emotions such as fear, a single term, 'emotion', will be used in the remainder of the text, to refer both to notions of emotion and attitude, without further distinction. For the present study, however, we will rely upon an empirical definition of emotion being a selection of emotions that are fairly well identified by listeners among a range of different emotions. Moreover, though neutrality is used as a reference for other categories, it will also be referred to as an 'emotion'.

Good recordings of spontaneously produced emotional utterances are difficult to acquire. Speech material that is spontaneously produced often has a number of drawbacks. The recordings are usually not free of background noises. It is uncertain what emotion or mixture of emotions the speaker was experiencing while speaking, and there is no control of the semantic and phonetic content of such utterances. For the present study, the lack of control of the speech material was felt to be too much of a drawback. It was therefore decided to use simulations of emotions, recorded in a laboratory under controlled elicitation.

- ***Approaches***

Charles Darwin (1872), whose research has been of influence in this area, considered vocalizations to be, in addition to facial and bodily expression, one of the primary means to express emotion. Attention to vocal and, in particular, to bodily (especially facial) expression of emotion, aroused quite a bit of interest in literature. For instance, a recent investigation of Ellison and Massaro (1997) was concerned with the perception and recognition of facial affect. Extensive reviews of studies concerned with the vocal expression of emotion were given by Frick (1985) and Scherer (1986). Sundberg (1987) reviewed studies concerned with emotion in speech and singing. A more recent review was presented by Murray and Arnott (1993). Emotional states can cause changes in posture, muscle tension, respiration, salivation or in register of phonation. As vocalizations are produced by the combined action of a great variety of muscles in the chest, throat, and head, these changes may strongly affect the acoustic characteristics of these vocalizations. Different factors, such as physiological (e.g., Scherer, 1986; Davitz, 1964), psychological, and socio-cultural factors, also influence the voice and speech production. Van Bezooijen (1984) studied the recognizability of emotions in speech as a function of age, sex, and culture, showing the influence of these factors on the expression of different emotions in speech.

Two types of approach have been developed so far, with respect to the study of the vocal expression of emotion. Some perception oriented studies were concerned with the ability of listeners to identify the intended emotion (van Bezooijen, 1984; Cahn, 1990; Carlson, Granström, and Nord, 1992; Friend and Farrar, 1994; Protopapas and Lieberman, 1995; Laukkanen, Vilkman, Alku, and Oksanen, 1997; Morlec, Bailly, and Aubergé, 1997). Scherer (1989) reviewed studies of this type, concerned with natural speech recorded in the field or induced in the laboratory. He infers that "enough differentiated information is available in vocal expression to allow a clear distinction of a fairly large number of discrete emotions" (p. 137). Other studies, oriented on acoustic analysis, were concerned with the evaluation of acoustic or phonatory-articulatory features of speech produced in emotional states. Examples of these features are: mean F_0, signal amplitude, speech rate, pauses, breathiness, and laryngealization (Fairbanks et al., 1939; Williams and Stevens, 1972; van Bezooijen, 1984; Higuchi, Hirai, and Sagisaka, 1994; Selting, 1994; Hirose, Kawanami and Ihara, 1997). Using this approach, the main concern of these studies was with acoustic measurements. Few studies considered how the acoustic and perceptual descriptions relate to each other (Boves, 1984).

Most previous studies were concerned with the identification of dimensions such as intonation, loudness, precision of articulation, or voice source, that supplement the semantic and syntactic content of the discourse. These studies are mostly considered to be extra-linguistic studies, although this is a somewhat arbitrary position as it is not clear to what extent the expression of emotion is language specific and rule governed (van Heuven, 1994, p. 3). It is generally agreed that there are correlates of the physiological responses characterizing different emotional states. These are articulatory correlates, such as manner and place of articulation, and phonatory correlates, such as voicing and intensity. In fact, many studies (e.g., Carlson, Granström, and Nord, 1992; Cosmides, 1983; Fairbanks and Pronovost, 1939; Ladd, Silverman, Tolkmitt, Bergman, and Scherer, 1985; Williams and Stevens, 1972) agree that most prosodic features, i.e., suprasegmental features such as pitch, speech rate, rhythm and loudness, contribute to the expression of emotion in speech, and that there is not one unique acoustic correlate for a particular emotion. In an experiment designed to test whether different individuals produce similar 'voice patterns' when they read the same emotional passage, Cosmides (1983) concluded, however, that different individuals adhere to reasonably standard acoustic configurations in expressing particular emotions. Scherer (1986) reports that "there is now sufficient evidence that simulators seem to use fairly standard (and to a large extent universal) rules

to encode emotion labels into patterned muscle action". (By 'simulators', he means actors that imitate an emotional state.)

Relatively few studies additionally tried to quantify the relevant extra-linguistic dimensions, for the purpose of obtaining parameter values for generating emotional speech (e.g., Carlson, Granström, and Nord, 1992; Cahn, 1990; van Bezooijen, 1984; Williams and Stevens, 1972). Van Bezooijen (1984) described the characteristics of the vocal expression of emotions mostly in terms of perception and made an effort to relate the perceptual scores to acoustic values. Cahn (1990) developed a tool for exploring the effect of acoustic parameters on the perception of emotion in synthetic speech, making use of DECtalk3. She called it the 'Affect Editor', and described it as "a program that applies speech parameter values for an emotion to a pre-analyzed sentence in order to synthesize the sentence with the correct affect" (p. 53). It is unfortunately not clear how the speech parameter values compare to those in human speech. Murray (1989) developed a system that is also based on manipulations of the combination of parameters of the DecTalk synthesizer (version 2.0). This HAMLET system was incorporated in a communication prosthesis for disabled persons and was evaluated in listening tests. Granström and Nord (1991), using synthetic speech, asked subjects to adjust test stimuli to some internal reference, such as joy, anger, etc., and to verbally describe the emotions resulting in the synthesized sentences. Carlson, Granström and Nord (1992) tried to model extra-linguistic features in speech by overlaying pitch curves and sentence durations on natural utterances.

- ***Present approach***

The related studies mentioned above, concentrated either on the production or on the perception of emotion in speech. For the present study, a combination of both approaches was felt to be most appropriate. Indeed, the value of the acoustic data resulting from the analysis-oriented studies, appears to depend strongly on what is known about the correspondence between the production data and the perception of emotion. Establishing a relationship between acoustic and perceptual data allows a more systematic approach.

This study is explicitly intended, not only to describe speech variations occurring in the expression of emotion and attitude in speech, but also to quantify these variations by expressing them as parameter values for rule synthesis. Therefore, parameters relevant for conveying emotion in speech first need to be identified. As intonation and speech rate are

considered to be the most important conveyors of emotion in speech (e.g., House, 1990; Ladd, Silverman, Tolkmitt, Bergman, and Scherer, 1985; Cosmides, 1983; Williams and Stevens, 1972; Fairbanks and Pronovost, 1939), it was decided to start with an extended study of these parameters. As a consequence, variations other than the ones in pitch and speech rate, such as variations in voice source and in intensity, are not investigated in the present thesis. On the other hand, a point in favor of our approach is that the role of intonation patterns (i.e., legal sequences of perceptually distinct pitch movements) for the expression of emotion has, as yet, never been investigated, so that little is known about how pitch varies over time in emotional utterances. This study does not only consider the global measures that are traditionally used, such as mean F_0, it also attempts to describe the pitch curves in more detail, and therefore to have a closer look at the structure of pitch curves. To this end, we chose, among the various approaches of intonation, to exploit the knowledge and experience available at the institute where the research took place, and to use the experimental phonetic approach of intonation analysis and the description of pitch curves as is traditionally the case at IPO ('t Hart, Collier, and Cohen, 1990). Moreover, with a classification into intonation patterns, this study sets the first step in addressing the role of these intonation patterns for the expression of emotion. Such a classification could help in the understanding of the contribution of pitch in the vocal expression of emotion.

2. Outline

In this section, we will globally present the approach adopted in this study, which alternates: a production-oriented study investigating variations introduced by the speaker in the expression of emotion in speech, and a perception-oriented study investigating variations conveying emotion in speech to the listener. This approach underlines the complementarity of the production and the perception processes, that is the basis of the spoken communication process itself. At first, production data will be gathered that will serve as the inspiration for speech manipulations for a perception study. Natural speech will be analyzed at utterance level, i.e., at the global level of the whole utterance. This analysis will be carried out by means of statistical values concerning the entire utterances, such as mean F_0, standard deviation, and overall utterance duration. These production data will be related to the perception data. In a second cycle, we will turn to a more refined production analysis, below utterance level, i.e., considering variations occurring over time within the utterances, such as the variation in shape of the F_0 curves, F_0 values at specific points in the utterances, and the relative duration of accented and unaccented speech segments. The perceptual relevance of these observations will be investigated.

In Chapter II, the study first aims at identifying a restricted number of parameters that are of primary relevance for conveying emotion in speech, and that can be sufficient for conveying emotion in speech. Because of the relevance of the study for speech technology, it is important that these parameters can be easily manipulated in most synthesizers. The study further aims at quantifying these global parameters at utterance level. Speech material is selected from a database containing utterances spoken with intended emotions. As a result of this selection, the seven emotions: neutrality (as a reference), joy, boredom, anger, sadness, fear, and indignation, are involved in the present study. The semantic content of the sentences selected is first tested, in order to determine whether it is sufficiently unbiased for this study. Next, optimal values for the expression of specific emotions are experimentally investigated for the parameters pitch level, pitch range, and speech rate. The term 'pitch level' simply refers to how high or low the pitch is. At utterance level, different measures such as mean pitch, or end frequency of the utterance, can instantiate pitch level. 'Pitch range' refers to the magnitude of the F_0 fluctuations in the course of utterances, and can be instantiated by measures such as the standard deviation of mean F_0 or the distance between 'F_0 minima' and 'F_0 maxima'. A small sample from the database, containing fourteen utterances of a male speaker, is subjected to a preliminary analysis, in order to provide a rough approximation of initial parameter values for each emotion studied. An extended analysis is postponed until Chapter III. Six experiments are carried out to further specify the optimal parameter values for expressing emotion in speech. First, a frame of reference is established for the identification of emotions in the original speech. Then, speech is manipulated around the initial parameter values, and listeners indicate which versions they found best at conveying each particular emotion. Optimal values of these parameters are established on the basis of such perception experiments. In identification experiments, the optimal values are perceptually tested for the emotions studied, first in resynthesized speech, then in synthetic speech.

Subsequent chapters are concerned with understanding, in greater depth, how the variability associated with emotion in speech is realized in utterances. A discussion is opened about the accuracy that can be achieved by an approach at utterance level, and its sufficiency for the study of emotion in speech. These chapters are also concerned with identifying other features of pitch and duration relevant for conveying emotion in speech. These parameters should provide a greater level of detail than the global parameters at

utterance level, and allow the local changes in the course of utterances to be considered. Finally, the features are described and quantified within models, as far as possible.

In the first part of Chapters III and IV, analyses of emotional speech are carried out at utterance level, and the results of these production studies are related to those from the perception study reported in Chapter II. Chapter III is concerned with approximations of pitch level and pitch range by means of different measurements, and Chapter IV with overall speech rate.

Once parameters have been identified at utterance level and optimal parameter values have been established in Chapter II, a second cycle of the complementary studies of production and perception can be started. A refined production study is now carried out, considering local variations within utterances, i.e., below utterance level. Chapter III focuses exclusively on pitch variations. This refined analysis involves measurements of fundamental frequency (F_0), at fixed places in the utterances. The results of this refined analysis are discussed in the framework of a two-component model of intonation. Two main types of intonation models try to account for those aspects of the F_0 curves that are relevant to the speech communication. In the first type of models, F_0 curves are considered as the superposition of relatively fast pitch movements on a slowly declining line, i.e., the declination line or 'baseline'. Models of this type are called two-component models of intonation. One component represents the global aspect of the pitch curve; it represents the declination line and relates to pitch level. The other component represents pitch-movement variations and relates to pitch range. The second type of intonation models considers F_0 targets rather than pitch movements. Differences can be observed between analyses, interpreted in terms of one or the other type of model. The F_0 measurements, at fixed places in the utterances, lead to observations concerning the final lowering and the relative height of the pitch accents realized on lexically stressed syllables. The perceptual relevance of these observations is tested in a listening experiment. Chapter IV is concerned with the shape of the pitch curves over time, as described by means of the abstract intonation patterns. After making an inventory of the intonation patterns used in the emotional speech studied, an investigation is carried out as to whether the choice of intonation patterns is perceptually relevant for conveying emotion in speech and whether some patterns are more suited than others for conveying particular emotions. In Chapter V, an investigation is carried out as to whether temporal differences occurring in emotional speech are realized by linear stretching or shrinking, or

whether accented speech segments (i.e., accented lexically stressed syllables) and unaccented speech segments (i.e., one or several syllables in succession, realized without accentuation and being either lexically stressed or unstressed), are affected differently as a function of different emotions. As the overall speech rate varies as a function of emotion, it is necessary to verify whether temporal changes below utterance level occur as a function of the emotions themselves, or as a function of the overall speech rate varying with the emotions. Again, the analysis of speech is complemented by a listening test.

Chapter VI concludes this study by summarizing the main findings, and by explaining how the present study contributes to our understanding of conveying emotion in speech, of speech variability, and of modeling these variations. The advantage of studying speech both from a production and perception point of view will be emphasized.

Chapter II

Perceptually based optimal values for pitch level, pitch range and speech rate

ABSTRACT

This study focuses on the perception of emotion in speech. The identifiability of emotions in speech material was investigated. Systematic perception experiments were carried out to determine optimal values for the acoustic parameters: pitch level, pitch range and speech rate. Speech manipulations were realized, varying these parameters around the values found in a subset of the speech material, i.e., one utterance of two sentences by a male speaker acting out seven emotions: neutrality, joy, boredom, anger, sadness, fear, and indignation. Listening tests were carried out with this speech material, and optimal values for pitch level, pitch range and speech rate were derived for the generation of emotional speech from a neutral utterance. These values were perceptually tested in re-synthesized speech and in synthetic speech generated from LPC-coded diphones.

I. INTRODUCTION

Spoken communication involves more than just conveying the syntactic and semantic content of sentences. Prosody can add information, or modify the strictly linguistic content. Indeed, prosody not only carries information on word stress, phrasing and emphasis, but is additionally thought to be strongly related to speaker specific characteristics, and factors such as the expression of the speaker's emotions. Extra-linguistic information, given voluntarily or involuntarily by the speaker, is contained in prosodic features such as intonation, tempo, rhythm, precision of articulation and voice quality. Nevertheless, quantitative details of the correspondence between prosodic features and emotion, are still poorly understood.

Related studies concerned with the vocal expression of emotion have reported qualitative analyses of variations occurring in the expression of emotion in speech. Relatively few studies have tried to quantify the relevant dimensions in the speech signal for the purpose of obtaining parameter values for generating emotional speech (e.g., van Bezooijen, 1984; Cahn, 1990; Carlson, Granström, and Nord, 1992; Williams and Stevens, 1972). Furthermore, quantitative control over relevant prosodic features allowing the expression of emotion, could prove a powerful means for making synthetic speech sound more natural. Modeling this variability is expected to have an impact on the quality of synthetic speech, and therefore, to enhance the usability of speech technologies. However, the effect of such parameters has not yet been quantified.

The focus of this chapter is primarily oriented towards perception of emotion in speech. Indeed, if we want the results to be usable for synthesizing emotional speech by transformation of neutral speech, it is important to be concerned with the impression on the listener, especially since the computer does not experience any emotion. The present study is largely concerned with the ability of listeners to identify the emotion intended. As intonation and speech rate are considered to be highly relevant parameters for conveying emotion in speech, the focus in this chapter will be narrowed, in the present perception study, to three parameters: one for the end of the declination baseline, instantiating pitch level in this perception study, one for the excursion size of the pitch movements, i.e., distance between declination baseline and topline which is instantiating pitch range in this perception study, and one for the duration of utterances relative to neutrality, instantiating overall speech rate (more details concerning the notions

associated with pitch level and pitch range will be discussed in Chapter III). These parameters at utterance level (the global level of the utterance as a whole), are easy to control and to manipulate, also in Text-To-Speech systems.

Establishing a relationship between acoustic and perceptual data, requires a systematic approach. This study starts by examining the correspondence between the perception of emotion and the acoustic correlates of the expression of emotion. Indeed, initial values for the perceptive tests are inspired by an analysis of speech produced while expressing emotion vocally. Consequently, the procedure adopted here successively involves natural speech, manipulations of natural speech via analysis-resynthesis, and synthetic speech. An advantage of this systematic procedure is that the time course of the pitch parameters, as in natural speech, can be approximated by F_0 stylizations, i.e., straight-line approximations. These stylizations, when re-synthesized, result in speech that is perceptually identical to the speech re-synthesized with the original parameters. By using such analysis-resynthesis, one can then manipulate the parameters and study their perceptual importance. Next, if the stylizations are successful in conveying the emotion, optimal values can be deduced for rule-based synthesis.

A very general problem with this kind of investigation, regardless of whether one focuses on perceptual aspects through listening tests or on phonatory/articulatory aspects through acoustic analysis, is the question of ecological validity. How does one know that actors, when asked to express a certain emotion, do not simply reproduce something they acquired in training? How does one know that participants in listening experiments do not simply learn to assign a certain label to a perceived utterance, without their answer having anything to do with the real perception of a sign of emotional arousal in the speech? The true answer is that we never know for sure. We can, however, take measures in the design of experiments, that keep possibilities for artificial behavior to a minimum. The combined approach of eliciting emotions in actors, and independently testing the identification of these emotions by listeners, is such a measure. Furthermore, subjects in listening tests were never given feedback about correctness or consistency of their responses. In identification experiments there were never more than four repetitions of the same emotion expressed in the same utterance, and stimuli were presented in different, random orders to each listener. Subjects took listening-test runs independently of one another, while for each experiment, a new group of listeners was engaged. Without giving an absolute guarantee, all these measures should increase the likelihood

that, as far as pitch and speech rate variations are concerned, experimental data reflect the same perceptual and performance behavior as expression and perception of emotions in every-day life.

- ***Selection of acoustic parameters***

Even though earlier research (e.g., Fairbanks and Pronovost, 1939; Williams and Stevens, 1972; Bouwhuis, 1974; Ladd, Silverman, Tolkmitt, Bergman, and Scherer, 1985; Murray, 1989; Carlson, Granström, and Nord, 1992; Leinonen, Hiltunen, Linnankoski, and Laakso, 1997; Protopapas and Lieberman, 1997) has ascertained that parameters such as voice quality, loudness, rhythm, precision of articulation, and pause structure are relevant for the vocal expression of emotion, it would be convenient if, for the time being, the emotions could be modeled by a limited number of parameters.

Kitahara and Tohkura (1992) asserted that prosodic components, composed of pitch structure, temporal structure, and amplitude structure, contribute more to the expression of emotions than the spectral structure of speech. Frick (1985) reported that loudness was not an important means for communicating emotion. Lieberman and Michaels (1962) found that fundamental frequency (F_0) is very important, but that if it is the only information, it is insufficient to transmit full emotional content. House (1990) reported that F_0 is an important cue to the perception of mood, but that other cues such as the interplay between F_0, intensity dynamics, spectral characteristics, and voice quality are also crucial to the expression of emotion. Williams and Stevens (1972) determined that the aspect of the speech signal providing the clearest indication of the emotional state of the speaker is the 'contour of F_0 vs. time', i.e., the F_0 curve; emotion appeared to modify the pitch-curve shape generated for a breath group in several ways. Cosmides (1983) found that mean F_0 is one of the most consistently used parameters in the expression of emotion in speech. Cahn (1990) wrote that "the acoustic features affected by emotion are mostly the F_0 and duration correlates" (p. 38). Other studies have shown that average pitch level and average pitch range differ from one emotion to another (e.g., Fairbanks et al., 1939; Carlson, Granström, and Nord, 1992; van Bezooijen, 1984).

Because of all the interactions between the speech parameters, it seems sensible to initiate the study of emotion in speech by considering conveyors of major importance, that are preferably used with consistency in the expression of emotion in speech. This should provide a stable basis that could be extended later by studying more detailed

information on such parameters. Moreover, Ladd, Silverman, Tolkmitt, Bergman, and Scherer (1985) showed that pitch curves and voice quality have independent effects for conveying emotion in speech. This suggests that the study of pitch curves and that of voice quality can be conducted independently. Taking this last point into consideration, and on the basis of the findings of related studies summarized above, it was decided to concentrate in the present study on the pitch level, pitch range, and speech rate. These parameters are easy to control and to manipulate, also in a Text-To-Speech system, so that their communicative relevance can be checked. Moreover, the contribution of these acoustic parameters for the expression of various emotions, has already been assessed in the existing literature.

- ***Aim of the study***

The first goal in this chapter, is to verify whether emotional utterances can be synthesized from neutral utterances, by manipulating pitch level, pitch range, and speech rate at the rather global level of the full utterance. Controlling the intonation pattern might, thereby, help to give new insights.

The second goal is to determine to what extent the parameters pitch level, pitch range, and speech rate, are suitable for creating basic emotions in speech, by overlaying modifications on neutral utterances, and what their values should be. Basic rules about intonation and duration are formulated, quantifying acoustic parameters, and rule-driven expressions of emotion imposed on neutral natural and synthetic speech are tested in listening experiments.

The choice of generating emotions vocally, by converting neutral utterances into emotional ones, is based on the fact that, traditionally, most Text-To-Speech systems are designed to produce non-involved, neutral speech. Besides synthesizing emotion in Text-To-Speech systems, the present study offers the opportunity to investigate the contribution of prosodic attributes for conveying emotion in speech. Synthesis will, thus, not only constitute a goal, but also a tool for this investigation (Carlson, 1991; Carlson et al., 1992; Beckman, 1997).

II. SPEECH MATERIAL

A database, recorded at IPO in the context of the SOBU-project 92EA 'Emoties in spraak' (Eng.: Emotions in speech) was available. This database contains speech from three Dutch actors, two male speakers and one female speaker. They were instructed to elicit thirteen emotions: neutrality (Dutch category name: 'neutraliteit'), joy ('blijheid'), happiness ('geluk'), boredom ('verveling'), worry ('zorg'), anger ('boosheid'), sadness ('verdriet'), fear ('angst'), guilt ('schuld'), disgust ('walging'), haughtiness ('hautain'), indignation ('verontwaardiging'), and rage ('woede'). They were produced in evocative situations, on the eight following sentences: 'Zij hebben een nieuwe auto gekocht' (They have bought a new car), 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane), 'Hij is morgen naar Amsterdam' (He is in Amsterdam tomorrow), 'Het is bijna negen uur' (It is almost nine o'clock), 'Zij heeft gisteren gebeld' (She phoned yesterday), 'De lamp staat op het bureau' (The lamp is on the desk), 'Jan is naar de kapper geweest' (John has been to the hairdressers), 'Zij was aan het telefoneren' (She was making a phone call). The actors were native speakers of Dutch and expressed themselves in standard Dutch. The verbal content of the sentences was intended to be as semantically neutral as possible with respect to emotions. In order to evoke the emotions prior to reading out these sentences in a particular emotion, the actors uttered sentences of semantically emotional content, intended to evoke emotional situations, such as, 'How nice to see you here' for the expression of joy. In that mood, they spoke the fixed group of eight sentences. This was done for each emotion, and the procedure was repeated three times. The 936 speech samples of the database (3 speakers × 8 sentences × 13 emotions × 3 times), were digitized with 16 bit precision, at a sampling frequency of 10 kHz.

A preliminary perception experiment was conducted, in order to select seven specific emotions and five particular sentences for the present study. Utterances that presented no dis-fluencies were selected on the basis of identification performances, for being representative of the emotion expressed. The five sentences 'Zij hebben een nieuwe auto gekocht' (They have bought a new car), 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane), 'Het is bijna negen uur' (It is almost nine o'clock), 'De lamp staat op het bureau' (The lamp is on the desk), and 'Jan is naar de kapper geweest' (John has been to the hairdressers) in the expression of the seven emotions neutrality (Dutch category name: 'neutraliteit'), joy ('blijheid'), boredom ('verveling'), anger ('boosheid'), sadness ('verdriet'), fear ('angst'), and indignation ('verontwaardiging'), were selected for further use in the present study, which reduced the database to 315 utterances

(3 speakers × 5 sentences × 7 emotions × 3 times). A subset of 14 utterances (1 speaker × 2 sentences × 7 emotions × 1 time) was also selected, on the basis of identification performances, for use in the current chapter. It contained the sentences 'Zij hebben een nieuwe auto gekocht' (They have bought a new car) and 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane), spoken by the male speaker MR in the expression of the seven emotions mentioned above.

III. FRAME OF REFERENCE

1. Experiment 1: identification of the emotions in the original utterances

The first experiment served to provide a frame of reference for the previously selected utterances. Its purpose was to determine how well the emotion expressed in each selected original utterance could be identified, and how good an example of a particular emotion the subjects found each utterance to be.

a. Procedure

The previously selected fourteen utterances (1 speaker × 2 sentences × 7 emotions) served as stimuli. Ten volunteers agreed to participate as subjects in the experiment. They were either students or employees at IPO.

For each sentence, a block of fourteen trials was made, with each utterance being presented twice in a different random order to each subject. Sentence order was counterbalanced across subjects. This design resulted in twenty-eight stimuli. The entire test lasted about ten minutes. There was no training or feedback, which is also the case in all the following experiments reported in this study. After hearing an utterance over headphones, subjects decided which emotion was expressed; they were given the seven emotion labels to choose from. After this forced choice from the seven alternatives, subjects also attributed an adequacy rating for that expression, ranging between 1 (bad) and 5 (good).

b. Results

When the attributed emotion label corresponded with the emotion the actor intended to communicate, the response was considered to be correct. For each sentence, the proportion of correct responses was determined. Every time the subject gave a correct

Table 1: Mean proportion of correct responses and mean adequacy ratings on a 5-point scale (in parentheses) for Experiment 1: identification of the emotions in the original utterances selected

Emotion	Sentence 1: 'Zij hebben een nieuwe auto gekocht'	Sentence 2: 'Zijn vriendin kwam met het vliegtuig'
neutrality	1.00 (4.15)	.90 (4.06)
joy	.90 (3.11)	.55 (2.91)
boredom	.95 (3.95)	.95 (3.95)
anger	.50 (3.20)	.35 (2.71)
sadness	.95 (4.05)	.90 (4.39)
fear	.30 (3.83)	.35 (3.14)
indignation	.80 (4.06)	.95 (4.21)

Table 2: Confusion matrix for Experiment 1: identification of the emotions in the original utterances

Results are pooled across the two presentations of the two sentences and the ten subjects.

Intended emotion	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
neutrality	**38**	0	2	0	0	0	0	40
joy	7	**29**	0	2	0	0	2	40
boredom	0	0	**38**	0	1	0	1	40
anger	9	7	0	**17**	0	1	6	40
sadness	0	0	0	1	**37**	2	0	40
fear	2	8	0	1	2	**13**	14	40
indignation	0	5	0	0	0	0	**35**	40

response, the associated rating was taken into account in the computation of the mean adequacy rating. Mean proportions of correct responses and mean adequacy ratings are presented in Table 1, for both sentences separately. The confusion matrix of the results pooled over both sentences is given in Table 2.

Fear and anger were less successfully identified than the other emotions. There was a positive correlation between the adequacy ratings and the proportion of correct responses, which was found to be statistically significant [$r(14) = .69$, $p < .007$]. This indicates that utterances in which the emotion was easy to identify, were also considered to be more appropriate for that emotion. From the identification data in Table 2, estimates of stimulus entropy, response entropy and mutual information were calculated, as proposed by Shannon and Weaver (1949). As all stimuli were presented to the subjects an equal number of times, the stimulus entropy, for seven categories, is 2.81 bits per stimulus ($^2\log 7$); this measure serves as a reference. The estimate of mutual information provides information concerning the consistency of the responses given by the subjects and the quantity of information transmitted. Random responses of the subjects result in the mutual information being zero, while consistent responses of the subjects, associating a specific response exclusively with one specific stimulus, result in a mutual information equal to the stimulus entropy; note that this is the case even if the response does not correspond to a correct identification. The estimated mutual information transfer was found to be 1.71 bits per stimulus, which indicates that the subjects were rather consistent in assigning their responses. This corresponds well with the result of 74% correct identification of the intended emotions. The response entropy was found to be 2.69 bits per stimulus. This estimate is rather close to the stimulus entropy, which indicates that subjects were not particularly biased in their choices of labels in the experiment.

c. Discussion

The results form a frame of reference against which the results of the following experiments will be compared. Some emotions appeared to be more difficult to identify than others. The differences in adequacy ratings suggest that the speaker was not always equally successful in expressing some emotions, so that some utterances might constitute rather poor instances of a specific emotion. The least successful realizations of emotions, according to the adequacy ratings, were found to be the utterances of anger and joy. However, some emotions might be intrinsically more difficult than others to identify from the speech signal only. This seems quite reasonable considering that the physiological reactions caused by different emotions might be quite similar, which, in the acoustic space, could lead to a substantial overlap of the sets of acoustic correlates for different emotions. Emotions such as fear and anger can, for example, provoke dryness of the mouth (Williams and Stevens, 1972). Also, a specific emotion might provoke

different physiological states, which could lead to varying acoustic correlates. The respiration pattern, and any tremor can be indicators of an emotional conflict (Williams and Stevens, 1972). Different physiological states might also lead to acoustic correlates lying near each other in acoustic space. Fear appeared to be confused quite frequently with indignation, while no utterances intending to convey indignation were confused with fear. A possible explanation is that the set of speech correlates for indignation could be more scattered than the set for fear, with an overlap between the two sets. Neutrality and boredom, on the other hand, are hardly ever mistaken for any other emotion. Considering the fact that neutrality has been chosen as a reference to which all selected emotions relate, this observation concerning neutrality is reassuring. Confusion of other emotions with boredom is also very rare. This suggests that the acoustic correlates for boredom are distinctive. Furthermore, the positive correlation between identification rates and 'adequacy values' suggest that the identification rate is not merely a measure of stimulus discriminability, but also reflects the successfulness of the realization in terms of its expressivity.

2. Experiment 2: semantic content

Although acknowledging that there might always be a certain interaction between the linguistic content of a sentence and the conveyed emotion, an attempt was made to put an emotive overlay on a sentence which is as neutral as possible. It might still be that the semantic content of the utterances biases the responses of the subjects, who could experience greater facility or difficulty in associating specific emotions with particular sentences. The purpose of this control experiment was to test whether the semantic content of the utterances had any impact on the identification of the emotions.

a. Procedure

The two written sentences used in the present study formed the experimental material. Twenty volunteers participated in the experiment. The two sentences were presented to these subjects on paper. Subjects were asked to read them silently and to indicate, on a scale from 1 (bad) to 5 (good), how well the semantic content of each sentence fitted each of the seven emotions. The mean adequacy rating and its standard deviation were computed.

b. Results

The mean semantic adequacy ratings and its standard deviation are reported in Table 3. If we want the semantic content of the sentences to be neutral, while allowing the expression of the emotions studied, the expected results consist of a high score for neutrality, and equally low scores for all other emotions. It appears, indeed, that for both sentences, neutrality received the highest score on the semantic adequacy scale. On the other hand, the ratings obtained for the other emotions are not all equally low, which indicates that the emotions are not all equally likely to be conveyed with the sentences. The ratings obtained for joy, for instance, remain lower than the ones for neutrality, but are higher than the ratings for the other emotions. However, the most important point is whether the semantic content of the sentences influences the identification of the emotions. In order to test this point, the semantic adequacy ratings were compared with the results of Experiment 1. If the semantic content of the sentences is partially responsible for the differences in identification of the emotions in Experiment 1, one would expect the identification rates in Experiment 1 to positively correlate with the semantic adequacy ratings. On the other hand, if semantic adequacy and identification rates are independent from each other, correlation would be zero. It appears that the semantic adequacy correlated neither with the identification rates of Experiment 1 [$r(14) = .33$, NS], nor with the adequacy ratings of Experiment 1 [$r(14) = .00$, NS]. The

Table 3: Mean semantic adequacy ratings on a 5-point scale and standard deviation (in parentheses) for Experiment 2: semantic content

Means are pooled across the two sentences and twenty subjects.

Emotion	Sentence 1: 'Zij hebben een nieuwe auto gekocht'	Sentence 2: 'Zijn vriendin kwam met het vliegtuig'
neutrality	3.90 (1.41)	4.25 (1.12)
joy	3.30 (1.42)	2.80 (1.11)
boredom	1.50 (0.83)	1.45 (0.83)
anger	1.85 (0.99)	1.45 (0.76)
sadness	1.25 (0.44)	1.40 (0.82)
fear	1.15 (0.49)	1.80 (1.15)
indignation	2.35 (1.35)	1.95 (1.05)

two correlation coefficients do not significantly deviate from 0. This lack of correlation with the results from Experiment 1 shows that, despite the variations in the scores, the semantic content of the sentences had no substantial impact on the identification scores of the emotions in the perception experiment. Therefore, these sentences will be used in the following study.

3. Prosodic analysis of the speech material

A preliminary analysis of speech material was carried out on the fourteen utterances selected for use in the current chapter. An analysis of the whole database will take place in Chapter III. As a first approximation of the selected acoustic parameters at utterance level, measures for overall speech rate and measures for pitch range and pitch level (the latter as determined by subharmonic summation (Hermes, 1988)), were calculated. The speech rate was represented by the relative utterance duration, and, for this preliminary production study, the pitch level was represented by the overall mean F_0, and the pitch range by the standard deviation of this mean. The adequacy of these measures will be discussed in Chapter III.

In addition, a perceptual analysis of the pitch curves was carried out, for each utterance. This consisted of labeling each utterance according to the Dutch intonation grammar by 't Hart, Collier, and Cohen (1990). The elementary, basic units of this grammar are abstract pitch movements, i.e., categories of perceptually distinct pitch movements, such as 'early prominence-lending rise' ('1') or 'late prominence-lending fall' ('A'). In this intonation transcription, digits refer to rises and letters to falls. For conciseness sake, the abstract pitch movements will simply be referred to as 'pitch movements'. Where it is relevant, it will be made explicit that we refer to the concrete acoustic realization of pitch movements. An inventory of pitch movements and a set of rules allowing the generation of legal sequences of pitch movements constitute this grammar of intonation. A legal sequence of pitch movements for the whole utterance constitutes an intonation pattern. Any part of such an intonation pattern will be called a 'pattern of pitch movements'. Each utterance was attributed an intonation pattern in the intonation labeling according to this grammar. The results of this labeling are presented in Tables 4 and 5. Figure 1 shows schematized representations of intonation patterns occurring in the fourteen utterances. With only one exception ('14E'), all the observed sequences of pitch movements in Tables 4 and 5 constitute 'legal' combinations, i.e., they fit one of the attested intonation patterns that are covered by the grammar of Dutch intonation ('t Hart et al., 1990, p. 81).

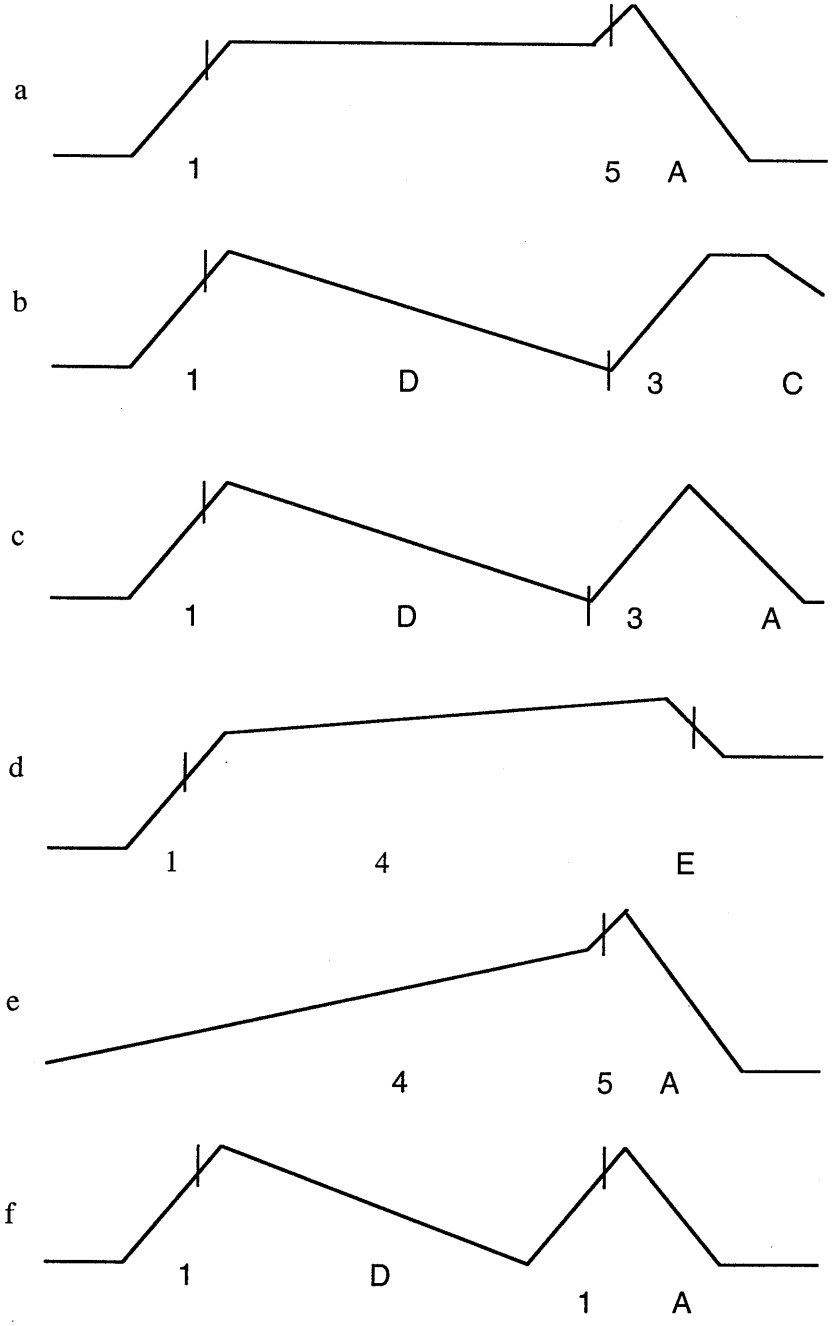

Figure 1: Schematized graphical representation of intonation patterns with their labels
Declination is not represented in these intonation patterns composed of standardized pitch movements. The vertical lines represent the vowel onsets.

Table 4: Duration, pitch, and intonation patterns of Sentence 1: 'Zij hebben een nieuwe auto gekocht'

Emotion	Duration (sec)	Duration relative to neutrality	Mean pitch in Hz	St. dev. pitch	Intonation pattern	
neutrality	1.66	1.00	132	18.7	15&A	see Fig. 1a
joy	1.58	.95	205	30.6	1D3C	see Fig. 1b
boredom	2.58	1.55	131	11.0	1D3&A	see Fig. 1c
anger	1.45	.87	201	31.6	15&A	see Fig. 1a
sadness	1.84	1.10	168	12.7	14E	see Fig. 1d
fear	1.40	.84	237	19.2	3C	see Fig. 1b
indignation	1.96	1.18	254	37.8	45&A	see Fig. 1e

Table 5: Duration, pitch, and intonation patterns of Sentence 2: 'Zijn vriendin kwam met het vliegtuig'

Emotion	Duration (sec)	Duration relative to neutrality	Mean pitch in Hz	St. dev. pitch	Intonation pattern	
neutrality	1.76	1.00	135	19.2	1D1D	see Fig. 1f
joy	1.78	1.01	205	39.3	1D3C	see Fig. 1b
boredom	2.35	1.33	133	13.2	1D1&A	see Fig. 1g
anger	1.58	.89	193	45.3	15&A	see Fig. 1a
sadness	2.12	1.20	170	19.2	1D1D	see Fig. 1f
fear	1.53	.86	230	27.2	1D3C	see Fig. 1b
indignation	2.35	1.33	245	53.3	45&A	see Fig. 1e

This suggests that no specific intonation patterns other than the ones included in the grammar are needed for the expression of emotion in speech. The same intonation pattern was used for the 'plane' and the 'car' sentence for the emotions anger ('15&A'), indignation ('45&A'), and joy ('1D3C'). In the expression of fear, both sentences ended in '3C'. On the other hand, the sequence '1D1D' is found for neutrality as well as for

sadness, '1D3C' for joy and fear, and '15&A' for neutrality and anger. A possible unique relationship between specific emotions and specific intonation patterns is not evident.

Tables 4 and 5 also present the absolute duration, the relative duration with respect to one of the neutral utterances of the corresponding sentence, and the mean pitch with its standard deviation. The emotions boredom and, to a lesser extent, indignation have a low speech rate, while fear and anger have a rather high one. The emotions neutrality and boredom have rather low average pitch values; fear and indignation have rather high ones. For indignation, anger, and, to a lesser extent, joy, the rather large standard deviation of the mean pitch seems an indication of rather large pitch variations in the expression of these emotions.

IV. OPTIMAL VALUES FOR PITCH LEVEL, PITCH RANGE, AND SPEECH RATE

Now that a reference has been established and parameters to be evaluated have been selected, the following series of experiments will determine optimal values for pitch level, pitch range, and speech rate, in order to proceed towards rule-based generation of emotional speech. The term 'optimal' is used for values that, among all values proposed in the experimental set-up, provide the best results in perceptual tests. In order to avoid caricatural emotions, the experimental set-up did not include parameter values giving an unnatural expression of emotion. In Experiment 3, an attempt will be made to determine optimal values for pitch level and pitch range that will be tested against the original pitch curves in Experiment 4. Experiment 5 will aim to determine optimal values for speech rate. In Experiment 6, the optimal values will be tested in re-synthesized speech. Finally, Experiment 7 will be an ultimate test for the optimal values in synthetic speech. These values will make up rules for the generation of emotional speech from neutral speech.

1. Experiment 3: optimal values for pitch level and pitch range

In the intonation grammar for Dutch by 't Hart et al. (1990), standard rules have been developed for the generation of synthetic pitch contours. A pitch contour is the rule-based straight-line manifestation of a specific intonation pattern, realized in a log F_0 versus time domain. The derivation of these generation rules is largely based on observations of neutral speech. The analysis of the fourteen emotional utterances selected for this study showed that, with one exception ('14E'), the combinations of pitch

movements realized were obeying this grammar. An investigation into the intonation patterns will be reported in Chapter IV. The present experiment intends to find optimal parameter values for pitch level and pitch range, for each emotion, when used in combination with the intonation patterns used by the speaker in original emotional utterances. If this appears to be possible, it will also confirm that the intonation patterns can be used for the generation of pitch contours adequate for emotional speech.

a. Procedure

The original F_0 curves of the emotional utterances were replaced by curves as in Figure 1, based on the rules of Dutch intonation. The pitch curves were synthesized by these rules, using dedicated software (Zelle, de Pijper, 't Hart, 1984). As input for this software, one has to supply the intonation pattern, number of pitch accents, and accent location. In this experiment, these data were derived from the actor's realizations. The number of accents, and the exact location of the vowel onset in the accented, lexically stressed syllables of the original utterances, were determined manually. The pitch contours for both the 'car' and the 'plane' sentences were synthesized using the same intonation patterns for the same emotion, i.e., the ones for the 'car' sentence reported for each emotion in Table 4. Since both sentences induce the realization of the same number of accents, namely two, this is not expected to favor the identification of emotions in the 'car' sentence. Nine synthetic F_0 curves were generated for each utterance, by varying the two other parameters: end frequency of the baseline in Hertz (Hz), which determines the pitch level, and excursion size of the pitch movements in semitones (s.t.), which determines the pitch range. The F_0 curves generated differed in three excursion sizes and three end frequency values. Excursion sizes were either 4, 6, and 8 semitones, or 6, 8, and 10 semitones. The variation in end frequency involved steps of 15 Hz around the values observed in the original utterances. As a reference, in standard applications for neutral speech synthesized in Dutch, the end frequency is fixed at 75 Hz, and the size of the pitch movements is fixed at six semitones (Collier, 1991). The declination line (the tilted baseline on which the pitch movements are superimposed), was automatically computed on the basis of the end frequency and the duration of the utterance according to Cohen, Collier and 't Hart (1982). The transplantation of the synthetic F_0 curves onto the original emotional utterances was based on the Time-Domain Pitch-Synchronous OverLap-and-Add (TD-PSOLA) algorithm (Charpentier and Moulines, 1989; Verhelst and Borger, 1991). This time-domain manipulation is performed directly on the continuous waveform, and allows a high quality of naturalness as long as the pitch

modification is not too substantial. The manipulation respected the segmental durations of the original utterances. The nine variants per emotion, resulting from this PSOLA manipulation, served as stimuli.

The two sentences were presented in blocks. Sentence order was counterbalanced across subjects. In a block, the stimulus order was randomly varied per listener. Ten subjects, different from the ones who participated in the previous experiments, volunteered to participate in the listening experiment. They were students or staff at IPO. For each emotion label, subjects could interactively listen to the nine variants (3 end frequencies × 3 excursion sizes). They could listen to the stimuli as often as they wanted. The stimuli were presented over headphones. The task was to choose the three variants that best express the given emotion label, and to rank these variants in first, second, and third choice.

The rank-order values of the three best variants were transformed into a score in which the very best variant received three points, the second best two, and the third selection one point. The mean score for each intonation variant of both sentences was calculated. The pitch curve that received the highest mean score was considered to be optimal for the specific emotion. The corresponding optimal parameters and their mean score are presented, per emotion, in Table 6. When the sentences are compared, a good agreement in the optimal values is shown. Spearman's rank-correlation test yields a high correspondence between the pitch levels (rs = .9821) and between the pitch ranges (rs = .9196) in both sentences.

b. Results

The results suggest that variations of the parameters pitch level and pitch range, parameters that are usually attributed standard values in speech synthesizers, are an important part of the generation of vocal emotions. The same intonation pattern can apparently express different emotions, depending on its pitch range and/or level. According to the information in Table 6, it appears that the '15&A' intonation pattern, associated with an end frequency of 65 Hz and a range of 6 semitones, is appropriate for neutrality. The same intonation pattern, associated with an end frequency of 110 Hz and a range of 10 semitones, is suitable for anger. On the other hand, the intonation pattern may also be relevant: joy and indignation in Sentence 1 have the same end frequency (155 Hz) and excursion size (10 s.t.), but these emotions were produced by the speaker

Table 6: Optimal parameter values for the rule-based pitch curves with corresponding scores and their standard deviations

Note that a score of 1 means on average a third place out of nine variants, while a score of 3, i.e., the maximum score, means the best choice out of the nine variants.

Emotion	Intonation pattern	Sentence 1: 'Zij hebben een nieuwe auto gekocht'			Sentence 2: 'Zijn vriendin kwam met het vliegtuig'		
		Pitch level (End frequency)	Pitch range in semitones	Score (s.d.)	Pitch level (End frequency)	Pitch range in semitones	Score (s.d.)
neutrality	15&A	65	6	1.90 (1.20)	65	4	2.10 (0.99)
joy	1D3C	155	10	2.40 (1.26)	155	10	2.90 (0.32)
boredom	1D3&A	65	4	3.00 (0.00)	65	4	2.20 (1.14)
anger	15&A	110	10	2.70 (0.95)	110	10	2.30 (1.25)
sadness	14E	95	6	1.30 (1.25)	110	8	1.60 (1.51)
fear	3C	200	8	2.30 (0.95)	200	7	1.80 (1.03)
indignation	45&A	155	10	2.40 (1.26)	185	10	2.50 (1.08)

with structurally distinct intonation patterns ('1D3C' and '45&A', respectively). It thus suggests that the three parameters, pitch level, pitch range and intonation pattern, play a role in the expression of emotions.

It also has to be noted that, in many cases, the versions that received the highest scores have extreme values (highest versus lowest frequency, smallest versus largest excursion). The fact that extreme values were selected by the subjects raises the question whether the range of the values for pitch level and pitch range presented to the subjects were large enough, or whether, for some emotions, subjects would have chosen even more extreme values if those had been included in the set of stimuli. On the other hand, including more extreme values in the stimuli set could have lead to the generation of emotions that could be perceived as exaggerated. Such cases have been avoided on the basis of the limited range of variations introduced by the manipulations around the values produced in the natural speech analyzed (steps of 15 Hz for pitch level, and 2 semitones for pitch range). Furthermore, it can be observed that subjects agreed more on the optimal realization of some emotions than on the optimal realization of some others. This is represented by the standard deviation of the adequacy ratings (Table 6).

2. Experiment 4: test of the optimal pitch curves

This experiment investigates the effect of standardizing the pitch curve. To this end, the rule-based pitch curves, generated with the optimal values for pitch level and pitch range obtained from Experiment 3, are compared with the pitch curves of the original emotional sentences. The recognizability of the vocally expressed emotion in the standardized utterances should be high if the rule-based pitch curves are perceptually close to the original, emotional ones.

a. Procedure

Test utterances were made by transplanting the synthetic pitch curves found in Experiment 3 to be optimal onto the original emotional utterances, i.e., the stylizations made of straight lines in the log F_0 vs. time domain with the values for pitch range and pitch level that received the highest scores in the previous experiment. In order to be sure that stylizing is not a factor, control utterances also had stylized pitch curves, but these were close-copies (de Pijper, 1983) of the original utterances. Indeed, a close-copy stylization is a straight-line approximation of the original F_0 curve which, in re-synthesis, is perceptually indistinguishable from the original. All the manipulations were again done by means of the PSOLA techniques. The temporal structure and the voice quality of the original emotional utterances was respected.

Ten new subjects from IPO participated voluntarily in the experiment. For both sentences, two blocks of fourteen utterances were successively presented to the subjects. Each block contained seven close-copy stylizations of the original utterances, and seven utterances with a rule-based F_0 curve in a different random order. The sentence order was counterbalanced across the subjects. After listening to an utterance once, subjects had to choose one of the seven labels corresponding to the emotion they thought was expressed in the utterance. In contrast with previous experiments, subjects were not asked to supply adequacy ratings, because utterances with stylized pitch curves were also involved. These utterances were perceptually indistinguishable from the original, thereby inducing a strong adequacy bias. As a consequence, the adequacy ratings for the manipulated versions were postponed until the next experiment.

Table 7: Mean proportion of correct responses for the original and the rule-based pitch curves

Pitch curves	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
Original	.85	.62	.92	.32	.97	.60	.85	.73
Rule-based	.67	.72	.85	.42	.75	.42	.77	.66

Table 8: Confusion matrix for Experiment 4: test of the optimal pitch curves

Results are pooled across two presentations of the two types of pitch curves, on the two sentences for the ten subjects.

Intended emotion	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
neutrality	**61**	0	10	3	3	0	3	80
joy	10	**54**	1	3	1	5	6	80
boredom	3	0	**71**	6	0	0	0	80
anger	31	15	0	**30**	0	0	4	80
sadness	5	0	3	2	**69**	0	1	80
fear	2	7	0	0	10	**41**	20	80
indignation	0	1	0	6	5	3	**65**	80

b. Results

For each listener, the proportion of correct responses was computed. A two-way analysis of variance (ANOVA), with the seven emotions and the two pitch curves as within-subjects variables, showed that the effect of pitch curve was not significant [$F(1,9) = 2.17$, $p = .175$]. The emotions in the utterances with the rule-based pitch curve were recognized almost as well as the ones with the close-copy stylization of the original utterances. The effect of emotion was very significant [$F(6,54) = 9.76$, $p < .001$]. An interaction was found between pitch and emotion [$F(6,54) = 2.58$, $p = .029$]. The mean proportion, reported in Table 7, was pooled over the two sentences. The confusion matrix is presented in Table 8. All emotions were identified well above chance level (.14), though anger (.42) and fear (.42) were less well recognized than the other emotions. As can be seen in Table 8, utterances expressing fear were often labeled as

indignation, and anger was very frequently confused with neutrality. Estimates of stimulus entropy, response entropy and mutual information were calculated (Shannon and Weaver, 1949) from the identification data in Table 8. The estimated mutual information was found to be 1.49 bits, stimulus entropy was 2.81 bits per stimulus, response entropy was 2.75 bits, and 70% of the emotion expressions were correctly identified. The mutual information and the response entropy indicate that subjects were not biased in their responses and that they were fairly consistent in their labeling.

c. Discussion

The results in Table 7 show that the difference in identification of the emotions in the test and the control utterances was rather small; the Spearman's rank-correlation [rs = .8036] indicates a high correspondence between the results obtained with both types of stimuli. This suggests that the manipulation had a small, though significant (t_s = 1.976, df = 158, $p < .05$), deteriorating effect. The results show that the complex original F_0 curve of an emotional utterance can be replaced by a simple approximation which only requires a limited number of parameters: intonation pattern, excursion size, which determines the pitch range, and end frequency which determines the pitch level. Moreover, the intonation patterns seem to be suitable for the generation of emotional speech. All emotions were identified well above chance level, in fact, fairly well, but it should be remembered that the utterances had the duration and the voice quality from the original utterances, so that the fairly good identification of the emotions is partly due to the presence of these other features of speech.

3. Experiment 5: optimal speech rate

This experiment is concerned with overall speech rate. Temporal variations can be realized as variations in articulation rate, rhythm, pause structure, or other temporal features. Clearly, the simplest manipulation of speech rate consists of a linear time compression or expansion of whole utterances, modifying the overall duration of the utterances, including both speech and any pauses within utterances, without affecting the fine temporal structure of the speech. If this linear approach, in which speech rate is inversely proportional to the overall utterance duration, is not too global for the expression of emotion in speech, optimal speech rates relative to neutrality may be determined for each emotion. These rates could then also be applied in rule-based diphone speech.

a. Procedure

The two original neutral utterances served as a starting point. First, they received the time-aligned F_0 curve from the emotional speech samples, via the previously mentioned TD-PSOLA algorithm (Verhelst and Borger, 1991). The resulting utterances, thus, had the same pitch curve as the emotional ones (i.e., the 'best' attainable pitch curve), but voice quality, energy, and all other micro-features of duration, were copied from the neutral utterance. In order to copy the time-aligned F_0 curve, the optimal Dynamic Time Warping (DTW) path was calculated, between the emotional and neutral utterance, so that the temporal correspondence was preserved. The F_0 curve was then copied from the emotional utterance to the neutral one by means of PSOLA. Second, the utterance created this way was made equal in duration to the original emotional one by linear compression/expansion via the PSOLA technique. The precision of this time-domain manipulation is limited to an integer number of pitch periods. Now the utterance had the same overall duration as the original emotional one, but the fine temporal structure was still from the neutral utterance. Starting from this situation, seven temporal variants were created by compressing/expanding the overall utterance by 70, 80, 90, 100, 110, 120, and 130 percent. These variants served as stimuli.

Ten subjects, working or studying at IPO and different from the previous subjects, agreed to participate in the experiment. The experiment took place in exactly the same way as Experiment 3; the only difference is that the subjects now had to choose and rank the three versions they found to be the best among the seven duration versions. The versions resulting from the overall speech rate manipulations were organized into two blocks (one for each text) and randomized. The sentence order was counterbalanced across listeners.

b. Results

As in Experiment 3, on the basis of the rank-ordering, a score of three points was assigned to the very best variant, two points were assigned to the second best variant and one point to the third best. The mean score of each variant was computed. The variant that received the highest mean score was considered to be optimal for that particular emotion. The corresponding optimal relative sentence duration, with score and standard deviation are reported in Table 9. For the expression of anger and joy, for instance, a

Table 9: Optimal sentence duration (speech rates) relative to neutrality and scores
Note that a score of 1 means on average a third place out of nine variants, while a score of 3, i.e., the maximum score, means the best choice out of the nine variants.

Emotion	Sentence 1: 'Zij hebben een nieuwe auto gekocht'		Sentence 2: 'Zijn vriendin kwam met het vliegtuig'		
	Relative sentence duration (speech rate)	Score (s.d.)	Relative sentence duration (speech rate)	Score (s.d.)	Mean relative sentence duration (mean speech rate)
neutrality	100% (1.00)	2.00 (1.00)	100% (1.00)	2.40 (0.92)	100% (1.00)
joy	85% (1.18)	2.30 (0.64)	80% (1.25)	2.00 (1.18)	83% (1.20)
boredom	154% (0.65)	1.70 (1.42)	145% (0.69)	1.90 (0.94)	150% (0.67)
anger	78% (1.28)	1.40 (0.92)	80% (1.25)	1.90 (1.14)	79% (1.27)
sadness	126% (0.79)	1.80 (0.92)	131% (0.76)	2.20 (0.87)	129% (0.78)
fear	92% (1.09)	1.50 (1.02)	85% (1.18)	1.50 (1.02)	89% (1.12)
indignation	129% (0.78)	1.80 (0.87)	106% (0.94)	1.40 (1.20)	117% (0.85)

speech rate higher than for neutrality was judged to be appropriate, while a lower speech rate seems to suit the expression of boredom and sadness.

c. Discussion

For all emotions except indignation, the speech rate found to be optimal for each of the two sentences differed by less than 10 percent. For indignation, the optimal overall duration varied from 106% to 129% (see Table 9). For this emotion, the subjects showed a tendency to select rates around the mean value. These results suggest that a variation in speech rate is easily tolerated for the expression of indignation, possibly because speech rate is not a very important attribute for the expression of indignation. In addition, the results correspond rather well with the speech rates of the original samples (see Tables 4 and 5) which corroborates the adequacy of the values found here.

The logical next step would be to test the speech rate values just obtained. It did not seem very interesting, however, to test speech rate values prior to testing both intonation and speech rate optimal values together. Indeed, letting subjects attribute an emotional label to neutral utterances on which only a manipulation of speech rate is imposed, did not seem sensible; under neutral circumstances, the most important factor determining

speech rate is time pressure. Another possibility was to test the optimal speech rate values by imposing them on emotional utterances, and comparing the identification performance against the one obtained with the original speech rate values. This option did not seem useful either, especially as the optimal values were only searched around the values produced by the speaker in the original speech. This would have resulted in rather small differences between the two conditions. Such a test would therefore only serve the purpose of confirming the role of speech rate for the expression of emotion in speech. Moreover, identification of the emotions in the original emotional utterances was already so high that a ceiling effect could have occurred. For this reason, rules concerning pitch and duration were tested together in the following experiment. The mean optimal values for pitch level and pitch range found for each emotion in Experiment 3, and the mean optimal values for speech rate found in the last experiment, were used simultaneously in an attempt to generate emotional speech from neutral speech.

4. Experiment 6: testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech

This experiment investigated whether it is possible to generate emotional speech from a neutral utterance, by applying the optimal values found for each emotion in the previous experiment for pitch level, pitch range, and speech rate. The adequacy of these values for conveying emotion in speech was also tested.

a. Procedure

Using PSOLA, a neutral utterance of both sentences was linearly compressed or expanded according to the optimal values deduced from Experiment 5. The places of the vowel onset, in the syllables to be accentuated, were determined by listening (Hermes, 1990). Using the appropriate software (Zelle, de Pijper, 't Hart, 1984), the rule-based F_0 curve with the optimal pitch level and the optimal pitch range from Experiment 3, was generated for each particular emotion. Using PSOLA, the neutral utterances with the rule-based overall speech rate were then provided with these F_0 curves. All the signals had the same voice quality and micro-duration structure, i.e., those of the original neutral sentences.

Ten new listeners recruited from IPO volunteered as subjects for the experiment. For each text, two blocks of fourteen trials were constructed. Per block, the seven emotional

utterances were presented twice in a different random order. The sentence order was counterbalanced across the subjects. The task of the subjects was to choose from the set of seven labeling alternatives, the emotion they thought was represented by each utterance they heard. They also gave a adequacy rating on a scale from 1 (bad) to 5 (good) for the chosen emotion.

b. Results

The mean proportions of correct responses and the mean adequacy ratings, pooled across the two sentences and the ten subjects, are presented in Table 10. On average, 48% of the emotions were correctly identified. In order to facilitate comparison, the reference results obtained in Experiment 1 are also presented in Table 10. An ANOVA, with the seven emotions as within-subject variables, showed that the differences between the emotions were significant [$F(6,54) = 9.76$, $p < .001$]. Sadness and fear were clearly identified less well than the other emotions. This was due to the confusion of fear with indignation and with sadness, and to a major confusion of sadness with boredom (see Table 11). The dispersed confusions concerning fear suggest that this emotion was insufficiently modeled. The modeling of sadness was even more problematic. The confusion matrix presented in Table 11 further shows that anger, that was recognized moderately well, was equally often labeled as anger and as neutrality. Joy was confused with indignation. Boredom and neutrality were very well identified. Estimates of stimulus entropy, response entropy and mutual information were calculated from the confusion matrix in Table 11. The estimated mutual information was 1.04 bits, stimulus entropy was 2.81 bits per stimulus, and the response entropy was 2.65 bits. This indicates that the consistency of the subjects in attributing their responses was not very good. This corresponds with the rather mediocre identification rate. The subjects were not particularly biased, but their confusions were not very systematic.

A comparison of the results of the present experiment, with the outcome of Experiment 1 that was concerned with the identification of emotion in natural speech (see Table 10), shows that the identification of the emotions boredom (.90), neutrality (.80), and anger (.38), is rather good in this experiment, if we compare it to the identification in natural speech (.95, .95, and .43, respectively). Although indignation was identified better than

anger in the present experiment, the identification performances for this emotion compare less well to those obtained with natural speech in Experiment 1. Other emotions such as joy (.38) and sadness (.20) also compare less successfully to the identification results obtained with natural emotional speech (.73 and .93, respectively). Considering the overall picture, Spearman's rank-correlation indicates a significant but small

Table 10: Mean proportion of correct responses and mean adequacy ratings for the pitch curves based on the optimal pitch level, pitch range, and speech rate, in comparison with the original utterances

Rule-based results (Exp. 6)

	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
Proportion correct	0.80	0.38	0.90	0.38	0.20	0.23	0.50	0.48
Adequacy ratings	3.87	3.47	4.14	3.33	3.75	3.44	3.05	3.58

Reference results (Exp. 1)

	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
Proportion correct	0.95	0.73	0.95	0.43	0.93	0.33	0.88	0.74
Adequacy ratings	4.11	3.02	3.95	2.96	4.22	3.47	4.14	3.70

Table 11: Confusion matrix for Experiment 6: testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech

Results are pooled across the two sentences and ten subjects.

Intended emotion	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
neutrality	**32**	0	4	3	0	0	1
joy	2	**15**	0	3	1	6	13
boredom	2	0	**36**	1	1	0	0
anger	16	4	1	**15**	0	1	3
sadness	1	0	28	0	**8**	0	3
fear	2	5	0	2	9	**9**	13
indignation	4	3	2	0	4	7	**20**

difference ($p > .05$) between both series of results for correct identification [rs = .6250], and a low correlation of the adequacy ratings in both experiments [rs = .2143]. This suggests that the identification of emotions, in emotional re-synthesized speech generated from neutral speech, is relatively successful, although the adequacy ratings are attributed differently by the subjects in both experiments.

c. Discussion

Despite the difficulties encountered with some emotions, the results for all emotions stayed above the chance level of 14.3%. The results clearly show that the rules based on the optimal values found for each emotion are successful to a certain extent, but there are costs associated with this rule-based generation of emotional-sounding speech from neutral speech.

In particular, the expression of sadness, using the selected parameters, seems to raise difficulties. This emotion seems to have a distinctive set of acoustic correlates, as suggested by the high identification rate and the infrequent confusions obtained with the original utterances in Experiment 1 (see Table 3). Furthermore, in the present experiment, the voice source and micro-duration features from the originally neutral utterances were used. This may mean that the voice source information, the micro-duration features, or both, are required to convey this emotion more clearly. In fact, simply listening to the original utterances expressing sadness, a remarkable element is the voice quality, different from that in other emotions. Fear was poorly recognized in the rule-generated utterances, but this was also the case in the original speech samples, which suggests that this speaker did not attribute acoustic cues that are distinctive enough for the expression of fear. The recognition of the rule-based utterances of anger was only moderate, but it was also moderate in the original utterances. So, at this stage, values for pitch and speech rate found for anger are not necessarily inappropriate, even though they do not lead to satisfyingly high identification scores. Identification of joy and anger was reasonably good in the original utterances, but it dropped in the rule-based versions to a moderate level, which is in fact what could be expected. The consideration of additional speech parameters may account for this difference. Finally, neutrality and boredom were recognized well above chance level, in all cases.

5. Experiment 7: rule-based generation of emotions from diphone-concatenated synthetic speech

The previous experiments demonstrated that it is possible to generate emotional speech by means of pitch and speech rate manipulations imposed on a neutral utterance. The aim of the present experiment is to investigate whether the rules based on these optimal values convey emotion in speech if imposed on synthetic speech. The goal is to generalize the findings to speech other than the one of the original speaker. Therefore, the values for pitch level, pitch range, and speech rate were applied to diphone-speech. Sufficient identification of the emotions from diphone speech would confirm that the parameters used are powerful cues for the expression of emotion in speech.

a. Procedure

Utterances were synthesized from a phonetic description of each of the two previously used sentences, using the IPO Text-To-Speech system (Van Rijnsoever, 1988). In this system, two LPC-coded diphone sets were available: one contained about 2000 diphones recorded using the voice of a male speaker (Zelle), the other contained about 1600 diphones recorded using a different male speaker (Bloemendal). The first set of diphones was coded in LPC with 12 poles, the second set with 18 poles. The phonemes were converted into LPC-coded diphones, using both diphone sets. The samples were synthesized at a sampling frequency of 10 kHz. The duration module of the Text-To-Speech system was kept active, so that the synthetic utterances could be considered adequate neutral expressions. These utterances were then, for each emotion, linearly compressed or expanded using the previously determined mean optimal speech rates (see Table 9). To generate monotonous utterances, the intonation module of the Text-To-Speech system was switched off, yielding a constant F_0. The location of the vowel onset in the lexically stressed syllables, was determined by listening (Hermes, 1990). An appropriate pitch contour was computed (see Table 4 for the corresponding intonation pattern) with the pitch range and the pitch level that were selected as the best ones in the former experiments (see Table 6). Finally, the resulting F_0 curves were imposed on the monotonous synthetic utterances. The voice quality and the fine temporal structure of the diphone-speech were left unchanged.

Twelve new subjects agreed to participate in the experiment. For each sentence, two blocks were realized, one for each set of diphones. The order of the emotions was different in each block. Two blocks, each in a different random order, made with the same sentence and the same set of diphones, were presented in succession. The order of presentation of the sentences and the sets of diphones was counterbalanced across subjects. Before starting the test with the diphone speech of one speaker, the listeners could get accustomed to the diphone speech of that speaker by listening to a passage of 40 seconds of instructions, synthesized with the corresponding set of diphones. This experiment was once again based on a seven-alternative forced choice paradigm concerning the seven emotion labels. The speech samples were recorded on a digital audio tape for presentation to listeners.

b. Results

The mean proportions of correct responses, pooled across the two sentences, are presented separately for each diphone set in Table 12. The confusion matrix of the responses, pooled across sentences and diphone sets, is presented in Table 13. On average, 63 percent of the emotions were correctly identified. An ANOVA, with seven emotions and two speakers as within-subject variables, yielded an insignificant scale value difference between the two diphone sets [60% vs. 67%, $F(1,11) = 2.47$, $p = .14$]. There were significant differences between the emotions [$F(6,54) = 11.45$, $p < .001$], and an interaction between the emotions and the diphone sets [$F(6,66) = 3.72$, $p < .005$]. Inspection of Table 12 suggests that joy was better identified with diphone set 1, whereas neutrality, anger, and sadness were better identified with diphone set 2. All emotions were recognized far above the chance level of 14.3%. The emotions boredom, neutrality, indignation, and joy were identified without problem. Anger was clearly identified, but was frequently confused with neutrality. In Experiment 4 'Test of the optimal pitch curves', anger was the less well identified emotion. It is possible that the intonation pattern that has been used for this emotion was not really optimal. Due to the fact that the identification of anger was 43% correct in the utterances with original pitch curve and 51% correct in the utterances with rule-based pitch curve, the possibility that the actor himself was not really efficient in expressing anger also has to be considered. Sadness was frequently confused with boredom, just as in the previous experiment. Though fear was the least successfully modeled emotion, with 41% correct identification, it was still clearly above chance level; most confusions occurred with indignation, as in previous experiments. Estimates of stimulus entropy, response entropy and mutual information

were calculated from the confusion matrix in Table 13. The estimated mutual information was 1.28 bits, the stimulus entropy was 2.81 bits per stimulus, and the response entropy was 2.76 bits. The estimates indicate no particular bias of the subjects and an acceptable consistency of their responses.

The most important result, however, is that the rather basic prosodic rules found in this study allow the generation of recognizable emotions in synthetic speech when applied to

Table 12: Mean proportion of correct responses for the pitch curves based on optimal pitch level, pitch range, and speech rate applied to synthetic speech

Results of Experiment 7	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
Diphone set 1	0.73	0.73	0.90	0.42	0.40	0.35	0.69	0.60
Diphone set 2	0.92	0.50	0.98	0.60	0.54	0.46	0.67	0.67
Overall results	0.83	0.62	0.94	0.51	0.47	0.41	0.68	0.63
Results of other experiments								
Results of Exp. 6	0.80	0.38	0.90	0.38	0.20	0.23	0.50	0.48
Reference results of Exp. 1	0.95	0.73	0.95	0.43	0.93	0.33	0.88	0.74

Table 13: Confusion matrix for Experiment 7: rule-based generation of emotions from speech synthesized by diphone concatenation

Results are pooled across two presentations of the two diphone sets on the two sentences to twelve subjects.

Intended emotion	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
neutrality	**79**	0	7	7	2	0	1	96
joy	5	**59**	0	7	0	12	13	96
boredom	1	0	**90**	2	2	0	1	96
anger	24	8	1	**49**	0	5	9	96
sadness	8	2	38	1	**45**	2	0	96
fear	4	16	0	4	12	**39**	21	96
indignation	0	2	1	6	12	10	**65**	96

a conventional synthesizer by diphone concatenation. Another interesting point is that the emotions in Experiment 7, involving synthetic speech, were better identified than in Experiment 6, involving re-synthesized speech, containing the voice quality and micro-duration structure of the original sentence intended to express neutrality. The difference could be due to the fact that the voice source and the temporal fine structure of neutrality can negatively influence the identification of the other emotions, which is in agreement with the findings of Carlson, Granström and Nord (1992). This suggests that these supplementary parameters are also of relevance for the vocal communication of emotions.

V. DISCUSSION

In the present study, most of the intended emotions were recognized far above chance level by the subjects. Siegwart and Scherer (1995) report that, in studies where the subjects' task is to infer the underlying emotion only by listening to natural speech, the accuracy of identification of the emotions was found to be approximately 50 or 60%, which is about five times higher than the chance level of 10 or 11%. This corresponds well with the identification rate of 74% for a chance level of 14%, found with the original natural speech in our frame of reference (Experiment 1: identification of the emotions in the original utterances). Identification rates also remained well above chance level in Experiment 4: test of the optimal pitch curves (58%), in Experiment 6: testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech (48%), and in Experiment 7: rule-based generation of emotions from speech synthesized by diphone concatenation (63%). In comparison with the findings of Siegwart and Scherer (1995), the acoustic cues used in this study (the type of intonation pattern, the F_0 level, the F_0 range, and the mean speech rate) seemed to allow the listeners to reliably identify most emotions. Stylized pitch curves obeying the intonation grammar for Dutch were found, as well as values found perceptually optimal for the other parameters studied concerning pitch and speech rate. Despite the fact that an eventual partial contribution of the PSOLA-manipulations to the identification of emotions cannot be excluded, such an effect seems quite unlikely, as the range of the manipulations was kept such that it did not introduce clearly conspicuous distortions in the speech signals.

Some emotions appeared to be more difficult to identify than others. For instance, fear was clearly less well identified than boredom, in all experiments. Various explanations for these differences in score can be proposed. The lower adequacy ratings for anger and joy in Experiment 1 indicate that some of the original utterances from the speaker might be sub-optimal realizations of a particular emotion. In fact, for an emotion such as anger, for which the emotion as expressed by the speaker was not satisfyingly well identified by the subjects, the values found optimal, on a perceptual basis, lead to an increase in emotion identification. Another explanation is that some emotions may intrinsically be more easily confused with each other, than other emotions, at least on the basis of the speech signal only. This is especially true with speech which excludes elements such as sighs, smacking sounds, or dis-fluencies, as is the case in the present study. Furthermore, cues other than the ones studied here might be especially relevant for some emotions; voice quality in particular may be a relevant cue (e.g., Cummings and Clements, 1995; Klasmeyer and Sendlmeier, 1995; Scherer, Ladd, and Silverman, 1984; Laukkanen, Vilkman, Alku, and Oskanen 1997).

The present findings seem to support the idea that emotions are signaled by a complex interaction of prosodic cues which, in principle, can be controlled in synthesized speech (e.g., Murray and Arnott, 1993). Despite some limitations of the present study (a limited number of emotions, only two sentences, a single speaker, relatively few listeners per experiment, a limited number of acoustic parameters), the results obtained are quite encouraging. Quantitative values for possible modeling of emotions have been proposed, that are based on, and have been evaluated through, perceptual measurements. Because of their relative simplicity, the parameters can, in principle, be manipulated in a broad range of synthesizers for the purpose of generating emotional speech.

One interesting finding of the present study has been that emotional speech can be generated with intonation patterns in concordance with the rules for the Dutch intonation grammar described by 't Hart et al. (1990). In this grammar, one has the option to choose from an array of structurally distinct intonation patterns in which one may vary the pitch level and the pitch range. No specific intonation pattern, other than these, were needed for a recognizable expression of emotion in speech. The sufficiency of intonation grammars as a basis for expressing emotions has not been discussed in any other literature as far as we know. No unique correspondence between emotion and intonation pattern was found. Depending on pitch range and/or pitch level, one intonation pattern

can apparently be used in the expression of different emotions. On the other hand, structurally distinct intonation patterns can be relevant for two emotions expressed with the same pitch range and pitch level. The perceptual element that the variability in intonation pattern may introduce was controlled in this study, but not investigated as such. The possible role of the intonation patterns in the expression of emotion in speech will be investigated in Chapters III and IV.

It has to be kept in mind that a unique modeling of emotion is probably not possible. One knows from everyday life experiences, that a particular emotion can be expressed in many ways, depending on the situation and on the cultural context. Anger, for example, can be expressed openly, freeing one's mind, or can be more or less repressed. Besides, individual speakers can prefer the use of different acoustic parameters to communicate the same emotion. Furthermore, as reported by Carlson, Granström and Nord (1992), extra factors such as sighs, linguo-dental smacks, voice breaks, and jitter can also contribute decisively to the vocal expression of emotion.

An important concern is with the possible generalization of the present findings. In the experimental design, precautions were taken in order to maximize validity; a possible training of the subjects was prevented as far as possible, the subjects took the tests independently of each other, they listened to the stimuli in different random orders, and a different group of subjects was used for each experiment. Subjects' answers appeared to be quite consistent, in spite of these precautions. The mutual information calculated in the identification experiments 1, 4, 6, and 7 supports this conclusion. The response entropies of respectively 2.69, 2.75, 2.65, and 2.76 bits, which compare to a stimulus entropy of 2.81 bits, confirm that subjects' responses were not particularly biased. People seem to agree quite well on how the expression of specific emotions sounds in Dutch. It does not seem unreasonable to generalize the perceptual behavior of our subjects to typical behavior of Dutch listeners. Moreover, the results could also be generalized to experimental settings in which other intonation grammars and other types of synthesizer would be used.

Part of the same validity issue is the potential difference between acted and spontaneous emotions. As a first step in research of emotional speech, the advantages of high-quality recordings, with control over the acoustic environment and over the content of utterances offered by acted speech, are obvious. Although Williams and Stevens (1972), who

Table 14: Parameters found optimal in the present study (in the left hand columns) and calculations on the speech material of Experiment 7: rule-based generation of emotions from synthetic speech averaged over two sentences and two diphone-sets (in the right hand columns)

Emotion	*Optimal values from Chapter II*			*Calculations on speech material of Experiment 7*	
	End frequency (Hz)	Excursion size of pitch movements (s.t.)	Duration relative to neutrality (%)	Mean F_0 (Hz)	'Declination. range'* (s.t.)
neutrality	65	5	100	94.0	6.5
joy	155	10	83	246.2	6.1
boredom	65	4	150	93.8	8.0
anger	110	10	79	185.6	6.0
sadness	102	7	129	160.3	7.3
fear	200	8	89	277.0	6.5
indignation	170	10	117	279.0	7.1

* 'Declination range' is here defined as the mean distance between begin and end frequency.

compared acted and spontaneous emotional speech, concluded that data obtained in spontaneous speech are not inconsistent with data obtained from acted emotions, further study involving spontaneous speech will become more attractive once more detailed rules have been determined. The same argument holds for potential differences in expression of emotion by different actors. The fact that a single speaker was used in this study forms a limitation that should be considered in future research.

The modeling proposed here seems, at the moment, to be quite acceptable for male speech in Dutch. It is, furthermore, methodologically founded. The approach not only made it possible to qualify the differences between emotions, but also quantified them in terms of deviation of the parameters from the neutral setting.

The parameter settings found to be optimal for conveying the emotions considered in the present study, are reported on the left hand side of Table 14, averaged over the two sentences, while they were given separately for each sentence in Table 6, with the intonation patterns. To facilitate comparison with the results of related studies, the mean

pitch and the mean distance between initial and final frequency, referred to as 'declination range', were calculated for the speech material of Experiment 7. The values averaged over both sentences and both diphone sets are reported on the right hand side of Table 14. These measures are convenient because they conform to those in related studies. From the two values for pitch range: excursion size of the pitch movements (simply called pitch range) given in column 3 and 'declination range' given in column 6, the highest of the two indicates the largest variation range, and is taken as a basis for comparison with other studies.

This study represents a step towards understanding how emotion is transmitted through prosody. The extent to which the present results can be generalized needs to be investigated. In the following chapters, the correspondence of the values found optimal in the current chapter, with production values measured in all the speech material selected, will be investigated. Further study will involve more sentences and different speakers, both male and female. Presently, there is little known about the settings that would be adequate for female speech. The role of intonation patterns for the expression of emotions, investigating which patterns are most adequate for specific emotions, and which meaning is most likely attached to a specific pattern, also deserves further study. Such an investigation concerning these intonation patterns will be reported in Chapter IV. Moreover, the controlled variation of additional parameters requires further study. The fluctuation of parameters during the utterance, for instance, might be interesting. Therefore, an investigation of pitch and speech rate variations occurring within utterances will be done in the next chapters.

Additionally and beyond the scope of the present thesis, the importance of the voice source also remains a subject demanding attention. Furthermore, a necessary step is to assess a widely accepted definition and taxonomy of emotion. That would permit more systematic future research, and would facilitate comparison of results. Efforts to consider the prosodic aspects of the signal in relation with the linguistic aspects should be rewarding. The scope of the use of the parameters in various languages, spoken in various cultures, is also a subject of interest.

Chapter III

F_0 fluctuations and pitch variations

ABSTRACT

A production study was conducted which focused on the mean and the standard deviation of the fundamental frequency (F_0) of the speech of three Dutch speakers, two males and one female. The informants spoke five sentences three times while expressing each of the following seven emotions or attitudes: neutrality, joy, boredom, anger, sadness, fear, and indignation. The results were compatible with those in Chapter II, in which optimal values for conveying the same seven emotions were derived on the basis of perceptual experiments. Because the description in global terms, such as mean F_0 and its standard deviation, may obscure more detailed characteristics of the speech produced that might be distinctive for the vocal expression of emotion, a more refined analysis was carried out. First, the pitch curves of each utterance of the database were labeled in terms of the intonation grammar for Dutch ('t Hart et al., 1990). It appeared that the '1&A' pattern, composed of a rise followed by a fall and lending prominence to the syllable on which it is realized, was frequently used, and could be used in the expression of each of the emotions studied. Some other patterns of pitch movements appeared to be used most by the speakers in the expression of some specific emotions. Second, a detailed analysis took place that was concerned with F_0 values at some fixed positions within the emotional utterances. The results of the refined description are first presented in the framework of a model of intonation. It appeared that some information on the pitch curve of emotional utterances could not be captured in the model. A perception experiment was conducted in order to test whether the information that could not be described by the model was perceptually relevant, and whether it contributed to the perception of emotion and attitude in speech. The information appeared to be relevant, but modeling it did not appear to consistently and substantially improve the identification of all emotions and attitudes. Even though modeling this variability improves the identification performances for indignation, it is detrimental to the identification of boredom and neutrality.

I. INTRODUCTION

It has been shown, both in production studies (Fairbanks and Pronovost, 1939; Williams and Stevens, 1972; van Bezooijen, 1984; Scherer, 1989; Kitahara and Tohkura, 1992) and in perceptual experiments (Lieberman and Michaels, 1962; van Bezooijen, 1984; Ladd, Silverman, Tolkmitt, Bergman, and Scherer, 1985; Cahn, 1990; Carlson, Granström, and Nord, 1992; previous chapter) that pitch level, pitch range and speech rate constitute strong prosodic cues relevant for the identification of a number of emotions in speech. In the current chapter, the work concentrates exclusively on one of the most promising speech parameters: pitch.

In the experiments described in Chapter II, utterances were re-synthesized with varying pitch level, pitch range, and speech rate, and listeners selected and ranked variants which they found best at expressing a given emotion. On the basis of these experiments, values found optimal for pitch level, pitch range and speech rate were found for each of the seven emotions studied. Synthesizing speech with values optimal for these emotions led to 63% correct responses in an identification test, which compares well with 74% correct identification in natural utterances produced by an actor. Although these results are quite promising, some questions remain. The first question is whether the values found to be optimal in the perception experiments of Chapter II correspond well with the range of values found in natural emotional speech of Dutch speakers. In order to investigate this point, the perception results from Chapter II will be compared with results from a production study carried out in the current chapter. This analysis is global, i.e., conducted at the level of the utterance as a whole, without taking variations within utterances into account. The adequacy of the measurements will be discussed. Additionally, the results, as far as pitch is concerned, will be compared with those of related studies. Those studies are briefly discussed below. The second question relates to the fact that pitch level and pitch range are generally determined at the utterance level. It is likely that information about the emotion expressed by the speaker is present in more detailed properties of the utterances. In this study, a more refined description of the course of F_0 through time will be presented, that is concerned with variations within utterances, i.e., below utterance level. This description is carried out in two ways. It combines a description of F_0 values at different positions in the course of the utterances, with a description of the course of pitch by means of labeling pitch curves into intonation patterns according to the IPO-model of Dutch intonation ('t Hart, Collier and Cohen, 1990).

Related studies

Some studies related to the present one are described in this section. The focus is on those parameters and emotions that are common to those studies and the present one. Other parameters, used in the various related studies but not in the present one, are mentioned in the text to indicate which other parameters are also considered to be important. In order to ease comparison of our findings with those of these related studies, an overview of their results concerning pitch in the emotions investigated in the present study, is reported in Table 1 together with our results from Chapter II.

Van Bezooijen (1984) studied the nine acted emotions disgust, surprise, shame, interest, joy, fear, contempt, sadness, and anger, often supplemented with neutrality. She furthermore considered thirteen parameters by means of rating scales. These parameters are not all physical parameters; they are: lip rounding, lip spreading, laryngeal tension, laryngeal laxness, creak, tremulousness, whisper, harshness, pitch level, pitch range, loudness, tempo, and precision of articulation. Measures of tempo, central tendency of F_0 (F_0 median), F_0 variation, F_0 perturbations and spectral slope, were made to serve as potential correlates for perceived tempo, pitch level, pitch range, harshness, loudness, tension, and laxness. Her study, like the present one, was concerned with spoken Dutch. Parameter values for pitch level and pitch range, for male speech, are given in Table 1. As mentioned above, only those emotions common to her study and ours are reported in this table. She aimed at establishing characteristics of vocal expression of emotion, and the recognizability of emotions as a function of sex, age, and culture. She described the characteristics of the emotions in terms of global perceptual ratings on her thirteen vocal parameter scales. She made acoustic measurements of potential correlates of some of the perceptual parameters, and observed the correlations between acoustic measures and their perceptual counterparts. She considered a broader range of characteristics than in most other studies. However, as in other studies, she did not study the intonation patterns.

Cahn (1990) studied the six emotions: anger, disgust, fear, gladness, sadness, and surprise, using seventeen parameters to synthesize sentences with these emotions with DECtalk3. The parameters: accent shape, average pitch, final lowering, pitch range, reference line, speech rate, breathiness, brilliance, laryngealization, and loudness, were represented in the synthesizer settings and organized around a zero norm (a value of 0 representing a neutral setting). The additional parameters 'contour slope', i.e., downward

Table 1: Results of related studies

This table only includes results concerned with the study of pitch. Only those emotions investigated in the related studies, that were also investigated in the present one, are mentioned here.

Present study (Chapter II)		
	end frequency as a measure for pitch level	*excursion size of pitch movements as a measure for pitch range*
neutrality	65 Hz	5 s.t.
joy	155 Hz	10 s.t.
boredom	65 Hz	4 s.t.
anger	110 Hz	10 s.t.
sadness	102 Hz	7 s.t.
fear	200 Hz	8 s.t.
indignation	170 Hz	10 s.t.
van Bezooijen (1984)		
	pitch level	*pitch range*
neutral	104 Hz	5.9 s.t.
joy	231 Hz	12.6 s.t.
anger	208 Hz	16.5 s.t.
sadness	158 Hz	8.1 s.t.
fear	203 Hz	9.2 s.t.
Cahn* (1990)		
	average pitch	*pitch range*
gladness	-3	10
anger	-5	10
sadness	0	-5
fear	10	10
Carlson, Granström, and Nord (1992)		
	mean pitch	*pitch movements*
neutrality	139 Hz	
happiness	154 Hz	high dynamic
anger	148 Hz	
sadness	131 Hz	monotonous
Kitahara & Tohkura (1992)		
	pitch structure	
joy	$(F_0(t) - F_0min) \times 1.4 + F_0min + 30$	
anger	$F_0(t) + 30$ (except for end of sentence)	
sadness	$(F_0(t) - F_0min) \times 0.6 + F_0min$	
Fairbanks & Pronovost (1939)		
	median pitch level	*total pitch range*
anger	228.8 Hz	20.6 s.t.
grief	135.9 Hz	18.0 s.t.
fear	254.4 Hz	22.4 s.t.
Scherer (1986, 1989)		
	mean F_0	*F_0 range*
joy	increase	increase
boredom	decrease	
anger		decrease
sadness		decrease
fear	increase	increase

* Cahn used a scale of values from -10 to +10.

tendency of the F_0 curve, stress frequency, exaggeration, fluent pauses, hesitation pauses, pause discontinuity, and precision of articulation, were, on the other hand, not related to the settings for neutrality. Her optimal parameter settings are reported in Table 1. She explored the effect of the acoustic parameters on the perception of synthesized emotion, varying the synthesizer settings and applying them to a pre-analyzed sentence. However, it remains unclear how the settings relate to acoustic values, and even more unclear how acoustic values should relate to those in human speech.

Carlson, Granström, and Nord (1992) studied the emotions happiness, sadness, anger, and neutrality, produced by actors, considering mean pitch and sentence duration. The parameter values they reported are presented in Table 1. They used analysis-by-synthesis to explore the possibilities offered by modeling aspects of speech that are more subtle than the usual parameters on the utterance level, in a parametric synthesizer. This approach provides information on the relative salience of the parameters, but again, the point of the relationship between the synthesizer values and the acoustic values remains unsolved.

Kitahara and Tohkura (1992) considered the emotions anger, joy, and sadness, relative to neutrality, and they investigated pitch structure, amplitude structure (stress and prominence), and temporal structure. Their temporal values are given in Table 1 in terms of relative duration, together with their recipes for pitch structure, to convert neutral speech into emotional speech. In their investigation of prosodic parameters conveying emotions, they carried out an analysis of emotional speech and performed listening experiments using synthetic speech. They tried to identify features of speech conveying the emotions studied and features conveying the degree of intensity of the emotion expressed.

Fairbanks and Pronovost (1939) conducted a production study, analyzing pitch curves, pitch level, pitch range and rate of change of F_0 during the simulation of the five emotions contempt, anger, fear, grief, and indifference. Their parameter values are reported in the lower part of Table 1. In this older study, authors went further than a description at the global level of the whole utterance and considered variations within utterances.

Scherer (1986, 1989) predicted changes in voice type for the emotions enjoyment/happiness, elation/joy, displeasure/disgust, contempt/scorn, sadness/ dejection, grief/desperation, anxiety/worry, fear/terror, irritation/cold anger, rage/hot anger, boredom/indifference, shame/guilt. In an effort to be explicit, Scherer specified the emotions he studied as 'couples'. In this way, he differentiated, for example,

enjoyment/happiness (a quiet bliss) from elation/joy (a bubbling elation). He considered F_0 perturbation, F_0 mean, F_0 range, F_0 variability, 'F_0 contour', i.e., F_0 curve, F_1 mean, F_2 mean, formant bandwidth, formant precision, intensity mean, intensity range, intensity variability, frequency range, high frequency energy, spectral noise, and speech rate. He developed a theoretical approach to the measurement of the vocal expression of emotion, and used analysis-resynthesis to study the influence of experimental variations of parameters on the inference of speaker affect. The general changes in parameter values he proposed are presented in Table 1.

Besides the interesting comparison of emotional speech produced by actors in a controlled situation and emotional speech recorded from a real-life situation, Williams and Stevens (1972) conducted a quantitative analysis of emotional speech at utterance level, as well as a qualitative analysis allowing more detail. They looked at differences among speakers and among emotions. They found that the clearest indication of emotion is the 'contour of F_0 vs. time', i.e., the F_0 curve. Since this emphasizes the relevance of a less global analysis, the need for a quantitative detailed analysis is felt even more.

Summarizing, the above presented studies give a good background for the present study, but do not always care to relate production data to perception data, do not generally quantify speech parameters below utterance level, and do not consider intonation patterns. A comparison with these studies will be done after presenting our own data.

The present study

The present study concentrates exclusively, but in much more detail than Chapter II, on the production of fundamental frequency variations and on the perception of pitch variations in speech. It would be interesting to find out how the perception results correspond with production data and to what extent the results of Chapter II can be generalized. An analysis of speech produced by three speakers might provide data to address these questions. In this sense, the present investigation is a follow-up on the perception study in Chapter II. The present production study was conducted on the totality of the speech material that was selected in Chapter II, while in that chapter, a preliminary analysis of a subset of fourteen utterances (1 male speaker × 2 sentences × 7 emotions) of this same speech material was conducted, in order to guide the manipulations of speech to be used as stimuli in the perception experiments. The same seven categories of emotions were studied, as is the case throughout this thesis: neutrality (Dutch category name: 'neutraliteit'), joy ('blijheid'),

boredom ('verveling'), anger ('boosheid'), sadness ('verdriet'), fear ('angst'), and indignation ('verontwaardiging'). Thanks to a common basis of speech material (same sentences, same emotions, a common speaker) the present analysis allows a proper comparison between production and perception of these emotions in speech.

The purpose of the production part of the study is to extract regularities and systematic differences in F_0 from the production data of the three speakers, and to investigate whether it is possible to distinguish each emotion on this basis. A global description of the data, in terms of mean and standard deviation, will be presented, and the results will be compared with the optimal values found for each emotion in the perception study described above. Indeed, considering how the production and the perception data relate, gives an added value to the results.

Next, it is challenging to find out how even better identification scores than in Chapter II can be obtained. The pitch parameters that were used in Chapter II (mean pitch and its standard deviation) had a global character, focusing on utterance level. A more detailed analysis might reveal whether F_0 curves contain more information that play a role in the identification of the emotions, than is present in mean F_0 and standard deviation. This analysis comprised in the first place a description of the F_0 curves produced, in terms of intonation patterns as presented in the IPO grammar of Dutch intonation ('t Hart et al., 1990). It appeared that in about all cases, this intonation grammar is adequate in this respect, i.e., the pitch movements of the grammar were adequate for describing the pitch curves produced in the large variety of emotional utterances studied here.

In the second place, the pitch values of the F_0 curves were analyzed at certain positions within the utterances. This will hopefully give better indications as to the pitch level and pitch range used by the speakers in different emotions. It is of interest to consider such an analysis of emotional speech in terms of some model, since it is uncertain whether a specific type of model of intonation can cope with such extreme data as the one collected under the expression of emotion. The adequacy of the type of model chosen will thus be discussed. A specific intonation model, i.e., the IPO-model, was chosen: it had been used in the previous chapter on perception, and the grammar of Dutch intonation ('t Hart et al., 1990) was used for the labeling of the pitch curves in terms of intonation patterns. Some of the data obtained could not simply be described in the IPO-model of intonation. The discrepancy will be described on the basis of an extension of this model. This study ends with an evaluation of the perceptual relevance of this discrepancy.

II. GLOBAL ANALYSIS: PITCH LEVEL AND PITCH RANGE

Figure 1 gives a schematized representation of concepts related to intonation. The *floor* and *ceiling* represent, respectively, the lowest and highest pitch the speaker is able to produce when speaking. The space between this floor and this ceiling is indicated with *tonal space*. In a two-component intonation model, a pitch curve consists of relatively fast pitch movements that are superimposed on a slowly declining line: the declination baseline. The end point of this declination line represents pitch level, while a measure for the excursion size of the pitch movements (i.e., distance between declination baseline and topline) represents pitch range. Equivalently, an F_0 curve can be represented within two slowly declining lines between which the concrete pitch movements are realized. The lower declination line is then referred to as the *baseline*, while the upper declination line is referred to as the *topline*. In this alternative, it is the distance between declination baseline and topline which represents pitch range. Since, in various studies (Maeda, 1976; 't Hart et al., 1990), it was concluded that the end point of the declination line is quite stable for the neutral speech of one speaker, the end point of the baseline is considered to represent the pitch level of the utterance within the two-component model used here, i.e., the IPO-model.

a. Speech material

The database contains speech from three Dutch speakers: the two male speakers MR and RS, and the female speaker LO, all native-speakers of Dutch and studying at a drama school at the time of the recordings. A subset of the database recorded by the male speaker MR was already used in Chapter II.

The same seven emotions as in the previous chapter were used, neutrality (as a reference), joy, boredom, anger, sadness, fear, and indignation. In order to elicit the emotions, the speaker first spoke a semantically emotional sentence, such as the sentence 'How nice to see you here', for the expression of joy. Once in the aroused mood, the speaker realized the five different sentences listed below with that same emotion. All emotions were elicited in this way, one after the other. This whole procedure was repeated three times. The utterances were recorded on DAT-tapes and, digitized with 16 bit precision at a sample frequency of 10 kHz, stored in computer sound files.

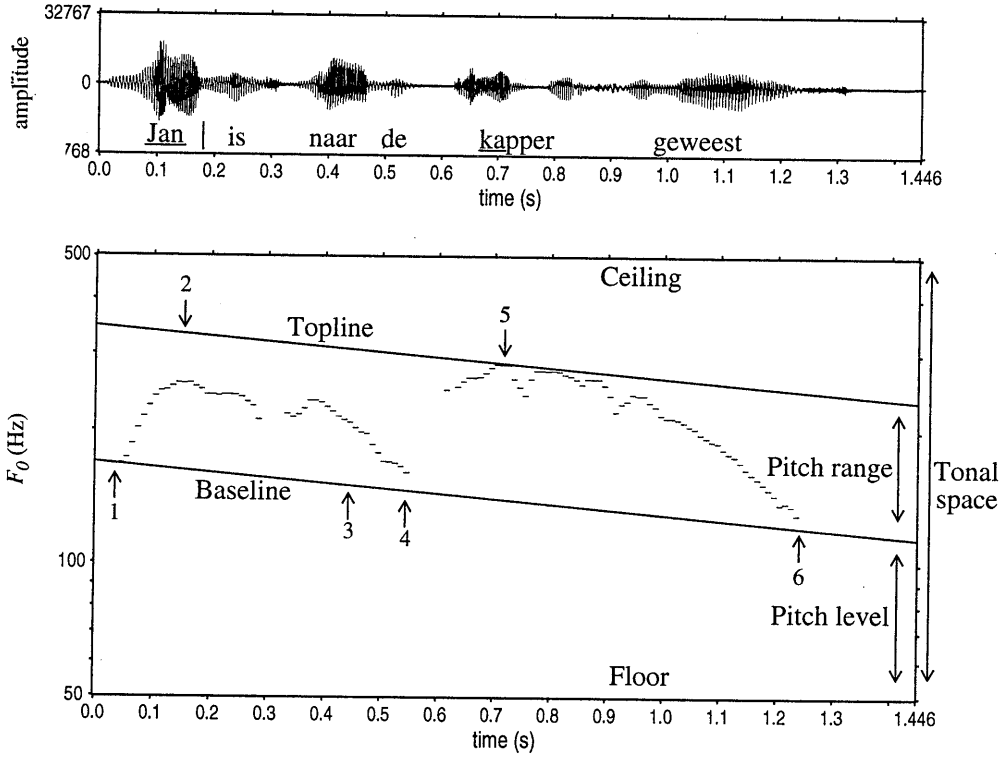

Figure 1: Representation of notions associated with pitch level and pitch range

The semantic content of the sentences was shown to be neutral in Chapter II. The utterances were, in most cases, realized with an accent on at least two syllables. Naturally, all accents were realized on a lexically stressed syllable. In the list of sentences presented below, lexically stressed positions on which an accent was realized in utterances of the database, are underlined. One accent was realized in the 'initial part' of the sentences, before the vertical slash. This first accent corresponds to what we will call the 'initial pattern', which can, in fact, either be the first pattern of pitch movements occurring in the utterance, or only the first pitch movement, if the pattern of pitch movements stretches over both parts of the utterance, such as '1A' or '1EA'. In the final part of the sentence, it was possible to realize two accents for the sentences 'auto' and 'negen uur'; in those cases the pattern we call 'final' is indeed the last realized combination of pitch movements, or only the last pitch movement, such as 'A' or 'EA', if this so-called 'final pattern' was initiated in the initial part of the sentence. It is, thus, not always the second pattern in the utterance. The parts carrying the 'initial' and the 'final pattern' are separated by a vertical slash in the sentences as follows:

'auto': Zij hebben | een nieuwe auto gekocht (They have bought a new car).
'vliegtuig': Zijn vriendin | kwam met het vliegtuig (His girlfriend came by plane).
'kapper': Jan | is naar de kapper geweest (John has been to the hairdresser).
'lamp': De lamp | staat op het bureau (The lamp is on the desk).
'negen uur': Het is bijna | negen uur (It is almost nine o'clock).

b. Procedure

The issue addressed now is how the main parameters of the two-component intonation model, pitch level and pitch range, can be determined from the F_0 curve realized in an utterance. In the previous perception study (Chapter II, Experiment 7), synthetic F_0 curves were generated with the end frequency of the baseline and the excursion size of the pitch movements as input parameters. In F_0 curves of natural utterances, it is not generally possible to determine the baseline, its end frequency, and the excursion of the pitch movements. These quantities have to be estimated on the basis of measurable parameters, such as mean pitch for estimating pitch level. As for estimating the pitch range, the best measure would be to consider the pitch variation around the declination line. Because of the difficulty of actually determining this declination line in natural speech, it was decided instead, to use a measure that considers the pitch variation around the mean pitch, i.e., the standard deviation of the mean pitch. Hence, as a first approximation, in order to evaluate this frequently used estimation of pitch range, it is assumed that pitch level and pitch range

are monotonically related to the mean pitch of the F_0 curves and its standard deviation, respectively. The adequacy of this estimation will be discussed later on. For the moment, we will present the mean pitch of the natural pitch curve and twice its standard deviation, and compare this with the mean pitch of the synthetic utterances and twice its standard deviation. (Anticipating later results, the reader can assume that this mean F_0 value minus twice its standard deviation, is the best estimate for the end frequency of the baseline.)

So, for each utterance of the three speakers, the mean F_0 of the F_0 curve and its standard deviation were measured. The choice of frequency scale is a relevant matter for the comparison of pitch ranges of women and men, and for the comparison of pitch ranges at high and low pitch levels (Terken and Hermes, forthcoming). Unfortunately, there is, as yet, no agreement about the adequacy of different scales. Rietveld and Gussenhoven (1985) found that their results best fit the Hertz scale. Hermes and van Gestel (1991) are in favor of the equivalent-rectangular-bandwidth (ERB) rate scale, while Traunmüller and Eriksson (1995) favor the semitone scale. The ERB scale is about logarithmic above 500 Hz, while for lower frequencies, it is intermediate between the linear and the logarithmic frequency scale. We decided to restrict ourselves to the two 'extreme' scales, thereby neglecting the 'intermediate' alternative. Therefore, the results will be presented both in Hertz (Hz), and in semitones (s.t.) relative to the arbitrary reference of 50 Hz. Although these two measures are different, calculations of the end frequency value yield very similar results, independently of the measure used. Pitch measurements were conducted by subharmonic summation (Hermes, 1988).

c. Results: mean and standard deviation

Observed means and standard deviations were subjected to an analysis of variance (ANOVA) with emotion (7), speaker (3) and sentence (5) as within-subject variables. As far as the mean is concerned, when presented in semitones, there were main effects of speaker [$F(2,290) = 483.7$, $p < .0001$], emotion [$F(6,290) = 129.6$, $p < .0001$], and interaction between speaker and emotion [$F(12,290) = 10.5$, $p < .0001$]. A minor but nevertheless significant effect of sentence was also found [$F(4,290) = 4.2$, $p < .01$]. Other interactions were non-significant ($p > .05$).

Table 2: Mean fundamental frequency averaged over three trials in five sentences, presented in Hertz per speaker and per emotion, and twice its standard deviation in Hertz

Measurements in speech synthesized using the values perceptually found optimal in Chapter II are presented in the right hand columns.

Emotion	Speaker MR		Speaker RS		Speaker LO		Synthetic speech	
	mean	2 × sd	mean	2 × sd	mean	2 × sd	mean	2 × sd
neutrality	129.6	37.4	98.2	30.2	167.2	70.4	95.4	29.0
joy	203.1	74.4	145.5	81.2	263.5	132.6	244.4	131.8
boredom	125.9	32.8	95.2	22.6	169.7	67.2	95.0	26.2
anger	191.6	77.0	168.2	90.6	252.6	125.4	180.3	102.0
sadness	192.0	48.2	137.0	52.6	282.7	117.0	165.9	54.2
fear	239.8	68.6	134.1	54.2	388.7	159.2	280.6	90.6
indignation	241.1	106.0	184.4	100.2	269.9	201.4	269.8	124.2

Table 3: Mean fundamental frequency averaged over three trials in five sentences, presented in semitones (s.t.) above 50 Hz, per speaker and per emotion, and twice its standard deviation in s.t.

Measurements in speech synthesized using the values perceptually found optimal in Chapter II are presented in the right hand columns.

Emotion	Speaker MR		Speaker RS		Speaker LO		Optimal values for synthetic speech	
	mean	2 × sd	mean	2 × sd	mean	2 × sd	mean	2 × sd
neutrality	16.3	5.3	11.4	5.4	20.4	8.3	11.0	5.4
joy	23.9	6.8	17.6	9.8	28.0	9.9	26.8	8.6
boredom	15.8	4.7	10.9	4.0	20.6	7.2	10.9	5.0
anger	22.8	7.8	19.8	10.0	27.1	9.8	21.5	9.9
sadness	22.9	4.4	16.7	6.9	29.3	8.1	20.5	5.7
fear	26.9	5.5	16.4	6.9	34.8	8.2	29.6	5.0
indignation	26.6	9.3	21.4	10.1	27.9	12.1	28.6	9.2

As far as the standard deviation in semitones is concerned, there were main effects of speaker [$F(2,290) = 47.4$, $p < .0001$], emotion [$F(6,290) = 35.4$, $p < .0001$], and a smaller but still significant effect of sentence was also found [$F(4,290) = 7.5$, $p = .0001$]. A small interaction between speaker and emotion [$F(12,290) = 2.1$, $p < .02$] was also found. Other interactions were non-significant ($p > .05$).

The significant effect of sentence indicated that the mean and standard deviation of one sentence individually were not reliable estimates of the mean and standard deviation of F_0 curves in general. The data of the five sentences were pooled so that the result represented a more reliable estimate of the mean and the standard deviation of F_0 curves in general. Hence, for each speaker and each emotion, the mean and standard deviation of the F_0 curves were estimated by averaging fifteen measurements (3 trials × 5 sentences). The thus obtained mean pitch and twice the standard deviation for the three speakers and the seven emotions are presented in Table 2 (in Hertz) and in Table 3 (in semitones above 50 Hz). In both tables, the first three double columns supply the results for the F_0 curves of the utterances produced by the three speakers, while the last double column presents the results for the synthetic F_0 curves, to be discussed later. (Whether the calculations of the mean pitch and its standard deviation are carried out on a scale of Hz or on an ERB-rate scale, the results are almost identical.)

Two issues will now be discussed. The first issue is to consider to what extent the emotions are characterized by the mean F_0 and the standard deviation of F_0. The second issue is to consider whether these values correspond with the optimal values found in the perception study as presented in Chapter II. 't Hart (1981) showed that only differences of more than three semitones play a part during communication. Therefore, it is proposed here to only pay attention to differences between groups of emotions which are larger than three semitones. For the three speakers (see data in Table 3), boredom and neutrality can then be characterized by a comparatively low mean F_0, which is associated with a small standard deviation. The core of a second group of emotions is formed by joy and sadness, emotions with a medium mean pitch, to which are added: anger for speaker MR, fear for speaker RS, and anger and indignation for speaker LO. Note that anger, fear and indignation can also be characterized by a high pitch for other speakers. Anger, joy and indignation are additionally characterized by a large standard deviation. For all three speakers, these three emotions show the largest standard deviations.

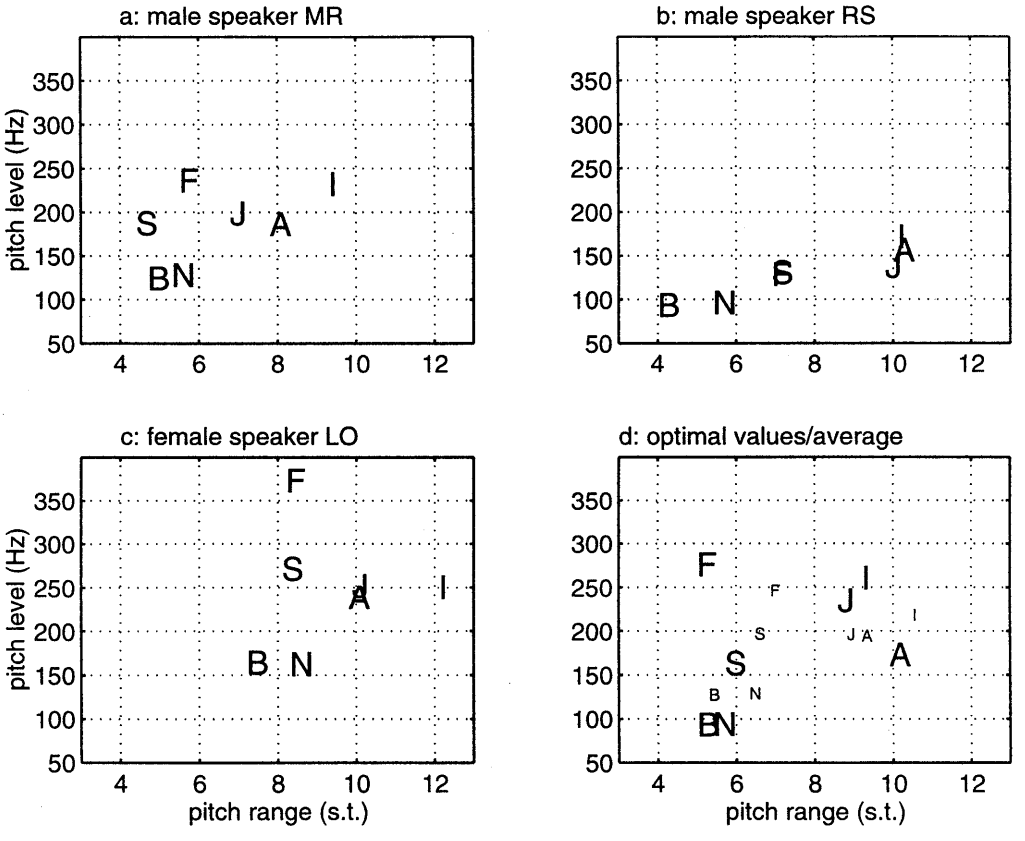

Figure 2: Representation, for seven emotions, of mean F_0 and its standard deviation for the three speakers (panels a, b, and c) and the perceptually optimal values from Chapter II (panel d). The average data over the three speakers are also represented in this last panel (in small symbols)

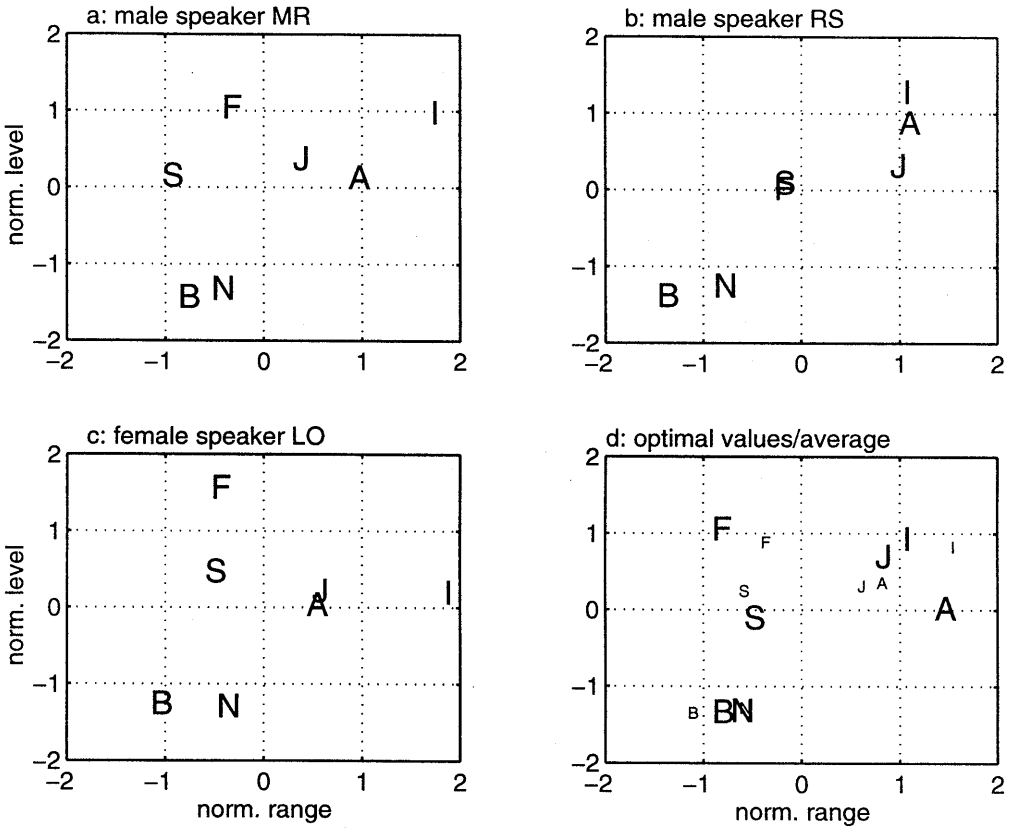

Figure 3: Normalized representation of pitch level and pitch range for the speakers, the seven emotions, and the perceptually optimal values

Normalization was carried out by subtracting the average value from each data point, followed by a division by the standard deviation of the data points. The normalized data points $\{n_i\}$ are calculated from the original data points $\{x_i\}$ according to $n_i = (x_i - \bar{x}) / s$, in which $\bar{x}$ is the average of $\{x_i\}$ and s the standard deviation of $\{x_i\}$. This was done for the ordinate (pitch level) and for the abscissa (pitch range). The result is that, for the three speakers and for the optimal values, the averages of the data are equal (= 0) and have the same standard deviation (= 1).

An overview of the estimated pitch levels and pitch ranges is plotted in Figure 2. Panel a, b, and c present the data for the speakers MR, RS and LO, respectively. In panel d, the values found perceptually optimal in Chapter II are plotted, together with the average data of the three speakers. The range of the data can vary strongly from speaker to speaker. Not only are the pitch levels lower for the male speakers MR and RS than for the female speaker LO, it appears, for example, that the variation in pitch level of RS is much smaller than that of MR. Apparently, RS makes less use of pitch level to express certain emotions than MR. The pitch range is represented by one standard deviation. In order to normalize for this variation, the data points of the three speakers, and also the perceptually optimal values, have been normalized to an average of 0 and a standard deviation of 1. This normalization, correcting for the differences in level and range among speakers, was done with values expressed in semitones. This transformation is not linear, but differences from a normalization in Hertz are very small. The results are plotted in Figure 3. Panel a, b, and c present the normalized pitch levels and ranges for the speakers MR, RS and LO, respectively. For the sake of clarity, note that negative values of normalized ranges correspond to ranges smaller than the average range of the seven emotions. Panel d presents, in larger font size, the normalized perceptually optimal pitch level and range, while the letters in smaller font size in panel d represent the average of the three speakers. In this panel, it can be seen that, as far as clustering in three groups of emotions is concerned, the two arrangements of the emotions compare rather well (see Figure 3, Panel d). Neutrality and boredom are characterized by a low pitch level and a small pitch range. Sadness and fear have a small pitch range, but the pitch level for fear is higher than for sadness. Joy, anger and indignation have an average pitch level but a high range. This description applies to the perceptually optimal values presented with larger font sizes as well as to the averages of the production of the three speakers presented with smaller font sizes. This is an indication, within the context of our study, of the agreement between production and perception.

In order to investigate whether the estimated pitch level is significantly different for the expression of various emotions, correlations were computed between the two sets of pitch-level values *in semitones* that were generated by the three speakers for each combination of two emotions (the computation involved each combination of two rows in Table 3). There are correlations of .80 or more among the emotions joy, sadness, fear, boredom and neutrality. The highest correlation (.92) is found between neutrality and boredom. The emotions indignation and anger have lower correlations with other emotions and with each other; $r < .70$, with one exception, namely $r = .74$, between anger and fear. Indignation

and anger appear to distinguish themselves from other emotions on the basis of pitch level. For all combinations of two emotions, correlations between estimated pitch ranges in semitones were rather low, the highest correlation being r = .65, between the pitch ranges for fear and for sadness. Correlations with r < .25 occur between the pitch ranges of anger and that of all other emotions, indicating that pitch range is distinctive for anger. Other correlations with r < .25 are found between joy and boredom and indignation, respectively, between indignation and sadness, and between boredom and sadness and neutrality, respectively. On the other hand, the highest correlations are found between fear and neutrality, and between sadness and neutrality, fear, and joy, respectively. Pitch range, thus, seems to be least distinctive for sadness.

Making similar computations *in Hz* (see Table 2) showed that emotions with a low mean pitch often had a small standard deviation, while emotions with a higher pitch level showed a larger standard deviation. Fear is an exception, as it combines a high mean pitch with a rather small standard deviation. If there is a positive relation between pitch level and pitch range, we might question the value of the two independent measures. Computing correlations between mean F_0 and its standard deviation, it appears that the correlation between the mean F_0 in Hz and its standard deviation in Hz is, with r = .742, much higher than the correlation between the two measures in s.t., with r = .298. This only corroborates the idea that, for our purpose, these measures are best considered in semitones.

d. Results: end frequency and excursion size

Now, the production data will be compared with the perception-based optimal values found in Chapter II. In that perception study, end frequency of declination baseline featured pitch level, while excursion size, i.e., the distance between declination baseline and topline, featured pitch range. In the current production study, the end frequency of the baseline and the excursion size cannot easily be measured directly, but they can be derived from the natural F_0 curves.

As for *pitch level*, in order to test to what extent the end frequency of the baseline can be estimated from mean and standard deviation, the end frequency of the synthetic F_0 curve, which was exactly known (see Table 4), was estimated from the mean and its standard deviation. It appeared that about the best fit could be obtained by subtracting twice the standard deviation from the mean. (The best fit was found for the mean minus 2.012 times

the standard deviation, both expressed in semitones.) In the following, the mean minus twice the standard deviation in the F_0 curves will be used to estimate the end frequency of the F_0 curves. Statistically, this means that only about 2.3% of the end frequencies will fall below this value. The results of this comparison are presented in the first three columns of Table 4. In all cases, the differences between the actual end points and the estimated end points was less that three semitones, though for sadness, a difference of 2.5 semitones approximated this level of communicative significance of three semitones, as defined by 't Hart (1981). In our data, where two accents were realized on the large majority of utterances, these results show how well pitch level can be estimated on the basis of measurements of mean F_0 and its standard deviation.

As for *pitch range*, 2.2 times the measured standard deviation appeared to yield the best estimate. For only two out of seven emotions, the approximation of the pitch range, based on standard deviation of the mean, differ from the input values for excursion size by more than one semitone. The largest difference is 2.5 semitones for fear. An ANOVA concerning the estimated standard deviation in semitones, based on the mean from which twice the standard deviation was subtracted, showed that there were main effects of

Table 4: Pitch level and estimated pitch level (Hz), pitch range and estimated pitch range (s.t.)

The optimal values derived from the perception experiments in the previous study, were used to synthesize speech. The optimal values are reported here together with estimated values resulting from measurements made on this rule-based synthesized speech. The estimated pitch range is twice the standard deviation measured in synthetic speech.

	Pitch level (Hz)		Difference (s.t.)	Pitch range (s.t.)		Difference (s.t.)
Emotion	Optimal end freqs	Estimated end freqs	Difference optim./estimat. end freqs	Optimal pitch range	Estimated pitch range	Difference optim./estimat. pitch Range
neutrality	65	69.2	1.1	5	5.9	0.9
joy	155	138.9	-1.9	10	9.5	-0.5
boredom	65	70.9	1.5	4	5.5	1.5
anger	110	97.4	-2.1	10	10.9	0.9
sadness	102	117.8	2.5	7	6.3	0.7
fear	200	203.5	0.3	8	5.5	-2.5
indignation	170	159.5	-1.1	10	10.1	0.1

speaker [$F(2,290) = 203.1$, $p < .0001$], and emotion [$F(6,290) = 41.3$, $p < .0001$], and smaller effects of sentence [$F(4,290) = 8.1$, $p = .0001$], and interaction between speaker and emotion [$F(12,290) = 6.53$, $p < .0001$]. Other interactions remain non-significant ($p > .05$). Summarizing, the same effects were identified as with the analysis for the standard deviation in s.t. reported above.

e. Discussion

The rules based on the values found optimal in Chapter II were in fact determined on the basis of male speech. Therefore, the following comparison between rules and production data will be restricted to male speech. The male speakers use low pitch for the emotions neutrality and boredom, and higher pitch for joy, anger, and indignation. For neutrality and boredom, the rule-based values are just as low or even lower than the ones used by these speakers. For the emotions joy, anger, and indignation, the rules recommend a pitch as high as, or higher than, the one realized by these speakers. The range of variations in pitch level is clearly larger in the rules than in the actors' speech. The rule-based values are generally in the same range or more extreme than the ones measured in the speech of the (male) speakers.

The results show, furthermore, that the standard deviation may be distinctive for some pairs of emotions that are realized at a similar pitch level, as is the case for sadness and anger. Apparently, both level and range are distinctive features of the emotion conveyed.

Some emotions are, more than other emotions, frequently confused with each other on the basis of auditory perception only, also when natural utterances are involved. Such confusions occur especially for emotions which have very similar mean pitch and standard deviations, like joy and anger, or boredom and neutrality. However, these emotions are still discerned far above chance level. Nevertheless, this global approach, considering quantities at utterance level, does not seem to be fully satisfactory: it does not allow a description that distinguishes between all emotions, but only distinguishes groups of emotions.

The results show that the estimation of pitch level and pitch range, on the basis of measurements at utterance level, is rather crude. For pitch level, a difference of more than one semitone between estimated end frequency (mean F_0 minus twice the standard deviation) and actual end frequency has been found for six of the seven emotions. This

difference should be noticeable, but not of significance for the communication ('t Hart, 1981). For pitch range, a difference of at least 1.5 semitones has been found for two emotions between the estimated value and the actual value. Rietveld and Gussenhoven (1985) found that pitch differences of 1.5 semitones result in reliable differences in prominence perception.

f. Conclusion

Despite the individual differences in realizations, the overall picture concerning subgroups of emotions realized with similar pitch level and/or pitch range, is quite consistent. The results of the current production study corroborate those of the perception study. In the speech of the three subjects, as well as in the rules based on the values found optimal in Chapter II, three groups of emotions can be distinguished. A rough ordering in terms of emotions with high or low pitch level and narrow or broad pitch range is quite similar in the perception study and in the current analysis of production data, even if the exact values are not always equal and if individual speakers might make more or less use of either pitch level or pitch range. The measurements at utterance level (mean F_0 and standard deviation) do not show obvious discrepancies for the production and the perception processes. As the value of acoustic data largely depends on what is known concerning their relationship to perception, as well as to physiology, this compatibility confirms the value of the estimation of the pitch parameters studied.

Additionally, some results of related studies are presented. Since comparison between the results of different studies is a difficult task, as long as no standards are widely used, this section can only give a global comparison. In order to keep this procedure simple, only values found optimal in the previous chapter are compared with results of the above mentioned related studies (see Table 1). A clear resemblance can be observed with findings from van Bezooijen (1984). The pitch level versus mean F_0 found for neutrality, joy, anger, and sadness, is similar in both studies. For fear, the mean F_0 reported in the present study is clearly higher than in her findings. The pitch ranges for neutrality, joy, sadness, and fear are comparable to ours. The pitch range she used for anger is even higher than the one found in this study. A comparison with the results of Cahn (1990) remains difficult, due to the rating parameters she used. The increase of pitch level she reported for fear, compared with neutral, agrees with our findings. But the decrease of pitch level she reported for gladness and anger is not in agreement with the present results, nor with those of other studies mentioned here. Her reports about pitch range seem to fit with the present

ones for the emotions joy, anger, and fear. For sadness, she used a pitch range smaller than for neutrality, while in the present study (and the study of van Bezooijen), neutrality was found to have a smaller range than sadness. Though the pitch levels reported by Carlson, Granström and Nord (1992) are in a narrower and lower range (from 131 to 154 Hz) compared with ours (from 94.0 to 246.2 Hz for the corresponding emotions), the ranking of the emotions from lower to higher pitch is the same as in the present study, except for neutrality which they reported higher. The findings of Kitahara and Tohkura (1992) also seem to be in agreement with the present ones. The highest pitch which they confer to joy, and the pitch for anger which is higher than for sadness, agrees with our findings. The pitch levels which Fairbanks and Pronovost (1939) reported match quite well the ones reported here, but they describe much wider pitch ranges. Considering the order of the emotions (from narrow to wide pitch range), they report a relatively larger pitch range for fear than in the present results. Comparing the results of the present study to the predictions of Scherer (1989), there is agreement about the reports of pitch level for all emotions considered. A discrepancy is noted concerning his data predicting a decreased F_0 range for anger and sadness, while all other results are in agreement. In conclusion, however, the results found do not, generally, disagree with those found in these related studies.

Furthermore, measuring mean pitch and its standard deviation appears to be a reasonable and informative, but not very accurate way, of estimating pitch level and pitch range. The complexity of the issue has already raised quite some interest, especially for an approximation of the baseline (Maeda, 1976; Lieberman, Katz, Jongman, Zimmerman, and Miller, 1985).

In the current section, the correspondence of the values found optimal in the experiments of Chapter II, with the range of values produced by Dutch speakers in the expression of emotion, has been investigated at utterance level. The question still remains whether variations occurring in the course of utterances are relevant for conveying emotion in speech. As this global approach at utterance level proved to be rather crude, a more detailed approach will be explored in the next analysis, consisting of a more refined description of the course of F_0 through time. This analysis will be concerned with variations within utterances, i.e., below utterance level. This should make it possible to capture more distinctive elements between the different emotions: pitch elements distinguishing, for instance, sadness from fear, where the distinction is not obvious from measurements at the utterance level. Making a distinction between a pitch curve at a high pitch level with a large

pitch range, and a pitch curve at an intermediate pitch level with a large pitch range, might not be enough. Therefore, we will now look into utterances, first at the intonation patterns realized on the utterances of various emotions (section III); then at F_0 values measured at anchor points, i.e., fixed positions through the sentences (section IV), and we will combine this information in a below utterance level description.

III. REFINED ANALYSIS: INTONATION PATTERNS

In this section, an overview of the patterns of pitch movements, realized on the utterances of the database produced by the three actors, will be presented. Fourteen of these utterances (from one male speaker) have already been labeled for the preliminary analysis in Chapter II. We will now determine which intonation patterns were used in the totality of the speech material (3 speakers × 5 sentences × 3 trials × 7 emotions = 315 utterances) and present those results in the current chapter. More details will be presented in Chapter IV, in which the focus is specifically on the study of intonation patterns in emotional utterances.

a. Procedure

The Dutch grammar of intonation by 't Hart, Collier, and Cohen (1990), allows a description of complete pitch curves in terms of abstract intonation patterns. In general, the term 'pattern' is used in this thesis for legal sequences of pitch movements, so that the term is used for an 'intonation pattern' realized on the whole utterance, but also for a 'pattern of pitch movements' realized on any part of an utterance. The basic pitch movements for Dutch are represented, in this grammar, by numbers for rises and letters for falls: '1' is an early rise lending prominence to the syllable, '2' a very late non prominence-lending rise, '3' is a late prominence-lending rise, '4' is a slow rise extending over various syllables, and '5' is a prominence-lending half rise, also called overshoot, that occurs after a rise. 'A' is a late prominence-lending fall, 'B' is an early non prominence-lending fall, 'C' is a very late non prominence-lending fall, 'D' is a slow fall extending over various syllables, and 'E' is a half fall that may be prominence-lending. In these descriptions, early and late refer to positions relative to vowel onset, within the syllable. Additionally, '0' and '⌀' stand for the pitch level on the lower and the upper declination lines, respectively. If two pitch movements occur on a single syllable, the symbols are linked with an ampersand, for example, '1&A' or '5&A'. In Figure 4, schematized examples of intonation patterns are given; vertical markers indicate vowel onsets.

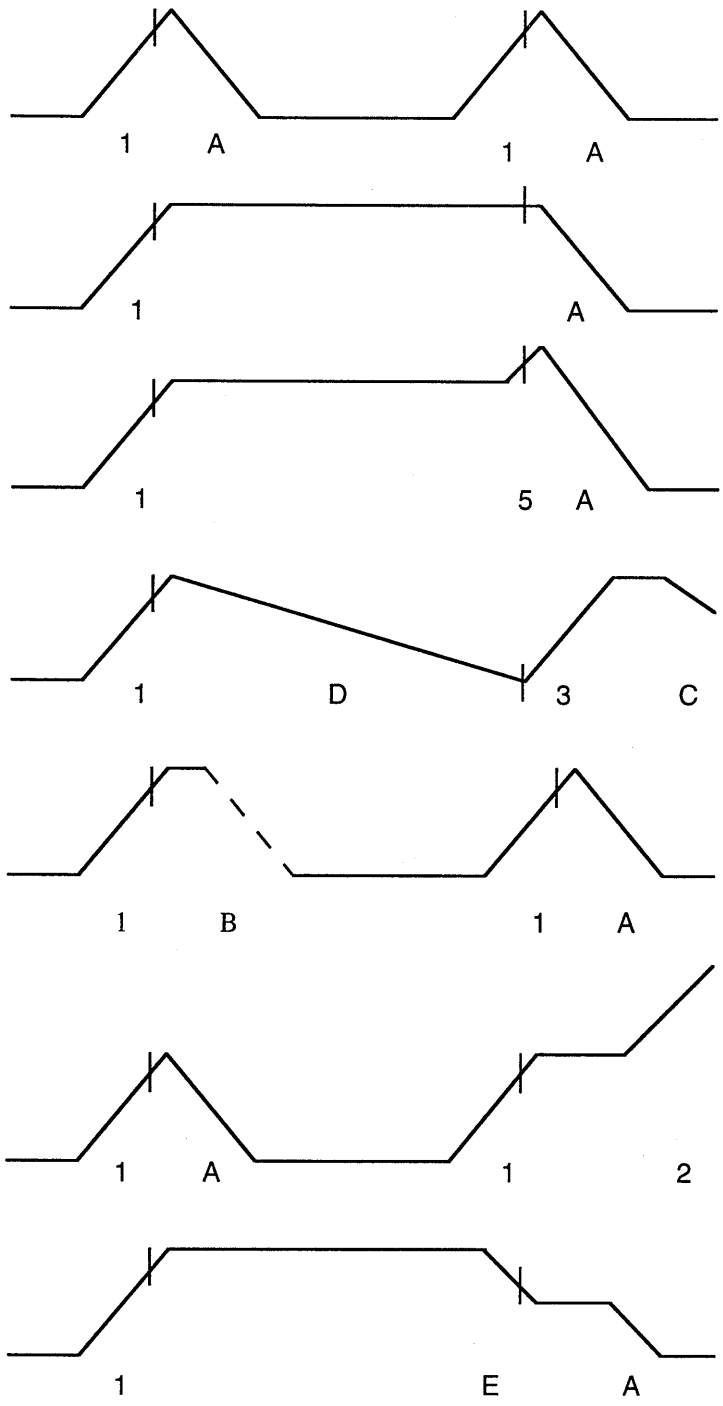

Figure 4: Schematized examples of intonation patterns with labels indicating the type of pitch movements

Two experts in intonation research were asked to listen to the utterances and to label the realized pitch curves of these utterances according to this grammar of intonation. They were instructed to discuss their judgment until they could agree on a single label.

Then, the labels attributed to the utterances by the experts, were distributed over a limited number of patterns of pitch movements. The labels assigned to the initial pattern belonged, for 94.6%, to the categories: '1&A', '1', '1B', '1D', '45', '1&2'/ '1&2B', and 'ø'. The ones assigned to the final pattern belonged, for 96.8%, to the categories: '1&A', 'A'/ '1A', '5A'/ '5&A', 'EA', '12'/ '1&2', and 'C'/ '1C'/ '3C'. As indicated by the slashes, some categories were considered in combination because, on the basis of our knowledge of Dutch intonation, it is not expected that the patterns that are grouped in a single category would be linguistically very different. Additionally, patterns of pitch movements of the following types were assigned to the 'rest' category: utterances lacking fluency or including a pause which may have affected the course of pitch; utterances whose label, such as 'øD3' or '1D32', could not unambiguously be assigned to one of the grammatical categories listed above; utterances with legitimate patterns of pitch movements other than the selected ones, such as '1&E', if this pattern occurred only once or twice in the database.

b. Results

Within the total set of 315 utterances, only five were realized without a pattern of pitch movements in the first part of the utterance. For the remaining 310 utterances, both first and second parts carried at least one pattern of pitch movements; some utterances of the sentences 'Ze hebben een nieuwe auto gekocht' and 'Het is bijna negen uur' were realized with one accent in the first part of the sentence and two in the second part. Over the whole database, 90.8% of the utterances carried at least two pitch accents and were satisfactorily labeled with the categories of patterns of pitch movements mentioned above. The remaining 9.2% were the five utterances realized without a single pattern of pitch movements in the initial part of the sentence and, therefore, labeled with 'ø' in that part of the sentence, and all utterances in the rest category for at least the initial or the final part of the sentence. The distribution of the patterns of pitch movements, over the various categories, is presented in Table 5 for the initial pattern, and in Table 6 for the final pattern in the utterances, pooled over speakers.

Table 5: Frequencies of occurrence of the INITIAL patterns of pitch movements for the three speakers

The patterns of pitch movements realized in the first part of the utterances are reported here.

Emotion	Patterns of pitch movements								
	1&A	1	1B	1D	45	1&2/1&2B	ø	rest	*total*
neutrality	7	12	15	3	-	5	-	3	45
joy	23	5	8	5	-	-	2	2	45
boredom	8	15	16	2	2	1	-	1	45
anger	18	6	16	1	-	2	1	1	45
sadness	19	3	16	1	1	1	-	4	45
fear	20	6	14	2	-	-	-	3	45
indignation	20	9	4	7	-	-	2	3	45
total	115	56	89	21	3	9	5	17	315

Table 6: Frequencies of occurrence of the FINAL pattern of pitch movements for the three speakers

The patterns of pitch movements realized in the final part of the utterances are reported here. If two patterns of pitch movements occur in this part of the sentence, only the last one is reported here.

Emotion	Patterns of pitch movements							
	1&A	A/1A	5A/ 5&A	EA	12/ 1&2	C/ 1C/ 3C	rest	*total*
neutrality	27	10	2	5	-	-	1	45
joy	32	2	2	-	-	9	-	45
boredom	15	12	3	5	1	7	2	45
anger	31	5	4	2	-	2	1	45
sadness	29	8	1	-	-	7	-	45
fear	27	3	2	-	-	11	2	45
indignation	9	-	9	-	13	10	4	45
total	170	40	23	12	14	46	10	315

The '1&A' pattern of pitch movements is the one the most frequently realized, in emotional speech as well as in non-emotional speech. It can apparently be used in the expression of all emotions under study. It does not mean that this pattern is the best choice for the realization of a particular emotion, but that it may be possible, using this pattern, to express each of the emotions studied here. With the exception of one speaker for indignation (the data will be presented in Chapter IV), the '1&A' pattern is the only one that has been used by all speakers in expressing each emotion, both as the initial and final pattern of pitch movements. However, for the expression of indignation, for instance, the results suggest that for some emotions, speakers more often use another pattern. More details concerning the intonation patterns will be presented in the next chapter.

c. Conclusion

There appears to be no one-to-one relationship between intonation pattern and emotion. The most frequent pattern of pitch movements, '1&A', can be used in the expression of all emotions under study. However, this does not mean that this pattern is automatically the best choice for the realization of a particular emotion, but it shows that it is possible, using this pattern of pitch movements, to express each of the emotions studied here. Furthermore, the choice of certain intonation patterns may be relevant for the expression of some specific emotions. The patterns of pitch movements were not equally frequently used in all the emotions. This matter will be addressed in Chapter IV.

IV. REFINED ANALYSIS: F_0 VALUES AT ANCHOR POINTS

a. Introduction

Obviously, synthetic F_0 curves generated by rule within a two-component model differ substantially from original F_0 curves in natural emotional speech. In this section, the focus will be on such deviations from original F_0 curves. Some conspicuous differences, relating to differences in shape of the F_0 curves can, to a large extent, be captured by describing the pitch curves in terms of standard intonation patterns as dealt with above. Other conspicuous differences remain, however. These concern the relative height of the peaks, and the level at the end of the speech which, in the original F_0 curves, is often lower than the baseline, as estimated on the basis of the first part of the utterance. In the two-component model, the end point must be in line with the baseline. Considering the F_0 curves, either within the model or in natural utterances, can lead to different estimations of pitch level and pitch range. For instance, considering an emotion realized in a natural

utterance with very high pitch peaks and a moderate end frequency clearly lower than the frequency in the previous part of the utterance, would suggest a moderate pitch level and a very large pitch range, whereas on the basis of measurements of mean F_0 and its standard deviation, the pitch level would, in comparison, be estimated to be higher. The pitch range estimated on the basis of pitch variations from the mean would be smaller than the one based on variations from the end frequency. A more detailed description of the original pitch curves was, in this respect, deemed necessary.

In this approach, we attempted to get a better description of the course of F_0 through time, by taking measurements at various well-defined points within the F_0 curves of the utterances. Analyzing the F_0 curves below utterance level is a way to explore more detailed distinctive features that allow the discrimination of an F_0 curve produced in a specific emotion, from F_0 curves produced in other emotions, even if the curves have similar mean F_0 and standard deviation. This analysis will be formulated in terms of the two-component model of intonation, as presented in Figure 1. As some information contained in the description cannot be captured in this two-component model, the relevance of the deviation from this model will be discussed.

b. Procedure

Measuring the pitch at structurally relevant fixed points in each utterance could provide a more accurate description of the F_0 curves. These fixed points, i.e., anchor points, should yield information about intonation concepts that are relevant to the communication process such as prominence and accentuation. On one hand, points were measured in the accented lexically stressed syllables (Anchor points 2 and 5, represented in Figure 1). On the other hand, points were chosen with the intention of capturing the position of the baseline. To this end, the frequency at the beginning of the F_0 curve was measured (Anchor point 1, see Figure 1) because it can give an indication of the position of the baseline at speech onset. Points were also measured at unaccented positions between the accents (Anchor points 3 and 4, see Figure 1) because they can be relevant as intermediary values for estimating the baseline; these unaccentend positions can be either lexically stressed or untressed. Then, the final point was located (Anchor point 6, see Figure 1). As a result of these measurements, information relevant to the strength of the accents or the prominence of accented syllables will be obtained, since these perceptual quantities appear to depend on how the peak relates to the preceding and the following low positions (Terken, 1991; Ladd, 1993; Repp, Rump, and Terken, 1993; Terken, 1994).

In this study, six anchor points per utterance were, thus, measured that are now described in more detail. The first point is located at the onset of the voiced speech: the low position preceding the initial-accent peak. In principle, this point indicates where the baseline starts. The second point is the initial-accent peak. The third and fourth points are two points in the interval of lower pitch between the two accent peaks, i.e., one point after the initial-accent peak and the other before the final-accent peak. The interval with a lower pitch, between the accent peaks, that will be referred to as 'valley', is a region in the utterance that is supposed to be on the intonation baseline. The fifth point is the final-accent peak, and the sixth point is the F_0 at the end of voiced speech in the last syllable of the utterance. It was decided to measure F_0 at such fixed syllables in the sentences, in parts of the vowels where the pitch is stable, which means that the measurements were not necessarily taken at the lowest positions in the valley. Fixed positions were chosen within the vowels in predetermined unaccented syllables of the sentences, because pitch is well defined at these instances (Hermes and Rump, 1994; House, 1990, d'Alessandro and Mertens, 1995). The measurements at the beginning and the end of speech were not taken in the stable part of the vowels but at the beginning and end of voicing, respectively. This was done because, in the first and the last syllable, the pitch in the stable part of the vowel may not represent the start and the end of the baseline well.

As has been shown, utterances were realized with different patterns of pitch movements. It seems desirable to only take utterances into account that are phonologically comparable, which means considering utterances that are all realized with the same pattern and excluding all utterances having any other pattern. As the '1&A' is the most frequently used pattern of pitch movements and this pattern can be utilized in the expression of each of the seven emotions, it was decided to investigate the correspondence of the results averaged only over the utterances realized with a '1&A' pattern of pitch movements in the final position, with the results averaged over the whole speech material.

c. Results

The average F_0s at the six anchor points are plotted for the three speakers and the seven emotions in Figures 5 and 6. Figure 5 includes all utterances; measurements on three repetitions of five sentences are averaged per emotion and per speaker. Figure 6 includes only the utterances realized with the final '1&A' pattern of pitch movements. The data points are connected for each emotion for the sake of clarity, but the resulting graphs do not represent real F_0 curves, not even stylized ones. These data should give information as

to pitch level and pitch range, but relevant information as to, for example, the timing of the pitch movements, is to a large extent removed from this representation.

The number of utterances involved in Figure 6, including only the '1&A' patterns, varies per emotion and speaker; the exact number can be found in Table 6 in the '1&A' column. For the female speaker LO, no comparative plot could be made for indignation, because she did not produce any utterance of indignation using this final pattern of pitch movements, whereas the line plotted for boredom represents a single utterance only in her case.

The results of Figures 5 and 6 compare rather well with each other. For speaker MR, there is a very good match between the schematized pitch curves concerned with all intonation patterns and those only involving the '1&A' pattern. Most intonation patterns he used involved '1&A'. For RS, the match is also rather good. In indignation and fear there is no good match for the last anchor point; anticipating results of Chapter IV that will show that ten and seven patterns out of fifteen per emotion, respectively, ended in a 'C' for this particular speaker, the fact that the F_0 curve for these emotions ends up much higher when all patterns are involved, does not come as a surprise. The good match on the third and fourth anchor points is due to the fact that this speaker used few intonation patterns in which the pitch between the accents corresponds to the topline, such as '1A' or '1EA' (See Figure 4). For the female speaker LO, the match between both figures is also not at all bad, but these two plots compare less well than was the case for the two other speakers. This is understandable, as she used more patterns of pitch movements other than '1&A', when compared to the other speakers. The shape of the pitch curves differs for some emotions as in sadness; this is due to the fact that, in this emotion, the speaker used '1A' intonation patterns, in which the pitch between the accents corresponds to the topline. Also, in certain utterances of sadness, LO realized a double focus involving the word preceding the one on which the final-accent fell. In joy, the pitch curves involving all patterns remain higher at the end because of the use of pitch movement 'C'. For anger, there is a shift in pitch level between the two plots. The differences between Figures 5 and 6, though smaller or larger for different speakers, show that the use of different intonation patterns induces a considerable amount of variation in the F_0 curves. It will be shown, in the next chapter, that this variation is perceptually extremely significant. The '1&A' pattern of pitch movements appears to be the best suited to control this variation, since this pattern can apparently be used in all emotions in both initial and final positions.

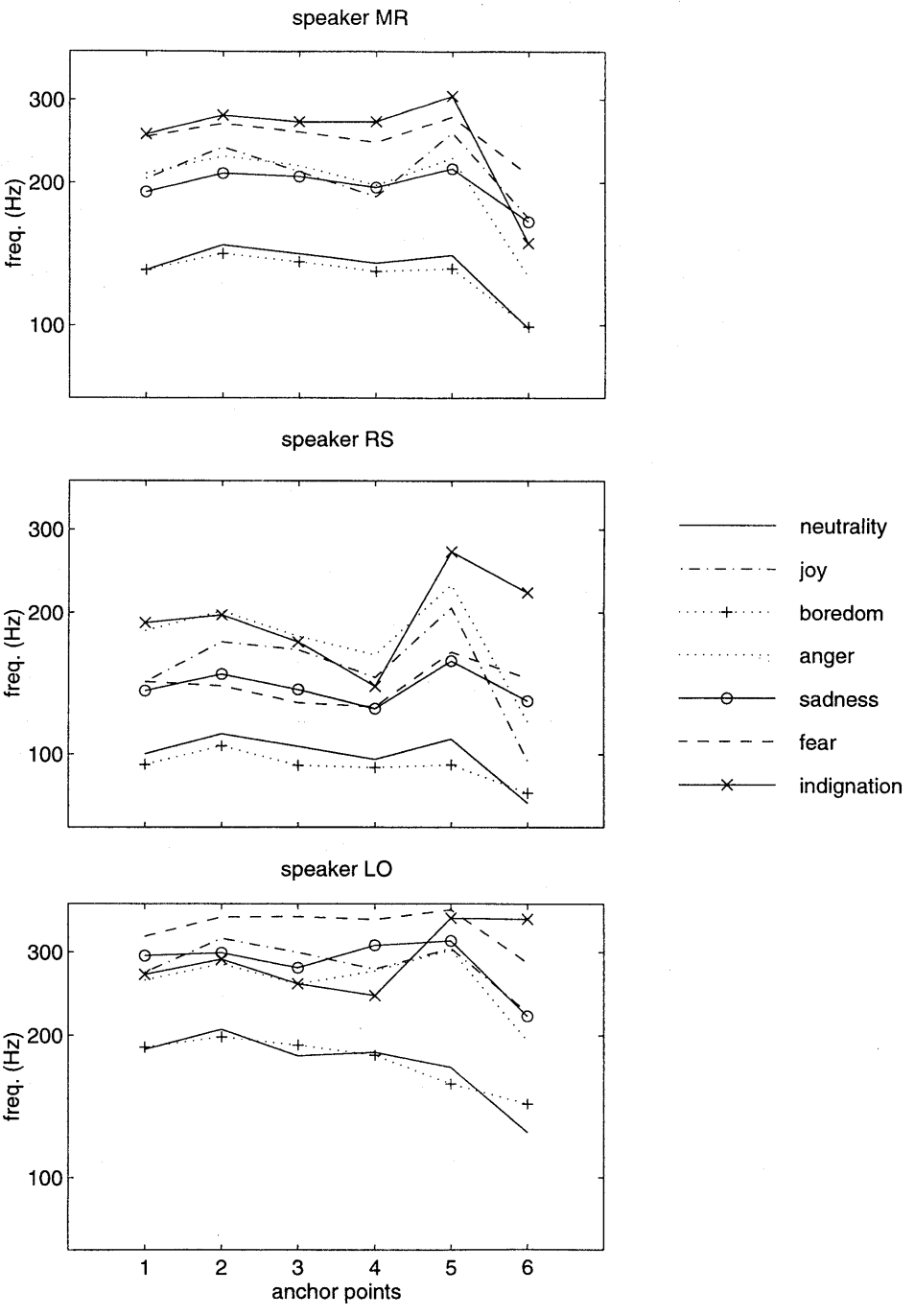

Figure 5: Schematized pitch curves averaged over all utterances per speaker and per emotion

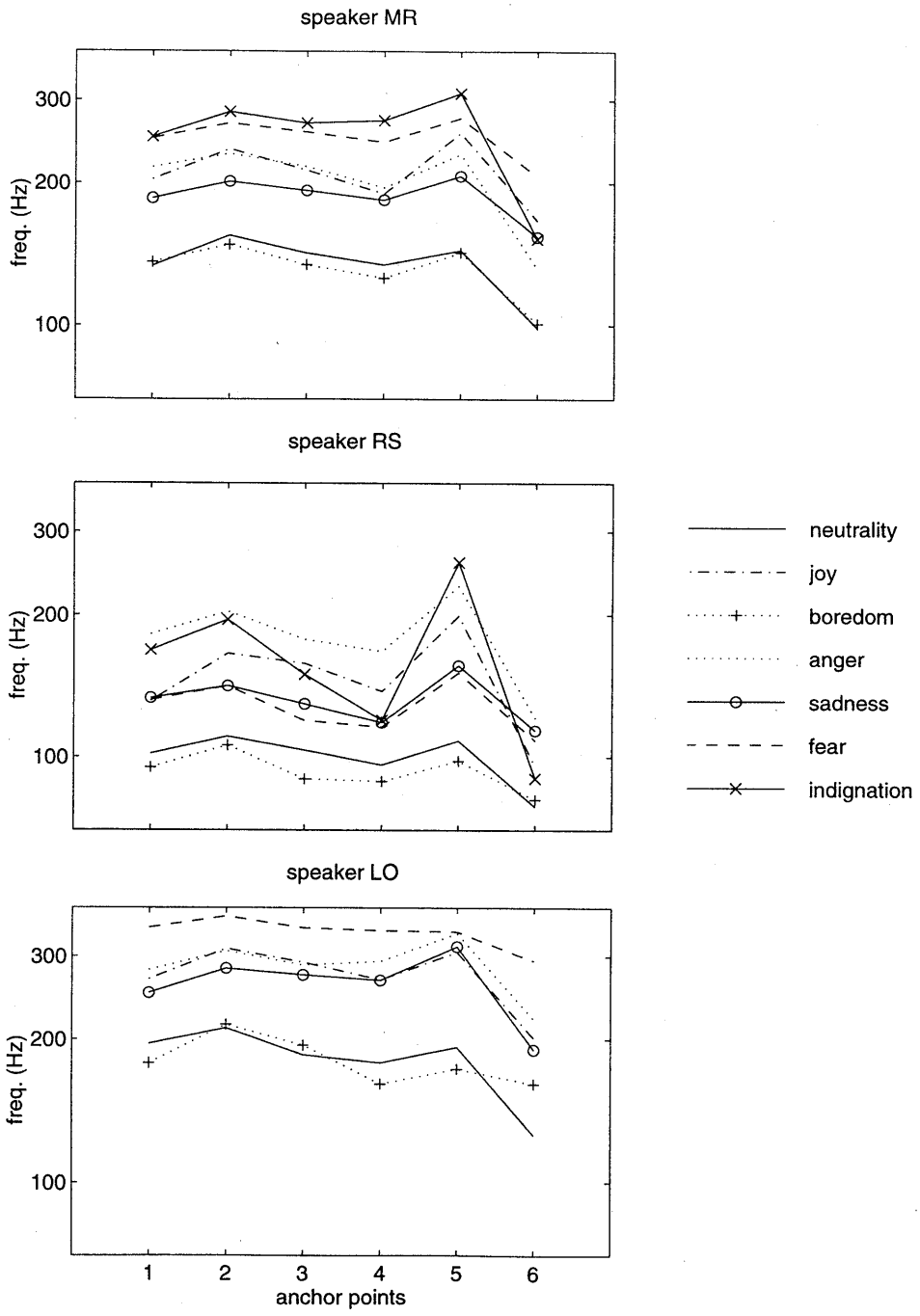

Figure 6: Schematized pitch curves averaged over the utterances in which the '1&A' pattern of pitch movements was produced

d. Description in terms of a two-component intonation model

Intonation model

In a two-component model, one component relates to variations in pitch level. The notion of pitch level represents how high or low an utterance is produced in the speaker's overall range. This notion of pitch level is represented in the model by the end-point of the baseline. The slope of this lower declination line depends on the duration of the utterance ('t Hart et al., 1990, p. 128). The other component of the model relates to variations in pitch range. The size of the pitch variations is given in terms of the distance between the baseline and the pitch values that are on the topline. For simplicity's sake, the pitch range will be assumed to be constant throughout the utterance. Although this is not strictly necessary within the model, the same was assumed in the previous chapter. In this study, the classical two-component model was extended with the notion of 'floor', the basic pitch level at which a particular speaker feels at ease when speaking. This supplementary notion would appear to be necessary for a description of those parts of the pitch curve which are lower than the baseline. This extended model of tonal space used here may very well be too simple in general, since other studies have shown that no simple relationship can be found between successive pitch peaks and between pitch peaks and baseline (Gussenhoven, Repp, Rietveld, Rump, and Terken, 1997; Terken, 1991, 1994).

Description of the data with the model

On the basis of the perceptually optimal values we found in Chapter II (see the right-hand double column in Table 2), the emotions boredom and neutrality can, in a classical two-component model of intonation, be represented as having a low pitch level and pitch movements of small excursion size. Fear and sadness can be represented by raising the baseline, which means that these emotions are represented at a higher pitch level. For anger and indignation, the topline should also be raised, resulting in a larger pitch range, and the baseline should be moderately raised. A combination of raised baseline and increased distance between baseline and topline would then suit joy.

In the extended model of tonal space used here, the following tentative description can be given, again on the basis of the optimal values found in Chapter II. For some emotions such as, for instance, boredom, F_0 at the end of speech goes all the way down to a final low value as low as in neutrality, while in other emotions such as, for instance, fear, F_0

would remain higher. For some emotions, such as joy and fear, the intermediate F_0 values between the accent peaks come down so that they are aligned with the supposed baseline. Accordingly, different sorts of pitch curves may be distinguished that are used for different emotions. A normally low pitch level, and a relatively narrow pitch range, suit the expression of neutrality and boredom. In fear and sadness, the pitch curve stays high at the end of the utterance, i.e., the position of the final low point is raised. For indignation, the accents are realized high in the tonal space which corresponds with a large pitch range and the pitch curve ending in a normal final low position. For anger and joy, the position of the final low pitch is moderately raised and the pitch range moderately enlarged.

Inspection of Figures 5 and 6 indicates that there are two major sources of deviation between the original F_0 curves of the natural utterances and the rule-based F_0 curves synthesized according to the classical two-component model. Firstly, all minima in the rule-based F_0 curves are on a single declining baseline, whereas, in the original F_0 curves, the utterance-final low pitch does not simply fall in line with the minima in the earlier part of the utterance. Instead, the final pitch can be considerably lower than expected on the basis of the earlier part of the F_0 curve. This is in agreement with the findings of Shriberg, Ladd, Terken, and Stolcke (1996). Secondly, whereas the pitch range in the rule-based F_0 curves is the same throughout the utterance, this is not necessarily the case in the original F_0 curves. Instead, there is a tendency that, when pitch curves get higher in pitch level, they display a final pitch-accent which becomes increasingly larger than the initial accent. These observations of emotional speech are not in agreement with Pierrehumbert's conclusions (1979, 1981) concerning neutral declarative speech, that the F_0 range narrows over the course of the phrase. According to Pierrehumbert (1979), and to Sorensen and Cooper (1980), the increase in pitch range affects the beginning of an intonation group more than the end. The different behavior concerning the relative peak height in emotional speech is striking.

In fact, the only emotions for which the original F_0 curves compare well to the rule-based versions are neutrality and boredom. This is not surprising for neutrality, since the model was largely developed on the basis of neutral, uninvolved utterances.

It appears that, in the two-component model, it is the last part of the sentence, including the sentence accent and the final low, which corresponds least well with natural F_0 curves. In order to better describe this part of the F_0 curve, three measures were considered, which should provide an estimation of pitch range. In all three measures, the F_0 at the fifth anchor

point, i.e., the F_0 peak of the final accent, is compared with the F_0 at another anchor point of the utterances. The fifth anchor point was chosen, since it appeared to best represent the deviation from the simple model and since it represents, in many cases, the position of the F_0 maxima, for which listeners are sensitive with respect to prominence (Sluijter, 1991). The F_0 at this fifth anchor point was compared with three points that are expected to represent the position of the baseline. In the first measure, M_{5-1}, this other point is the first anchor point, i.e., the start of the voiced speech; in the second measure, M_{5-4}, it is the fourth anchor point, i.e., the F_0 in the preceding valley; in the third measure, M_{5-6}, it is the sixth anchor point, i.e., the F_0 at the end of the voiced speech of the utterance. The second and third anchor points were not suited for this comparison; the second anchor point because it is an accented position; the third anchor point because the F_0 value at this position might still be influenced by the concrete realization of a pitch movement, such as 'D', and therefore lies higher than the baseline. The results as presented in Hertz in Table 7, and in semitones in Table 8, are averaged over the three sentences: 'Jan is naar de kapper geweest', 'Zijn vriendin kwam met het vliegtuig', and 'De lamp staat op het bureau'. The other two sentences could not be included in the investigation; 'Het is bijna negen uur' because the two accents were quite near each other, and 'Zij hebben een nieuwe auto gekocht' because the final accent was not always strictly restricted to the word 'auto' but could also involve the word 'nieuwe' in a double focus.

The question now is which of these three measures best represents the pitch range. Correlating the data, as presented in Table 7, yields very low coefficients when M_{5-6} is correlated with M_{5-1} ($r = .11$) and with M_{5-4} ($r = .34$). Correlating M_{5-1} with M_{5-4} gives a much higher coefficient ($r = .89$). So it appears that M_{5-6} contains different information than M_{5-1} and M_{5-4}, which corroborates the fact that the end point is not aligned with the other anchor points. The two-component model does not allow the description of these results. A separate treatment of final lowering appears to be necessary. The distance between the final-accent peak and the final low pitch is not a reliable estimate of pitch range, most probably because of the influence of final lowering. The distance between this accent peak and the start of the utterance or the preceding valley, is a better candidate. That makes Anchor point 1, i.e., the beginning of speech, and Anchor point 4, i.e., the point in the valley before the final-accent peak, the most representative of all anchor points for the baseline. This estimation of the baseline is rather rough, however, which is corroborated by the suspiciously small values found in Table 8 for neutrality and boredom.

Table 7: Distances between Anchor point 5 (the final-accent peak) and other anchor points, in Hertz, averaged over three sentences for each of the three speakers

Emotion	M_{5-1} Anchor 5 - Anchor 1			M_{5-4} Anchor 5 - Anchor 4			M_{5-6} Anchor 5 - Anchor 6		
	MR	RS	LO	MR	RS	LO	MR	RS	LO
neutrality	9.8	7.6	-15.7	5.3	10.2	-13.3	42.1	28.9	46.5
joy	51.4	62.3	33.9	68.1	59.2	29.3	86.5	107.9	81.6
boredom	0.9	-0.1	-31.1	1.9	1.2	-24.0	32.5	12.2	14.6
anger	16.1	46.5	41.7	27.8	67.2	29.5	98.7	112.3	108.7
sadness	23.2	21.3	22.0	18.7	32.6	6.5	48.6	27.8	97.0
fear	25.4	22.2	45.3	32.2	38.5	16.7	67.4	19.6	84.1
indignation	52.4	79.3	85.5	36.4	130.2	111.4	156.5	48.7	1.5

Table 8: Distances between Anchor point 5 (the final-accent peak) and other anchor points, in semitones, averaged over three sentences for each of the three speakers

Emotion	M_{5-1} Anchor 5 - Anchor 1			M_{5-4} Anchor 5 - Anchor 4			M_{5-6} Anchor 5 - Anchor 6		
	MR	RS	LO	MR	RS	LO	MR	RS	LO
neutrality	1.2	1.3	-1.5	0.7	1.7	-1.3	6.2	5.4	5.5
joy	3.9	6.3	2.0	5.4	5.9	1.7	7.2	13.0	5.4
boredom	0.1	0.0	-3.1	0.2	0.2	-2.4	4.9	2.4	1.7
anger	1.3	3.9	2.6	2.3	6.0	1.8	10.0	11.6	7.7
sadness	2.0	2.5	1.2	1.6	4.0	0.4	4.5	3.4	6.3
fear	1.7	2.5	2.3	2.1	4.6	0.8	4.9	2.2	4.5
indignation	3.3	6.0	4.8	2.2	11.4	6.5	12.4	3.5	0.1

This conclusion is reinforced when we compare the data of Table 8 with those in Table 3. In both the production and the perception study, it was found that joy, anger, and indignation, had larger standard deviations than the other emotions. With one exception, this arrangement of data reoccurs in the first three columns of Table 8, where M_{5-1} is presented for MR, RS, and LO. (The data presented for MR in the first column are the

exception, since figures for sadness and fear are larger than for anger.) This arrangement of data also reoccurs for M_{5-4}, this time with no exceptions. Finally, for M_{5-6}, this arrangement of data is only present for speaker MR.

Finally, when the emotions are put in order, from the ones involving small ranges of pitch variations to the ones involving large pitch ranges, considering the distances between the final peak and the preceding valley as presented in Table 8, the following ordering follows: boredom, neutrality, sadness, fear, anger, joy, and indignation. Looking at the pitch range by considering the standard deviation of the mean F_0 in Hz as reported in Table 2, results in the same ordering, except that the emotions joy and anger would swap places. Note that only the speaker MR expresses anger with a larger pitch range than he does joy, while the two other speakers produce similar pitch variations for both emotions.

In summary, the estimations of pitch range by means of measurements at Anchor points 4 and 5 are in good qualitative agreement with the other perceptual and production data we had gathered. The question remains, however, whether the F_0 measurement at point 4 is a good estimate of the baseline.

V. PERCEPTUAL SIGNIFICANCE OF DETAIL INFORMATION FROM THE REFINED APPROACH

a. Aim

The aim of this experiment is to determine to what extent the modeling of the discrepancies between the description of the F_0 curves for the various emotions, as it is obtained by means of measurements at anchor points, and the description obtained with the two-component model, can lead to improved identification of these emotions. In other words, this experiment is meant to investigate whether the information obtained on the relative height of the two pitch peaks and the final lowering of the F_0 curves, information that was not captured in the classical two-component model, have any perceptual significance for the identification of the seven emotions studied.

b. Speech material

A neutral utterance of the male speaker MR was selected as a single carrier sentence for the speech manipulations. A single text was considered sufficient here because in Chapter II, the semantic content of the utterances was found to have no substantial impact on

identification scores of the emotions. The sentence selected was: 'Zijn vriendin kwam met het vliegtuig'. It was decided to control the variation in intonation patterns, while investigating the relevance of final lowering and relative height of the peaks. The analysis of the shape of the pitch curves had shown that the '1&A' pattern of pitch movements was the best suited choice for this purpose. Therefore, the originally neutral utterance was provided with a standard '1&A 1&A' intonation pattern by means of a PSOLA-manipulation. The resulting utterance serves as a target utterance. It was re-synthesized with five combinations of pitch level and pitch range found optimal for each of the seven emotions, representing five conditions. Condition 1 models the rules based on the values found optimal in the previous perception study (Chapter II). For Conditions 2 to 5, it was decided to model characteristics of the F_0 curves as generated by speaker MR and displayed in Figure 6. This choice is based on the fact that, in the previous study, MR was the male speaker whose emotions were best recognized. Moreover, this speaker used the '1&A' pattern more often than the other two speakers. So, each pitch curve has a fixed intonation pattern ('1&A' on both pitch accents). Furthermore, baseline declination was fixed at 3.5 semitones per second, the duration was a constant, namely, 1.77 second, and the pulse train used as a voice source in all signals was a constant.

Conditions 1 and 2 constitute a frame of reference for comparison with the other conditions. The five conditions are described as follows:

- *Condition 1: Rule values within the classical two-component model.*

F_0 curves were generated using values for pitch level and pitch range that had been obtained in the previous perception study (Chapter II). The end frequency and the size of the concrete pitch movements vary as given in Table 9. The F_0 curves are produced with equal peak height and no final lowering.

- *Condition 2: Production values within the classical two-component model.*

For Conditions 2 to 5, F_0 curves were generated using values for pitch level and pitch range that were chosen to get as close a fit, to the original F_0 curves produced by speaker MR, as possible. In Condition 2, this was done within the limits of the two-component model. The values used are shown in Table 9. Because the final low point appeared not simply to be representative for the position of the baseline, a standard baseline, with constant declination, was anchored at utterance onset rather than offset, using frequencies as produced by speaker MR. In order to allow comparison between Conditions 1 and 2, within Table 9, the table shows end frequencies instead of onset frequencies. The value for the overall pitch range, i.e., the pitch range for both pitch accents, was chosen to be the distance (in semitones) between the final-accent peak and the standard baseline that was

Table 9: Parameter values per emotion for synthetic F_0 curves of Conditions 1 and 2

Emotion	Condition 1		Condition 2	
	End Frequency (Hz)	Pitch range (s.t.)	End Frequency (Hz)	Pitch range (s.t.)
neutrality	65	5	95	6
joy	155	10	125	12
boredom	65	4	100	5
anger	110	10	145	8
sadness	103	7	125	8
fear	200	8	160	9
indignation	170	10	180	9

adjusted in pitch level. This choice was inspired by the fact that the pitch range of the final-accent peak, relative to the baseline, varied much more in relation to emotion than that of the initial-accent peak. Summarizing, the size of both accent peaks was thus equal and based on the production data of speaker MR for the final accent. The end of speech was aligned with the declination baseline.

- *Condition 3: Peak modeling as extension to the model.*

Starting from the utterances in Condition 2, both accent peaks were independently modeled from the F_0 curves in Figure 6 for speaker MR, which means that the excursion size of the first pointed hat was no longer equal to that of the second pointed hat, as was the case in Condition 2. Instead, the first peak was modeled as the distance (in semitones) between the initial-accent peak and the standard baseline, the second peak as the distance (in semitones) between the final-accent peak and the standard baseline.

- *Condition 4: Modeling final lowering as extension to the model.*

Starting from the utterances in Condition 2, the utterance-final low pitch for each emotion was modeled from the F_0 curves in Figure 6 for speaker MR. The final low pitch was modeled as the distance (in semitones) between the standard baseline and the final low point.

- *Condition 5: Modeling relative peak height and final lowering as extensions to the model.*

The effects of Conditions 3 and 4 were combined.

Since, for neutrality and joy, the production values used for Condition 2 correspond to an accurate modeling of final lowering, the stimuli used in Conditions 2 and 3 (in which there is no modeling of final lowering) happen to be the same as those used for Conditions 4 and 5 (modeling final lowering), respectively. Hence, there were only 31 test utterances instead of 35 (7 emotions × 5 conditions).

c. Design and procedure

A series of 55 stimuli was presented to the 16 subjects who participated in this experiment. The first 19 stimuli gave an idea of the kind and amount of pitch variations allowed in the stimuli, the next 31 were the actual test stimuli presented in random order, and the last five stimuli were end-of-list fillers. The subjects were not informed of this fact and did not get any feedback concerning their performance. The experiment involved a seven-alternative forced choice paradigm with the seven emotion labels. The subjects performed individual interactive listening tests. They listened, only once, to each stimulus, over headphones, and decided which emotion had been expressed. The 31 test stimuli were presented to different subjects in different random orders.

d. Results

Table 10 gives the number of subjects (maximally 16) correctly identifying each emotion in the five conditions. It can be seen that the number of correct responses for neutrality and joy in Conditions 2 and 3 are in parentheses. This is because the F_0 curves for neutrality and joy, in Conditions 2 and 3, had the same utterance-final low pitch as the original F_0 curves, so that the F_0 curves generated in Conditions 2 and 3 for these two emotions actually are the same as for Conditions 4 and 5. To compare an equal number of judgments for each condition, the results for these stimuli were also included in Conditions 2 and 3.

Per condition, there was a total number of 112 stimulus presentations (16 subjects × 7 emotions). The chance level for correct responses was 2.3 per cell, i.e., 16 per condition. The total number of correct responses was not very different from one condition to the next: 30 for Condition 1, 31 for Condition 2, 36 for Condition 3, 39 for Condition 4, and 33 for Condition 5.

In order to comprehend the perceptual effect of the modeling of relative peak height and final lowering, it would seem more interesting to consider this effect on individual emotions, rather than to consider the overall effect on all emotions. Indeed, averaging the

results over all emotions can obscure the effect of a parameter for a particular emotion. In other words, the effect of modeling final lowering and/or the relative height of the peaks can be different for conveying one emotion than for another one. This is illustrated by a different distribution of the correct responses over the various conditions for different emotions, as can be seen in Table 10. For instance, for final lowering in Condition 4, modeling final lowering seems to *in*crease performances for indignation and to *de*crease performances for boredom. Another noticeable point is the tendency for the combined modeling of final lowering and relative height of the peaks to decrease performances when compared to the modeling of either final lowering or relative height of the peaks. This tendency is present, without exception, for each emotion studied here. A tentative interpretation is that the presence of variations that are not relevant to the expression of the emotion, actually introduces noise and therefore decreases performance.

Next, the results in the five different conditions will be discussed, per emotion. For neutrality and for joy, only the relevance of the relative height of the peaks could be investigated, as there was no final lowering in the production data. Indeed, the computation of a value for the end of speech, in Conditions 2 and 3, was based on: F_0 at the beginning of speech, the constant declination, and the constant duration of the sentence. For neutrality and joy, this computation resulted in the exact value produced by MR at the end of speech. For these two emotions, peak modeling degrades identification performance. For neutrality and for joy, the values used by the speaker lead to better results than the rules. Boredom seems to be characterized by monotony and best conveyed with a minimum of variations. It does not come as a surprise that, for this emotion, it is the absence of variations, and thus the absence of modeling, that is relevant, particularly the absence of final lowering. The rules best represent this monotony and, therefore, induce the best performances. Anger was hardly identified in the experiment; it was not correctly identified in three of the five conditions. Sadness was never correctly identified in Condition 2, the reference condition. Therefore, the significance of the increased identification of sadness with the modeling of relative peak height and final lowering is hard to estimate. Unexpectedly, the combined modeling of both features (Condition 5) seems to weaken the effect when compared to the individual modeling of each of the features. For fear, the modeling of either relative peak height or final lowering also tends to increase the identification rates, but the combined modeling tends to be detrimental when compared to the modeling of a single feature. Finally, for indignation, modeling the final lowering clearly tends to improve the identification rate, but modeling the relative height of the pitch accents does not bring any

Table 10: Number of correct responses per condition (C1-C5) and per emotion

The maximum per cell was 16, chance performance being 2.3. For neutrality and joy, the F_0 curves generated for Conditions 2 and 3 are actually the same as for Conditions 4 and 5. In order to compare an equal number of judgments for each condition, the results for these stimuli were also included in Conditions 2 and 3. They are shown in parentheses in this table.

Emotion	Condition 1: 2-comp. model representation of rule values	Condition 2: 2-comp. model representation of production values	Condition 3: Peak modeling	Condition 4: Final low modeling	Condition 5: Peak and final low modeling
neutrality	3	(10)	(9)	10	9
joy	3	(7)	(6)	7	6
boredom	11	6	6	2	5
anger	2	0	1	0	0
sadness	2	0	4	4	3
fear	6	5	7	7	5
indignation	3	3	3	9	5
total correct responses	30	31	36	39	33

improvement. Again, the combined modeling of both features seems to weaken the effect of the modeling of a single feature, namely, the modeling of final lowering.

e. Discussion

Although modeling either relative peak height or final lowering improves identification performances for particular individual emotions, the improvement is unexpectedly low for the combined modeling of these characteristics. The fact that the effect of the combined modeling of both characteristics fails to show for the emotions sadness, fear, and indignation, in Condition 5, may partly be due to the limited number of subjects involved in the experiment. In order to show such small effects, one would need considerably more subjects.

Moreover, the contribution of peak and final lowering modeling to the conveying of emotion in speech might also depend on the interaction with other speech parameters, such as loudness and voice quality. The rather low identification rate (33.1%, chance level being

14.3%) is probably due to the fact that no characteristics other than pitch have been manipulated; despite the relevance of duration to the expression of emotion in speech, all stimuli had the same overall duration. Anger and sadness gave particularly poor performances. For sadness, this may be due to the fact that voice source is an important component for this emotion, as was suggested by previous investigations (Chapter II).

From the identification data, estimates of stimulus entropy, response entropy and mutual information were calculated (Shannon and Weaver, 1949) in the same way as in Chapter II. Mutual information, expressed in bits per stimulus, is a measure of the consistency with which responses are assigned to particular stimuli. It is zero when all responses are completely random, and equals the amount of stimulus entropy if each response is tied uniquely to a specific stimulus. For sets of seven different stimuli, each stimulus being presented equally frequently, as was the case in the identification experiments, the stimulus entropy is 2.81 bits/stimulus. Mutual information and percentage correct identification were calculated from the confusion matrices obtained under the five conditions. Mutual information was .78 bits/stimulus for Condition 1, .59 for Condition 2, .57 for Condition 3, .70 for Condition 4, and .50 for Condition 5. The respective percentages of correct identifications were: 27, 28, 32, 35, and 30%. The values for response entropy were found to be: 2.73, 2.61, 2.62, 2.57, and 2.67 bits/stimulus, for Conditions 1 to 5, respectively. The rather low values of mutual information, ranging between .50 and .80 bits/stimulus compared with the stimulus entropy of 2.81 bits/stimulus, indicate a rather low transmission of information from the stimuli to the subjects. Moreover, the level of information transfer does not clearly increase in any particular condition. The response entropy is a measure providing information about the response bias of the subjects; a response entropy that would reach the value of the stimulus entropy, i.e., the maximum possible value for a response entropy, would indicate a total absence of response bias in the subjects. The values found indicate that subjects were not very biased in their choices of emotional labels in the test.

Computations of stimulus entropy, mutual information, or probability of correct response from empirical data organized in some confusion matrix, always yield estimates which are noisy and can be biased, depending on the number of trials and the size of the confusion matrix. The proportion of correct responses, measured along the diagonal of the matrix, typically yields an unbiased estimate. Observed stimulus entropy for a limited set of trials yields a biased estimate that is too low, whereas observed mutual information yields an estimate that is biased too high. For the observed bias to become negligible, the rule of

thumb is that there should be at least five times as many identification trials as there are cells in the matrix (Miller, 1954). In our case, there were 49 cells (7 × 7) in each matrix, and 112 trials per condition (16 subjects × 7 emotions). This is less than half the number of trials needed to eliminate bias in the obtained estimates of mutual information.

Bias and variability of the computed estimates can easily be obtained through a simulation experiment (Houtsma, 1983) in which the identification test is computer simulated. Two identification models were simulated. One model assumes that the stimuli have only nominal properties; it yields a confusion matrix where correct responses fall on the diagonal and incorrect responses are uniformly distributed off the diagonal. The other model assumes at least ordinal properties in the stimulus set, and yields confusion data that are clustered around the diagonal. Simulation runs were performed with both models, and estimates were obtained for a set of seven stimuli (7 × 7 matrix), an amount of internal noise in the model that yielded about the same percentage correct scores as were found in the actual listening experiments, and a number of 112 and 1000 runs. Simulation runs were repeated many times to empirically observe the statistical spread in the results. For the 112-trial runs, it was found that, for the 'nominal' model, the coefficient of variation, i.e., the standard deviation divided by the mean expressed as a percentage, is 12% for percentage correct, and for mutual information this coefficient is 28%. For the 'ordinal' simulation model, the respective coefficients of variation were found to be 11% and 17%. Comparison of model-simulated percentage correct and mutual information of the two models with the percentage of correct identification and mutual information of the listening experiments, leads to the conclusion that the ordinal model provides a much better description of the data than the nominal model. This implies that the incorrect responses were not randomly distributed over the response categories. Finally, staying with the ordinal simulation model, a comparison of percentages correct simulation scores, for 112 and 1000 trials, yielded virtually identical scores (33.4 vs. 32.4%), whereas mutual information estimates were .76 and .51 bits/stimulus, respectively. This shows the amount of bias in the mutual information estimates when the number of trials is too small, and gives an indication of the approximate correction factor.

Furthermore, for a total of 112 trials per condition, the 5% confidence limit can directly be calculated for the five conditions. It is 5.4 for Condition 1 with 30 correct responses, 5.6 for Condition 2 with 31 correct identifications, 6.5 for Condition 3 with 36 correct responses, 7.1 for Condition 4 with 39 correct responses, and 6.0 for Condition 5 with 33 correct identifications. The difference in overall identification between Condition 1 and

Condition 3, between Condition 1 and Condition 4, and between Condition 2 and Condition 4, reaches significance at $p < .025$. On the basis of pitch alone, emotions were identified better than chance, but modeling the details of the F_0 curves did not lead to a really substantial overall improvement, at least not in Condition 5 which produces the closest match to the original F_0 curves. The identification score for Condition 5 was not significantly higher than that in Conditions 1 and 2, providing frames of reference. However, the higher results in Conditions 3 and 4 might indicate that the score in Condition 5 could be due to a statistic based on a small number of observations. There seems to be an overall effect of final lowering and a small overall effect of relative peak height. Although these effects are not of primary importance, they might contribute to conveying emotion in speech.

Summarizing, pitch variations produced in emotional speech were described in the framework of a classical two-component model of intonation, namely in the IPO model, or in terms of F_0 values such as onset value (which represents pitch level), final-accent peak value (which, in relation to onset, indicates the pitch range) and final F_0 value (which relates to the final lowering). Neither of the formulations are mutually exclusive. Some of the local information described in terms of F_0 values, could not easily be captured in a classical two-component model. The perceptual relevance of this information below utterance level was tested in this perception experiment. A substantial overall improvement over all emotions was not shown. Nevertheless, considering some emotions separately, it appeared that either the presence or the absence of peak modeling or final lowering modeling can be relevant. Results suggest that the simplifications imposed by the model involve pitch characteristics that are perceptually relevant for the expression of specific emotions. On the other hand, neglecting these characteristics does not yield unacceptable results. Refraining from modeling these characteristics even appears to provide better performances for the expression of neutrality and joy. As overall results were not conclusive, the necessity to extend this two-component model, for its most common use, could not be proven.

VI. GENERAL DISCUSSION

Emotional speech produced by three individuals, was analyzed. These speakers made a fairly consistent use of the same speech parameters in the expression of emotion. To a certain extent, this corroborates the possible generalizibility of the results obtained from the production study. These results are in agreement with the conclusion of Cosmides (1983)

that different individuals adhere to reasonably standard 'acoustic configurations' in expressing particular emotions. It seems likely that various cues can contribute to the expression of a certain emotion. Apparently, speakers have personal preferences for certain cues, in expressing a specific emotion. A particular speaker would then prefer to use one cue, and rely less on the other possible cues, without necessarily showing discrepancies with other speakers, at least not if considered in a qualitative way when considering a relative ordering of the emotions according to the use of a specific cue.

The production data give a description of pitch phenomena that is also compatible with the representation obtained from perception tests; despite the quantitative individual differences in the expression of emotion observed in the speech of different subjects, the overall qualitative picture shows consistencies converging with the overall picture resulting from the perceptual study.

Three groups of emotions can be distinguished on the basis of an analysis of the data carried out at utterance level, and based on mean pitch and standard deviation; the estimation of pitch level and pitch range by means of these quantities appeared to be informative though not very accurate. Evaluating the precision of the estimation by means of mean and standard deviation showed that deviations from the exact value often reached a perceptible level, and sometimes approached a level considered to be relevant to speech communication. Furthermore, in order to distinguish between all emotions studied, a more detailed analysis was desirable.

The information obtained, by means of the refined approach, allows the course of F_0 through time to be better described. The height of the final peak and the end of speech seem to be particularly valuable supplementary information. The level of the baseline was not very well estimated using measurements of F_0 at local points on the pitch curve. Although, in the present study, the effect of the slope of the declination line, and its relevance for the expression of emotion in speech, have not been studied, it seems of interest to investigate these issues in further study. A secondary advantage of the refined approach is that the information about exact positions in the tonal space was able to refine an estimation of pitch level and pitch range. Such an approach including more detail revealed characteristics in the pitch curve that are obscured by an approach at utterance level.

An advantage of the study of such extreme data as the ones produced in the expression of emotion, is that it provides the opportunity to consider the adequacy of models and explore

the necessity of extending the models. The production data were described within a two-component model of intonation. Some details observed with the refined approach could not be captured in this two-component model. The perceptual evaluation of the relevance of this detailed information, as performed so far, did not show that information below utterance level was significant. Therefore, the difficulty of representing the detailed information should not be considered a major problem. However, the model appears to provide a simplification of the pitch phenomena. This simplification seems to remove information that might not be of primary importance, but still seems to be relevant to the expression of particular emotions in speech. The limited data that was collected here did not allow an unequivocal interpretation of the adequacy of the model for non-neutral speech. Nevertheless, the detailed information can still help us to understand the production of emotional speech, and it can be decisive in order to distinguish between certain emotions. Furthermore, although, among the two-component models, only the IPO-model was tested in the present study, it seems quite reasonable to extrapolate the results and to assume that other two-component models (Fujisaki, 1993; Taylor, 1994) would yield similar results.

Labeling the pitch curves with sequences of pitch movements, i.e., intonation patterns, appeared to be an interesting source of information. In almost all cases, the labeling of the emotional utterances yielded patterns which were grammatically correct within the IPO grammar for Dutch intonation ('t Hart et al., 1990). Even though some patterns may convey an emotion better than others, no specific pattern appeared to be strictly necessary for signaling emotion in speech. This means that the intonation grammar by 't Hart et al. (1990) does not require to be extended when considering emotional speech. Among the standard patterns of pitch movements, the '1&A' proved to be a pattern that can potentially be used in the expression of all emotions studied. It does not mean that '1&A' is the most appropriate pattern to express all emotions studied; another pattern can be more effective in signaling a particular emotion. However, all emotions could be identified in utterances re-synthesized using only the '1&A 1&A' intonation pattern. Whether the use of a specific pattern might yield better results for particular emotions, has not been tested. However, if one requires different intonation patterns not to contribute to experimental variability, the '1&A' pattern of pitch movements appeared to be best suited for controlling this variability. Further study of the relevance of intonation patterns for the expression in speech seems promising.

Finally, it is clear that a high identification rate of emotions cannot be obtained through pitch manipulation only. Lieberman and Michaels (1962) already found that pitch is very

important to the expression of emotion in speech, but that the full emotional content cannot be conveyed only by pitch. Other aspects, such as speech rate and voice quality must also be taken into account. Additionally, we saw in this chapter, that all emotions were identified in utterances generated with the '1&A' pattern of pitch movements, but that the identification was not satisfyingly high in those experiments in which the intonation patterns were kept constant in all the emotions. It also appeared that speakers did not use all patterns of pitch movements as frequently for the expression of all emotions. Therefore, the next step we will take, in Chapter IV, will be to further investigate the role of the intonation patterns in conveying emotion in speech. Moreover, such an analysis of pitch curves in terms of underlying intonation patterns, constitutes information that is complementary to the data obtained in the present study by measurements at anchor points.

Chapter IV

A study of intonation patterns

ABSTRACT

In a production study and a perception study, the relationship between the intonation pattern realized on an utterance, and the emotion expressed in that utterance, was investigated. In the production study, the pitch curves in the speech material, composed of 315 emotional utterances, were labeled in terms of the IPO intonation grammar. One pattern of pitch movements, the '1&A', i.e., a rise and a fall realized on a single syllable and lending prominence to this syllable, was produced in all emotions studied. Some other patterns, i.e., patterns ending with pitch movements 'C' or '2', were specifically used in expressing some emotions. In the perception study to check the perceptual relevance of these findings, the role of the patterns of pitch movements present in the database was tested. A listening test provided converging evidence on the contribution of specific patterns in the perception of some of the emotions studied. Some intonation patterns including a final '2' or a final 'C', which were specifically produced in some emotion, e.g., indignation, also introduced a perceptual bias towards that emotion. In that sense, the results from the perception study supported the results from the production study. Finally, clusters of intonation patterns were derived from the results of the perception test. This clustering corroborates the perceptual distinction among intonation patterns as defined in the IPO intonation grammar for Dutch ('t Hart et al., 1990).

I. INTRODUCTION

Speech not only conveys the strictly linguistic content of sentences but also the expression of emotions of the speaker. Prosody plays a role in this, which may result in adding information to the linguistic content and/or its modification. Pitch level, pitch range, and speech rate, are known to be important prosodic cues for, among others, the expression of emotions in speech (e.g., Williams and Stevens, 1972; Cahn, 1990; Carlson, Granström and Nord, 1992; Kitahara and Tohkura, 1992). Related studies were reviewed in Frick (1985) and in Murray and Arnott (1993). These studies rarely related production data to perception data, and did not consider intonation patterns. It is likely that information on the emotion expressed by the speaker is not only present in global properties of the utterances such as pitch level and pitch range, but also in more local properties involving changes of pitch over time.

Therefore, in the present study, the role of these intonation patterns in conveying emotions in speech is further investigated. First, the pitch curves in the speech produced by three speakers, that were described in Chapter III according to the IPO-model of intonation for Dutch ('t Hart, Collier and Cohen, 1990), are presented here in more detail. The distribution of the patterns of pitch movements over the emotions is analyzed. The results form the basis of a perception experiment in which pitch contours that correspond to a series of different intonation patterns found in these production data are synthesized. Originally neutral utterances are provided with these pitch contours and are presented to listeners, who indicate the emotion they think is being conveyed by the utterance. This is done for all seven values of pitch level and pitch range found optimal for each emotion in Chapter II, so that the intonation pattern is varied independently of the optimal pitch level and pitch range of an emotion. It is found that the intonation pattern has a highly significant effect on the distribution of the subjects' responses to the seven emotions.

In this study, it has been shown that the IPO-model of Dutch intonation is adequate for describing the vast majority of all pitch curves found in all the utterances produced in expressing emotion. No extensions appeared to be necessary. Nevertheless, in the process of synthesizing the experimental utterances, it appeared that various basic pitch movements, which form the basis of the model, were not well specified. It was necessary to have a closer look at the exact realizations of the various basic pitch movements defined in the IPO-model. In some cases, the details appeared to be extremely important.

Therefore, a detailed specification of all ten basic pitch movements, as used in the synthesis of the emotional utterances, is included in the present study.

II. PRODUCTION OF EMOTION: AN ANALYSIS OF SPEECH MATERIAL

a. Speech material

The speech material that was selected in Chapter II was used again in the current chapter. It includes 315 utterances (3 speakers × 7 emotions × 5 sentences × 3 trials). The lexically stressed positions on which pitch accents were realized by the speakers in the five sentences were given in Chapter III.

b. Procedure

In the Dutch grammar of intonation by 't Hart, Collier and Cohen (1990), pitch movements, besides being a rise or a fall, are characterized by the following features ('t Hart et al., 1990, p. 153): 1. their timing in the syllable, 2. their spread over one or more than one syllable, 3. their size. Their size, full or half, is defined as the whole distance between baseline and topline for a full movement, and as a smaller distance for a 'not-full' movement. A functional characteristic is, furthermore, that a pitch movement may or may not lend prominence to a syllable. This leads to the ten basic pitch movements of this grammar, as described in Chapter III. If two movements occur on a single syllable, the symbols are linked with an ampersand, for example, '1&A' or '5&A'. Not all combinations of pitch movements are grammatically correct. The main grammatically correct patterns of pitch movements for Dutch are the following: '1&A' (rise-fall on a single accented syllable) which is the most typical 'pointed hat', '1A' (rise on an accented syllable and fall on the next accented syllable) also called 'flat hat', '1B' (rise on an accented syllable and early fall, generally before a phonological boundary), '1D' (rise on an accented syllable and gradual fall), '5A' (overshoot after a rise, and full fall), '1EA' (rise, half fall, and fall), '12' (rise followed by a very late rise), '1C' (rise and very late fall), '3C' (late rise and very late fall), and '45&A' (slow rise with overshoot and a full fall on a single syllable). Some schematized examples of intonation patterns were given in the previous chapter.

As explained in Chapter III, two experts in intonation research labeled the realized pitch curves of these utterances according to this grammar of intonation. It was uncertain

Table 1: Distribution of the INITIAL patterns of pitch movements for the three speakers

Emotions	Patterns of pitch movements						
	1&A	**1**	**1B**	**1D**	**45**	**1&2/1&2B**	**rest**
speakers	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO
neutrality	**7**	**12**	**15**	**3**	**-**	**5**	**3**
	3 / 2 / 2	5 / 2 / 5(1)	5 / 4 / 6	2 / 1 / -	- / - / -	- / 4 / 1	- / 2 / 1
joy	**23**	**5**	**8**	**5**	**-**	**-**	**4**
	6 / 6 / 11	2 / - / 3	6 / 2 / -	- / 4 / 1	- / - / -	- / - / -	1 / 3 / -
boredom	**8**	**15**	**16**	**2**	**2**	**1**	**1**
	6(1)/ 1 / 1	5(3)/ 5 / 5	4(2)/ 8 / 4	- / - / 2	- / - / 2	- / 1 / -	- / - / 1
anger	**18**	**6**	**16**	**1**	**-**	**2**	**2**
	6 / 7 / 5	3 / - / 3	5 / 5 / 6	- / 1 / -	- / - / -	- / 1 / 1	1 / 1 / -
sadness	**19**	**3**	**16**	**1**	**1**	**1**	**4**
	9 / 5 / 5	- / 2 / 1	6 / 3 / 7	- / 1 / -	- / - / 1	- / 1 / -	- / 3 / 1
fear	**20**	**6**	**14**	**2**	**-**	**-**	**3**
	7 / 7 / 6	- / 4 / 2	6 / 2 / 6	- / 1 / 1	- / - / -	- / - / -	2 / 1 / -
indignation	**20**	**9**	**4**	**7**	**-**	**-**	**5**
	5 / 5 / 10	7 / 2 / -	1 / 1 / 2	- / 5 / 2	- / - / -	- / - / -	2 / 2 / 1
total over all emotions	**115**	**56**	**89**	**21**	**3**	**9**	**22**
	42/ 33/ 40	22/ 15/ 19	33 / 25 / 31	2 / 13 / 6	- / - / 3	- / 7 / 2	6 / 12 / 4

whether the pitch movements available in the grammar would permit a satisfactory description of all pitch curves realized by the speakers. Therefore, the experts were asked to explain, wherever necessary, in which way the concrete pitch movement realized by the speaker was deviant from the perceptually nearest pitch movement available in the grammar.

c. Results

The first results were already presented in the previous chapter. The five utterances which were realized with only one accent (accent realized on the second part of the utterances),

Table 2: Distribution of the FINAL pattern of pitch movements for the three speakers

Emotions	Patterns of pitch movements						
	1&A	**A/1A**	**5A/ 5&A**	**EA**	**12/ 1&2**	**C/ 1C/ 3C**	**rest**
speakers	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO	MR/ RS/ LO
neutrality	**27**	**10**	**2**	**5**	**-**	**-**	**1**
	9 / 13 / 5	4 / - / 6(2)	1 / - / 1	1 / 2 / 2(1)	- / - / -	- / - / -	- / - / 1
joy	**32**	**2**	**2**	**-**	**-**	**9**	**-**
	11 / 13 / 8	- / - / 2	- / 2 / -	- / - / -	- / - / -	4 / - / 5	- / - / -
boredom	**15**	**12**	**3**	**5**	**1**	**7**	**2**
	8(1)/ 6 / 1	4(4)/ 5(1)/ 3	- / - / 3(1)	3(1)/ 2 / -	- / - / 1	- / 2 / 5	- / - / 2
anger	**31**	**5**	**4**	**2**	**-**	**2**	**1**
	11 / 12 / 8	1 / 1 / 3	- / 2 / 2	2(1)/ - / -	- / - / -	- / - / 2	1 / - / -
sadness	**29**	**8**	**1**	**-**	**-**	**7**	**-**
	12 / 11 / 6	1(1)/ 1 / 6(2)	- / 1 / -	- / - / -	- / - / -	2 / 2 / 3	- / - / -
fear	**27**	**3**	**2**	**-**	**-**	**11**	**2**
	14 / 6 / 7	- / 1 / 2(1)	- / - / 2	- / - / -	- / - / -	1 / 7 / 3	- / 1 / 1
indignation	**9**	**-**	**9**	**-**	**13**	**10**	**4**
	6 / 3 / -	- / - / -	9 / - / -	- / - / -	- / - / 13	- / 10 / -	- / 2 / 2
total over all emotions	**170**	**40**	**23**	**12**	**14**	**46**	**10**
	71 / 64 / 35	10 / 8 / 22	10 / 5 / 8	6 / 4 / 2	- / - / 14	7 / 21 / 18	1 / 3 / 6

are now included in the 'rest' category for the first pattern of pitch movements. The distribution of the patterns of pitch movements over the various categories is presented separately for each speaker: in Table 1 for the initial patterns of pitch movements, and in Table 2 for the final patterns of pitch movements in the utterances.

In general, the experts were satisfied with the description in terms of the labels available in the grammar. On twenty-three occasions in the speech material of the three speakers, however, the experts specified that the pitch movement 'A' or '1' was realized with a delay, without being as late as a 'C' (very late fall) or a '2' (very late rise). These realizations involving a delay are indicated in Tables 1 and 2 by a number between

parentheses indicating how many of the realizations of the specific pattern of pitch movements were realized with delay. All the concrete delayed pitch movements occurred in sixteen different utterances, i.e., 7.3% of the database. Eight of the sixteen utterances were expressing boredom. Speaker MR realized most delayed pitch movements. In particular, he produced twelve delayed movements in six different utterances when expressing boredom. This can be considered as a personal preference, since speaker RS, for example, only produced a single delayed movement. It does suggest, however, that at least in the speech of speaker MR, the timing of the pitch movement may be of particular relevance to the expression of emotion.

From Tables 1 and 2, it appears that there is no one-to-one coupling between single emotions and isolated patterns of pitch movements. An emotion can be realized with different patterns of pitch movements, and a pattern can be used in different emotions; this is not to say that patterns of pitch movements are evenly distributed over all emotions. The most frequently used pattern of pitch movements is the so-called pointed-hat '1&A' ('t Hart et al., 1990), which is the only pattern of pitch movements that was produced by every speaker in expressing each emotion, both in the initial and final position, with the exception of one speaker, for indignation. It does not mean that this pattern of pitch movements is the best choice for the realization of a particular emotion, but that it is possible, using this pattern, to express each of the emotions studied here. This was the reason why it was used for the re-synthesis in Chapter III.

As mentioned before, most of the realizations of the *initial* pattern of pitch movements (Table 1) were so-called 'pointed hats', '1&A', or variants of the 'pointed hat', '1B' and '1D'; all together, 225 of the 315 utterances analyzed were realized with one or the other 'pointed hat'. Another 56 utterances were realized with a prominence-lending rise '1'. The remaining 34 utterances were distributed over the other categories. A direct, one-to-one relationship between pattern of pitch movements and emotion cannot be observed. Even the '1&A' pattern was not evenly distributed over the emotions ($\chi^2_6 = 14.5$, $p < .025$). Also, the patterns '1' and '1B' tended not to be equally distributed over the emotions. For '1', this was significant ($\chi^2_6 = 14.5$, $p < .05$), but this did not reach significance for '1B' ($\chi^2_6 = 10.8$, $.05 < p < .10$). This can, in general, be ascribed to the fact that the '1&A' pattern is less often realized in neutrality and boredom, for which emotions the early rise '1' is over-represented. For '1B', the tendency might be ascribed to the under-representation of the '1B' pattern in indignation. The other patterns of pitch movements were not realized often enough to show clear statistically significant effects.

For the *final* pattern (Table 2), some relationships are more direct. The 'C'/ '1C'/ '3C' and the '12'/ '1&2' patterns of pitch movements were never realized in neutrality. The '12'/ '1&2' patterns in the final position, were mainly used by the female speaker LO when expressing indignation. Other significant effects are that the '1&A' pattern of pitch movements ($\chi^2_6 = 19.0$, $p < .005$) is under-represented in boredom and indignation, that the 'A'/ '1A' ($\chi^2_6 = 20.5$, $p < .0025$) is over-represented in boredom and absent from indignation, and perhaps (expected values are smaller than 5, $\chi^2_6 = 13.2$, $p < .05$) that the '5A'/ '5&A' patterns are over-represented in indignation.

For the expression of indignation, the speakers used a specific final pattern of pitch movements more often than the, presumably more 'standard', '1&A' pattern; MR usually used the '5&A' pattern, RS the final pitch movement 'C', and LO the '12' or the '1&2' pattern. This suggests that the choice of intonation patterns may be of particular relevance for the expression of some emotions. Moreover, the patterns of pitch movements '12', '1&2', and 'C', were not used a single time by any of the speakers in the final position in neutral speech. A final '2' was only used by the female speaker LO, mainly in the expression of indignation. In the final position, patterns ending in the pitch movement 'C' are the second most commonly used patterns in emotional speech, just after the ones involving '1&A'. This pitch movement 'C' could be a good choice for signaling emotionality in general.

It is also quite interesting to notice that the distribution for the utterance final pattern, pooled over the three speakers, over all emotions, and over the few selected categories that were used in the present study as well as in the study reported by Collier (1972, p. 136), compare quite well to one another. How the percentages compare can be seen in Table 3.

d. Conclusion

As was already concluded in Chapter III, there is no direct one-to-one relationship between pattern of pitch movements and emotion. It appears that it is possible to use the pattern of pitch movements '1&A', which is the pattern most frequently used in Dutch, in the expression of each of the seven emotions. This does not mean that this pattern is the best possible choice for expressing each of these emotions.

Table 3: Percentage of type of configurations in final position used in two studies

	Categories of patterns of pitch movements				
	1&A/ A/ 1A	EA	C/ 3C/ 1C	12/ 1&2	others
Present findings	66.7	3.8	14.6	4.4	10.5
	Categories of patterns of pitch movements				
	1A/ A	1E	3C	2	others
Findings by Collier (1972)	68.8	3.8	12.6	1.3	13.5

Caspers (1997) reports converging results; she investigated the meaning associated with intonation patterns in Dutch and also found that the '1&A' pattern of pitch movements is 'unmarked' compared with other patterns ('A', '1&E' and '12'). In her study, the '1&A' pattern simply signals the standard situation in which the information provided in the sentence about the subject being discussed is new; other patterns can indicate a more particular situation, such as when the information is meant to be already known by the listener. In her study, as in the present one, the '1&A' is reported to be the most likely default pattern of pitch movements.

The data of the present study suggest two hypotheses. The first hypothesis concerns the possibility of conveying the seven emotions studied using the '1&A' pattern of pitch movements. As for this hypothesis, it is already supported by the results of the listening experiment of Chapter III, also reported by Mozziconacci (1995). These results suggest that the emotions studied can be conveyed using a sequence of '1&A' patterns of pitch movements as a constant phonological structure. A second hypothesis is that emotions can be conveyed to the listener by using specific patterns, such as patterns ending with '2' or 'C'. This follows from the finding that speakers can prefer a pattern of pitch movements in the expression of some specific emotion, that is a pattern they use more often than any other one when they express that emotion. For the expression of indignation, each speaker had a different preferred pattern of pitch movements for the final accent, all different from the more 'standard' '1&A' pattern. This suggests that the choice of intonation patterns may be relevant for the expression of some emotions. The patterns '12', '1&2', 'C', '1C' and '3C', were not used a single time by any of the speakers for neutrality. A final '2' was only used by the female speaker LO in the expression of indignation. A final 'C' is the second most used pitch movement in

emotional speech, just after the '1&A' pattern. This pitch movement might be a good choice in signaling emotionality in general.

Finally, the labeling of the final pattern of pitch movements in the utterances seems to show, more clearly than the labeling of the first pattern of pitch movements, the specificities of the use of intonation patterns in the expression of emotion in speech. This suggests that the end of the utterance is of particular importance to the characterization of the intonation pattern.

III. PERCEPTION OF EMOTION: AN EXPERIMENT

a. Aim

The previous analysis of emotional speech shows that, in this recorded speech database, the '1&A' pattern of pitch movements was produced in the expression of all emotions, and that patterns ending with '2' or 'C' were produced in emotional speech but not in neutral speech.
Here, we investigate whether specific patterns of pitch movements affect the perception of emotion in speech. Therefore, the identification of intended emotions, encoded in speech by varying the patterns of pitch movements, is investigated. If the distribution of the subjects' responses differs significantly for different patterns, the relevance of the choice of intonation patterns for conveying emotion in speech can be established.

b. Speech material

Neutral utterances of both Sentence 1 'Ze hebben een nieuwe auto gekocht' (They have bought a new car) and Sentence 2 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane) were manipulated by analysis-resynthesis based on the PSOLA technique (Verhelst and Borger, 1991). The manipulations combine different intonation patterns with a pitch level and pitch range that was found to be adequate for each of the seven emotions in Chapter II. The values for pitch level and pitch range used in the manipulations are reported in Chapter II, in Table 14. Among all intonation patterns occurring in the speech database, only those patterns occurring more than a couple of times and considered 'legal', according to the intonation grammar for Dutch ('t Hart et al., 1990), were retained for testing in the listening experiment. They comprised 11 intonation patterns: '1&A 1&A', '1B 1&A', '1D 1&A', '12 1&A', '1 5&A', '1A', '1EA', '1&A 3C', '1B 3C', '1D 3C', and '1&A 12'.

Distinctive features of the pitch movements constituting these patterns are, among others, their timing in the syllable and their excursion size. While synthesizing the test-utterances, information such as the exact moment the rise or fall should start, how long the movement should last, and how much it should rise or fall, appeared to be very important perceptually. Small differences of less than 30 ms in timing and duration of a pitch movement, could sometimes mean the absence or presence of an accent.

In 't Hart, Collier and Cohen (1990, p. 153), the start and the end of voicing are suggested as the perceptual reference points for timing pitch movements within the syllables. At other instances (e.g., p. 73), 't Hart et al. (1990) present the vowel onset, instead of the voicing onset, as a syllable boundary relevant in this respect (p. 73). Collier (1991) also presents a description of rise-fall patterns of pitch movements in various languages, the so-called 'hat patterns', in which the timing of the pitch movement is specified with respect to the vowel onset. Moreover, in other studies, it has been argued that the vowel onset is a better candidate than the start of voicing (e.g., d'Alessandro and Mertens, 1995). This is based on a model of tonal perception, in which pitch in spectrally stable parts plays a far more important role in speech intonation than pitch in other parts of speech (House, 1990). Since the spectral content of the consonant cluster of the syllable onset varies very rapidly while the vowel is relatively stable, the vowel onset is a more likely reference point, where perceptually relevant pitch information in the syllable starts to be processed linguistically. This is confirmed by further studies (Hermes, 1997; Hermes, Beaugendre and House, 1997). On the other hand, it was also shown that pitch in voiced parts of the syllable *after* the vowel, can play a role in linguistically relevant distinctions (House, 1996; Hermes, Beaugendre and House, 1997). Therefore, vowel onsets and end of voicing, measured in the carrier utterance, served as points of reference when synthesizing the pitch movements.

The timing and the duration of the pitch movements in the synthesis of '1&A' and '1A' were derived from Collier (1991). The realization of the other movements was partly based on a description of representative pitch movements given in 't Hart and Collier (1975). For some pitch movements, this description was not exhaustive for timing, duration and excursion size. In these cases, the F_0 variations of the natural realizations by the actors were taken as the point of departure to find appropriate specifications. The exact timing, duration and excursion of each movement was consequently fine-tuned by trial and error. The realizations of the pitch movements that were finally adopted are

Table 4: Description of the timing, duration and excursion size of the pitch movements as realized in the pitch contours synthesized

vo : vowel onset: if not explicitly mentioned, the vowel onset in the accented syllable is meant.
eov: end of voicing of the first accented syllable.

Pitch movement	Beginning	Duration	Excursion size
1	vo - 70 ms	120 ms	Pitch range
2			
- in initial pattern	eov - 120 ms	100 ms	Pitch range + 3 s.t.
- in final pattern	vo of last syllable + 20 ms	80 ms	Pitch range × 2
3	vo	80 ms	Pitch range
4	end of preceding fall	till start of next rise (various syllables)	Pitch range
5	vo - 70 ms	60 ms	Pitch range × 0.5
A			
- in 1&A	vo + 80 ms	120 ms	Pitch range
- in 1A	vo - 20 ms	120 ms	Pitch range
- in 1EA	vo of last syllable - 70 ms	120 ms	Pitch range × 0.5
- in 15A	vo + 40 ms	120 ms	Pitch range × 1.5
B	vo of next syllable - 110 ms	120 ms	Pitch range
C	vo of last syllable + 70 ms	120 ms	Pitch range
D	top of preceding rise	till start of next rise (various syllables)	Pitch range
E	vo - 70 ms	120 ms	Pitch range × 0.5

described in Table 4. Schematized standardized versions of relevant intonation patterns were given in Chapter III (Figure 4).

The fall 'D' was simply realized as a gradual movement, progressively declining from the top of the first peak until the beginning of the next rise. The excursion size of the half fall 'E' was chosen to be half of the standard excursion size. In order to produce the fall 'B' relatively early, this movement was timed to end around the vowel onset of the syllable

following the pitch accent, so that this movement would not be perceived on that next syllable. The timing of the very late pitch movements '2' and 'C', were especially difficult to deal with. The rise '2' in the initial pattern of pitch movements was timed to fully develop just before the end of voicing of the first part of the sentence. In the final pattern of pitch movements, the '2' and the 'C' were timed relative to the vowel onset of the last syllable of the utterance, in order to realize the pitch movement as late as possible. For the 'C' in particular, it was important not to realize the pitch movement too soon: that would lend prominence to the final syllable, which had to be avoided. In fact, this fall remained virtual (i.e., the fall was not completed before the end of the utterance, which means that F_0 did not reach its minimum value) in Sentence 1, 'Ze hebben een nieuwe auto gekocht', because it could not begin earlier without lending undesired prominence to the last syllable (cf. 't Hart and Collier, 1975). The phonetic content of the last syllable of Sentence 2, 'Zijn vriendin kwam met het vliegtuig', allowed a complete fall without undesired prominence. Contrary to the indications of 't Hart and Collier (1975), who found rise '3' to have a larger excursion than rise '1', the rise '3' was here realized with a standard excursion size. This choice to keep the pitch range constant was made for methodological reasons; if pitch range and intonation patterns had both been varied, it would have been impossible to know whether the results were to be attributed to the difference in pitch range or to the difference in intonation pattern. The excursion size of the rises '2' and '5' were chosen in such a way that these movements would still be perceivable when the optimal pitch range is small, but not exaggerated when the range is large. Therefore, for the pitch movement '2' in the initial part of the sentence, three semitones were simply added to the value for pitch range.

For each intonation pattern, the utterances were re-synthesized with each of the seven combinations of pitch level and pitch range considered optimal for one of the seven emotions (Chapter II, See Table 14 of that chapter). The manipulation was based on the PSOLA technique (Moulines and Laroche, 1995). Segmental durations were left unchanged. The resulting 154 variants (11 intonation patterns × 2 sentences × 7 combinations of pitch level and pitch range) served as stimuli.

c. Design and procedure

Each sentence was presented in a separate block. The order of presentation of the two blocks was counterbalanced across the subjects. Within a block, the stimuli were

presented to each subject in a different, random order. In total, 24 subjects (12 female, 12 male) participated in the listening experiment. Half of them were either students or staff members at IPO, the others were paid for their participation. None of them had any particular knowledge of phonetics. Each subject took the test individually, using headphones and an interactive computer program. Subjects could only listen once to each stimulus. Their task was to assign one of the seven emotion labels to the utterance they heard. They did not get any feedback on their performance. They had a short break before the second block, splitting the test into two periods of about 10 to 15 minutes each.

d. Results

The responses of the subjects were pooled into a three-dimensional table with pitch-level and pitch-range combination as one dimension PITCHCOMBI, intonation pattern as a second dimension PATTERN, and the response of the subjects as a third dimension RESPONSE. So the first two dimensions represented independent variables, while the third represented a dependent variable. The two sentences were analyzed both separately and collapsed. These data were subjected to a three-way loglinear analysis (Fienberg, 1980). The analysis showed that our data can be fitted into the two loglinear models described below (the data did not deviate significantly from these two loglinear models). Such a loglinear model has the advantage of making no assumption of normal distribution of the data, and allows the information contained in confusions in the responses of subjects to be taken into account, instead of only considering the correct responses. In the first model, there were significant interactions between PITCHCOMBI and RESPONSE, between PATTERN and RESPONSE, but not between PITCHCOMBI and PATTERN ($\chi^2_{420} = 391.7$, $p > .16$ for Sentence 1; $\chi^2_{420} = 374.6$, $p > .05$ for Sentence 2; $\chi^2_{420} = 459.7$, $p > .9$ for the results collapsed over the two sentences). In the second model, there were interactions between all three variables (Sentence 1: $\chi^2_{360} = 365.6$, $p > .5$; Sentence 2: $\chi^2_{360} = 357.4$, $p > .7$; both sentences: $\chi^2_{360} = 419.6$, $p > .9$). All other models yielded expected values that differed significantly from the values obtained at the $p < .001$ level. (It should be mentioned that these models yielded expected values that were lower than 5 in many cells of the three-dimensional table. This indicates that presented p-values must be interpreted with caution. Furthermore, loglinear models assume independence of every observation, whereas observations came from 24 subjects, so that within-subject correlations might have obscured some results.) In view of the

Table 5: Response distribution per pattern

Patterns	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
1&A 1&A	136	53	48	25	28	16	30
1B 1&A	132	54	43	19	35	25	28
1D 1&A	130	52	47	30	42	15	20
12 1&A	126	70	32	21	28	22	37
15&A	90	69	31	52	20	16	58
1A	94	19	36	76	26	18	67
1EA	85	33	47	62	39	8	62
1&A 3C	73	44	50	20	55	26	68
1B 3C	79	64	64	18	42	16	53
1D 3C	62	41	58	29	55	24	67
1&A 12	30	27	29	29	25	48	148
total	1037	526	485	381	395	234	638

Table 6: Response distribution per combination of pitch level and pitch range optimal for each emotion

Optimal combination of pitch level and pitch range	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
Combination for neutrality	**195**	28	156	45	56	9	39	528
Combination for joy	139	**93**	16	63	51	45	121	528
Combination for boredom	169	11	**215**	36	57	7	33	528
Combination for anger	165	136	23	**47**	27	16	114	528
Combination for sadness	237	68	54	44	**41**	14	70	528
Combination for fear	46	73	9	70	100	**98**	132	528
Combination for indignation	86	117	12	76	63	45	**129**	528
total	1037	526	485	381	395	234	638	3696

Table 7: p-values associated with the chi-squares (χ^2) used for measuring dissimilarities between patterns for Sentence 1: 'Ze hebben een nieuwe auto gekocht'

All values > .05 are shown.

	Intonation patterns										
	1&A1&A	1B1&A	1D1&A	121&A	15&A	1A	1EA	1&A3C	1B3C	1D3C	1&A12
1&A 1&A	-	.908	.596	.416							
1B 1&A	.908	-	.888	.763					.061		
1D 1&A	.596	.888	-	.156							
12 1&A	.416	.763	.156	-	.264						
15&A				.264	-						
1A						-	.680				
1EA						.680	-	.183	.238	.177	
1&A 3C							.183	-	.931	.840	
1B 3C		.061					.238	.931	-	.796	
1D 3C							.177	.840	.796	-	
1&A 12											-

good fit of the first loglinear model, especially for the data pooled over the two sentences, and the negligible improvement of the fit obtained when applying the second model, we will adopt the first model; this means that both the interaction between PITCHCOMBI and RESPONSE and between PATTERN and RESPONSE are significant, but that the interaction between PITCHCOMBI and PATTERN is not significant. The implication of this is that, within each response class, PITCHCOMBI and PATTERN can be assumed to be independent factors. In the first instance, this may seem trivial since, in the experimental set-up used here, PITCHCOMBI and PATTERN are varied independently. On the other hand, it shows, for example, that one specific intonation pattern does not exclude a particular response of the subject or uniquely determine his or her response, whatever the combination of pitch level and pitch range. A practical consequence is that, within one response class, we can collapse the results over each of the independent variables. The results are presented in Tables 5 and 6.

Table 8: p-values associated with the chi-squares (χ^2) used for measuring dissimilarities between patterns for Sentence 2: 'Zijn vriendin kwam met het vliegtuig'
All values > .05 are shown.

	Intonation patterns										
	1&A1&A	1B1&A	1D1&A	121&A	15&A	1A	1EA	1&A3C	1B3C	1D3C	1&A12
1&A 1&A	-	.609	.916	.825	.191		.148	.188	.778	.203	
1B 1&A	.609	-	.902	.967	.092			.162	.208		
1D 1&A	.916	.902	-	.865	.163		.085	.247	.322	.099	
12 1&A	.825	.967	.865	-	.243		.064	.152	.317	.080	
15&A	.191	.092	.163	.243	-	.207	.622	.458	.135	.659	
1A					.207	-	.649				
1EA	.148		.085	.064	.622	.649	-	.083		.250	
1&A 3C	.188	.162	.247	.152	.458		.083	-	.295	.932	
1B 3C	.778	.208	.322	.317	.135			.295	-	.361	
1D 3C	.203		.099	.080	.659		.250	.932	.361	-	
1&A 12											-

As can be seen in the last row of Table 5 or 6, the responses of the subjects were not equally distributed over all emotion categories. For five of the seven emotions, the responses differ significantly ($p < .005$) from the mean value (528) corresponding to an equal distribution of responses; only the number of responses in the category joy was similar to the mean value ($p = .93$), and the responses for boredom differ from the mean but to a lesser extent ($p = .006$). Over-represented response categories were neutrality and indignation, under-represented response categories were fear, anger, and sadness.

Table 5 forms the basis of a cluster analysis in which we attempt to find out whether intonation patterns can be combined into groups. The members of such groups give rise to similar responses. The composition of these clusters may provide us with information about which properties of the intonation patterns are essential in conveying a certain emotion. Indeed, each possible combination of two rows results in a two-dimensional

Table 9: p-values associated with the chi-squares (χ^2) used for measuring dissimilarities between patterns for both sentences

All values > .05 are shown.

	Intonation patterns										
	1&A1&A	1B1&A	1D1&A	121&A	15&A	1A	1EA	1&A3C	1B3C	1D3C	1&A12
1&A 1&A	-	.920	.843	.680							
1B 1&A	.920	-	.733	.800							
1D 1&A	.843	.733	-	.226							
12 1&A	.680	.800	.226	-							
15&A					-						
1A						-	.333				
1EA						.333	-				
1&A 3C								-	.437	.948	
1B 3C								.437	-	.280	
1D 3C								.948	.280	-	
1&A 12											-

(2 × 7) table. If the two intonation patterns corresponding with the two rows induce a different distribution of the responses over the emotions, differences will be observed between the two rows. If, on the other hand, the distribution of the responses over the emotions is similar for the two intonation patterns, the two rows will be the same except for statistical fluctuations. In other words, in the latter case, the composition of the rows will be independent of the intonation pattern. Therefore, for each possible combination of two rows, a (loglinear) test of independence was applied. The computation of chi-squares (χ^2's), determining the deviation from this model, was assumed to provide a measure of association for the two intonation patterns corresponding with the two rows. This χ^2 was obtained with six degrees of freedom, and the corresponding p-values were obtained. The smaller the χ^2, the more the two intonation patterns are associated, and inversely for the p-values. In Tables 7, 8 and 9, the p-values larger than .05 are presented for the χ^2's with 6 degrees of freedom. In Table 7, the results are presented for Sentence 1, in Table 8 for Sentence 2, and in Table 9 for the collapsed results. These combined results are the most

clear. There are three clusters of intonation patterns. The first cluster is composed of '1&A 1&A', '1B 1&A', '1D 1&A' and '12 1&A', and will further be called '1...1&A'. The second cluster is composed of '1A' and '1E' and will be referred to as '1...A' in the following. The third cluster which is composed of '1&A 3C', '1B 3C' and '1D 3C' will be referred to as '1...3C'. The remaining intonation patterns, '1&A 12' and '1 5&A', differ from each other and from all the other patterns. Note that the members of these clusters correspond, to a large extent, to the last pattern of pitch movements occurring in the utterance. All intonation patterns ending in '1&A' form a cluster, as well as the three intonation patterns ending in '3C'.

For the two sentences separately, there are more cells where the χ^2 reaches the level where it becomes significant (i.e., $p < .05$), and this occurs more often in Sentence 1 than in Sentence 2. In Sentence 2, these cells where significance level is not reached, involve either a final 'C' or the '1 5&A' pattern. For the '15&A' pattern, this can, to a large extent, be attributed to the segmental composition of its last part 'vliegtuig': /vlix tœyx/ (the vowel of the accented syllable is underlined), which consists of only two syllables and contains relatively more unvoiced parts than Sentence 1, while the listener has to extract all pitch information from these final syllables. Sentence 1 ended in 'nieuwe auto gekocht': /ni wə o̲ː to ɣə kɔxt/, where more pitch information is available and confusions are less likely to occur. (Recall that, in order to avoid learning effects, listeners could listen to each sentence only once.) Additionally, for the patterns with a 'C' in the final position, it should be recalled that the 'C' was a fully realized fall in Sentence 2, while the pitch contour remained higher in Sentence 1, i.e., the fall remained virtual. This difference occurred because, in Sentence 1, the fall had to be realized as virtual in order not to lend prominence to the last syllable; this risk of lending prominence to the last syllable of Sentence 2 was not present because there was only one syllable left after the last accent peak. This observation about the exact realization of this pitch movement suggests that the 'C' is more characteristic, i.e., more distinct from an 'A' when the fall is not complete.

Finally, the responses, collapsed over PITCHCOMBI as presented in Table 5 were summed over all members of each cluster. The result is presented in Table 10. Deviations from a loglinear model in which perceived emotion and intonation pattern are independent, are marked with ↑ if the obtained value is higher than expected ($\chi^2 > 3.84$, $p < .05$), and with ↓ if the obtained value is lower than expected. The results show, for instance, that in the response class 'neutrality', stimuli with intonation patterns of the

Table 10: Response distribution per cluster of intonation patterns

Clusters	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
1...1&A	524↑	229↑	170	95↓	133	78	115↓	1344
15&A	90	69↑	31↓	52↑	20↓	16	58	336
1...A	179	52↓	83	138↑	65	26↓	129	672
1...3C	214↓	149	172↑	67↓	152↑	66	188	1008
1&A 12	30↓	27↓	29↓	29	25	48↑	148↑	336
total	1037	526	485	381	395	234	638	3696

Table 11: Correct identification of emotions per cluster of intonation patterns

Clusters	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
1...1&A	90↑	45↑	75	11	7↓	38	26↓	292
15&A	24	13	20	4	3	8	12	84
1...A	37	12	34	20↑	6	12	33∧	154
1...3C	34↓	20	69∧	8	21↑	30	33	215
1&A 12	10∨	3∨	17	4	4	10	25↑	73
total	195	93	215	47	41	98	129	818

'1...1&A' cluster are significantly over-represented, while stimuli of the '1...3C' cluster and the '1&A 12' intonation pattern are under-represented. Another example is that within 'indignation' '1&A 12' is over-represented and '1...1&A' is under-represented. All other over-representations and under-representations of responses can be seen in Table 10.

A similar cluster analysis (but now for considering how emotions cluster) was carried out on the basis of the data that was presented in Table 6, where, collapsed over all intonation patterns, the distribution of the responses is presented over the seven emotion categories for the combinations of pitch level and pitch range found optimal for each of these emotions. The result was that, considering all utterances, only two clusters of emotions

could be formed, one for neutrality and boredom (p = .116), the other for joy and indignation (p = .159). All other χ^2's were so high that their corresponding p-values were lower than .005. Considering the sentences separately, three additional groupings were found, one for Sentence 1 concerning indignation and fear, two for Sentence 2 concerning anger and joy, and indignation and anger. Not surprisingly, these clusters correspond to the groupings that could be found in the previous chapter on the basis of pitch level and pitch range.

Additionally, if one considers an emotion to be 'correctly identified' when an utterance that had received the pitch level and pitch range adequate for that particular emotion was labeled with that same emotion by a subject, independently of the intonation pattern used in the utterance, then emotions were correctly identified in 818 utterances (see Table 11); according to the previous definition, 22.1% of the cases were correctly identified. Although low, this percentage of correct identification is higher than a chance level of 14.3%. This percentage can naturally not be compared with the usual percentage of correct identifications because, in the present study, the stimuli were not prepared to instantiate a specific emotion. Each of the eleven intonation patterns was indeed imposed on each of the seven pitch curves made with optimal pitch level and pitch range. In other words, the choice of intonation pattern could very well provide information conflicting with the information provided by pitch level and pitch range. The number of correct identifications is reported in bold on the diagonal of Table 6. Obviously, all emotions were not equally well identified; 40.7% of the utterances with pitch level and pitch range optimal for boredom were identified as such (for the number of responses corresponding with the percentages mentioned here, see the diagonal of Table 6), 36.9% for neutrality, 24.4% for indignation, 18.6% for fear, 17.6% for joy, 8.9% for anger and 7.8% for sadness.

In Table 11, the number of correct identifications is reported separately for each cluster of intonation patterns (see Tables 7, 8, and 9). The data presented in this table deviated significantly from a loglinear model in which perceived emotion and intonation pattern are independent ($\chi^2_{24} = 95.16$, $p < .0001$). Data points which deviate significantly from such a model ($\chi^2_1 = 3.84$, $p < .005$) are marked with ↑ if the obtained value is higher than expected, and with ↓ if the obtained value is lower than expected ($\chi^2_1 = 3.84$, $p < .005$). In order to compare the results with Table 10, significance levels of $.05 < p < .01$ are also presented by means of the symbols ∧ and ∨. The arrows and the symbols make over-representation and under-representation of responses easily visible in the tables. For

Table 12: Percentage correct identification over all patterns and only over the '1&A 1&A' pattern

	Emotion						
	neutrality	joy	boredom	anger	sadness	fear	indignation
Over all patterns (%)	36.9	17.6	40.7	8.9	7.8	18.6	24.4
Over 1&A 1&A (%)	45.8	10.4	47.9	10.4	0.0	16.7	18.7

instance, the identification of neutrality was particularly high with the use of intonation patterns from the '1...1&A' cluster, and anger was best identified with intonation patterns of the '1...A' cluster.

Table 12 presents the percentage of correct identification obtained by using either all the intonation patterns (corresponding to the total number of correct identifications in the last row of Table 11) or only the '1&A 1&A' intonation pattern (corresponding to a subset of the number of correct identifications for the '1...1&A' cluster in the first row of Table 11). As mentioned above, using the '1&A 1&A' intonation pattern leads to an over-representation of neutrality as a response. The increased percentage of correct identification for pattern '1&A 1&A' is consistent with this over-representation of the response neutrality, as well as the decreased percentage for the emotions joy, sadness, fear, and indignation. For boredom and anger, there is also an increase in percentage of correct identification associated with the use of the '1&A 1&A' intonation pattern. This indicates that the '1&A 1&A' pattern is well suited for conveying these two emotions, even if another intonation pattern is even better suited.

Considering confusion matrices where the data are pooled over the clusters (see Table 11) confirms that, for some clusters, the identification of specific emotions is better than for others; different clusters of intonation patterns infer different confusions. Confusions with neutrality are greater in number with the use of the intonation patterns of the cluster '1...1&A', and are less with the use of '1&A 12' and '1 5&A'. Most confusions with joy also occurred with the use of intonation patterns of the cluster '1...1&A'. The use of '1 5&A' and '1...A' resulted in the smallest number of confusions with fear. Most confusions with anger occurred when intonation patterns of the cluster '1...A' were used. Most confusions with sadness occurred when patterns of the cluster '1...3C' were used. Most confusions with indignation occurred when patterns of the cluster '1...3C' or

intonation pattern '1&A 12' were used. Confusions with boredom occurred mostly with the two clusters of intonation patterns '1...1&A' and '1...3C'.

e. Discussion

The choice of intonation pattern clearly appears to be relevant for the perception of emotion in speech. Some patterns appear to be more suited to convey certain emotions than others. The hypothesis that the '1&A' pattern of pitch movements is adequate for conveying all emotions studied was corroborated, although this pattern is not especially the best choice for each of the emotions. The hypothesis that some patterns can convey emotion is also supported by the data. Patterns involving a fall 'C' in the final position appear to be perceived as a signal that the utterance was not expressing neutrality. The bias introduced by the use of a final '2' is even stronger. The data showed that specific patterns introduce a perceptual bias towards the perception of particular emotions.

Furthermore, intonation patterns could be clustered. The clusters of intonation patterns appeared to be perceptually relevant to the expression of emotion in speech. The data show that these clusters have a communicative function; these clusters of functionally similar intonation patterns can be considered as melodic families of intonation patterns. This notion is closely related to, but not necessary identical with, what 't Hart et al. call '*basic* intonation patterns'. How the melodic families, as manifested in terms of clusters of intonation patterns, relate to *basic* intonation patterns is an empirical question which will be briefly discussed. This point seems particularly relevant since 't Hart et al. distinguish six *basic* intonation patterns occurring in Dutch: /1A/ also called 'hat pattern', /1E/, /4A/ also called 'valley pattern', /3C/ also called 'cap pattern', /1/, and /2/, and that the suggestion was made ('t Hart and Collier, 1975; 't Hart et al., 1990) that *basic* intonation patterns can carry different attitudinal and/or emotional connotations. Tentatively, the clusters '1...1&A', '1...3C', and '1&A 12' correspond to the *basic* intonation patterns /1A/, /3C/, and /2/, respectively. The '1...A' cluster and the '1 5&A' cluster also correspond best to the /1A/ *basic* intonation pattern. The /1A/ *basic* intonation pattern is possibly too large a category and may eventually have to be split into sub-categories. However, the correspondence, even if partial, seems to corroborate quite well the suggestion mentioned above.

IV. GENERAL DISCUSSION

The first important conclusion of this study is that the intonation pattern realized on an utterance is a very important determinant of the emotion conveyed in an utterance. Besides speaking rate, pitch level, pitch range, and other features of speech that were not investigated in this study, utterances, as produced by actors expressing an emotion, can involve a wide variety of intonation patterns. Although, in the production study, no clear-cut, one-to-one relationship between intended emotion and intonation pattern was found, some clear relationships could be distinguished. The pattern of pitch movements '1&A' occurred most often, both in the initial and in the final position of the sentences used in this study, and was produced in all emotions. It has, therefore, been concluded that, in situations in which one does not want to introduce variability by using different intonation patterns, the '1&A 1&A' pattern is the most suited intonation pattern for controlling this variation. This was the rationale for using this intonation pattern in the previous chapter on the role of pitch in conveying emotion in speech. For the same reason it is used in a follow-up study on the role of duration in conveying emotions in speech (Chapter V).

This does not mean that the '1&A' pattern of pitch movements is the most appropriate one for the expression of all the emotions; another pattern could be even more effective at signaling a particular emotion. Furthermore, some intonation patterns, in particular those ending in 'C', were not used in neutral utterances in the database. These patterns could serve as a signal of emotionality. One speaker frequently used the final 'C' in the expression of indignation and of fear. All utterances ending with '2', except one, expressed indignation.

The perception experiment confirmed that the intonation pattern is a relevant cue in signaling an emotion. Furthermore, the cluster analysis showed that it was predominantly the last part of the intonation pattern which affected the listener's response. The suggestions, based on the production study, that the '12' pattern of pitch movements was associated with indignation, was strongly confirmed. Another suggestion that the '12' and the '3C' patterns of pitch movements are negatively associated with neutral speech was also confirmed. The suggestion that '1&A' would lead to a reasonable identification of all emotions was also confirmed. Therefore, if one requires different intonation patterns not to contribute to experimental variability, a sequence of '1&A' patterns of pitch movements is found to be best suited for controlling this variability. It can be

concluded that the production and the perception study support the same conclusions. Moreover, the perception study revealed many more interdependencies between intonation pattern and conveyed emotion. The reason might be that the number of speakers who participated was restricted to three, and these three showed clear differences in the frequency with which certain intonation patterns were used. If more speakers could participate, more combinations of conveyed emotion and intonation pattern may have resulted.

A number of reasons can be given to account for the low proportion of correct identifications found in this study. First, it has been shown that speaking rate (e.g., Kitahara and Tohkura, 1992; Cahn, 1990; Scherer, 1989) and voice quality (e.g., Laukkanen, Vilkman, Alku, and Oskanen, 1997; Cummings and Clements, 1995; Scherer, Ladd, and Silverman, 1984) are important determinants of the emotion conveyed. These two factors were kept constant. Second, the two factors varied in this study were the optimal pitch range and pitch level of an emotion, and the intonation pattern. These were varied independently, as all possible combinations of pitch-level and pitch-range were presented with every intonation pattern studied. So, for very many stimuli, optimal pitch level and pitch range were combined with an intonation pattern that did not correspond with the emotion associated with this pitch level and pitch range. It is unclear what the strategy of the subjects will have been in such a situation; clearly, stimuli were carrying conflicting information. The data presented may indicate that, in some cases, the pitch-level pitch-range combination was the stronger factor, while in other cases the intonation pattern may have been stronger. The third factor which may have kept the proportion of correct identifications low, is the fact that subjects could only listen once to the stimuli and were not given any feedback as to the correctness of their response. This was necessary to prevent any learning effect and to guarantee that the subjects' responses were determined by their first impression of the emotional content of the stimulus. They should not be given the opportunity to listen to details which might provide them with a cue on the basis of which they may develop a strategy. The subjects' choices should be based on their own mental image of emotional speech. In spite of these various reasons for a low proportion of correct identifications, the results show a very significant effect of intonation pattern, and very many large effects were obtained.

A second important conclusion concerns the description of Dutch intonation as presented in 't Hart, Collier and Cohen (1990). The results show that this description was adequate in describing the vast majority of pitch curves produced by the speakers acting out the

emotions: 'neutrality', 'joy', 'boredom', 'anger', 'sadness', 'fear', and 'indignation'. The only deviant pitch movements indicated by the listeners were a prominence-lending rise and a prominence-lending fall. The rise was too late to be a '1', but certainly too early to be a '3'. In all descriptions of Dutch intonation, two prominence-lending rises have, at most, been distinguished. In auto-segmental phonology, the early rise is represented as L+H*, while the late prominence-lending rise is represented by L*+H (Pierrehumbert, 1980). In the speech material studied here, these deviant rises were most often used by one speaker (MR) in the expression of boredom. In the absence of more information, it seems best to characterize these rises as early rises, which have been delayed somewhat by the speaker for expressive means.

For the moment, it seems best to draw similar conclusions for the deviant fall, which the labelers indicated to be too early to be a proper prominence-lending 'A'. First, its occurrence was very often coupled with a deviant rise preceding this fall. Furthermore, there is no clear agreement in the literature concerning the existence of either one single fall or two different ones. Swerts (1994) shows that the position in the syllable of an 'A' bears a relationship with the 'finality' of the syllable on which the fall is positioned. In late positions, the 'A' is positioned earlier. This strengthens the argument for gradual differences among falls. On the other hand, Pierrehumbert (1980) describes two prominence-lending falls for English, indicated by H*+L and H+L*, while Gussenhoven and Rietveld (1992/1993) give H*L and !H*L for Dutch, in which the ! indicates the presence of a down-step (Pierrehumbert, 1980). Moreover, Hermes (1997) showed that Dutch listeners were able to distinguish a 'low' accent lent by an early fall, from a 'high' accent lent by a late fall. More data need to be gathered, about these observed falls. For now, we will assume only one basic prominence-lending full fall for Dutch, which the speaker may time earlier in the syllable to signal finality or for some other expressive means.

Hence, the ten basic pitch movements described by 't Hart et al. (1990) for Dutch seem to be sufficient to describe the pitch curves of the emotional utterances used in this study. In addition, the distinctive features of intonation presented in the Dutch grammar appear to be satisfactory. These features consist of the direction of the pitch movement, its timing with regard to syllable boundary, its spreading over one or several syllables, its rate of change, and its size. As to the details of the realizations, 't Hart, Collier and Cohen are not very detailed, probably because, in real speech, realizations of pitch movements can vary significantly and are very often underspecified. In 't Hart and Collier (1975) the first

specifications are presented. Collier (1991) presents more details for Dutch and three other West-European languages about the so-called 'hat pattern', a succession of a rise and a fall most often used in neutral, read speech. As to the timing of the pitch movements, 't Hart and Collier (1975, p. 241) mention in a footnote that "the position of the pitch movement in the syllable is defined with respect to the vowel onset moment". For the non-prominence-lending rise '2', they describe its timing as "as late as possible in the syllable", which indicates that this refers to the end of voicing in the syllable.

In synthesizing the pitch contours for this study, it was not only the timing of the pitch movements that appeared to be important; the duration of the pitch movements also had to be selected with care. It is emphasized, here, that in the process of synthesizing these pitch contours, the presence of a rapid pitch movement within the syllable appeared to be an extremely strong cue for accentuation. The presence of a rapid pitch movement in a non-accented syllable had to be avoided as far as possible. One specific problem was the realization of the non-prominence-lending fall 'C', which is normally preceded by a late prominence-lending rise '3'. 't Hart and Collier (1975, p. 241) present the following description: "In some contours, during the last 20 to 50 ms of phonation in the utterance, F_0 goes down rapidly to an immaterial value. This movement, although probably a mere relaxation phenomenon, is perceptually relevant: its omission in the stylized contour is readily noticed. Its position in the final syllable should be very late to avoid undesired prominence of the syllable." This description is very precise. The problem, in Sentence 1 of this study, was that the presence of just about any rapid pitch movement on the last lexically unstressed syllable of the last word 'gekocht' /ɣə kɔxt/ induced an unnatural accent on this syllable, which had to be overcome. Therefore, this pitch movement had to remain virtual in this sentence.

In one aspect, the synthesized '3C' pattern of pitch movements deviated from specifications in the literature. In 't Hart and Collier (1975, p. 241) it is mentioned that the excursion size of the '3' is often larger than for the '1'. This may indeed be so, however, in this study we decided to keep the excursion size constant, so as to guarantee that possible differences between utterances containing an early prominence-lending rise '1' and a late prominence-lending rise '3', could be unambiguously attributed to the timing differences between the pitch movements and not possibly be due to differences in excursion size. It is likely that the difference in excursion size between a '1' and a '3' may enhance the differences already found in this study between the final '1&A' and the final '3C' pattern of pitch movements.

In view of these new results, the following summary can now be given. Although duration was kept constant in the present study, it is known to be relevant for the expression of emotion in speech. Therefore, results obtained from Chapter II about overall duration, will be included in the results obtained here. It has been shown that neutral speech can be characterized by a low pitch level, a small pitch range, a moderate speaking rate, and the final pattern of pitch movements '1&A'; the intonation patterns should not end in '3C' or '12'. As far as pitch level and pitch range is concerned, boredom is closest to neutral speech. It has a slower speaking rate, and is optimally realized with '3C' as the final pattern, and not with '5&A' or '12'. Sadness has a medium pitch level, a moderate pitch range, a slower speaking rate than neutrality, and preferably '3C' as a final pattern. Like boredom, it should not end in '5&A'. Fear has a high pitch level, a medium pitch range, a quicker speaking rate than neutrality, and ends in '12', but not in a single 'A' or 'EA'. Fear was very badly identified in these experiments, probably due to the fact that voice quality is an important factor in expressing this emotion, and this factor is not varied in this study. The last three emotions: joy, anger and indignation, have large pitch ranges and, except anger where it is somewhat lower, high pitch levels. The speaking rate is quicker than in neutrality for joy and anger, and slower for indignation. While joy preferably ends in a '1&A' or a '5A' pattern, but not a single 'A', an 'EA' or a '12', anger ends in '5&A' or a single 'A' or 'EA', and not in '1&A' or '3C'. Finally, indignation is very strongly associated with the '12' pattern of pitch movements, and not with the '1&A' pattern.

It has already been mentioned that voice quality has not been considered in these studies, but may also play an important role in conveying emotion in speech. The relative durations of different phonetic segments may also play a role in this respect. The temporal variations at utterance level (i.e., at the global level of the utterance as a whole) and below utterance level (i.e., within utterances, considering local fluctuations) will be investigated in the next chapter.

Chapter V

Temporal variations

ABSTRACT

The present chapter is concerned with temporal features in the production and perception of emotion in speech. First, a production study involving the whole database was conducted, in order to extract regularities and systematic differences in speech rate from production data. The results are compared to those of Chapter II in which, on the basis of perceptual experiments, optimal values were derived for conveying the same seven emotions. The perception and the production data appeared to correspond quite closely. The description in global terms, such as speech rate, may obscure more detailed characteristics in the analysis of the speech produced. Because these detailed characteristics might be distinctive for the vocal expression of emotion, a more refined analysis was carried out. The detailed analysis was concerned with relative duration of accented and unaccented speech segments of the emotional utterances. The differences found in the relative duration of accented and unaccented speech segments may be due to the expression of emotion itself, but could also simply be due to the differences in speech rate that are observed in the expression of emotion in speech. A reference was needed. Therefore, neutral utterances spoken by one of the male speakers at speech rates varying over the range of speech rates used in the emotional speech observed, were analyzed in the same way as for the emotional speech. Finally, a perception experiment was conducted to investigate whether the differences in relative duration of accented and unaccented speech segments between the emotional speech and the neutral speech, are relevant for conveying emotion in speech. The differences in relative duration of accented and unaccented speech segments that are associated with speech rate appeared, not to be perceptually relevant. On the other hand, the differences in relative duration of accented and unaccented speech segments that are associated with the expression of emotion, appeared to be perceptually very relevant for the expression of neutrality and indignation. Relationships with the production data are discussed.

I. INTRODUCTION

As mentioned in Chapter II, the importance of F_0 and duration phenomena for the expression of emotion in speech, is considered to be paramount. Therefore, the present study, combining a perception oriented and a production oriented approach in the investigation of the prosodic variations due to emotion, focuses on these variations as they can be quantified in acoustic parameters corresponding with pitch and tempo. In Chapter II, optimal values for pitch level, pitch range and overall speech rate were estimated experimentally for the seven emotions studied. Chapters III and IV were concerned with the role of pitch and F_0 in the expression of emotion in speech. The present chapter concentrates exclusively on the role of temporal variations. In order to understand how temporal features contribute to the expression of emotion in speech, temporal variations across emotions are described, as well as the variability in temporal realizations between three speakers.

The notion of speech rate provides a measure of how rapidly the speech was produced. As speakers increase or decrease their speech rate, they modify their rate of articulation (which can be measured, for instance, in syllables per second), as well as the number and duration of pauses they produce within utterances. In the present study, the issue of the number of pauses and their duration will not be taken into account separately. The notion of speech rate is defined in the most simple way, as inversely proportional to the overall utterance duration. This simplified definition remains quite acceptable for our purposes in the present study, because the relatively short sentences in the speech material did not induce pauses within utterances. If, however, incidental pauses occur, the total utterance duration is subject to that influence.

In the first part of the present chapter, a production study was conducted. First, it was carried out at utterance level, investigating overall speech rate in the expression of each emotion studied. Additionally, the values for overall speech rates observed in this analysis were compared with the values that were found to be optimal for the overall speech rate in Chapter II. These optimal values were based on perceptual tests involving a subset of the database used in the present study. These perceptual data and the data available from the perception test in the present chapter allow a proper comparison to be made between production and perception. It is possible, however, that speech rate does not vary equally in different parts of the utterance. Since overall speech rate may be too rough a measure to

capture the details of the speech production that might be distinctive for the expression of various emotions, a more refined analysis took place concerning the relative duration of accented and unaccented parts of speech. In order to determine whether these variations in relative duration of accented and unaccented speech segments were associated with the expression of emotion in speech, or with the differences in overall speech rate associated with the expression of emotion in speech, results were compared with those of an analysis of neutral speech produced with speech rates covering the range observed for the various emotions.

Then, the perceptual relevance of differences observed between emotional and neutral speech was tested in a perception experiment. If these variations appear not to be perceptually relevant for conveying emotion in speech, a description in terms of overall speech rate can be considered to be sufficient for the expression of emotion in speech. In addition, the average speech rate measured over the complete utterance can be considered adequate for a description of the temporal characteristics of emotional speech. If, on the other hand, variations within the utterances appear to be perceptually relevant for the identification of emotions in speech, it can be concluded that simply stretching the speech signal linearly does not constitute a satisfying approach for generating emotion in speech. Indeed, the temporal details below utterance level cannot be described in such a linear model and these details would have to be taken into account for a proper description of temporal features conveying emotion in speech.

The ultimate aim is to understand which detailed characteristics of the temporal structure of emotional utterances are transmitting the emotion in speech, to quantify these characteristics, and to consider to what extent the modeling of these details into re-synthesized speech leads to improved identification of the emotions.

II. ANALYSIS OF OVERALL SPEECH RATE IN EMOTIONAL SPEECH

a. Speech material

The speech material consisted of 315 utterances of one female and two male Dutch speakers who produced three times five sentences of semantically neutral content while expressing the emotions: neutrality (as a reference), joy, boredom, anger, sadness, fear, and indignation. These same utterances served as speech material in Chapters III and IV, and a

subset of this material, i.e., 14 utterances of the male speaker MR, was already used in Chapter II.

b. Procedure

The overall sentence duration was measured in the speech material (3 speakers × 7 emotions × 5 sentences × 3 trials). For each speaker individually, the mean overall sentence duration, per emotion, and its standard deviation were determined. The sentence duration relative to neutrality was calculated for each emotion, by dividing the mean overall duration for a specific emotion by the sentence mean for neutrality. This measure was transformed into speech rate relative to neutrality.

c. Results

The mean overall sentence duration per emotion is presented for each speaker in Table 1, together with its standard deviation. The speech rate relative to neutrality is reported in Table 2 for each speaker, and the sentence duration relative to neutrality is given in parentheses. In order to ease comparisons, the values that were experimentally found to be perceptually optimal in Chapter II are also presented in the right most column of Table 2. In the experiments described in Chapter II, utterances were re-synthesized with varying pitch level, pitch range, and speech rate, and listeners selected and ranked variants in which they found a given emotion best expressed. Subsequent experiments in Chapter II tested the

Table 1: Mean overall sentence duration per emotion in seconds and its standard deviation

Emotion	Speaker MR		Speaker RS		Speaker LO		*average*	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
neutrality	1.49	0.21	1.70	0.20	1.74	0.24	1.64	0.22
joy	1.50	0.18	1.69	0.27	1.68	0.18	1.62	0.21
boredom	2.05	0.30	1.95	0.32	2.04	0.29	2.01	0.30
anger	1.28	0.21	2.01	0.70	2.11	0.33	1.80	0.41
sadness	1.59	0.23	1.96	0.55	1.80	0.31	1.78	0.36
fear	1.30	0.19	1.36	0.26	1.85	0.31	1.50	0.25
indignation	1.73	0.37	1.69	0.30	2.54	0.33	1.99	0.33
mean	1.56	0.24	1.77	0.37	1.97	0.28	1.76	0.30

Table 2: Overall speech rate per speaker and, in parentheses, sentence duration relative to neutrality

Emotion	Speaker MR	Speaker RS	Speaker LO	*mean*	Optimal values
neutrality	1.00 (100%)	1.00 (100%)	1.00 (100%)	1.00 (100%)	1.00 (100%)
joy	0.99 (101%)	1.00 (100%)	1.03 (97%)	1.01 (99%)	1.20 (83%)
boredom	0.73 (137%)	0.87 (115%)	0.85 (117%)	0.82 (122%)	0.67 (150%)
anger	1.16 (86%)	0.84 (119%)	0.82 (122%)	0.94 (106%)	1.27 (79%)
sadness	0.93 (107%)	0.87 (115%)	0.96 (104%)	0.92 (109%)	0.78 (129%)
fear	1.15 (87%)	1.25 (80%)	0.93 (107%)	1.11 (90%)	1.12 (89%)
indignation	0.86 (116%)	1.00 (100%)	0.68 (146%)	0.85 (118%)	0.85 (117%)

identification of emotion in re-synthesized and synthetic speech, generated using these parameter values.

Considering the mean overall sentence duration for neutral speech, as well as the mean overall sentence duration in all the utterances, MR speaks most quickly and LO most slowly. Individual differences between speakers will be discussed in the next section.

d. Discussion

The results of this analysis provide information on the way each of the three speakers exploits overall speech rate for conveying emotion in the recorded speech. The results are compared among speakers, and are also related to the optimal values for overall speech rate obtained in Chapter II. Furthermore, a global comparison can be done between these values found optimal in Chapter II, and results of the related studies that were discussed in Chapter III. In order to ease comparisons, the results from these related studies, as well as the optimal values found in Chapter II, are presented in Table 3 which shows notions comparable with overall speech rate.

A clear resemblance can be observed with findings of van Bezooijen (1984). The relative durations in the present study and in hers are exactly the same for joy, and quite similar for anger and fear. The only difference concerns the duration of sadness, which is much longer in the present study. A noticeable difference with the results of Cahn (1990) is also the shortest duration Cahn reports for sadness, while we found that this emotion requires a

Table 3: Speech rates relative to neutrality found optimal in present and related studies

	Chapter II relative duration	van Bezooijen relative tempo	Cahn* speech rate	Carlson et al. relative duration	Kitahara et al. relative tempo	Scherer speech rate
neutrality	100%	100%	-	100%	-	-
joy/gladness happiness	83%	83%	2	130%	100%	increase
boredom	150%	-	-	-	-	-
anger/rage	79%	86%	8	135%	70%	increase
sadness	129%	80%	-10	151%	115%	decrease
fear	89%	101%	10	-	-	increase
indignation	117%	-	-	-	-	-

* Cahn used a scale of values from -10 to +10.

sentence duration longer than neutrality. The main difference between our findings and those of Carlson, Granström and Nord (1992) is that, in their data, the neutral utterances were the most rapid ones. Taking this fact into consideration, comparisons in duration relative to neutrality are rather difficult, but joy and anger are uttered using comparable speech rates, and sadness is expressed with a lower speech rate than these two emotions, which corresponds fairly well with our findings. Kitahara and Tohkura (1992) propose relative tempo values, corresponding relatively well with what we found. A discrepancy with the predictions of Scherer (1989) is the decreased speech rate he predicts for anger, while all other results are in agreement.

Our three speakers agree quite well on the relative speech rate for the expression of joy (see Table 2). Joy and neutrality lie, from a temporal point of view, very near each other. For the expression of joy, the optimal values found through perceptual experiments are higher than those produced by the speakers. The optimal values are in agreement with findings by van Bezooijen (1984), and the speakers' data are in agreement with findings by Carlson, Granström and Nord (1992), Kitahara and Tohkura (1992), and Scherer (1989). There is also agreement amongst the speakers on a reduced speech rate in the expression of sadness (see Table 2). The optimal values even show a clearer reduction of speech rate. It might be interesting to note here that the rate experimentally found to be perceptually optimal for sadness results in a longer overall duration than in most other comparable studies; only

Carlson et al. (1992) propose an ever slower speech for this emotion. In boredom, the speakers and the optimal values indicate a lower speech rate. MR slows his speech more than the two other speakers, and the optimal values show an even more reduced speech rate.

For other emotions, speakers do not seem to agree with each other. For the expression of anger, MR speaks much quicker than the two other speakers (see Table 2); relative to neutrality, he realizes time reductions where other speakers choose for time expansion. Hereby, he agrees with most related studies, while other speakers are in agreement with findings by Carlson et al. (1992). For the expression of fear and indignation, LO realizes durations quite different from the two other speakers. In both cases she speaks more slowly than the two male speakers, and more slowly than was found optimal in the previous study. She is, thereby, in agreement with findings by Cahn (1990) and Scherer (1989). The optimal values generally indicate a slightly more extreme speech rate than the one realized by MR, whose speech was used in Chapter II.

Ordering the emotions from the emotion expressed with the highest to the one with the lowest speech rate (see Table 2) results in a different list for each speaker:

- For MR: anger and fear; neutrality and joy; sadness; indignation; and boredom.
- For RS: fear; neutrality, joy, and indignation; boredom and sadness; and anger.
- For LO: anger; joy; fear; neutrality; indignation; sadness; and boredom.

The optimal values are often more extreme than the ones observed in the emotional speech produced by the three speakers. A possible explanation for the deviation of the optimal values from the production data, is that other components of the re-synthesized and/or synthetic speech used in the perceptual tests were not powerful enough. In fact, in order to find the optimal speech rate, the only parameter that was manipulated was the overall duration, keeping the intonation and the voice source constant. Consequently speech rate was more relied upon to convey the emotion. It seems reasonable that the values found optimal are more extreme values than the ones observed in the production data: in perception experiments, only one component of speech is used to convey the emotion instead of more components and their combined effects in natural speech. This does not mean that using these optimal values in synthesis would result in generating caricatural emotional speech, as the values tested in perception experiments were varied in a reasonable range around the values observed in production data, i.e., a range of values resulting in utterances that did not sound unnatural to the experimenter. As an alternative, although less

probable explanation for the deviation of the optimal values from the production data, it could be speculated that speakers and listeners do not need to follow the same rules; production and perception, as related as they are, are two different processes. However, these deviations are not that substantial.

Relating the results with those of Chapter III, it seems noteworthy that, for instance, in the expression of boredom, the speaker MR realizes pitch curves that are less distinctive than those of the other speakers. Indeed, monotony is a characteristic of boredom, and MR's F_0 curves are less monotonous than those produced by the two other speakers in this emotion. In contrast, he uses the most distinctive speech rate. This may illustrate the fact that a speaker may rely more on speech rate than on pitch variations in order to convey a particular emotion. This corroborates the idea of mutual compensation of components of speech: if one component carries the emotion, the other components can carry less weight.

The standard deviations reported in Table 1 might provide information concerning the consistency of the speaker over 15 utterances (5 sentences × 3 trials), in the usage of a specific speech rate for a particular emotion. Of course, a relatively small standard deviation is expected for a relatively low mean, but it also seems reasonable to suppose that a high relevance of the temporal component for the expression of a particular emotion will result in smaller variations in overall duration, and thus in a smaller standard deviation. On average, joy shows the least temporal variation for all the speakers. The variations in speech rate are moderate in neutrality, and, to a lesser extent, in fear. More variation in speech rate can be observed in boredom, and the most variations in speech rate occur in indignation, sadness, and anger.

- MR' s speech shows the highest variability in speech rate for the expression of indignation and boredom: the two emotions in which he produces the slowest speech. Referring to the analysis of pitch level and pitch range (Chapter III), these two emotions are the most extreme and represent opposites. This suggests that for the expression of indignation and boredom, this speaker makes such a use of intonational factors that the temporal aspect is not the one of primary significance for him: even when compared to other speakers he makes relatively more use of speech rate and may be less use of pitch variations.
- RS allows large speech rate variations in the expression of sadness and anger. For him, speech rate might be an element of lesser relevance to the expression of sadness and anger. He might make more intensive use of intonation and/or voice quality in order to convey these emotions. Sadness, anger, but also boredom, were produced with relatively slow

speech by this speaker, but the standard deviation for boredom is clearly the smallest of the three; this suggests that duration is more relevant to boredom than it is to sadness or anger.

- The deviations in speech rate presented by LO in different emotions are not very different from one another. It suggests that, for her, variations in speech rate are relevant to the expression of all emotions, without really being more relevant to the expression of specific emotions. She might prefer to use a combination of factors. She indeed makes use of speech rate in the utterance of emotionality, while simultaneously relying on other components of speech.

e. Conclusions

The values experimentally found to be optimal, appear, in many cases, to be more extreme than the regularities found in the production data, i.e., where speakers reduce their speech rate, the values found optimal correspond to an even greater reduction. This does not come as a surprise, as these optimal values were obtained in a set-up where only duration was varied.

Speakers seem to rely more on speech rate for the expression of some emotions than for the expression of others. For some emotions, there is a good agreement about speech rate among the three speakers; boredom and sadness are expressed with a speech rate lower than in neutrality, joy with a speech rate very similar to neutrality.

In the expression of a particular emotion, speakers might have a personal preference for the use of either intonational elements, temporal factors, vocal effects, and/or a combination of the previous components of speech. The disagreements between the speakers probably reflects the diversity of strategies that can be successful when expressing a single emotion in speech.

III. ANALYSIS OF RELATIVE DURATION OF ACCENTED AND UNACCENTED SPEECH SEGMENTS

a. Problem statement and aim

Although it has been shown in the perception study (Chapter II) that a linear manipulation of the utterances can influence the perception of emotion, it is well worth considering whether a more subtle approach than simply stretching or shrinking the utterances linearly would be preferable. Moreover, if the neutral speech rate is relatively steady, as suggested

by the small standard deviation of the mean duration, it could mean that a deviating speech rate can signal the presence of an emotion expressed by the speaker. Klatt (1976) states that so-called extralinguistic factors, such as speaker mood and physical conditions, influence the overall speech rate. On the other hand, a particular emotion might also determine how the difference in speech rate is realized within the utterance. This would be in agreement with Lehiste (1970, p. 51-52) who wrote that, "Changes of the relative durations of linguistic units within a sentence [...] convey something about the mood of the speaker or about the circumstances under which the utterance was made."

In order to allow more precision in the manipulation of speech rate, temporal variations occurring within utterances are investigated in this section. A study below utterance level can involve temporal variations at different levels; different types of speech segments are potential candidates to see whether a non-linear time distribution occurs in emotional speech, and such temporal variations have been investigated in more general speech studies. The lengthening of the last syllable in an utterance is a well known temporal phenomenon (e.g., Oller, 1973; Klatt, 1975; Beckman, 1990). One option, for instance, could be to describe the way the phrase-final lengthening behaves in emotional speech. Since, when speakers slow down their speech, a good portion of the extra speaking time is spent on pauses (Goldman-Eisler, 1968), a second option could be to consider the way a decreased speech rate induces more silent parts in the utterances. By contrast, in the present study, temporal variations occurring in the speech itself are investigated, as no pauses were realized in our speech material. Furthermore, as increases in speech rate are accompanied by phonological and phonetic simplifications, as well as shortening of vowel and consonants (Klatt, 1976), another option could be to observe the temporal variations affecting the different phonemes. Finally, the simplest option would be to consider the relative length of accented and non-accented speech segments. This last option allows an analysis below utterance level, without involving too many details. Furthermore, it has been shown that listeners have a well-defined internal representation of the appropriate duration of the vowel of an accented syllable (Nooteboom, 1973). In this last option, it is possible to exclude the effect of final lengthening, simply by not including the last syllable of the utterances in the measurements. The other factors listed above are expected to be kept constant. Indeed, the speakers were instructed to realize similar reductions or simplifications in all utterances on the same text, and not to realize any pauses in the utterances, which was possible because the speech material consisted of short sentences. Adopting this last option for the present study, relative durations of accented and unaccented speech segments were examined. An accented speech segment is composed of

one lexically stressed syllable on which a pitch accent was realized, whereas an unaccented speech segment is composed of one or several syllables in succession, on which no pitch accent was realized.

In order to distinguish the effect of emotion from the effect of speech rate variation, in emotional utterances, it was necessary to have a reference concerning the relative durations of accented and unaccented syllables in neutral utterances. A satisfactory, reliable reference could not be found in the literature. For a reduced speech rate, it has been reported that the extra speaking time is distributed unevenly on accented and unaccented syllables (Peterson and Lehiste, 1960; Kozhevnikov and Chistovich, 1965; Miller, 1981). For an increased speech rate, den Os (1988, p. 50) finds that, in Dutch (a stress-timed language), the unaccented syllables tend to be relatively more shortened than the accented ones. Peterson and Lehiste (1960) reached the same conclusion for English. Lehiste (1970), on the other hand, specifies that in some languages, an increased speech rate is realized by shortening unaccented syllables, while in other languages, the decrease in duration is equally distributed over the whole utterance. Moreover, there were, to our knowledge, no studies available about the distribution of accented and unaccented syllables for a range of progressively increasing/decreasing values of speech rate. Therefore, in order to obtain an initial reference, an analysis of neutral speech at various speech rates is carried out in this study. This neutral speech covers the whole range of speech rates used in the emotional speech observed.

The purpose of this analysis is to investigate whether the temporal distribution over accented and unaccented syllable types is similar in emotional speech and in neutral speech. If this were the case, a linear model would suffice for the description of the temporal phenomena relevant for conveying emotion in speech. Otherwise, an investigation should be carried out as to whether and how the temporal distribution of both types of syllables follows emotion-specific rules.

b. Speech material

- ***Emotional speech***

In the analysis of overall speech rate, measurements were carried out on all five sentences used for the recordings of the database described previously. For the present analysis, two of these five sentences could not be used; one because the speakers did not always realize

one single accent on the same syllable, and the other one because the second accent was realized on the last syllable of the sentence; the effect of final lengthening then interferes with the effect of accentuation. The three other sentences were selected for the measurements for the present analysis. These three sentences contained two lexically stressed syllables, on each of which a pitch accent was realized. These lexically stressed syllables are underlined in the list below, and the accented and unaccented segments are separated by a vertical slash. The sentences used in the investigation are the following:

1. Zijn vrien|din | kwam met het | vlieg|tuig (His girlfriend came by plane).
2. Jan | is naar de | ka|pper ge|weest (John has been to the hairdressers).
3. Het is | bij|na | ne|gen | uur (It is almost nine o'clock).

In Sentence 1 (vliegtuig), the accented syllable /dɪn/ contains a short vowel and the accented syllable /vlix/ a long vowel. In Sentence 2 (kapper), both accented syllables /jɑn/ and /kɑ/ contain short vowels while in Sentence 3 (uur), both accented syllables /bɛi/ and /ne:/ contain long vowels.

In total, 189 candidate utterances were considered (3 speakers × 7 emotions × 3 sentences × 3 trials). Seven of these utterances were disregarded, either because the occurrence of lengthening after a prosodic boundary would make the comparison with other utterances difficult, or because the speaker was not perfectly fluent or did not produce two accents in the utterance. The seven disregarded trials are the following: two utterances by speaker MR on Sentence 1 (vliegtuig), one in anger, one in sadness; two utterances by speaker RS on Sentence 2 (kapper), one in sadness, one in indignation; and three utterances by speaker LO, one on Sentence 3 (uur) in fear, and two on Sentence 1 (vliegtuig), one in boredom, one in anger. Since these seven utterances were expressions of different emotions by different speakers, it can be assumed that disregarding these utterances will not significantly have influenced the results.

- ***Neutral speech***

In order to get an impression of the variation of relative duration of accented and unaccented syllables in neutral speech, the male actor MR participated in a new series of recordings on the three sentences 'vliegtuig', 'kapper' and 'uur'. The recordings took place in the same recording booth and under similar conditions as for the emotional speech. Speaker MR was asked to say these sentences in a neutral way at progressively increasing speech rates. He was instructed not to include any silence time in his utterances, so that speaking slowly

should mean using a lower articulation rate rather than including pauses in the utterances. The recordings resulted in a total of 171 utterances, namely 57 for each of the three sentences.

c. Procedure

In order to investigate the temporal structure of emotional speech, and to compare the results with those obtained in neutral speech covering the whole range of speech rate variations realized in emotional speech, measurements of the duration of accented and unaccented speech segments were carried out on the 182 emotional utterances and on the 171 neutral utterances. Note that the 182 'emotional' utterances included 27 utterances spoken in the expression of 'neutrality'.

The factual presence of a pitch accent, realized on each lexically stressed syllable labeled as accented, was checked in all utterances. Per utterance, the durations of the segments (in seconds) were determined. Measurements were carried out from the syllable onset of a speech segment to the syllable onset of the next segment. Depending on the sentence, a single unaccented speech segment can be composed of several successive syllables (for instance 'is naar de' in Sentence 2). The last syllable of the utterances (respectively 'tuig', 'weest', and 'uur') were disregarded, to exclude the effect of final lengthening in the present investigation.

For each utterance, the durations of the two accented segments were summed. The durations of all unaccented segments, apart from the last syllable, were also summed separately. The ratio of the duration of the accented segments and the total duration of both accented and unaccented segments, was computed. This proportion will be referred to as 'durational proportion of accented speech segments'. Considering the three sentences separately, regression lines were calculated, and fitted through the data points representing the durational proportion of accented speech segments of all the utterances of the neutral speech recorded at increasing speech rates.

d. Results

- ***Neutral speech recorded at increasing speech rates***

Per sentence, the durational proportion of accented speech segments of all neutral utterances of speaker MR, recorded at progressively increasing speech rates, are represented as points

in Figure 1. The lines represent the regression lines fitted through these data. For Sentence 2 (kapper) and Sentence 3 (uur), considering two regression lines instead of one, clearly increased the explained variation, which was significant for the sentence 'uur'; the root mean square distance between regression line and utterances reduced from .0243 for a single regression line to .0227 for two lines for the sentence 'kapper', and from .0229 to .0173 for the sentence 'uur' (F_s = 1.74, $F_{.05\ [56,56]}$ = 1.56, $p < .05$). For Sentence 1 (vliegtuig), the root mean square distance of .0174 for one line did not get much better with .0171 for two lines; a single regression line was, therefore, considered to suffice for this sentence. The functions describing the regression lines are the following, where x is the overall utterance duration in seconds and y the durational proportion of accented speech segments:

- Sentence 1 (vliegtuig) : For all x, $y = 0.4206 - (0.0388 \times x)$
- Sentence 2 (kapper) : For $x < 1.83$, $y = 0.3246 + (0.0387 \times x)$
  else $y = 0.4845 - (0.0487 \times x)$
- Sentence 3 (uur) : For $x < 1.09$, $y = 0.3012 + (0.1026 \times x)$
  else $y = 0.4906 - (0.0712 \times x)$

- ***Emotional speech***

The durational proportion of accented speech segments in the emotional speech of the three speakers is presented separately for the three sentences, Sentence 1 (vliegtuig) in Figure 2, Sentence 2 (kapper) in Figure 3, and Sentence 3 (uur) in Figure 4. In these figures, the durational proportion of accented speech segments realized by each speaker in each emotional utterance is presented, together with the regression lines corresponding to the neutral speech of speaker MR at various speech rates, as in Figure 1. A different symbol is used for utterances produced by different speakers.

In Sentence 1 (vliegtuig), the durational proportion of accented speech segments for neutrality, joy, and boredom, more or less follows the regression line, representing this durational proportion as a function of overall sentence duration, though in boredom most points are above the line. This is even stronger in the other four emotions: anger, sadness, fear, and indignation, where just about all points are above the regression line. This may indicate that an increased lengthening of the accented syllables may signal an emotion. This slight tendency to produce relatively longer accented syllables seems to correspond with lower overall speech rates than the one optimal for neutrality.

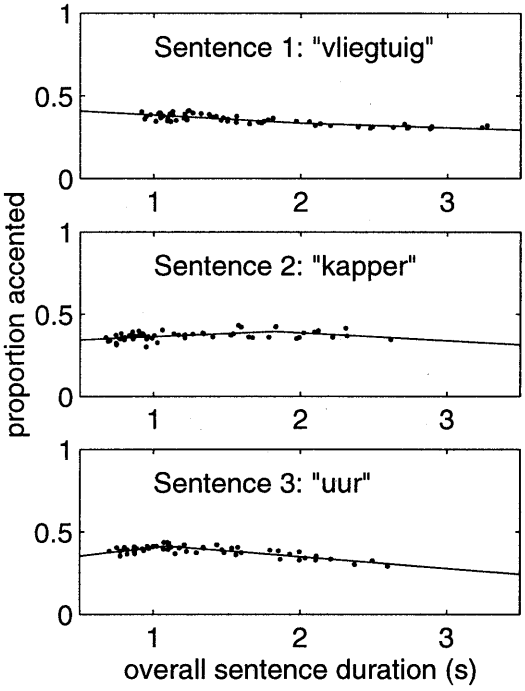

Figure 1: Proportion accented speech segments in neutral speech of speaker MR, at progressively increasing sentence duration

In Sentence 2 (kapper), all data points, except one, quite closely follow the regression lines. Note, furthermore, that both in Sentence 1 and 2, the durational proportion of accented speech segments by Speaker MR (represented by circles in the figures) for neutrality, in the two recording sessions, closely match. As previously mentioned, in the first recording session, aimed at gathering emotional speech, neutral utterances were produced: 'neutrality' being a reference category for the other emotions. In the second recording session, about 4 years later, neutral utterances were produced by Speaker MR at various speech rates (regression lines in Figures 1, 2, 3, and 4). In Sentence 3 (uur), the relative lengthening of the accented syllables produced by MR in the two recording sessions, does not match, as shown in Figure 4, where all circles representing MR are above the regression lines. It is noted, however, that this holds for all emotions and not only for neutrality. The data points for the other speakers, RS and LO, are more or less close to the regression lines, though for neutrality all, except one, are below the lines.

e. Discussion

Although the analysis of speech produced by the three speakers shows the presence of differences in the durational proportion of accented speech segments, it does not allow the formulation of hypotheses concerning the exact realization of these variations for the expression of specific emotions. If we allow the consideration of slight tendencies in the emotional speech, there seems to be a tendency to stretch accented speech segments relatively more than non-accented ones, for overall speech rates lower than the rate for neutrality. This is best shown in Figure 2 concerning the sentence 'vliegtuig'. For overall speech rates higher than the ones used in neutrality, the time distribution over accented and unaccented parts of speech is less consistent. In increased speech rates, time distribution over accented and unaccented segments might primarily depend on the speech rate, or might be influenced by the combination of emotion and speech rate. A specific stretching of accented syllables could not be associated with a particular emotion, on the basis of the analysis of production data. This could mean that no spectacular improvements, in terms of identification of an emotion, can be expected from a modeling of temporal variations below utterance level, taking into account the distribution of time over accented and unaccented segments of speech, when compared to a simple modeling of overall speech rate according to the linear approach.

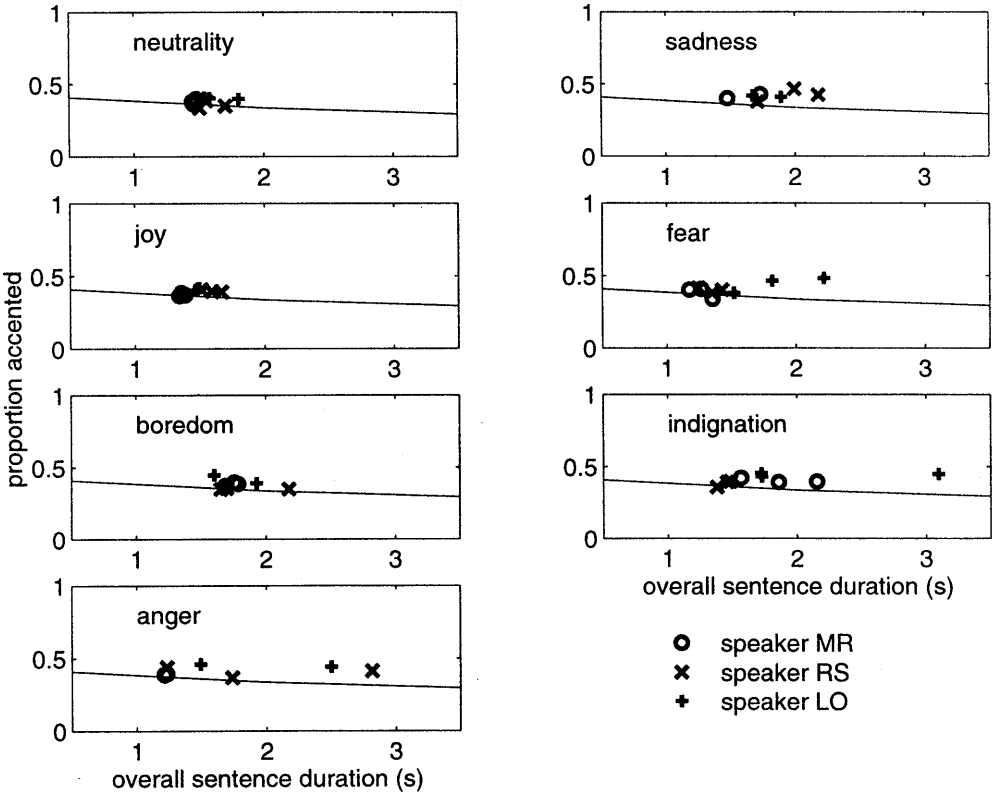

Figure 2: Proportion of accented speech segments in emotional speech of the three speakers, for Sentence 1 'vliegtuig'

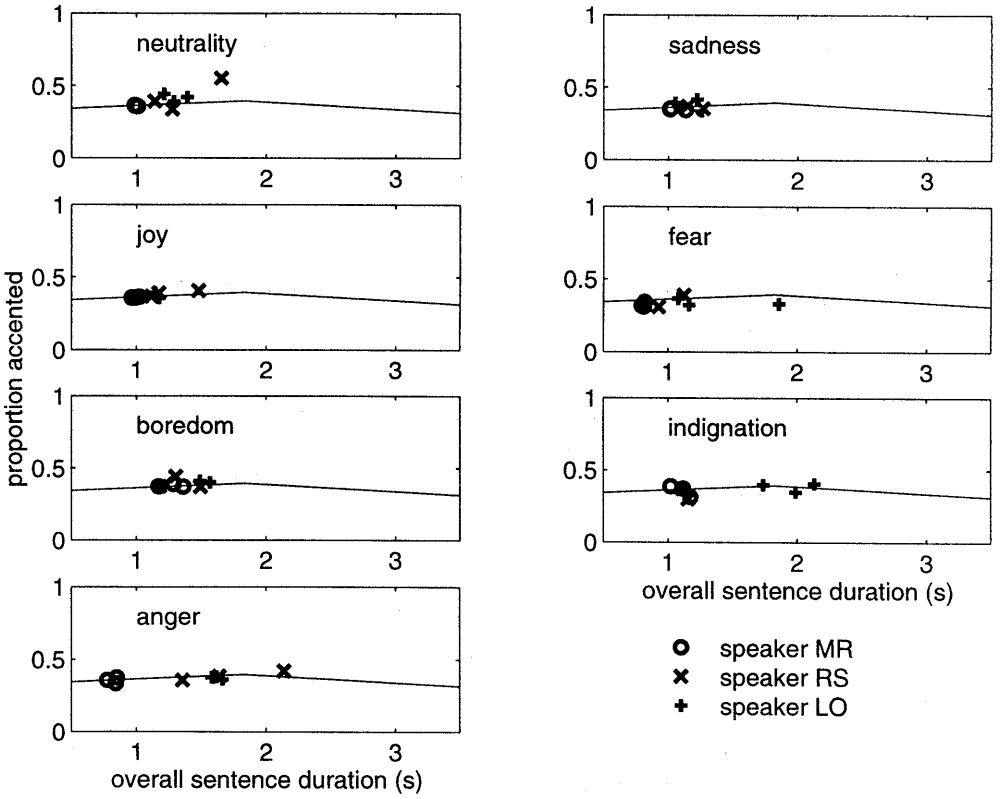

Figure 3: Proportion of accented speech segments in emotional speech of the three speakers, for Sentence 2 'kapper'

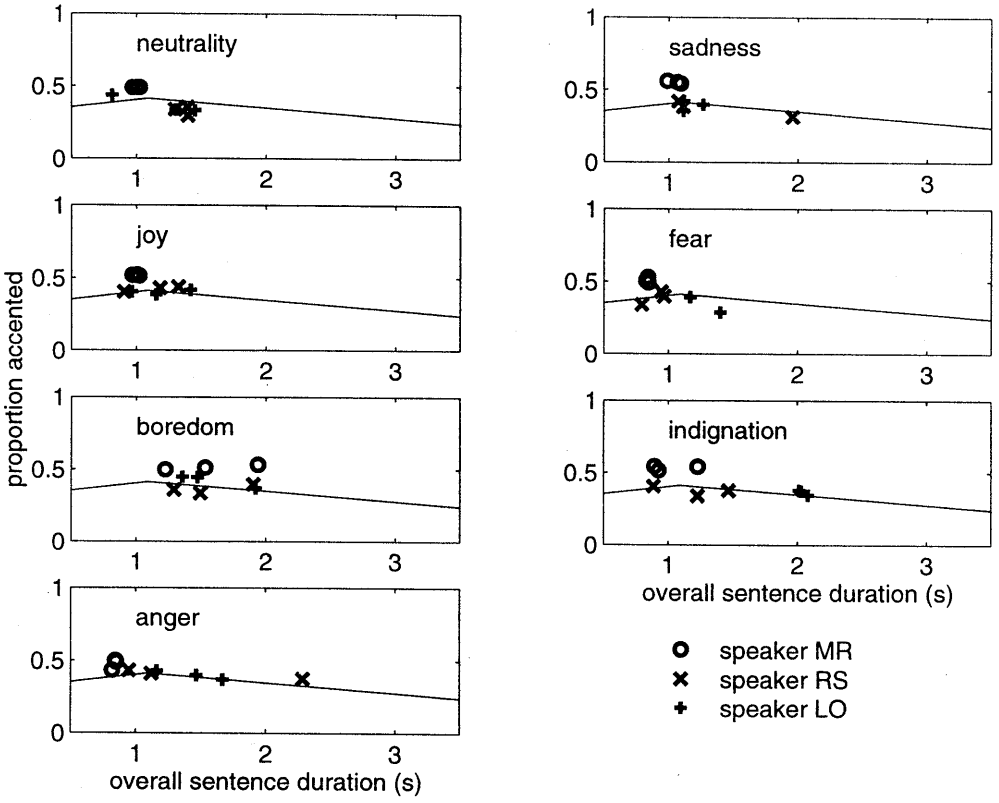

Figure 4: Proportion of accented speech segments in emotional speech of the three speakers, for Sentence 3 'uur'

However, an important limitation in the description of the relative duration of accented and unaccented segments of emotional speech is that the durational proportion of accented speech segments was computed on the basis of the neutral speech of a single speaker. It is to be expected that this obscures some of the information in the data. Indeed, large differences in speech rate are reported between speakers of the same language (den Os, p. 53). Klatt (1976) reports that there is considerable inter-speaker and intra-speaker variability in durational studies. Moreover, in order to produce the speech analyzed in this study, speakers did not only use temporal features, but a combination of different features such as pitch level, pitch range, intonation pattern, and voice quality. Therefore, the speakers have the possibility to rely more on some of the features and less on others. The fact that this production study does not show the systematic use of a specific durational proportion of accented speech segments for the expression of a specific emotion, is not necessarily a sign that the variations in durational proportion of accented speech segments that are observed in the data are not perceptually relevant for the identification of emotion in speech. A perception experiment can provide evidence concerning the perceptual relevance of variations in durational proportion of accented speech segments, for the expression of emotion in speech.

IV. PERCEPTUAL SIGNIFICANCE OF THE RELATIVE DURATION OF ACCENTED AND UNACCENTED SEGMENTS FOR CONVEYING EMOTION IN SPEECH

a. Aim

Despite the presence of varying lengths of the accented segments in emotional speech, no conclusion could be drawn from the production study described above, as to whether the time distribution over accented and unaccented speech segments realized while expressing emotion in speech, simply depends on overall speech rate, just as it would in neutral speech, or whether it depends on the emotion expressed. In the latter case, emotion influences the distribution of time over the accented and unaccented speech segments, as well as the overall speech rate. If the use of different durational proportions of accented speech segments does not influence the identification of emotion in speech, a linear-stretch model, in which all speech segments are equally stretched or shrunk, should be sufficient to capture temporal variations as far as the expression of emotion in speech is concerned. However, if the durational proportion of accented speech segments is perceptually relevant to the expression of emotion in speech, it should be rewarding to model these variations.

The aim of this experiment is to investigate whether the durational proportion of accented speech segments is perceptually relevant for the expression of emotion in speech, and, if so, to find the optimal durational proportion of accented speech segments for each emotion.

b. Speech material

Three neutral utterances of speaker MR, one utterance of Sentence 1 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane), one of Sentence 2 'Jan is naar de kapper geweest' (John has been to the hairdressers), and one of Sentence 3 'Het is bijna negen uur' (It is almost nine o'clock) were manipulated by analysis-resynthesis. The manipulations combine the following factors: pitch level and pitch range, overall speech rate, intonation pattern, and durational proportion of accented speech segments. The values for pitch level and pitch range used in the manipulations were those found in Chapter II to be optimal for each of the seven emotions (see the left hand side of Table 14 in Chapter II). The values for overall speech rate that were also found to be optimal for each of the seven emotions in Chapter II (and that are presented here in the right most column of Table 2), were used in a first set of six conditions. The same intonation pattern was used across emotions (see below). This resulted in a series of manipulations in which the overall speech rate optimal for each emotion varied according to that emotion, as did pitch level and pitch range. In a second set of six conditions, the overall speech rate was kept constant, i.e., all stimuli had the same total duration as the original neutral utterance.

In the first of the six conditions (see Tables 4 and 5), no manipulation of durational proportion of accented speech segments was carried out, leaving the lengthening of the accented and the unaccented syllables as in the original utterance. In the other five conditions of both sets, the durational proportion of accented segments was systematically varied. In the second condition, the lengthening of the accented syllables was manipulated in such a way that it co-varied with the lengthening of the accented segments found at varying overall speech rates in neutral speech. The value for the corresponding 'stretching factor' was given, per sentence, by the regression lines in Figure 1. In the third condition, the durational proportion of accented speech segments was manipulated in such a way that the accented speech segments were 20% shorter than, in neutral speech, at an overall speech rate prescribed by the regression lines. In the fourth and fifth conditions, manipulations in durational proportion of accented segments resulted in accented speech segments 20% and 40% longer than would be the case in neutral speech, according to the regression lines in Figure 1, respectively. Note that the difference between the first

Table 4: The experimental conditions in brief

Unless explicitly mentioned, the intonation pattern '1&A 1&A' was used in the manipulations.

Conditions	*Sets of conditions*	
	Overall speech rate optimal for the respective emotions	Overall speech rate kept constant
1: linear	No local stretching	No local stretching
2: ref.	Stretching factor = proportion accented as in neutral speech at the corresponding overall speech rate	Stretching factor = proportion accented as in neutral speech at the corresponding overall speech rate
3: -20%	Stretching factor - 20%	Stretching factor - 20%
4: +20%	Stretching factor + 20%	Stretching factor + 20%
5: +40%-1&A	Stretching factor + 40%	Stretching factor + 40%
6: +40%-1B	Stretching factor + 40% with intonation pattern '1B 1B'	Stretching factor + 40% with intonation pattern '1B 1B'

condition corresponding with a linear manipulation, and the second condition in which the variation in lengthening of accented syllables is represented as a function of overall sentence duration, as shown in the data of Figure 1, is less important than the variation of 20% increase/decrease or 40% increase used in the manipulations for the other conditions.

In those five conditions, all utterances were re-synthesized with the '1&A 1&A' intonation pattern, as the '1&A' was indeed found to be suitable in the expression of all emotions studied (Chapter III). In situations in which one does not want to introduce variability by using different intonation patterns, a combination of '1&A' patterns of pitch movements was found to be the most suitable one for controlling this variation. Stretching the accented syllables does, however, not only increase the duration of the vowels of these syllables. The fall 'A' of the '1&A' pattern of pitch movements canonically starts 80 ms after the vowel onset. This means that when the vowel is stretched, a larger portion of the fall 'A' is contained within the vowel. As a consequence, the 'A' of the '1&A' pattern of pitch movements becomes more audible when the vowel is stretched. In synthesizing the stimuli, it was intuitively felt that this increased audibility of the 'A' in the '1&A' pattern of pitch

Table 5: Proportion accented/unaccented speech segments per sentence and per condition

Conditions	*Sentence 1: 'vliegtuig'*						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear	0.362	0.362	0.362	0.362	0.362	0.362	0.362
ref.	0.362	0.374	0.331	0.377	0.342	0.370	0.350
-20%	0.290	0.299	0.265	0.302	0.273	0.296	0.280
+20%	0.435	0.449	0.397	0.452	0.410	0.444	0.420
+40%-1&A	0.507	0.524	0.463	0.528	0.478	0.518	0.490
+40%-1B	0.507	0.524	0.463	0.528	0.478	0.518	0.490
Conditions	*Sentence 2: 'kapper'*						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear	0.385	0.385	0.385	0.385	0.385	0.385	0.385
ref.	0.385	0.393	0.361	0.395	0.371	0.390	0.376
-20%	0.308	0.314	0.289	0.316	0.297	0.312	0.301
+20%	0.462	0.471	0.433	0.474	0.445	0.468	0.452
+40%-1&A	0.538	0.550	0.505	0.553	0.519	0.546	0.527
+40%-1B	0.538	0.550	0.505	0.553	0.519	0.546	0.527
Conditions	*Sentence 3: 'uur'*						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear	0.410	0.410	0.410	0.410	0.410	0.410	0.410
ref.	0.410	0.392	0.377	0.388	0.393	0.398	0.402
-20%	0.328	0.313	0.302	0.310	0.314	0.319	0.322
+20%	0.492	0.470	0.452	0.465	0.471	0.478	0.482
+40%-1&A	0.574	0.549	0.528	0.543	0.550	0.558	0.563
+40%-1B	0.574	0.549	0.528	0.543	0.550	0.558	0.563

movements might influence the perception of some emotion. Therefore, in order to separate the effect of the lengthening of the vowel from the increased audibility of the fall in the '1&A' pattern of pitch movements, a sixth condition was included, in which the lengthening of the accented syllables was not only increased by 40%; the '1&A 1&A' intonation pattern was also replaced by the '1B 1B' intonation pattern. The non-lending pitch movement 'B' starts very early in the syllable following the accented one. It was synthesized so that it started at the end of the vowel of the accented syllable. As a result, it was hardly, if at all, audible as a pitch movement (cf. House, 1990; Hermes, 1997). Table 4 summarizes the six conditions of both sets. The exact values used for the durational proportion of accented speech segments are reported per condition for each sentence in Table 5.

The locations of the vowel onsets in the accented syllables were determined by listening (Hermes, 1990). The rule-based pitch contour was generated for each emotion with the '1&A 1&A' or the '1B 1B' intonation pattern and with the pitch level and pitch range found optimal in the previous study. Due to the possible interference with final lengthening, the last syllable of the utterances was not affected by the manipulations of lengthening of the accented and unaccented syllables, in any condition in either of the sets. The last syllable was thus kept as it was in the original utterances.

The manipulations were based on the PSOLA technique (Moulines & Laroche, 1995). In all conditions, duration manipulations were carried out first. The pitch contour corresponding to the intonation pattern was then applied to the result. The resulting 252 variants (2 sets with a different manipulation of overall speech rate × 3 sentences × 7 combinations of pitch level and pitch range × 6 conditions) served as stimuli. All the signals had a constant voice quality, namely that of the original neutral utterances.

c. Design and procedure

Each of the three sentences was presented to the subjects in a separate block. The order of presentation of these three blocks was counterbalanced across the subjects. Within a block, the stimuli were presented to each subject in a different random order.

A total of twenty-four subjects (12 female, 12 male) participated in the listening experiment. Half of them were either working or studying at IPO, the other half came from outside the institute. None of them had any particular knowledge of phonetics. The subjects took the

test individually, in an isolated booth, using headphones and an interactive computer program. The experiment was based on a seven-alternative forced choice paradigm; subjects were provided with the seven emotion labels, and they were instructed to listen to each utterance and then to assign one of the emotion labels to this utterance (i.e., the label corresponding with the emotion they thought was expressed in the utterance), before proceeding to the next utterance. Subjects could listen only once to each utterance, and were not provided with any feedback on their performance. They had two short breaks between the blocks, splitting the test into three periods of about 10 to 15 minutes each.

The differences induced in the responses of the subjects by the re-synthesized utterances representing different conditions provided information about the perceptual relevance of the deviations from a linear model description.

d. Results

Since our results are identification data that cannot be assumed to be normally distributed and since the identification confusions are informative in the present study, the results can pre-eminently be subjected to a categorical, cross-classified loglinear analysis (Fienberg, 1980). To that end, the results were pooled into two three-dimensional tables. The first table contains the responses of the six conditions of the first set, in which the stimuli were not only provided with the optimal pitch level and pitch range of a certain emotion, but also with the corresponding optimal overall speech rate. The second table contains the responses of the six conditions of the second set, in which the overall speech rate of the sentences was kept constant. Optimal overall speech rate was not varied independently of optimal pitch level and pitch range. Therefore, these two sets of six conditions were subjected to two separate loglinear analyses.

The dimension with the optimal values, i.e., the combination of pitch level and pitch range optimal for the seven emotions, will be indicated with COMBI. This dimension contains seven categories, each corresponding to one of the seven emotions studied. The other dimensions are COND, which has six categories corresponding to the six conditions, and RESP, the response of the subjects, which has the same seven categories as COMBI. The first two dimensions, COMBI and COND, represent independent variables, while the third dimension RESP represents a dependent variable.

The three sentences were analyzed both separately and collapsed. In all cases, the three-way loglinear analysis showed that from all relevant loglinear models, three did not deviate significantly from the data. The chi-squares (χ^2's), degrees of freedom, and p-values of the two simplest of these models, are presented in Table 6 for the three sentences separately, and for the collapsed results. In the most simple model, Model 1, only COMBI and RESP have significant effects, while there are no significant interactions. So, in this model there was no significant effect of COND. In the second model, Model 2, there are significant effects of COMBI, COND and RESP, and significant interactions between COMBI and

Table 6: Chi squares, degrees of freedom and corresponding p-values for the two loglinear models fitting the data

Results are presented for the three sentences separately, and collapsed over the three sentences.

	Conditions with *optimal* overall speech rate			Conditions with *constant* overall speech rate		
Model	χ^2	*df*	*p*	χ^2	*df*	*p*
			Sentence 1: 'vliegtuig'			
Model 1	158.08	281	1.000	218.12	281	0.998
Model 2	111.25	210	1.000	127.47	210	1.000
Δ models	46.82	71	0.988	90.66	71	0.058
			Sentence 2: 'kapper'			
Model 1	161.96	281	1.000	222.43	281	0.996
Model 2	116.14	210	1.000	117.68	210	1.000
Δ models	45.82	71	0.991	104.76	71	0.006
			Sentence 3: 'uur'			
Model 1	153.80	281	1.000	150.22	281	1.000
Model 2	128.59	210	1.000	121.24	210	1.000
Δ models	25.21	71	1.000	28.99	71	1.000
			All sentences			
Model 1	257.05	281	0.844	302.00	281	0.186
Model 2	168.05	210	0.985	161.67	210	0.994
Δ models	89.00	71	0.073	140.33	71	< 0.001

RESP and between COND and RESP. Subtracting χ^2's and degrees of freedom (Fienberg, 1980, p. 57) shows that for Sentence 2 (kapper), and for all three sentences taken together, this leads to a significant improvement, i.e., $p < .05$, for the set of the conditions in which the overall speech rate was kept constant. In these conditions, the second loglinear model also tends to result in a better model for the sentence 'vliegtuig' ($.05 < p < .10$). For the conditions in which overall speech rate was varied with pitch range and pitch level, there was only a tendency towards a significant improvement for the collapsed results.

The third, and most complex model, is characterized by significant interactions between all three variables. However, this did not lead to any significant improvement compared to Model 2, in any occasion. It is, therefore, concluded that the second loglinear model best fits our data.

In summary, interactions of COMBI and RESP, and of COND and RESP, appear to be significant, while the interaction between COMBI and COND is not. The implication of this is that within each category of RESP, i.e., within each response class, COMBI and COND can be assumed to be independent factors. In the first instance, this may seem trivial since, in the experimental set-up used here, COMBI and COND are varied independently. On the other hand, it shows, e.g., that one specific combination of pitch level and pitch range does not exclude or uniquely determine the response of the subject, whatever the condition. A practical consequence is that within each response class, the results over each of the independent variables can be collapsed. The collapsed results are presented in Tables 7 and 8 for the six conditions in which the overall speech rate was perceptually optimal, and in Tables 9 and 10 for the six conditions in which overall speech rate was kept constant. These tables are described further, in the following section.

Tables 7 and 9 form the basis of a cluster analysis, in which we attempt to find out whether the various conditions can be combined into groups of conditions which give rise to the same distribution of responses over the various response classes. A (loglinear) test of independence (see Chapter IV for a description of this cluster analysis) was applied to each possible combination of two rows in these tables. The smaller the χ^2 resulting from the analysis, the more the two conditions are associated. The p-values corresponding with this χ^2, for six degrees of freedom, are presented in Table 11 for the conditions with optimal overall speech rate, and in Table 12 for the conditions with constant overall speech rate. In Table 11, it can be seen that the three conditions in which the durational proportion of

Table 7: Number of responses within each response class, for each of the six conditions with OPTIMAL overall speech rate, collapsed over optimal pitch level and pitch range corresponding with the emotions

Conditions	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
linear	99	72	109	57	70	46	51	504
ref.	94	88	114	50	54	40	64	504
-20%	87	76	105	72	60	50	54	504
+20%	65	80	124	52	50	48	85	504
+40%-1&A	50	60	110	59	59	57	109	504
+40%-1B	54	83	104	63	51	54	95	504
total	449	459	666	353	344	295	458	3024

Table 8: Number of responses within each response class, for the seven combinations of optimal pitch level and pitch range corresponding with the emotions, collapsed over the set of six conditions of proportional temporal manipulations of accented and unaccented speech segments with OPTIMAL overall speech rate

The categories figuring in the first column feature the factor COMBI; they correspond to the combinations of pitch level and pitch range optimal for each emotion that are, for convenience sake, named after these emotions. For this set of conditions, the stimuli were not only provided with the optimal pitch level and pitch range, but also with the overall speech rate optimal for each emotion. The list of perceived emotions figures in the second row.

COMBI	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
neutrality	248	15	44	50	29	4	42
joy	40	126	1	96	11	48	110
boredom	9	3	345	16	47	2	10
anger	66	120	4	130	4	30	78
sadness	30	8	236	6	103	12	37
fear	26	131	0	40	28	116	91
indignation	30	56	36	15	122	83	90
total	449	459	666	353	344	295	458

Table 9: Number of responses within each response class, for each of the six conditions with CONSTANT overall speech rate, collapsed over optimal pitch level and pitch range corresponding with the emotions

Conditions	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear	189	99	23	27	58	49	59
ref.	199	97	20	29	71	48	40
-20%	178	94	20	42	58	49	63
+20%	163	94	27	24	73	46	77
+40%-1&A	125	82	38	26	56	67	110
+40%-1B	99	83	37	42	77	46	120
total	953	549	165	190	393	305	469

Table 10: Number of responses within each response class, for the seven combinations of optimal pitch level and pitch range corresponding with the emotions, collapsed over the set of six conditions of proportional temporal manipulations of accented and unaccented speech segments with CONSTANT overall speech rate

The categories figuring in the first column feature the factor COMBI; they correspond to the combinations of pitch level and pitch range optimal for each emotion that are, for convenience sake, named after these emotions. For this set of conditions, the stimuli were not generated for all emotions with the same constant overall speech rate. The list of perceived emotions figures in the second row.

COMBI	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
neutrality	239	16	50	43	38	4	42
joy	67	111	9	29	72	59	85
boredom	235	10	64	47	39	7	30
anger	146	106	11	14	34	29	92
sadness	204	47	29	15	63	19	55
fear	22	107	1	19	92	111	80
indignation	40	152	1	23	55	76	85
total	953	549	165	190	393	305	469

Table 11: p-values corresponding to the χ^2_6 for each two of the rows presented in Table 7

Conditions	linear	ref.	-20%	+20%	+40%-1&A	+40%-1B
linear	1.000	0.793	0.932	0.114	0.003	0.016
ref.	0.793	1.000	0.693	0.546	0.008	0.103
-20%	0.932	0.693	1.000	0.240	0.013	0.109
+20%	0.114	0.546	0.240	1.000	0.507	0.880
+40%-1&A	0.003	0.008	0.013	0.507	1.000	0.823
+40%-1B	0.016	0.103	0.109	0.880	0.823	1.000

Table 12: p-values corresponding to the χ^2_6 for each two of the rows presented in Table 9

Conditions	linear	ref.	-20%	+20%	+40%-1&A	+40%-1B
linear	1.000	0.838	0.917	0.761	0.000	0.006
ref.	0.838	1.000	0.540	0.207	0.000	0.000
-20%	0.917	0.540	1.000	0.555	0.000	0.008
+20%	0.761	0.207	0.555	1.000	0.013	0.130
+40%-1&A	0.000	0.000	0.000	0.013	1.000	0.300
+40%-1B	0.006	0.000	0.008	0.130	0.300	1.000

accented speech segments was kept constant, was varied with overall speech rate, and was decreased by 20%, respectively, formed one cluster. In addition, the two conditions in which the lengthening of accented speech segments was increased by 40%, formed one cluster. The last condition, in which durational proportion of accented speech segments was increased by 20%, was, as expected, somewhere in between. It is closest to the '+40%-1B' condition in Table 11, and closest to the condition in which overall speech rate was kept constant in Table 12.

In the next analysis, this '+20%' condition was grouped with the two '+40%' conditions. Grouping them with the other cluster does not change the conclusions in any significant way. Note that both '+40%' conditions also form one cluster: the stimuli with the '1&A 1&A' intonation pattern and those with the '1B 1B' intonation pattern do not elicit significantly different responses. This shows that the different responses generated by the

Table 13: Number of responses in each emotion category collapsed over the two clusters of conditions with OPTIMAL overall speech rate

Conditions	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear, ref., -20%	280 ↑	236	328	179	184	136	169 ↓
+20%, +40%1&A, +40%-1B	169 ↓	223	338	174	160	159	289 ↑
total	449	459	666	353	344	295	458

Table 14: Number of responses in each emotion category collapsed over the two clusters of conditions with CONSTANT overall speech rate

Conditions	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear, ref., -20%	566 ↑	290	63	98	187	146	162 ↓
+20%, +40%-1&A, +40%-1B	387 ↓	259	102	92	206	159	307 ↑
total	953	549	165	190	393	305	469

stimuli with 40% increase of the durational proportion of accented speech segments, can be attributed to the lengthening of the speech segment, and not to the concurring better audibility of the fall 'A' in the '1&A' pattern of pitch movements.

Finally, the responses collapsed over the two clustered groups are presented in Tables 13 and 14. Deviations from a loglinear model in which perceived emotion and condition are independent are marked with ↑ if the obtained value is higher than expected ($p < .01$), and with ↓ if lower than expected. The data presented in Table 13 and 14 are collapsed over the three sentences. Similar results were obtained for all sentences separately (with lower levels of significance), except for Sentence 3 (uur), for which not even a tendency towards a deviation from independence was found. Tables 13 and 14 show that significant deviations were only obtained for the response classes 'neutrality' and 'indignation'. In the response class 'neutrality', the normal durational proportion of accented speech segments was over-represented, while the increased durational proportion was under-represented. In the response class 'indignation', it was the other way around.

In order to find out whether the seven combinations of pitch level and pitch range found to be optimal for the seven emotions can be combined into clusters of such combinations yielding the same distribution of responses over the various response classes, a similar cluster analysis was carried out on the basis of the data presented in Table 8 and 10: the distribution of the responses over all seven combinations of optimal pitch level and pitch range (and overall speech rate) are presented collapsed over the six pertinent conditions. In the six conditions in which overall speech rate was kept constant, about the same results were obtained as in Chapter IV. Neutrality clustered with boredom ($p = .8$). At a significance level of .05, joy and indignation ($p = .054$) tended to belong to different clusters. All other χ^2's, except the one for fear and indignation ($p = .023$), were so high that their corresponding p-values were lower than .001. In the six conditions in which overall speech rate was optimally varied with pitch level and pitch range, all response distributions over the emotions differed significantly with p-values smaller than .001, except for 'joy' and 'anger', where the p-value was .041.

In addition, if we define a response as being correct when it corresponds with the emotion for which, independently of the lengthening of accented speech segments, the optimal values were used (for pitch level, pitch range, and, for the first set of conditions, overall speech rate), then the following is true: emotions were correctly identified in 1153 utterances in the first set of conditions with optimal overall speech rate, and in 687 utterances in the set of conditions with constant overall speech rate. The distribution of these correct responses are reported in Table 15, per category, for the first set of conditions, and in Table 16 for the second set of conditions; these correct responses constitute a subset of the total number of responses per category reported in Table 7 and Table 9, respectively. These identification results correspond to 38.1% overall correct identification for the first set of conditions using optimal values of overall speech rate, and to 22.7% overall correct identification for the second set of conditions with constant overall speech rate. Although both percentages of correct identification are low, they are higher than the chance level of 14.3%. The difference between those two percentages shows the importance of the overall speech rate for the identification of emotion in speech. When the overall speech rate was kept constant, results yielded a clearly lower identification of the emotions, particularly for boredom, anger, and, to a lesser extent, sadness. In fact, boredom was very well identified with the optimal overall speech rate, and badly identified with the constant overall speech rate. For this emotion, the information of the overall speech rate is, thus, of primary importance.

Table 15: Number of correct identifications of emotions per condition with OPTIMAL overall speech rate per emotion

Conditions	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
linear	53	20	55	21	22	18	10	200
ref.	55	26	61	22	13	18	9	204
-20%	39	21	55	24	20	21	11	185
+20%	40	25	64	19	16	14	12	190
+40%-1&A	25	14	55	20	18	23	26	181
+40%-1B	36	20	55	24	14	22	22	193
total	248	126	345	130	103	116	90	1153
percentage identification	57.4%	29.2%	79.9%	30.1%	23.8%	26.9%	20.8%	38.1%

Table 16: Number of correct identifications of emotions per condition with CONSTANT overall speech rate

Conditions	Responses of subjects							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*total*
linear	49	23	11	2	7	12	13	117
ref.	53	20	9	1	9	21	7	120
-20%	44	17	6	2	11	21	9	110
+20%	39	23	11	0	11	15	13	112
+40%-1&A	24	15	16	6	15	17	24	117
+40%-1B	30	13	11	3	10	25	19	111
total	239	111	64	14	63	111	85	687
percentage identification	55.3%	25.7%	14.8%	3.2%	14.6%	25.7%	19.7%	22.7%

Obviously, all emotions were not equally well identified (see Table 15 and Table 16). The mean percentage of correct identification over the two sets of conditions is 56.4% for neutrality, 27.4% for joy, 47.3% for boredom, 16.7% for anger, 19.2% for sadness, 26.3% for fear, and 20.3% for indignation.

The number of correct identifications collapsed over the three sentences and over the two clustered groups are presented in Tables 17 and 18. Deviations from a loglinear model in which perceived emotion and condition are independent are marked with ↑ if the obtained value is higher than expected ($p < .05$), and with ↓ if lower than expected. Note that the present analysis was done for χ^2 with only one degree of freedom. For neutrality, in the conditions with optimal overall speech rate, deviations were only approaching significance ($p = .07$) and are therefore marked with ∧/∨. Tables 17 and 18 show that significant deviations were only obtained for the response class 'indignation', and, in the condition with constant overall speech rate, for the response class 'neutrality'. Naturally, the effect of durational proportion of accented speech segments is easier to determine when the overall speech rate is kept constant. When the overall speech rate is varied, its effect overwhelms the smaller effect of the proportion of stretching of accented and unaccented syllables. Despite this fact, the deviations concerning both clusters of emotions were significant, with optimal overall speech rate ($p = .0043$), as well as with constant overall speech rate ($p = .0002$), showing that there is an effect of durational proportion of accented speech segments relevant for conveying emotion in speech. Moreover, for indignation, significantly more correct responses were given by the subjects when the lengthening of the accented speech segments was increased. For neutrality, an increase in lengthening of the accented syllables resulted in less correct identifications. Comparing these results with those of Tables 13 and 14 shows that, as expected, the higher number of identifications occur in the same response classes in which an over-representation of responses occur.

If confusion matrices, that are not shown here, are constructed for data collapsed over the two clusters of temporal conditions, consistent differences can be observed between the two sets of conditions with constant and optimal overall speech rate, respectively. On the one hand, consistencies among the confusions occurring in temporal conditions belonging in the same cluster confirm the clustering in two groups of temporal conditions. On the other hand, differences observed between the two sets of conditions once again confirm the great importance of overall speech rate. This can be illustrated with a few examples. Sadness, for instance, is confused with neutrality when the overall speech rate is kept constant, but it is confused with boredom when the speech rate is optimal. This is

Table 17: Number of correct identifications in each emotion category collapsed over the two clusters of conditions with OPTIMAL overall speech rate

Conditions	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear, ref., -20%	147 ∧	67	171	67	55	57	30 ↓
+20%, +40%-1&A, +40%-1B	101 ∨	59	174	63	48	59	60 ↑
total	248	126	345	130	103	116	90

Table 18: Number of correct identifications in each emotion category collapsed over the two clusters of conditions with CONSTANT overall speech rate

Conditions	Responses of subjects						
	neutrality	joy	boredom	anger	sadness	fear	indignation
linear, ref., -20%	146 ↑	60	26	5	27	54	29 ↓
+20%, +40%1&A, +40%-1B	93 ↓	51	38	9	36	57	56 ↑
total	239	111	64	14	63	111	85

consistent with the fact that both sadness and boredom are best expressed with a low overall speech rate. Another example is indignation, that is either confused with joy when the overall speech rate is kept constant, or with sadness when the overall speech rate is optimal. Indeed, indignation and sadness are optimally expressed with a decreased speech rate.

e. Discussion

In the most simple loglinear model, that did not differ significantly from the data obtained, there are only significant effects of COMBI and RESP. There is no significant effect of COND. The effect of COND could only be established by the significant improvement obtained by adopting the second loglinear model, in which there were significant effects of all three variables and significant interactions between COMBI and RESP, and between COND and RESP. This shows that the effect of varying the durational proportion of accented speech segments is relatively small. The effect of varying pitch level, pitch range, and overall speech rate is much stronger. In fact, this is as could be expected, since

variations in pitch level, pitch range and overall speech rate are known to have strong effects on conveying emotion in speech (e.g., Williams et al., 1972; Ladd et al., 1985).

Despite the fact that the effect of the durational proportion of accented speech segments is small, it should be noticed that the first two conditions, i.e., the condition in which this durational proportion was kept constant, and the condition in which the durational proportion of accented speech segments was varied with the overall speech rate, are both in the same cluster. This shows that this effect below utterance level is indeed due to the expression of emotion and not simply due to the variation in overall speech rate. The fact that both conditions in which the lengthening of accented syllables was increased by 40% also are in the same cluster, independently of the use of the pattern of pitch movements '1&A' or '1B', shows that the relevance of the durational proportion of accented speech segments for the perception of indignation is associated with the length of the accented speech segments and not with the fact that pitch movements are better audible when the syllable is longer. (This is explicitly specified for indignation and not for neutrality, simply because, for neutrality, the relevant fact is that the lengthening of accented and unaccented speech segments do not deviate significantly from linearity.) The fact that the two conditions with 40% increase are in the same cluster, also confirms the independence of the parameters intonation pattern and durational proportion of accented speech segments. It also corroborates the clustering of patterns obtained in the previous study (Chapter IV), in which these two patterns were members of the same cluster of patterns of pitch movements. It is, therefore, concluded that the differences in duration of the accented segments indeed induce the differential effect for neutrality and indignation. This not only confirms the finding by Nooteboom (1973) that listeners have a well-defined internal representation of the length of accented vowels, but also that varying the duration of accented speech segments can be used for communicative purposes, i.e., signaling a specific emotional state.

The results showed, additionally, that the effect of the durational proportion of accented speech segments only expressed itself in the response classes 'neutrality' and 'indignation'. In the other response classes, no significant effect could be found. Intuitively, we had expected that the emotion anger would be associated with a small durational proportion of accented speech segments. Although the largest number of responses was indeed obtained in the '-20%' condition in the response class anger (see Tables 7 and 9, Column 4), this did not reach a level of significance. So, it can only be speculated that the variability introduced by the large effect of pitch level, pitch range, and overall speech rate obscures smaller

effects as proposed here. In order to establish these smaller effects, it would seem necessary to do more dedicated experiments and/or to have more subjects participating in the experiment.

V. GENERAL DISCUSSION

The present chapter investigated temporal variations produced in speech conveying emotion, and the role of temporal variations in the perception of emotion in speech. First, the focus was on variations involving the whole utterance, i.e., variations in overall speech rate. Then, relative durations were considered below utterance level, i.e., within the utterance, in the production and in the perception.

1. Production

In the production study carried out at utterance level, an analysis of the speech produced by three speakers was carried out in order to extract regularities and systematic differences occurring in speech conveying the seven emotions studied. For each emotion, inter-speaker and intra-speaker variations were described. The three speakers agreed rather well on the relative speech rate for the expression of joy, sadness, and boredom. They showed less agreement for the expression of anger, fear, and indignation. Consequently, ranking the emotions on the basis of the overall speech rate used by the three speakers resulted in differently ordered lists of the emotions. However, a comparison of the values found in the production study, with those found optimal on the basis of perception experiments in Chapter II, showed that, despite the individual differences in the expression of emotions, the overall picture resulting from the production study at utterance level converges with the overall picture resulting from the perception study. Generally, optimal values are more extreme than the values measured in the speech produced by the speakers. In other words, when speakers expressing an emotion increase or decrease their speech rate accordingly, the optimal values correspond to an even more increased or decreased speech rate. Since, in the perception experiment, overall speech rate was the only parameter that could be relied upon to convey the emotion in speech, while in the production study, speakers expressing an emotion also rely on several other parameters and their interactions, it seems quite natural to obtain more extreme values as optimal values in the perception study.

The importance of overall speech rate for the expression of emotion in speech that was already established in related studies (e.g., Fairbanks and Pronovost, 1939; Ladd,

Silverman, Tolkmitt, Bergman, and Scherer, 1985; Carlson, Granström, and Nord, 1992) was corroborated by the results of the perception experiments reported in Chapter II. The relative contribution of overall speech rate to the expression of the emotion seems to vary from speaker to speaker. There seems to be mutual compensation of speech parameters; if one parameter conveys a specific emotion, other speech parameters can be less distinctive for that emotion. Some speakers seem to prefer the use of either pitch factors, temporal factors, other factors such as voice quality or loudness, or a combination of some of the previous speech parameters for the expression of specific emotions. This means that some speakers tend to rely more than others on the overall speech rate for conveying a particular emotion. For some emotions though, there seems to be a reasonable agreement among speakers; boredom and sadness are best conveyed with a low speech rate, joy with a speech rate similar to the one for neutrality. There is more disagreement for anger, fear, and indignation.

As temporal variations appear to be very important to the expression of emotion in speech, at the utterance level, it seemed relevant to determine exactly how these temporal variations are realized below utterance level, i.e., within the utterances. Indeed, if some of the emotional states influence the motor control, it has to be expected that the articulatory control and the breathing control are also affected by the expression of emotion. If the articulatory and breathing patterns are affected, it can be expected that the relative length of speech segments will vary accordingly. It, therefore, seems reasonable to expect an effect below utterance level in speech produced in emotional states.

For this reason, the variations in the durational proportion of accented speech segments were investigated, and some variations were observed in emotional speech. In order to determine whether these variations are due to emotion or are simply a function of varying overall speech rate, neutral speech was investigated at varying overall speech rates.

In addition to the results concerning emotional speech, the durational proportion of accented speech segments was thus described as a function of overall speech rate in neutral speech. It was shown that, at decreased speech rates, the accented speech segments are relatively less lengthened than the unaccented ones. At increased speech rates, results are not as clear. In Sentence 1 (vliegtuig), the lengthening of accented syllables increases with increasing overall speech rate. For both other sentences, there seems to be a turning point at a particular overall speech rate where the durational proportion of accented speech segments

is the highest. From that point, the durational proportion of accented speech segments decreases with increasing overall speech rate.

The results concerning the neutral speech at varying overall speech rates, and the emotional speech, were compared. The comparative results of the production part in this study below utterance level were not conclusive, but there seems to be a slight tendency, in emotional speech involving lower overall speech rates than in neutrality, to lengthen the accented speech segments relatively more than the non-accented ones. Results are less consistent for speech rates higher than in neutrality. Obviously, if such variations below utterance level play a role in the expression of emotion in speech, they constitute a rather small effect compared to the utterance level effect of the overall speech rate.

2. Perception

Finally, a perception experiment was carried out in order to investigate the perceptual relevance of the temporal variations below utterance level. These variations were tested both in the presence and in the absence of variations in overall speech rate.

Results show that temporal features contribute to the expression of emotion in speech. The largest contribution is clearly the one of overall speech rate. The difference in the results obtained in the presence or absence of utterance level variations, confirm the major importance of overall speech rate for conveying emotion in speech. For boredom, for instance, the information conveyed by the overall speech rate overwhelms other effects.

The differences in durational proportion of accented speech segments that were found as a function of overall speech rate in the neutral speech material appeared not to be perceptually relevant to the expression of emotion in speech. It is possible that these variations below utterance level provide a variability in the speech which enhances naturalness without conveying a specific meaning. Moreover, Kitahara and Tohkura (1990) found that temporal structure is relevant to listeners for recognizing words.

On the other hand, the differences in the durational proportion of accented speech segments associated with the expression of emotion proved to be very relevant for the perception of the emotions indignation and neutrality. When a variation in durational proportion of accented speech segments was used that was larger than the one varying with overall speech rate, the number of responses in the category neutrality was significantly lower.

This suggested that such a higher durational proportion of accented speech segments might signal an emotion. This was only confirmed for indignation, however; lengthening the accented syllables relatively more than in neutral speech with a correspondingly low overall speech rate, contributes to the perception of indignation.

Despite the fact that the relevance of durational proportion of accented speech segments for the expression of emotion in speech could not be established in the production study, evidence was obtained in the perception study. This shows the importance of both sorts of investigations, concerned with production and with perception. These two types of investigations provide complementary information. Some variations can be observed in the production data. It does not necessarily mean that they are relevant to the perception of emotion. On the other hand, as is the case in the present study, a specific type of variation may be perceptually relevant to the perception of emotion, while there may be no clear evidence of it in the limited set of production data studied. Again, we are satisfied that the range of variations was not so extreme as to induce caricatural expressions of emotions.

It appeared that the effect of proportionally lengthening the accented and the unaccented segments was, statistically, most significant for the conditions in which the overall speech rate was kept constant. When the overall speech rate was varied, the effect only tended to reach significance. The reason for this is probably that the effect of lengthening the accented speech segments is easier to assess when the overall speech rate is not varied. When the overall speech rate is varied, its effect obscures the smaller temporal effect of the durational proportion of accented speech segments. As the speakers expressed the emotions by using overall speech rate as well as other speech parameters, this might explain why the variation is less clear in the production study, compared to the perception study.

Finally, the perceptual test carried out in the present chapter in the presence of variations in overall speech rate, can be compared with a previous study (Chapter II, Experiment 6). In that study, the speech was manipulated using the same optimal values for pitch level, pitch range and overall speech rate, but the utterances were provided with the pitch contour corresponding to the intonation pattern originally used by the speaker in that emotional utterance. The same seven-alternative forced choice paradigm was used as in current chapter. The only difference in the preparation of the stimuli was that the same intonation patterns were used for all the emotions, in the present study, while, in the previous study, the intonation pattern of the corresponding original emotional utterance was used per emotion. The mean proportion of correct identification per emotion in the two clusters of

both sets of conditions in the present investigation, i.e., with and without manipulation of overall speech rate, is shown in Table 19 with the results of the previous study (Chapter II, Experiment 6). In that previous study, 48% of the subjects' responses were correct which compares with 38% correct identification in the present study, when the optimal overall speech rate was used. As the effect of intonation pattern has been shown to be very relevant for conveying emotion in speech (Chapter IV), such a difference in percentage of correct identification was to be expected. This difference confirms the relevance of the intonation patterns and shows the loss in percentage of correct identification occurring when this parameter is kept constant.

Summarizing the new results for the emotions studied, it appeared, on the basis of the perception experiment conducted in the present chapter (see Table 17 and Table 18), that a linear-stretch model can be considered sufficient for the five emotions joy, boredom, anger, sadness, and fear. For neutrality, it is even necessary for the durational proportion of accented speech segments not to deviate significantly from this linear model. On the other hand, a proportional temporal manipulation of accented and unaccented speech segments is

Table 19: Mean proportion of correct identification over clusters of conditions and comparative results from previous study

Set of conditions with OPTIMAL overall speech rate								
Conditions	Emotions							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
linear, ref., -20%	0.68	0.31	0.79	0.31	0.25	0.26	0.14	0.39
+20%, +40%-1&A, +40%-1B	0.47	0.27	0.81	0.29	0.22	0.27	0.28	0.37
Set of conditions with CONSTANT overall speech rate								
Conditions	Emotions							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
linear, ref., -20%	0.68	0.28	0.12	0.02	0.12	0.25	0.13	0.23
+20%, +40%-1&A, +40%-1B	0.43	0.24	0.18	0.04	0.17	0.26	0.26	0.23
Results of Chapter II, Experiment 6								
	Emotions							
	neutrality	joy	boredom	anger	sadness	fear	indignation	*mean*
	0.80	0.38	0.90	0.38	0.20	0.23	0.50	0.48

important for conveying indignation. For this emotion, the results of the perception experiment indicate that accented syllables should be stretched relatively more than the unaccented ones.

Furthermore, for neutrality, a moderate overall speech rate is found to be adequate. Perceptually, it is important that no significant deviation from a linear model occurs in durational proportion of accented speech segments (see Table 15 and Table 16). Joy, fear, and anger, were found to be optimally identified with an overall speech rate higher than the one used in neutrality. Sadness, on the other hand, is best conveyed with a lower overall speech rate than in neutrality. Boredom is mostly characterized by a very low overall speech rate. When this overall speech rate is sufficiently decreased, it is a distinctive parameter for this emotion. Indignation has a lower overall speech rate than in neutrality and is characterized by a high durational proportion of accented speech segments. For indignation, the durational proportion of accented speech segments contained in the utterance is relevant to the perception of the emotion.

Finally, it is important to recall that features of speech other than the temporal ones, are also relevant to the expression of emotion in speech. In particular, previous chapters were concerned with an investigation of F_0 variations in emotional speech and pitch phenomena conveying emotion in speech. In the final chapter of this thesis, the results found in the present chapter will be combined with those of the previous chapters.

Chapter VI

General conclusions and final discussion: Integration of findings

ABSTRACT

In this last chapter, the research area of the present study is demarcated and the findings resulting from this investigation are summarized. The results found to be specific to the expression of emotion in speech are given in the form of a series of rules for synthesizing speech in each of the emotions studied. Additionally, general results concerning the suitability of models for handling the extreme variations occurring in emotional speech are summarized. Finally, a few suggestions are given for future research on the expression of emotion in speech.

I. DEMARCATION OF THE PRESENT STUDY

Before summarizing the main results of the present study, the most relevant demarcations of the present study will be outlined. Although the expression of emotion usually involves different communication channels simultaneously, the present research focused exclusively on specific aspects of a single channel, namely the oral/auditory one. Bodily gestures and facial expressions were not included in the study. The semantic and syntactic aspects of the expression of emotion, as well as non-speech sounds, such as breathing sounds, and lip smacks, were also excluded from the investigation. The present study was strictly limited to the investigation of certain prosodic components playing an important role in conveying emotion in speech sounds, namely intonation and speech rate. Other components, such as the voice quality, and the intensity of the speech signal, were not included in the present study despite the fact that they are known to be relevant for the expression of emotion in speech. In summary, the work on this thesis was restricted to the study of variations in pitch and temporal variations associated with the expression of emotion in speech sounds.

In this investigation, variations in fundamental frequency were described in terms of models of intonation. For a model to be used in this investigation, it was expected that it would allow the generation of re-synthesized or synthetic speech, as well as the description of production data. In the present study, the F_0 curves produced by the speakers in emotional speech were analyzed in the framework of the IPO-model of intonation, and stimuli were also generated in the framework of this model of intonation. This was convenient as expertise on this model was available at the institute where the investigation took place. In a different approach, such as an auto-segmental approach, a different type of model of intonation would be exploited. However, the choice of this two-component model does not imply that another model of intonation would have been less suited for use in this study.

In order to control the variability of the speech material, use was made of recordings of pre-determined emotions, uttered by actors in predetermined sentences, and recorded under controlled conditions in an optimal acoustic environment, the drawbacks of using a database recorded in the laboratory, and containing speech that was not spontaneously produced in real-life situations were thereby accepted. As a consequence, the results of the

production part of the investigation might not be considered to be representative of the spontaneous expression of emotions. On the other hand, a study by Williams and Stevens (1972) in which spontaneous and acted emotional speech were compared, showed that data obtained from spontaneous speech were not inconsistent with data obtained from acted emotions. Moreover, as there are so many ways of expressing a single emotion depending on the personality and the physiology of the speaker, the language used, the circumstances, and so forth, it appears that, in the study of emotional speech, the generalizability of the findings is a difficult issue anyway, whether or not spontaneous speech is used.

The lack of agreement on a widely accepted definition and a taxonomy of emotion constitutes a drawback for the present study. The selection of categories of emotions for use in the present study was, therefore, exclusively based on identification rates. The limited number of emotions selected, namely: neutrality (as a reference), joy, boredom, anger, sadness, fear, and indignation, constitutes a limitation of the present study. Despite the fact that emotions are not well defined, speakers and listeners appear to be able to rely on empirically based notions of what emotions are. This follows from the fact that they are quite capable of performing the pertinent experimental tasks in a consistent way. Thanks to this ability, a study concerning the speech variations associated with the expression of emotion can be carried out. Such a study of emotional speech is mainly concerned with the variability occurring in speech when emotions are expressed. The results cannot reveal a direct relationship between this variability and the psychological or physiological aspects of emotions.

II. RESULTS OF THE PRESENT STUDY

1. Conveying emotion in speech

In this section, our qualitative results concerning the expression of emotion in speech will be summarized. The knowledge acquired about the use of optimal parameters for conveying specific emotions in speech will be summarized, first for parameters with a global character, at utterance level, then for parameters with a more local character, taking variations within utterances into account.

a. General aspects

Emotions are expressed by individuals in many different ways depending on factors such as age, sex, language, culture, and social background. Even the same speaker expresses emotions in various manners depending, among others, on the circumstances. Notwithstanding these intra-speaker and inter-speaker differences in expression, the resulting variability is certainly not random. Speakers appear to adhere to a reasonably limited set of 'acoustic configurations' in the expression of specific emotions. This is in agreement with findings by Cosmides (1983) who found that various cues can contribute to the expression of a particular emotion. The relative contribution of these cues conveying emotion in speech seems to vary. There seems to be a certain mutual compensation of speech parameters; apparently, if one cue is used by a speaker in an effective way, there is less need to rely on the other possible cues for conveying a specific emotion. This compensatory effect does not necessarily lead to discrepancies between speakers, at least, not if viewed in a qualitative way, considering a relative ordering of the emotions according to the use of a specific cue. Furthermore, Hoene (1996), inspired by the analyses in the present study, investigated variations in pitch level, pitch range, and speech rate in the production of emotional speech in Swedish, and also found that speakers express emotions using different strategies, by making relatively more or less use of variations in specific speech features. Heuft, Portele, and Rauth (1996) also found that speakers make more use of some cues than of others. Arndt and Janney (1991) also comment on different strategies of speakers, considering them in the context in which the communication takes place.

Although, in many cases, people are able to identify an emotion expressed by a speaker, not all emotions are equally easy to identify from the speech signal only. Some emotions may intrinsically be more easily confused with each other, than other emotions. The frequently reported confusions, also found in this study, between anger and joy, or between boredom and neutrality, may be due to the vicinity of optimal speech-parameter values for these pairs of emotions.

b. Optimal parameter values at utterance level

With the aim of properly synthesizing emotional speech, optimal parameter values for pitch range, pitch level, and speech rate were determined, on the basis of a male voice. These parameter values are based on the results of perceptual experiments and were also perceptually evaluated. Additionally, the speech of two male speakers and a female speaker

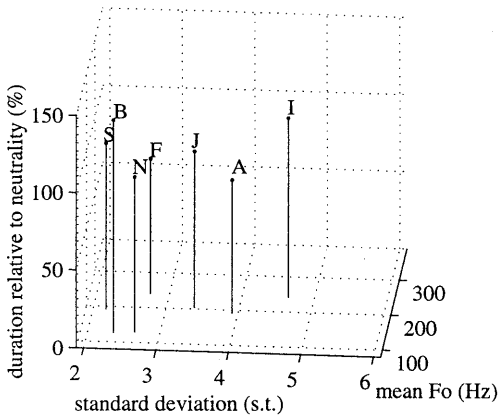

Figure 1: male speaker MR

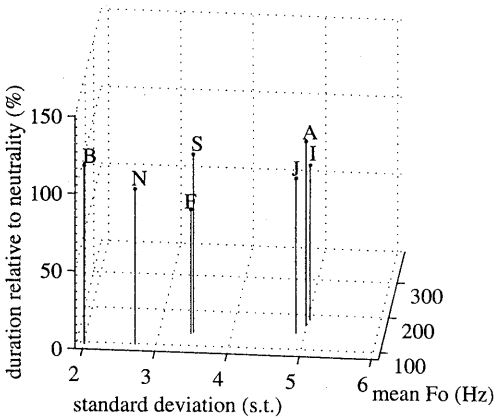

Figure 2: male speaker RS

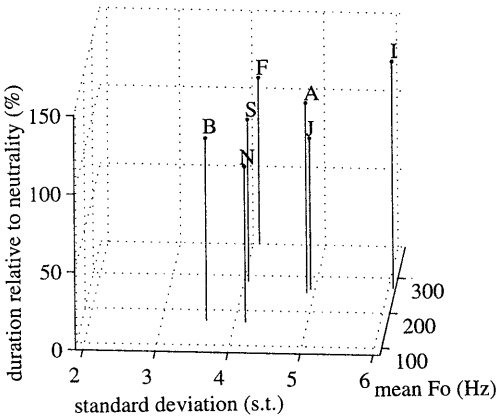

Figure 3: female seaker LO

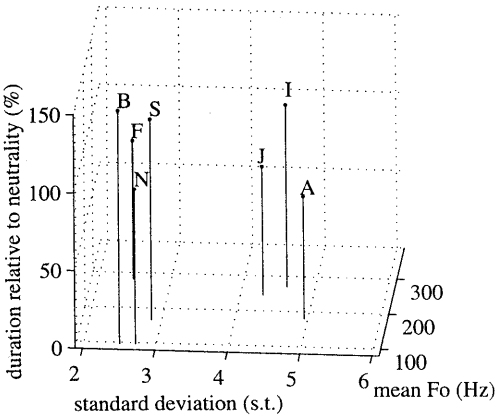

Figure 4: Perceptually optimal values

was analyzed, involving measures featuring pitch range, pitch level, and speech rate. An overview of the results of this production study is presented in Figures 1, 2, and 3, while the perceptually based optimal values, also given in Table 14 of Chapter II, are represented in Figure 4. These figures combine the results obtained for pitch level and pitch range presented in Tables 2 and 3 of Chapter III, with those obtained for speech rate presented in Table 2 of Chapter V. They allow a simple comparison between the values obtained for the different speakers in the production study, as well as between these values and the optimal values resulting from the perception study. Despite the individual differences in the expression of emotion observed in the speech of different subjects, the individual pictures show tendencies consistent with the overall picture resulting from the perceptual study. Some speakers appear to make relatively more use of one specific parameter, while other speakers rely more on the use of another parameter. For instance, to compare the two male speakers, it can be seen from the representation of the values of mean and standard deviation of F_0 (Figures 1 and 2), that the speaker RS makes relatively more use of pitch range, while the speaker MR relies more on pitch level to express the various emotions distinguishably. This exemplifies the mutual compensation between cues used for conveying emotion in speech, as mentioned earlier.

c. Parameters below utterance level

Within utterances, the prosodic features: shape of the pitch curves, relative height of the pitch accents, final lowering, and relative duration of accented and unaccented speech segments, appeared to be of some relevance for the expression of at least some emotion in speech.

Final lowering and relative height of pitch accents

The F_0 curves of the emotional speech produced by the three speakers were analyzed by means of pitch measurements at anchor points chosen in the utterances. Final lowering and relative height of pitch-accent peaks appeared to be two major sources of deviation between the rule-based pitch curves that were synthesized according to the IPO model of intonation, and the F_0 curves actually realized by the speakers. The relevance of these deviations was perceptually investigated.

When the relevance of the relative height of the pitch-accent peaks realized in the emotional utterances was tested separately, this cue appeared to be relevant for the expression of

sadness and fear, and relevant, but in a negative way, for the expression of neutrality and joy. Indeed, for speech conveying neutrality and joy, modeling a different height of the accents decreased the identification rates; apparently, it is the absence of differences in peak heights that is important for these emotions.

When investigated separately, the realization of final lowering appeared to be perceptually relevant for the expression of indignation. To a lesser extent, it is also relevant to the expression of fear and sadness. For boredom, it is of relevance not to realize any final lowering.

The findings are more confusing when the combined effect of both cues, i.e., relative height of the pitch accents and the final lowering, is considered. For sadness, fear, and indignation, the modeling of one of both cues separately was found to be relevant. Nevertheless, the combined effect of the relative height of the pitch accents and the final lowering came out as a small, non-significant effect, at least for most emotions.

- ***Intonation patterns***

The shape of the pitch curves has been studied in terms of intonation patterns, using the intonation grammar for Dutch by 't Hart, Collier and Cohen (1990). A result of this study is that almost all pitch curves occurring in emotional speech can be described with this grammar, and that emotional speech can be generated using the standard pitch movements presented in this description of Dutch intonation. Although no unique correspondence was found between a specific emotion and a particular pattern, clear relationships could be established between patterns and emotions: the choice of a specific intonation pattern was found to be relevant to the perception of specific emotions in speech. One pattern can convey various emotions, and one emotion can be conveyed by various patterns. Even though some patterns may convey an emotion better than others, no specific intonation pattern appeared to be strictly necessary for signaling emotion in speech. Furthermore, a cluster analysis of the intonation patterns that was carried out in this study on the perception data, showed that predominantly the final pattern of pitch movements plays a role in conveying the intended emotion. It appeared that, for some patterns, the identification of some emotions could increase, while for others, it could decrease. These relationships between final pattern of pitch movements and conveyed emotion are summarized in the table presented here. For instance, the final patterns of pitch movements '12' and '3C' were found to be inadequate for conveying neutral speech. Pattern '12' can signal the presence of

Table 1: Relationships established among emotions and patterns of pitch movements in final position

	Pattern(s) to be preferred	Pattern(s) to be avoided
neutrality	1&A	12 and 3C
joy	1&A and 5&A	A, EA, and 12
boredom	3C	5&A and 12
anger	5&A, A and EA	1&A and 3C
sadness	3C	5&A
fear	12 and 3C	A and EA
indignation	12 in particular, but also 3C, A and EA	1&A

indignation and fear, and pattern '3C' the presence of fear, indignation, boredom and sadness.

Furthermore, it was found that, in spite of decreasing the identification of anger and indignation, the '1&A' pattern was used by all speakers, at some time, for the expression of every emotion. Hence, if one wants to avoid the introduction of an extra source of variability due to the use of different intonation patterns, for instance while conducting an experiment, the '1&A' pattern of pitch movements is best suited for controlling this variability. Although this pattern is certainly not the best possible choice for expressing each emotion studied here, it leads to acceptable results concerning identification of the emotions studied, when used in combination with appropriate variations in pitch level and pitch range.

- ***Relative duration of accented and unaccented speech segments***

The simplest model of temporal speech variation in speech is a linear-stretching model, in which all speech segments are equally stretched or shrunk. In such a model, only the variations at utterance level are taken into account. In a production study of emotional speech it appeared, however, that, at overall speech rates lower than normal in neutrality, speakers showed a slight tendency to lengthen the accented speech segments more than the non-accented ones. Results concerning emotional speech produced at speech rates higher than in neutrality were less consistent. In order to investigate the perceptive relevance of the

relative duration of accented and unaccented speech segments within utterances, a listening test was carried out which showed that the relative duration of accented and unaccented speech segments was only relevant for the two emotions: neutrality and indignation. For indignation, lengthening the accented syllables relatively more than would be the case in neutral speech with a correspondingly low overall speech rate, contributes to the perception of this emotion. For neutrality, in order to avoid confusions with emotions, the durational proportion of accented and unaccented speech segments should not deviate significantly from the linear model. For the five other emotions, although it cannot be excluded that the effect of variations in overall speech rate obscures the smaller temporal effect of the durational proportion of accented speech segments, the effect of varying the length of accented speech segments will be small. The largest temporal contribution to the expression of emotion in speech is clearly that of overall speech rate, particularly for boredom. It is therefore concluded that, for the five emotions, joy, boredom, anger, sadness, and fear, the linear model can be considered sufficient, while, for neutrality, any deviation from this linear model must be avoided, and, for indignation, it is relevant to model the relative lengths of accented and non-accented speech segments.

2. General findings resulting from the present study

In addition to the results directly concerned with the expression of emotion in speech, the present study also sheds light on a few more general issues. Indeed, a study considering speech variations as extreme as the ones occurring in the expression of emotion, is a source of opportunities to confront measurement procedures and models commonly used in prosodic studies, with speech samples displaying a wide range of variations. If such a model is found to be adequate for the description of the variations perceptually relevant to the expression of emotion in speech and for the re-synthesis of the emotional speech, its adequacy for modeling the speech variations relevant to spoken communication is confirmed. On the other hand, if the model neither appears to be sufficient for describing speech variations nor for re-synthesizing emotional speech, then the consideration of how the model can be adapted or modified, can contribute to our understanding of speech variations.

When studying speech variability, such crude measures as mean and standard deviation of F_0 must be expected to obscure an important part of the variations present in the speech material, which is, in a study such as this one, precisely what we wish to take into consideration. Despite this fact, these measures are the ones most frequently used in this

type of studies. The main reasons for their frequent use are that they are conveniently easy to obtain, that their common use facilitates the comparison of results across studies, and that they provide parameters easy to manipulate in most synthesizers. We found that, though perhaps not very accurate, these measures remain quite informative: the information obtained by means of F_0 measurements at anchor points within the utterances does, however, allow a better description of the course of F_0 in time. These local measurements can provide information complementary to the rough estimation of pitch level and pitch range by means of measures such as mean and standard deviation of F_0. The consideration of the relative height of the final peak and the F_0 value at the end of speech, appears to provide valuable supplementary information for the analysis of intonation.

The production data used in the present study were first represented in a model of tonal space and then described within the IPO's two-component model of intonation. It appeared that some details that were observed with the tonal approach, involving pitch measurements at anchor points within the utterances, could not be captured in this model. As mentioned above, these details concern the relative height of the accent peaks and the final lowering. A perceptual evaluation of the relevance of these details below utterance level showed that they are not very important for the expression of most emotions in speech. Therefore, the difficulty to represent this detailed information should not be considered a major problem. This simply means that the model provides a simplification of the pitch phenomena on the basis of perceptual relevance, and does not undermine its adequacy for describing speech, even when speech involves a large range of variations. Moreover, it could be speculated that other two-component intonation models would also be adequate, although this was not tested in the present study.

The IPO intonation grammar for Dutch by 't Hart et al. (1990) was used in the present study for the description of the course of the fundamental frequency. The labeling of the pitch curves in terms of intonation patterns is presented at the 'third level of description in the IPO approach' ('t Hart et al., 1990, p. 80). In the present study, the domain of these intonation patterns is the full utterance. The intonation patterns consist of sequences of pitch movements which are defined at the 'second level of description' in the IPO approach ('t Hart et al., 1990, p. 78). In the intonation patterns found in the present study, we considered the initial and the final patterns of pitch movements. This approach appears to be adequate for the description of the large majority of concrete combinations of pitch movements produced in the emotional speech. This same approach was also used in

synthesis. Indeed, in order to study the relevance of the choice of intonation patterns for conveying emotion in speech, it was necessary to synthesize the corresponding pitch contours. It appeared that, in the existing literature, the timing and duration of some pitch movements were not precisely specified. Therefore, further specifications of some of the patterns of pitch movements in this intonation grammar had to be estimated, in terms of the exact timing and duration of some of the pitch movements. They are reported in Chapter IV (Table 4). It would be useful to systematically evaluate the adequacy of these specifications perceptually. However, for now, an extension of this grammar for Dutch did not appear necessary for the description or the synthesis of speech in different emotions. The distinctive features presented in this grammar appeared to be sufficient. Only the specifications given for standardized movements in neutral speech for excursion size and declination need adaptation. Excursion size and level of anchoring of the baseline differ from emotion to emotion. Declination still needs to be investigated, certainly, also in combination with the presence or absence of final lowering.

Perceptually based clusters of intonation patterns were derived from the results of the study. The intonation patterns '1&A 1&A', '1B 1&A', '1D 1&A', and '12 1&A', form one cluster. A second cluster contains the patterns '1&A 3C', '1B 3C', and '1D 3C', and a third cluster the '1A' and '1EA' patterns. Finally, both '1&A 12' and '1 5&A' each formed separate clusters by themselves; they did not cluster with any other intonation patterns. This clustering clearly shows that when considering the influence of the intonation patterns on the perception of emotions in speech, it is mostly the final part of these intonation patterns that is relevant to the listener. Furthermore, the partial correspondence of these clusters of intonation patterns with the *basic* intonation patterns for Dutch corroborates the idea that *basic* intonation patterns can convey different emotions in speech. Tentatively, clusters '1...1&A', '1...A', and '1 5&A', correspond to the /1A/ *basic* intonation pattern, while cluster '1...3C' corresponds to the /3C/ *basic* intonation pattern and cluster '1&A 12' with the /2/ *basic* intonation pattern.

Moreover, one can wonder whether making use of another approach to intonation research would have led to similar results. A simple comparison between intonation labeling in the IPO approach and in the auto-segmental phonological approach is, however, not obvious. A direct equivalent to the second level of description in the IPO approach is, to our knowledge, not given for the auto-segmental phonological description of Dutch intonation. There is, thus, no universally agreed rule for transcribing the three patterns of pitch

movements, '1&A', '1B', and '1D', into their auto-segmental counterparts. These patterns, which all correspond with the first accent produced in the utterances, could correspond, if simplification is allowed, with a bitonal H*L accent in the auto-segmental description of intonation. Since all intonation patterns that were different only in this first part were grouped into the same clusters, it is to be expected that, using an auto-segmental transcription system, even if it makes different distinctions in the group of '1&A', '1B', and '1D' patterns, will lead to the same general conclusions.

Another point is that, for the study of the relative duration of accented and unaccented speech segments in emotional speech, a reference was determined in terms of a description of the relative duration of accented and unaccented speech segments, as a function of a continuum of overall speech rates in non-emotional speech. Although the differences in durational proportion of accented and unaccented speech segments, as a function of overall speech rate, appeared not to be perceptually relevant to the expression of emotion in speech, it cannot be excluded that these variations below utterance level fulfill a function in communication. They may provide a variability in the speech which enhances naturalness without conveying a specific meaning. These variations in neutral speech are described in Chapter V.

Finally, the approach involving the successive analysis of natural speech, the re-synthesis of speech allowing manipulations of natural speech, and the rule-based synthesis of speech, constitutes a valuable methodological background for the present study. Complementary studies of production and perception were felt to be a necessary prerequisite for establishing the communicative significance of the investigated speech parameters. This need for a combination of both production and perception studies, in order to make progress in understanding emotional vocal communication, was already expressed by Scherer (1991). A certain type of variation can be perceptually relevant, while not systematically showing up in the production data. In this study, this was more or less the case for the relative duration between accented and unaccented speech segments. Indeed, the efficient use of a particular parameter can allow the relaxation of another one. That other parameter can be potentially relevant. On the other hand, in a perception study, the perceptual relevance of a cue observed in a production study may be obscured by the experimental set-up or the effect of another cue. Alternatively, the cue can simply be relevant to an aspect of the communication process other than the cue studied. On the other hand, the correspondence

of results obtained in production and perception studies firmly establishes the communicative importance of the parameters being studied.

III. RESEARCH PERSPECTIVES

The present study provides a few encouraging answers to the numerous questions concerning the expression of emotion in speech. Since this research topic involves, various disciplines such as psychology, linguistics, phonetics, engineering, and physiology, there is a great variety of interesting options for further research on this topic. We will restrict ourselves here to lines of studies directly suggested as a follow-up of the present study.

The first obvious investigation that comes to mind, as a continuation of the present line of research, would be a perception experiment testing the combined results of the thesis work in rule-based synthetic speech. One option for extending the present study would be to use other speech material, both in production studies and in a series of perception tests involving re-synthesized speech and, ultimately, synthetic speech. The size and the complexity of the speech material could be increased. This increased complexity could imply the study of longer sentences with a variety of pitch accents, where pauses within sentences are also included.

The study of speech parameters, such as voice quality and intensity, which are not considered in this study but are certainly relevant to the vocal expression of emotion, would also be an interesting follow-up. In order to allow a proper comparison of results across studies, it is important for a first estimation of optimal values, to analyze speech material corresponding to predetermined standards and to carry out investigations in a controlled methodological way. Such choices imply restrictions, as can be seen in the present study, and make it necessary to conduct follow-up studies for testing the generalization of results on other speech material.

Furthermore, determining the range of variations of the acoustic properties associated with the expression of emotion in speech, provides knowledge about the range within which the parameters studied can be varied *without* conveying emotion. It is expected that adding variability to synthetic speech would result in generating better sounding, more natural synthetic speech. This additional variability does not only consist of variations conveying emotion, but also includes variations within the range of neutral speech. The improvement

expected in terms of naturalness of the synthetic speech could be tested in a perceptual evaluation. Additionally, defining the range of variations of the acoustic properties associated with the expression of emotion also implies determining the maximal range of variations conveying emotion in speech. Emotions can be perceived as more or less intense, but there is a limit to this perceived intensity. An exaggerated expression will be perceived as a caricature. Such caricatured expressions may, however, be of interest for modeling, for instance, the speech of characters in animated cartoons. It should be considered that a turning-point might be found, after which variations might be experienced as so extreme, that speech carrying such variations might simply sound unnatural, instead of conveying emotion.

Investigating how the expression of emotion in speech differs in spontaneously expressed emotions and in acted emotions, is another direction of research. An intermediary alternative between spontaneous and acted speech would be to record the carefully elicited vocal expression of emotions in optimal situations.

The speech community will benefit from the development of a standard protocol for the recording of databases. Such a protocol will be particularly useful for recordings of emotional speech realized under comparable conditions. Including a description of the recording conditions themselves, as well as a description of elicitation procedures would certainly be useful. We could obtain comparable recordings of speech produced in elicited emotional states, by standardizing the way the emotions are elicited. The freedom of movements left to the speakers for bodily expression, and, as a consequence, practical details such as the position and orientation of the microphone towards the speaker should be described. Hopefully, a clear protocol could stimulate people to include emotional utterances in databases that are recorded for more general use.

Additionally, it is a fact that there is a lack of a generally agreed upon definition of emotion, and that there are many ways of expressing the same emotion depending on factors such as the personality of the speaker, the situation in which the spoken communication takes place, and the social background of the speaker. It would therefore be convenient to include in a single study several different renditions of the same emotion by the same speakers. It would be interesting to observe variations within a single emotion. This would lead us to consider how a speaker realizes less typical expressions of specific emotions, conveying,

for instance, boredom while speaking rather quickly. This could help us to shed some light on some of the correlations among parameters.

Furthermore, identifying and quantifying relevant parameters leads to formulating results in terms of sets of rules for the generation of synthetic speech, which seems to increase the potential usability of results. Moreover, such a formulation of our understanding of speech variability in an extended set of rules should ultimately allow us to cope with the synthesis of different voices, of males and females of different ages, varying in, for instance, regional and social backgrounds.

Involving numerous speakers in the investigations is more and more common in the field of speech research, and worthwhile for generalizing the findings. Whether and how the findings correspond or diverge among different languages and different cultures is an intriguing question. Wouldn't it be interesting, for instance, to know how the results of this study conducted in Dutch would compare to results from a study conducted in a Romance language? And what about investigations involving speakers and listeners not sharing a common language and/or cultural background?

In the present study, the sentences on which emotions were uttered were the same in all emotions. Moreover, the semantic content of the utterances was as neutral as possible. Involving sentences of various semantic contents would allow another type of study, concerned with the interference of linguistic content and prosodic variations. This could lead, for instance, to an investigation of the contrastive realization of a particular emotion on sentences of semantic contents associated with a different emotion.

Anticipating possible applications of the results obtained in studies that are concerned with the expression of emotion in speech, it seems important to consult potential users in order to better understand their needs, and to evaluate the acceptability of synthetic speech. In particular, people with a sensory handicap, who have specific wishes about the possibilities of systems, could provide usual information for us to produce knowledge and products that indeed meet their needs.

Finally, the increasing interest in a multi-modal interaction between humans and machines also opens opportunities for interesting interdisciplinary research. Studying the correspondence of the speech variations with other forms of expression of emotion can be

done in collaboration with colleagues in different fields. An interdisciplinary approach would allow investigating the correspondence of speech variations associated with emotion, with other types of changes, for instance: in body gestures, postures, facial expressions, respiration and heart rate. Collaboration with colleagues working in the field of speech recognition, speaker identification, and speaker verification, should also be very rewarding. A fascinating field of ambitious investigations lies ahead of us in these multiple directions.

Bibliography

Allessandro, C. d', and Mertens, P. (**1995**). "Automatic pitch contour stylization using a model of tonal perception," *Computer Speech and Language, 9*, 257-288.

Arndt, H., and Janney, R. W. (**1991**). "Verbal, prosodic, and kinesic emotive contrasts in speech," *Journal of Pragmatics, 15*, 521-549.

Beckman, M. E. (**1997**). "Speech models and speech synthesis". In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.) *Progress in speech synthesis*, Springer-Verslag, New-York, 185-209.

Beckman, M. E., and Edwards, J. (**1990**). "Lengthenings and shortenings and the nature of prosodic constituency". In: J. Kingston and M. E. Beckman (Eds.) *Papers in laboratory phonology I: Between the grammar and physics od speech*, Cambridge University Press, Cambridge, 152-178.

Bezooijen, R. A. M. G. van (**1984**). *The characteristics and recognizability of vocal expression of emotion.* Foris, Dordrecht, The Netherlands.

Bouwhuis, D. G. (**1974**). "The recognition of attitudes in speech," *IPO Annual Progress Report, 9*, 82-86.

Boves, L. (**1984**). *The phonetic basis of perceptual ratings of running speech.* Foris, Dordrecht, The Netherlands.

Cahn, J. E. (**1990**). *Generating expression in synthesized speech.* Technical report, MIT Media Lab., Boston.

Carlson, R. (**1991**). Synthesis: modelling variability and constraints, *Proceedings Eurospeech'91, Genova, Italy, 3*, 1043-1048.

Carlson, R. (**1994**). "Models of speech synthesis". In: D. B. Roe and J. G. Wilpon (Eds.) *Voice communication between humans and machines,* National academy of sciences, Washington D. C., 116-134.

Carlson, R., Granström, B., and Nord, L. (**1992**). Experiments with emotive speech: acted utterances and synthesized replicas, *Proceedings ICSLP 92, Banff, Alberta, Canada, 1*, 671-674.

Caspers, J. (**1997**). Testing the meaning of four Dutch pitch accent types, *Proceedings Eurospeech'97, Rhodes, Greece, 2*, 863-866.

Charpentier, F., and Moulines, E. (**1989**). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Proceedings Eurospeech'89, Paris, France, 2*, 13-19.

Cohen, A., Collier, R., and Hart, J. 't (**1982**). "Declination: Construct or intrinsic feature of speech pitch?," *Phonetica, 39*, 254-273.

Cole, R. , Hirschman, L., Atlas, L., Beckman, M., Biermann, A., Bush, M., Clements, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D. G., Ostendorf, M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., and Zue, V. (**1995**). "The challenge of spoken language systems: research directions for the nineties," *IEEE Tansactions on speech and audio processing, 1 (3)*, 1-20.

Collier, R. (**1972**). *From pitch to intonation.* Catholic University of Leuven, Belgium.

Collier, R. (**1991**). "Multi-language intonation synthesis," *Journal of Phonetics, 19,* 61-73.

Cosmides, L. (**1983**). "Invariances in the acoustic expression of emotion during speech," *Journal of Experimental Psychology: Human Perception and Performance, 9,* 864-881.

Cummings, K. E., and Clements, M. A. (**1995**). "Analysis of the glottal excitation of emotionally styled and stressed speech," *Journal of the Acoustical Society of America, 98 (1),* 88-98.

Darwin, C. (**1872**). *The expression of the emotions in man and animals.* John Murray, London (Reprinted, University of Chicago, Chicago, 1965).

Davitz, J. R. (**1964**). "Auditory correlates of vocal expressions of emotional meanings". In: J. R. Davitz (Ed.) *The communication of emotional meaning*, McGraw-Hill, NY, 101-112.

Eagly, A. H., and Chaiken, S. (**1993**). *The psychology of attitudes.* Harcourt Brace Jovanovich, London.

Ekman, P. (**1982**). *Emotion in the human face, second edition.* Cambridge University Press, New York.

Ellison, J. W., and Massaro, D. W. (**1997**). "Featural evaluation, integration, and judgment of facial affect," *Journal of Experimental Psychology: Human Perception and Performance, 23,* 213-226.

Eskénazi, M. (**1993**). Trends in speaking styles research, *Proceedings Eurospeech'93, Berlin, Germany, 1,* 501-509.

Fienberg, S. E. (**1980**). *The analysis of cross-classified categorical data, second edition.* The MIT Press, Cambridge, Massachusetts.

Fairbanks, G., and Pronovost, W. (**1939**). "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Speech Monographs, 6,* 87-104.

Frick, R. W. (**1985**). "Communicating emotion: the role of prosodic features," *Psychological Bulletin, 97,* 412-429.

Friend, M., and Farrar, M. J. (**1994**). "A comparison of content-masking procedures for obtaining judgments of discrete affective states," *Journal of the Acoustical Society of America, 96 (3),* 1283-1290.

Frijda, N. H. (**1986**). *The emotions.* Cambridge University Press, Cambridge, England.

Fujisaki, H. (**1993**). "Dynamic characteristics of voice fundamental frequency in speech and singing". In: P. F. MacNeilage (Ed.) *The production of speech*, Springer, New York, 39-55.

Goldman-Eisler, F. (**1968**). *Psycholinguistics: Experiments in spontaneous speech.* Academic, New York.

Granström, B., and Nord L. (**1991**). Ways of exploring speaker characteristics and speaking styles, *Proceedings XIIth ICPhS, Aix-en-Provence, France, 4,* 278-281.

Gussenhoven, C., and Rietveld, T. (**1992/1993**). The influence of syllable composition on the alignment of pitch targets, *Proceedings Dept. Language and Speech, Nijmegen, The Netherlands, 16/17,* 91-95.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H., and Terken J. (**1997**). "The perceptual prominence of fundamental frequency peaks," *Journal of the Acoustical Society of America, 102 (5),* 3009-3022.

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

Hart, J. 't, and Collier, R. (**1975**). "Integrating different levels of intonation analysis," *Journal of Phonetics, 3,* 235-255.

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

Heuft, B., Portele, T. , and Rauth, M. (**1996**). Emotions in time domain synthesis, *Proceedings ICSLP 96, Philadelphia, PA, USA, 3,* 1974-1977.

Hermes, D. J. (**1988**). "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America, 83,* 257-264.

Hermes, D. J. (**1990**). "Vowel-onset detection," *Journal of the Acoustical Society of America, 87 (2),* 866-873.

Hermes, D. J., and van Gestel, J. (**1991**). "The frequency scale of speech intonation," *Journal of the Acoustical Society of America, 90,* 97-102.

Hermes, D. J. (**1997**). "Timing of pitch movements and accentuation of syllables in Dutch," *Journal of the Acoustical Society of America, 102,* 2390-2402.

Hermes, D. J., Beaugendre, F., and House, D. (**1997**). "Individual differences in accentuation boundaries in Dutch," *IPO Annual Progress Report, 32,* 131-138.

Heuven, V. J. van (**1994**). "Introducing prosodic phonetics". In: C. Odé and V. J. van Heuven (Eds.) *Experimental studies of indonesian prosody,* Leiden University, Leiden, 1-26.

Higuchi, N., Hirai, T., Sagasika, Y. (**1994**). Effect of speaking style on parameters of fundamental frequency contour, *Proceedings ESCA/IEEE workshop on speech synthesis, New York, USA,* 135-138

Hirose, K., Kawanami, H., and Ihara, N. (**1997**). Analysis of intonation in emotional speech, *Proceedings ESCA workshop on intonation: theory, models and applications, Athens, Greece,* 185-188.

Hoene, I. (**1996**). "Effects of speaker attitude on fundamental frequency and duration: an acoustic description". *Intern rapport of the department of phonetics,* Institute of linguistics, University of Umeå, Sweden.

House, D. (**1990**). *Tonal perception in speech.* Lund University Press, Lund.

House, D. (**1996**). Differential perception of tonal contours through the syllable, *Proceedings ICSLP 96, Philadelphia, PA, USA, 4,* 2048-2051.

Houtsma, A. J. M. (**1983**). "Estimation of mutual information from limited experimental data," *Journal of the Acoustical Society of America, 74 (5),* 1626-1629.

Izard, C. E. (**1977**). *Human emotions.* Plenum Press, New York.

Izard, C. E., and Saxton, P. M. (**1988**). "Emotions". In: R. C. Atkinson, R. J. Herrnstein, G. Lindzey, and R. D. Luce (Eds.) *Stevens' handbook of experimental psychology, second edition, Vol. 1.* A Wiley-Interscience Publication, Chichester, 627-676.

Kitahara, Y., and Tohkura, Y. (**1990**). The role of temporal structure of speech in word perception and spoken language understanding, *Proceedings ICSLP 90, Kobe, Japan, 1*, 389-392.

Kitahara, Y., and Tohkura, Y. (**1992**). "Prosodic control to express emotions for man-machine interaction," *IEICE Transactions on Fundamentals of Electronics, communications and computer sciences, 75*, 155-163.

Klasmeyer, G., and Sendlmeier, W. F. (**1995**). Objective voice parameters to characterize the emotional content in speech, *Proceedings XIIIth ICPhS 95, Stockholm, Sweden, 1,* 182-185.

Klatt, D. H. (**1975**). "Vowel lengthenings is syntactically determined in connected discourse," *Journal of Phonetics, 3,* 129-140.

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

Kozhevnikov, V. A., and Chistovich, L. A. (**1965**). *Speech: articulation and perception* (Translation). Joint Publications Research Service, Washington D. C.

Ladd, D. R. (**1993**). "On the theoretical status of 'the baseline' in modelling intonation," *Language and speech, 36 (4),* 435-451.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergman, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation contour type, voice quality, and F_0 range in signalling speaker affect," *Journal of the Acoustical Society of America, 78*, 435-444.

Laukkanen, A.- M., Vilkman, E., Alku, P., and Oksanen H. (**1997**). "On the perception of emotions in speech: the role of voice quality," *Journal of Logopedics and Phoniatrics Vocology 22(4),* 157-168.

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

Leinonen, L., Hiltunen, T., Linnankoski, I., and Laakso, M.- L. (**1997**). "Expression of emotional-motivational connotations with a one-word utterance," *Journal of the Acoustical Society of America, 102 (3),* 1853-1863.

Lieberman, P., and Michaels, S. B. (**1962**). "Some aspects of fundamental frequency and envelope amplitude as related to emotional content of speech," *Journal of the Acoustical Society of America, 34*, 922-927.

Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., and Miller, M. (**1985**). "Measures of the sentence intonation of read and spontaneous speech in American English," *Journal of the Acoustical Society of America, 77 (2),* 649-657.

Liénard, J.- S. (**1995**). "From speech variability to pattern processing: a non-reductive view of speech processing". In: C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen (Eds.) *Levels in speech communication: relations and interactions*, Elsevier Science, Amsterdam, 137-148.

Maeda, S. (**1976**). *A characterization of American English intonation.* MIT Press, Cambridge, Massachussets.

Miller, G. A. (**1954**). "Note on the bias of information estimates". In: H. Quastler (Ed.) *Information theory in psychology,* The Tree Press, Glencoe, Ill.

Miller, J. L. (**1981**). "Effects of speeking rate on segmental distinctions". In: P. D. Eimas and J. L. Miller (Eds.) *Perspectives on the study of speech* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 39-74.

Morlec, Y., Bailly, G., and Aubergé, V. (**1997**). Generating the prosody of attitudes, *Proceedings ESCA workshop on intonation: theory, models and applications, Athens, Greece,* 251-254.

Moulines, E., and Laroche, J. (**1995**). "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication, 16*, 175-205.

Mozziconacci, S. J. L. (**1995**). Pitch variations and emotions in speech, *Proceedings XIIIth ICPhS 95, Stockholm, Sweden, 1,* 178-181.

Murray, I. R. (**1989**). *Simulating emotion in synthetic speech.* University of Dundee, Scotland, UK.

Murray, I. R., and Arnott, J. L. (**1993**). "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," *Journal of the Acoustical Society of America, 93*, 1097-1108.

Nooteboom, S. G. (**1972**). "The perceptual reality of some prosodic duration," *Journal of Phonetics, 1,* 25-45.

Oller, D. K. (**1973**). "The effect of position in utterance on speech segment duration in English," *Journal of the Acoustical Society of America, 54 (2),* 1235-1247.

Os, E. den (**1988**). *Rhythm and tempo of Dutch and Italian; a contrastive study.* Elinkwijk, Utrecht, The Netherlands.

Protopapas, A., and Lieberman, P. (**1995**). Effects of vocal F_0 manipulations on perceived emotional stress, *Proceedings ESCA-NATO workshop on speech under stress, Lisbon, Portugal,* 1-4.

Protopapas, A., and Lieberman, P. (**1997**). "Fundamental frequency of phonation and perceived emotional stress," *Journal of the Acoustical Society of America, 101 (4),* 2267-2277.

Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," *Journal of the Acoustical Society of America, 32*, 693-703.

Pierrehumbert, J. (**1979**). "The perception of fundamental frequency declination," *Journal of the Acoustical Society of America, 66 (2),* 363-369.

Pierrehumbert J. (**1980**). *The phonology and phonetics of English intonation.* Massachusetts Institute of Technology, Cambridge, Massachusetts.

Pierrehumbert, J. (**1981**). "Synthesizing intonation," *Journal of the Acoustical Society of America, 70 (4),* 985-995.

Pisoni, D. B. (**1997**). "Perception of synthetic speech". In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg (Eds.) *Progress in speech synthesis*, Springer-Verslag, New-York, 541-560.

Pijper, J.- R. de (**1983**). *Modelling British English intonation: an analysis by resynthesis of British English intonation.* Foris, Dordrecht, The Netherlands.

Plutchik, R. (**1980**). *Emotion: a psychoevolutionary synthesis.* Harper & Row, New York.

Pols, L. (**1998**). "Foreword". In: R. Sproat (Ed.), *Multilingual text-to-speech synthesis: the Bell Labs Approach*, Kluwer academic publishers, Massachusets, XXIII-XXIV.

Repp, B. H., Rump, H. H., and Terken, J. M. B. (**1993**). "Relative perceptual prominence of fundamental frequency peaks in the presence of declination," *IPO Annual Progress Report, 28*, 59-62.

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

Rijnsoever, P. van (**1988**). "A multilingual text-to-speech system," *IPO Annual Progress Report, 23*, 34-39.

Selting, M. (**1994**). "Emphatic speech style - with special focus on the prosodic signalling of heightened emotive involvement in conversation," *Journal of Pragmatics, 22*, 375-408.

Shannon, C. E., and Weaver, W. (**1949**). *The mathematical theory of communication.* University of Illinois Press, Urbana.

Scherer, K. R., Ladd, D. R., and Silverman, K. E. A. (**1984**). "Vocal cues to speaker affect: Testing two models," *Journal of the Acoustical Society of America, 76 (5),* 1346-1356.

Scherer, K. R. (**1986**). "Vocal affect expression: a review and a model for future research," *Psychological Bulletin, 99*, 143-165.

Scherer, K. R. (**1989**). "Vocal measurement of emotion". In: R. Plutchik and H. Kellerman (Eds.), *Emotion: theory, research, and experience, Volume 4.* Academic Press, San Diego, 233-259.

Scherer, K. R. (**1991**). "Emotion expression in speech and music". In: J. Sundberg, L. Nord, and R. Carlson (Eds.), *Music, language, speech, and brain*, Wenner-Gren Center International Symposium Series, MacMillan, London, 146-156.

Schlosberg, H. (**1954**). "Three dimensions of emotion," *Psychological review, 61*, 81-88.

Shriberg, E., Ladd, D. R., Terken, J., and Stolcke, A. (**1996**). Modeling pitch range variation with and across speakers: predicting F_0 targets when "speaking up", *Proceedings ICSLP 96, Philadelphia, PA, USA, Addendum*, 1-4.

Siegwart, H., and Scherer, K. R. (**1995**). "Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi," *Journal of Voice 9 (3),* 249-260.

Sluijter, A. M. C. (**1991**). "Een perceptieve evaluatie van een model voor alinea-intonatie met synthetische spraak" (A perception evaluation of a model for paragraph intonation with synthetic speech). *Intern IPO Report 801,* Institute for Perception Research, Eindhoven, The Netherlands.

Sorensen, J. M., and Cooper, W. E. (**1980**). "Syntactic coding of fundamental frequency in speech production". In: R. Cole (Ed.) *Perception and production of fluent speech*, Erlbaum Ass., Hillsdale, NJ, 339-440.

Sundberg, J. (**1987**). *The science of the singing voice.* Northern Illinois University Press, Dekalb, Illinois.

Swerts, M., Bouwhuis, D. G., and Collier R. (**1994**). "Melodic cues to the perceived "finality" of utterances," *Journal of the Acoustical Society of America, 96 (4),* 2064-2075.

Taylor, P. (**1994**). "The rise/fall/connection model of intonation," *Speech Communication, 15*, 169-186.

Terken, J. M. B. (**1991**). "Fundamental frequency and perceived prominence of accented syllables," *Journal of the Acoustical Society of America, 89*, 1768-1776.

Terken, J. M. B. (**1994**). " Fundamental frequency and perceived prominence of accented syllables II: Non-final accents," *Journal of the Acoustical Society of America, 95*, 3662-3665.

Terken, J., and Hermes, D. J. (**forthcoming**). "The perception of prominence". In: M. Horne (Ed.), *Prosody: Theory and experiment*, Kluwer Academic Publishers, Dordrecht.

Tischer, B. (**1995**). Acoustic correlates of perceived emotional stress, *Proceedings ESCA-NATO workshop on speech under stress, Lisbon, Portugal*, 29-32.

Traunmüller, H., and Eriksson, A. (**1995**). "The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations," *Journal of the Acoustical Society of America, 97(3)*, 1905-1915.

Verhelst, W., and Borger, M. (**1991**). Intra-speaker transplantation of speech characteristics: an application of waveform vocoding techniques and DTW, *Proceedings Eurospeech'91, Genova, Italy, 3*, 1319-1322.

Williams, C. E., and Stevens, K. N. (**1972**). "Emotions and speech: some acoustical factors," *Journal of the Acoustical Society of America, 52*, 1238-1250.

Zelle, H. W., Pijper, J.- R. de, and Hart, J. 't (**1984**). Semi-automatic synthesis of intonation for Dutch and British English, *Proceedings Xth ICPhS, Utrecht, The Netherlands, IIB*, 247-251.

Summary

Experiences in every-day life illustrate that the contents of spoken communication are not restricted to *what is said*, but also involve *how it is said*. A huge number of variations occur in speech, so that saying a sentence twice does never result in exactly the same acoustic realization. This might lead a listener to interpret the two utterances as two different messages. Speakers exploit this freedom to vary speech components in order to express themselves, and listeners take this variation into account when decoding the spoken message. Today's speech-synthesis systems do not compare with humans, even remotely, when it comes to exploiting prosodic variation. As a consequence, today's synthetic speech, despite the fact that it is considered reasonably intelligible, is also perceived as dull. It sounds rather unnatural and uninvolved. Modeling variability in synthetic speech is expected to enhance its quality and, therefore, to increase its potential use. The scale of variation involved in speech produced in emotional states, is wide. Acquiring knowledge concerning these variations is expected to make it possible to model speech variation associated with emotion, as well as to model more moderate variation that is not so much associated with emotional involvement, but rather with enhancing naturalness in neutral utterances.

In the present study, the variation of the prosodic elements: pitch level, pitch range, intonation pattern, and speech rate was investigated in the vocal expression of emotion. These parameters are considered to have a major contribution in conveying emotions. In order to be able to use the results of the present study in speech synthesis, it is of relevance not only to describe the speech variation qualitatively, but also to quantify it. Since utterances conveying neutrality are the usual output of speech-synthesis systems, it is also convenient to express variation in parameter values in terms of deviation from neutrality. In order to model only the speech variability as far as it is relevant to communication, the present investigations do not only include *production* studies, but also *perception* studies. An experimental approach is used, in which analyses of natural speech variation are carried out and perceptual tests involving synthetic or re-synthezised speech are performed, in order to test the relevance of the data found. Furthermore, the consideration of these variations in the framework of models commonly used in speech studies, allows the validity of these models to be tested.

In Chapter I, the problems at hand are described. The framework, in which studies concerned with the expression of emotion in speech are carried out, is depicted, approaches are discussed, and the approach adopted for the present study is presented. Finally, an outline of the investigation is given.

Chapter II deals with the selection of the speech material for use in the present study. The selection of 315 utterances (3 speakers × 5 sentences × 7 emotions × 3 trials) was based on appropriate emotion identifiability. A representative subset of these, consisting of 14 utterances (1 speaker × 2 sentences × 7 emotions × 1 trial), was intended for use in the preliminary analyses of Chapter II. The seven emotions: 'neutrality', 'joy', 'boredom', 'anger', 'sadness', 'fear', and 'indignation', were involved in the present investigation. The identification of these seven emotions in the original speech was tested in a perception test. The results form a useful basis for comparison with the results of later experiments. Next, the adequacy of the semantic content of the five sentences for use in this study was tested and confirmed. An analysis of the subset of fourteen utterances was then carried out at utterance level, by means of measurements of pitch level, pitch range, and speech rate. Additionally, these fourteen utterances were individually labeled in terms of intonation patterns, according to the Dutch grammar of intonation by 't Hart, Collier and Cohen (1990). A series of experiments was conducted in which pitch level, pitch range, and speech rate were systematically varied, per emotion, around the values found for these parameters in the original speech. The variation in intonation patterns was controlled by providing each test utterance with the same intonation pattern as in the original utterance of the corresponding emotion. Perception experiments were carried out, in which subjects ranked the utterances they found best for the expression of a specific emotion. On the basis of the results, optimal values for pitch level, pitch range, and speech rate were derived for the generation of emotional speech from a neutral utterance. These values were then perceptually tested, in experiments in which subjects labeled utterances with the name of one of the seven emotions. The first series of experiments involved resynthesized speech, while the last experiment involved rule-based synthetic speech. Applying the values that were found optimal, onto synthetic speech, lead to a good identification of the emotions, namely 63% correct identification. Although some emotions were less successfully identified than others, general results were quite encouraging. Results showed that pitch and speech rate are powerful cues for conveying emotion in speech.

In Chapter III, an extensive study was conducted, concerned with F_0 fluctuations produced in the expression of emotion, and with the relevance of perceived pitch variations for the identification of emotion in speech. Pitch level and pitch range were estimated on the basis of measurements of mean F_0 and its standard deviation in the 315 utterances in the database. It was shown that, after speaker normalization, the values found in natural utterances produced by the three speakers eliciting the seven emotions, closely matched the optimal values obtained in the perception tests of Chapter II. The course of pitch in all individual utterances was described in terms of the model of intonation by 't Hart, Collier and Cohen (1990), describing a pitch curve as a combination of a slowly decreasing component (the declination line) and relatively fast pitch movements, superimposed on this baseline. In this model, the end point of the declination line represents the pitch level, while the excursion size of the pitch movements represents the pitch range. In principle, this excursion size of the pitch movements is considered to be constant throughout the utterance, so that pitch curves could also be described with a lower declination line, or baseline, and an upper declination line, or topline, between which the pitch movements are realized. The overall excursion size of the pitch movements then equals the distance between the lower and the upper declination line. In Chapter III, the relationship was discussed between two ways of estimating pitch level and pitch range. One estimation was model-based, involving the end point of the baseline and the difference between baseline and topline, respectively. The other estimation, more strictly data oriented, was based on the average of F_0 in the utterances and the standard deviation of F_0, respectively. Furthermore, pitch level and pitch range can only be defined as properties over the whole utterance. In naturally produced pitch curves, many details can be distinguished which cannot be captured in such a model of intonation. In order to study the fluctuations of F_0 occurring within utterances, F_0 was measured at a number of fixed points in the utterances. Measurements were carried out in the first voiced part of the utterance, in the vowel of the first accent peak, in a vowel after the initial accent peak, in a vowel before the final accent peak, in the vowel of the last peak, and in the last voiced segments of the utterance. It appeared that utterances produced while conveying different emotions could vary considerably with regards to relative peak heights and the extent of final lowering. For instance, the F_0 measurements concerning the last accent peak often yielded a higher value than the measurements concerning the first peak, which cannot be accounted for on the basis of declination only. Especially for some emotions, the final measurement of F_0

yielded a lower value than could be expected on the basis of preceding measurement of F_0 that are expected to be representative of the baseline. In a perception study, the relevance of these differences was put to the test. Although some effects appeared to be significant, e.g., modeling final lowering appeared to increase the number of responses of the subjects indicating indignation, the effects found were relatively small.

The 315 utterances selected as speech material were labeled in terms of intonation patterns, and the distribution of the patterns of pitch movements over the various emotions was investigated per speaker. The results are presented in Chapter IV. It appeared that the patterns were not equally distributed over all seven emotions. The '1&A' pattern, a prominence-lending rise-fall, was the most often used pattern; it was regularly produced in all seven emotions. Therefore, the hypothesis emerged that this '1&A' pattern would be a good candidate to apply to all emotions, so that no variability is introduced by the realization of different intonation patterns. From the production study, however, it also appeared that many utterances were produced with other intonation patterns, and some intonation patterns seemed to be more characteristic for some emotions than for others. In particular, it was noticed that the patterns '12' (a rise followed by a very late rise) and '3C' (a late rise and a very late fall), were never used in final position in utterances expressing neutrality. A second hypothesis, therefore, emerged concerning the question of whether the two patterns '12' and '3C' could signal emotion in speech. A perception experiment was carried out, investigating the perceptual relevance of intonation patterns for identifying emotions in speech. This test provided converging evidence on the contribution of specific patterns in the perception of some of the emotions studied. Some intonation patterns introduced a perceptual bias towards a specific emotion. Finally, clusters of intonation patterns were derived from the results of the perception experiment. The last part of the pattern appeared to be of particular relevance. The clustering reflected the perceptual distinctions among intonation patterns.

In Chapter V, temporal variations conveying emotion in speech were investigated. First, an analysis of speech rate was performed at utterance level. Global measurements of overall sentence duration and its standard deviation were carried out on the 315 utterances selected as speech material. Averages per emotion were calculated for each speaker. It was investigated whether a linear approach, simply consisting of stretching or shrinking the

whole utterance linearly, i.e., manipulating the overall speech rate, is sufficient for expressing emotion in speech, or whether a more detailed approach would be necessary. To this end, an analysis was performed below utterance level. Measurements of relative duration of accented and unaccented speech segments (syllables or groups of syllbales) were made, in order to acquire some insight into the internal temporal structure of emotional utterances. Although differences are small and the analysis of production data did not provide conclusive evidence of the systematic use of variation in the internal temporal structure of utterances in speech conveying an emotion, some of the detailed information could not be described with a linear-stretch model. The perceptual relevance of separately stretching or shrinking speech segments within utterances was then questioned. The deviation from a linear model could either specifically be due to the expression of emotion, or simply due to the modification of overall speech rate and, therefore, be only indirectly related to the expression of emotion (i.e., only because emotion is conveyed with changes in overall speech rate). In order to obtain the reference required for deciding which interpretation is correct, the same measurements of relative duration of accented and unaccented speech segments were made in neutral speech, spoken at different overall speech rates, by one of the male speakers. The results of the measurements in emotional and in neutral speech were compared. The temporal structure appeared to change non-linearly and to vary with some of the emotions. An experiment was carried out in order to test the perceptual relevance of these variations. Speech manipulations were carried out in order to generate emotional speech, either by simply stretching or shrinking the whole utterance linearly, or by proportionally varying the duration of accented and unaccented speech segments. Values for relative durations tested in the experiment were inspired from the production data. The differences in relative duration of accented and unaccented speech segments that are *associated with speech rate,* appeared not to be perceptually relevant. On the other hand, the differences in relative duration of accented and unaccented speech segments that are *associated with the expression of emotion,* appeared to be perceptually very relevant for the expression of neutrality and indignation.

Finally, in Chapter VI, the limited research area of the present investigation is once again justified and the results of the study are summarized. It is concluded that an interaction of some prosodic cues permits the vocal expression of emotion, and that most emotions can be conveyed in synthetic speech by controlling the parameters studied here. For some

Relationships established between emotions and parameters, based on the production and/ or on the perception studies

Parameters	Emotions						
	neutrality	joy	boredom	anger	sadness	fear	indignation
pitch level	65 Hz	155 Hz	65 Hz	110 Hz	102 Hz	200 Hz	170 Hz
pitch range	5 s.t.	10 s.t.	4 s.t.	10 s.t.	7 s.t.	8 s.t.	10 s.t.
final lowering	-	-	no	-	yes	yes	yes
relative peak height	-	-	-	-	yes	yes	-
pattern(s) to prefer in final position	1&A	1&A and 5&A	3C	5&A, A and EA	3C	12 and 3C	especially 12, but also 3C
pattern(s) to avoid in final position	12 and 3C	A, EA, and 12	5&A and 12	1&A and 3C	5&A	A and EA	1&A
duration relative to neutrality	100%	83%	150%	79%	129%	89%	117%
durational proportion acc./unacc. segments	no deviation from linearity	-	-	-	-	-	stretch acc. segments 40% more than unacc.

emotions, however, this is less successful. For these emotions, other cues, such as voice quality, loudness or other properties of intonation, may be essential. The results that were found to be specific to the expression of emotion in speech are given as a series of rules for generating speech in each of the emotions studied. These rules are summarized in the table presented above , in which optimal values are mentioned for each emotion. A specification is also given of which patterns are preferred or should be avoided in the modeling of the emotions, and whether or not a modeling of final lowering and relative height of the peaks is expected to be relevant.

Additionally, general results concerning the suitability of models for handling the extreme variations occurring in emotional speech were summarized. The thesis is concluded by some suggestions of lines for future research concerned with the expression of emotion in speech.

Samenvatting

De dagelijkse ervaring leert dat de inhoud van gesproken communicatie niet beperkt is tot *wat er precies wordt gezegd*, maar ook afhangt van *hoe het wordt gezegd*. Natuurlijke spraak heeft talloze variatiemogelijkheden met als gevolg dat een zin die tweemaal wordt uitgesproken nooit resulteert in twee akoestisch identieke representaties. Dit kan ertoe leiden dat een luisteraar deze zinnen op verschillende wijze interpreteert. Deze vrijheid om onderdelen van spraak te variëren wordt door sprekers gebruikt om zich uit te drukken, en luisteraars houden hier rekening mee bij het decoderen van een gesproken boodschap. Moderne spraaksynthesesystemen lopen ver achter bij mensen in het exploiteren van dergelijke variaties, met name op het gebied van prosodie. Hedendaagse synthetische spraak wordt dan ook vaak als saai ervaren, ondanks het feit dat die spraak meestal redelijk goed verstaanbaar is. De spraak klinkt nogal onnatuurlijk en afstandelijk. Modellering van de variabiliteit kan de kwaliteit van synthetische spraak verbeteren en het gebruik ervan doen toenemen. Spraak geproduceerd onder emotionele omstandigheden wordt bij uitstek gekenmerkt door grote variatie. Kennis van deze variabiliteit maakt het mogelijk om zowel variatie geassocieerd met emoties in model te brengen, alsook variatie die geen verband met emoties heeft, b.v. variatie die de natuurlijkheid van neutrale uitingen bevordert.

Deze studie naar het vocaal tot uitdrukking brengen van emotie richt zich op de rol van de volgende prosodische elementen: toonhoogteniveau, toonhoogtebereik, intonatiepatroon, en spreeksnelheid. Deze elementen worden gezien als de belangrijkste dragers van emoties. Om resultaten van deze studie bruikbaar te maken voor spraaksynthese moet variatie in spraak niet alleen kwalitatief, maar ook kwantitatief worden beschreven. Bestaande spraaksynthesesystemen produceren doorgaans neutrale uitingen. Het ligt daardoor voor de hand om variatie in spraakparameters uit te drukken als afwijkingen ten opzichte van neutrale parameterwaarden. Om ervoor te zorgen dat alleen die spraakvariabiliteit die belangrijk is voor communicatie wordt gemodelleerd, zijn niet alleen spraak-*productie*analyses maar ook *perceptie*-experimenten uitgevoerd. Bij de experimentele aanpak worden variaties in natuurlijke spraak geanalyseerd, en worden waarnemingsexperimenten met synthetische of geresynthetiseerde spraak uitgevoerd om de perceptieve relevantie van de gevonden waarden te toetsen. Bovendien stelt de vergelijking

van gevonden variatie met bestaande, gebruikelijke spraakmodellen ons in staat de geldigheid van zulke modellen te toetsen.

Hoofdstuk I behandelt de algemene probleemstelling. Het raamwerk waarin studies over uitdrukking van emoties in spraak zijn uitgevoerd staat erin beschreven, alsook de aanpak, de uiteindelijk gekozen benadering, en een overzicht van de studie.

Hoofdstuk II beschrijft de selectie van het spraakmateriaal voor deze studie. Op basis van perceptieve identificatieresultaten werden 315 uitingen (3 sprekers × 5 zinnen × 7 emoties × 3 keer) geselecteerd. Een representatieve subset hiervan, bestaande uit 14 uitingen (1 spreker × 2 zinnen × 7 emoties), werd gebruikt voor een voorlopige analyse. De zeven emoties in kwestie waren: neutraliteit, blijheid, verveling, boosheid, verdriet, angst, en verontwaardiging. De perceptieve identificeerbaarheid van deze zeven emoties werd experimenteel onderzocht aan de hand van natuurlijke spraakuitingen. De resultaten vormden een goede vergelijkingsbasis voor experimenten die erop volgden. De semantische neutraliteit van de gebruikte zinnen werd onderzocht en bevestigd. Vervolgens werd op de 14 uitingen een globale analyse (over de gehele uiting) uitgevoerd, door het toonhoogteniveau, het toonhoogtebereik, en de spreeksnelheid te meten. Dezelfde 14 uitingen werden ook ingedeeld naar hun intonatiepatronen volgens de Nederlandse intonatiegrammatica van 't Hart, Collier en Cohen (1990). Een aantal experimenten werd uitgevoerd, waarin toonhoogteniveau, toonhoogtebereik en spreeksnelheid per emotie systematisch werden gevarieerd rondom de waarden van de natuurlijke spraak. De variatie in intonatiepatroon werd onder controle gehouden door iedere gemanipuleerde uiting hetzelfde intonatiepatroon te geven als in de oorspronkelijke uiting van de betreffende emotie. Perceptie-experimenten werden uitgevoerd, waarin proefpersonen een rangorde moesten bepalen voor de mate waarin uitingen een bepaalde emotie vertolkten. Aldus konden optimale waarden worden bepaald voor het toonhoogteniveau, het toonhoogtebereik en de spreeksnelheid. Deze waarden kunnen gebruikt worden voor het genereren van emotionele spraak uit neutrale uitingen. Deze optimale waarden werden perceptief getoetst d.m.v. experimenten waarin proefpersonen uitingen moesten identificeren m.b.t. de onderliggende emotie. De eerste serie experimenten betrof geresynthetiseerde neutrale spraak, de tweede serie regel gebaseerde synthetische spraak. Toepassing van deze optimale waarden op synthetische spraak leidde tot een gemiddelde

van 63% correcte identificaties. Ofschoon sommige emoties minder goed herkend werden dan andere, waren de resultaten over het algemeen bemoedigend. Ze toonden aan dat toonhoogte en spreeksnelheid belangrijke dragers voor emotie in spraak zijn.

In Hoofdstuk III wordt een uitgebreide studie beschreven van, enerzijds geproduceerde F_0 fluctuaties in het uitdrukken van emoties, en anderzijds de perceptieve relevantie van toonhoogtevariaties voor het identificeren van een uitgedrukte emotie. Toonhoogteniveau en -bereik werden benaderd op basis van metingen van gemiddelde en standaardafwijking van F_0 in de 315 uitingen uit de totale dataverzameling. Het bleek dat, na normalisatie m.b.t. de sprekers, de parameterwaarden die waren gevonden in de natuurlijke uitingen van de drie sprekers bij het uitdrukken van de zeven emoties heel dicht lagen bij de optimale waarden gevonden in de perceptietoets uit Hoofdstuk II. Het toonhoogteverloop in alle afzonderlijke uitingen werd beschreven in termen van het intonatiemodel van 't Hart, Collier en Cohen (1990). Hierin wordt het toonhoogteverloop beschreven als een combinatie van een langzaam dalende component (de declinatielijn) en snellere toonhoogtebewegingen die op deze lijn zijn gesuperponeerd. In dit model representeert het eindpunt van de declinatielijn het toonhoogteniveau, en representeert de excursiegrootte van de toonhoogtebewegingen het toonhoogtebereik. In principe wordt de excursiegrootte van toonhoogtebewegingen constant verondersteld tijdens de gehele uiting, zodat toonhoogtecontouren ook beschreven kunnen worden met een lage declinatielijn of basislijn, en een hoge declinatielijn of toplijn, waartussen toonhoogtebewegingen plaatsvinden. De globale excursiegrootte van de toonhoogte-bewegingen is dan gelijk aan de afstand tussen de lage en de hoge declinatielijn. Vervolgens worden twee manieren van schatten van toonhoogteniveau en toonhoogtebereik besproken. De ene manier van schatten is gebaseerd op een model en betreft achtereenvolgens het eindpunt van de basislijn en het verschil tussen bovenlijn en basislijn. De andere manier van schatten, die meer data-georiënteerd is, berust op de gemiddelde F_0 van de uiting, en op de standaardafwijking hiervan. Verder zijn toonhoogteniveau en toonhoogtebereik globale eigenschappen die de uiting als domein hebben. In natuurlijke spraak zijn er vele details in het gerealiseerde F_0 verloop te onderscheiden die onmogelijk gevat kunnen worden in zo'n intonatiemodel. Om F_0 fluctuaties binnen een uiting te kunnen bestuderen, werd de F_0 gemeten op een aantal vaste plaatsen binnen zo'n uiting. Metingen werden verricht in het eerste stemhebbende gedeelte van de uiting, in de klinker van de eerste piek, in een klinker na de initiële accentpiek, in een klinker voor de finale accentpiek,

in de klinker van de laatste piek, en in het laatste stemhebbende segment van de uiting. Het bleek dat uitingen gesproken onder verschillende emoties sterk van elkaar konden verschillen wat betreft de relatieve piekhoogte en de grootte van de toonhoogteverlaging aan het eind van de uiting. Zo leverden de F_0 metingen in de laatste accentpiek vaak een hogere waarde op dan F_0 metingen tijdens de eerste piek, hetgeen niet verklaarbaar is uitsluitend op grond van het concept van de declinatielijn. Voor sommige emoties leverde de eindmeting van F_0 een lagere waarde op dan verwacht kon worden op basis van voorafgaande F_0 metingen die representatief geacht konden worden voor de basislijn. De perceptieve relevantie van deze verschillen werd getoetst in een luisterexperiment. Ofschoon sommige verschillen belangrijk lijken, b.v. het feit dat toepassing van een eindverlaging van F_0 de respons 'verontwaardiging' lijkt te bevorderen, waren de effecten over het algemeen klein.

De 315 uitingen van het spraakmateriaal, werden vervolgens afzonderlijk gelabeld op grond van hun intonatiepatroon, en de verdeling van de patronen van toonhoogtebewegingen werd per spreker onderzocht voor iedere emotie. De resultaten zijn beschreven in Hoofdstuk IV. Het bleek dat de patronen niet gelijkelijk over de zeven emoties waren verdeeld. Het '1&A' patroon, een accent-verlenende stijging-daling, was het meest gebruikte patroon; bij alle emoties werd het regelmatig gebruikt. Dit leidde tot de hypothese dat het '1&A' patroon een goede kandidaat zou zijn om de variabiliteit die het gebruik van verschillende intonatiepatronen met zich meebrengt constant te houden. Uit de productiestudie bleek echter ook dat vele uitingen met een ander intonatiepatroon dan '1&A' waren geproduceerd, en sommige intonatiepatronen leken een bepaalde emotie beter te karakteriseren dan andere. In het bijzonder werd opgemerkt dat in neutrale uitingen de patronen '12' (een toonhoogtestijging gevolgd door een erg late toonhoogtestijging) en '3C' (een late toonhoogtestijging en een erg late toonhoogtedaling) nooit in het laatste deel van de uiting voorkwamen. Een tweede hypothese kwam daarom naar voren, namelijk dat deze twee patronen emotie in spraak kunnen signaleren. Er werd een perceptie-experiment uitgevoerd waarin de perceptieve relevantie van intonatiepatronen werd onderzocht voor het identificeren van emoties in spraak. Deze test toonde aan dat het gebuik van specifieke intonatiepatronen aan de waarneming van enkele van de emoties bijdraagt. Tenslotte werden er clusters van intonatiepatronen afgeleid uit de resultaten van het perceptie-experiment. Het laatste stuk van het patroon bleek van bijzondere betekenis te zijn. De clusters gaven een beeld van de perceptief distinctieve kenmerken van de intonatiepatronen.

Hoofdstuk V beschrijft de temporele variaties die emotie in spraak kunnen overbrengen. Ten eerste werd er een analyse van de spreeksnelheid uitgevoerd op het niveau van de uiting. Globale metingen van de gehele duur van een zin en de standaardafwijking ervan werden uitgevoerd op de 315 geselecteerde uitingen. Per emotie werden voor elke spreker de gemiddelden berekend. Er werd onderzocht of een lineaire benadering, eenvoudigweg bestaande uit het linear oprekken en inkrimpen van de uiting, d.w.z. het manipuleren van de globale spreeksnelheid, voldoende was voor het uitdrukken van emotie in spraak, of dat een meer gedetailleerde benadering nodig zou zijn. Hiertoe werd er tevens een analyse uitgevoerd op een niveau lager dan dat van de gehele uiting. Metingen van de relatieve duur van geaccentueerde en ongeaccentueerde spraaksegmenten (lettergrepen of lettergreepgroepen) werden uitgevoerd om inzicht te verkrijgen in de interne temporele structuur van emotionele uitingen. De verschillen waren klein en de analyse van productiegegevens toonde niet overtuigend aan dat de variatie van de interne temporele structuur van spraakuitingen systematisch werd gebruikt bij het overbrengen van emotie. Toch konden niet alle details beschreven worden binnen het lineaire model. De perceptieve relevantie van het lokaal rekken en krimpen werd ter discussie gesteld. Het afwijken van het lineaire model zou specifiek het gevolg kunnen zijn van de uitdrukking van een emotie, maar het zou ook een gevolg kunnen zijn van het veranderen van de algehele spreeksnelheid en daardoor alleen indirect gerelateerd zijn aan het uitdrukken van emotie (d.w.z. alleen omdat emotie wordt overgedragen door veranderingen in algehele spreeksnelheid). Om tot een referentie te komen aan de hand waarvan bepaald kan worden welke interpretatie juist is, werden dezelfde metingen van de relatieve duur van geaccentueerde en ongeaccentueerde spraak uitgevoerd in neutrale spraak die met verschillende snelheden werd uitgesproken door één van de mannelijke sprekers. Het resultaat van de metingen in emotionele en in neutrale spraak werd met elkaar vergeleken. De temporele structuur bleek bij sommige emoties niet lineair te veranderen en te variëren afhankelijk van de emotie. Een experiment werd uitgevoerd om de perceptieve relevantie van deze variaties te toetsen. Spraakmanipulaties werden uitgevoerd teneinde emotionele spraak te genereren door het eenvoudig lineair rekken en krimpen van de gehele uiting, alsook door de duur van geaccentueerde en ongeaccentueerde spraaksegmenten afzonderlijk te variëren. Waarden voor de relatieve duur werden ontleend aan de productiedata, en vervolgens verhoogd c.q. verlaagd. De verschillen in relatieve duur tussen geaccentueerde en ongeaccentueerde

Relaties gevonden tussen spraakparameters en emoties, gebaseerd op de productie en/of perceptie data

Parameters	Emoties						
	neutral.	blij.	verveling	boos.	verdriet	angst	verontw.
toonhoogte-niveau	65 Hz	155 Hz	65 Hz	110 Hz	102 Hz	200 Hz	170 Hz
toonhoogte-bereik	5 s.t.	10 s.t.	4 s.t.	10 s.t.	7 s.t.	8 s.t.	10 s.t.
finale daling	-	-	nee	-	ja	ja	ja
relatieve piekhoogte	-	-	-	-	ja	ja	-
patronen te verkiezen in laatste deel uiting	1&A	1&A en 5&A	3C	5&A, A en EA	3C	12 en 3C	in het bijz. 12, ook 3C
patronen te vermijden in laatste deel uiting	12 en 3C	A, EA, en 12	5&A en 12	1&A en 3C	5&A	A en EA	1&A
relatieve duur t.o.v. neutraliteit	100%	83%	150%	79%	129%	89%	117%
proportionele duur van geacc. en ongeacc. segmenten	niet van lineariteit afwijken	-	-	-	-	-	geacc. segmenten 40% meer rekken dan ongeacc.

spraaksegmenten die waren *geassocieerd met spreeksnelheid* bleken perceptief niet relevant. Daartegenover, bleken de verschillen in duur tussen geaccentueerde en ongeaccentueerde spraaksegmenten die waren *geassocieerd met het uitdrukken van emotie* perceptief wel zeer relevant bij het uitdrukken van neutraliteit en verontwaardiging.

Tenslotte wordt in Hoofdstuk VI het begrensde onderzoeksgebied van de huidige studie nader toegelicht, en werden de resultaten ervan samengevat. Er wordt geconcludeerd dat het zich vocaal uitdrukken inclusief emotie mogelijk is dankzij een interactie van prosodische

kenmerken, en dat de meeste emoties in synthetische spraak overgebracht kunnen worden door het aansturen van de alhier bestudeerde parameters. Voor enkele emoties is dit minder succesvol. Voor deze emoties zouden andere kenmerken, zoals stemkwaliteit, luidheid, of andere intonatie-eigenschappen, wel eens essentieel kunnen zijn. De gevonden resultaten die specifiek zijn voor de uitdrukking van emotie in spraak, worden gepresenteerd als een serie regels om synthetische spraak met elk van de bestudeerde emoties te genereren. Deze regels die overeenkomen met de relaties die gevonden zijn tussen spraakparameters en emoties zijn samengevat in de bovenstaande tabel, waarin voor elke emotie de optimale waarden vermeld worden, en waarin tevens gespecificeerd wordt welk patroon de voorkeur verdient of vermeden dient te worden bij het modelleren van emoties, en wanneer het modelleren van finale daling van de toonhoogte en de relatieve hoogten van de pieken verwacht wordt echt van belang te zijn.

Bovendien wordt in het laatste hoofdstuk een samenvatting gegeven van de algemene resultaten met betrekking tot modellen waarmee de met emotie samenhangende extreme spraakvariaties beschreven kunnen worden. Het proefschrift wordt afgesloten met enkele suggesties voor toekomstige onderzoekslijnen betreffende de uitdrukking van emotie in spraak.

Résumé

L'expérience quotidienne nous permet de constater que, dans la communication orale, le message passe non seulement par *ce qui est dit,* mais aussi par *la façon dont c'est dit.* La parole naturelle est si riche en variations que le même texte énoncé deux fois résulte systématiquement en deux réalisations acoustiques différentes, ce qui peut porter un interlocuteur à les interpréter comme porteuses de deux messages différents. Les locuteurs exploitent, pour s'exprimer, cette liberté de varier les composantes de la parole et les interlocuteurs tiennent compte de ces variations lorsqu'ils décodent le message parlé. Actuellement, la parole synthétique est encore loin d'égaler la parole humaine en ce qui concerne l'exploitation des variations prosodiques. Cette parole de synthèse, bien que raisonnablement intelligible, est ennuyeuse et artificielle. La modélisation de la variabilité dans la parole de synthèse devrait en améliorer la qualité, et favoriser ainsi son utilisation. La parole produite en situation émotionnelle présente un très large éventail de variations. Comprendre le fonctionnement de ces variations devrait nous permettre d'obtenir, d'une part, un modèle de variations associées à l'émotion, d'autre part, un modèle de variations plus modérées, indépendantes de l'émotion, rendant, par exemple, les phrases synthétiques neutres plus naturelles.

Dans la présente étude, le rôle des paramètres prosodiques : hauteur mélodique, dynamique fréquentielle, configuration intonative et débit de parole a été étudié pour l'expression de l'émotion dans la parole. Ces paramètres sont considérés contribuer largement à la communication des émotions. Il est souhaitable, afin d'obtenir des résultats utilisables en synthèse de parole, de décrire les variations non seulement de façon qualitative, mais aussi de façon quantitative. Puisque les systèmes de synthèse de parole produisent généralement des phrases à connotation neutre, il importe d'exprimer les variations des paramètres en termes de déviation par rapport à une norme correspondant à l'expression neutre. Afin de ne prendre en compte que la variabilité pertinente dans la communication, la présente thèse comprend aussi bien des études de *production* que des études de *perception.* Une approche expérimentale a été utilisée, consistant à effectuer des analyses portant sur les variations présentes dans la parole naturelle, et à vérifier la pertinence des données ainsi obtenues en effectuant des expériences de perception impliquant soit de la parole synthétique, soit de la parole manipulée par re-synthèse. De plus, le fait de considérer ces variations dans le cadre

de modèles communément utilisés dans les études sur la parole, nous permet de tester la validité de ces modèles.

Le Chapitre I traite de la problématique, et cerne le cadre de référence des recherches portant sur l'expression dans la parole. Différentes approches sont présentées et celle adoptée dans la présente étude est commentée. Finalement, les grandes lignes de notre étude sont exposées.

Dans le Chapitre II, la sélection de 315 phrases (3 locuteurs × 5 textes × 7 émotions × 3 fois) a été effectuée pour notre étude, en prenant comme critère de sélection l'identification de l'émotion dans les phrases. Parmi ces 315 phrases, 14 (1 locuteur × 2 textes × 7 émotions × 1 fois) ont été retenues pour des analyses préliminaires. Sept émotions ont été considérées : 'la neutralité', 'la joie', 'l'ennui', 'la colère', 'la tristesse', 'la peur' et 'l'indignation'. L'identification de ces sept émotions dans la parole naturelle a été testée expérimentalement. Les résultats forment une base de comparaison utile pour les expériences ultérieures. L'adéquation du contenu sémantique des phrases utilisées a ensuite été testée et confirmée. Les 14 phrases ont été soumises à une analyse globale, consistant à mesurer, au niveau de la phrase entière, la hauteur mélodique, la dynamique fréquentielle et le débit de parole. Ces 14 phrases ont, en outre, été étiquetées en termes de configurations intonatives selon la grammaire intonative pour le Néerlandais de 't Hart, Collier et Cohen (1990). Une série d'expériences a ensuite été effectuée. Dans ces expériences, pour chaque émotion, la hauteur mélodique, la dynamique fréquentielle et le débit de parole ont été variés systématiquement autour des valeurs trouvées pour ces paramètres dans la parole naturelle. Pour éviter la variation en configuration intonative, toutes les phrases-test correspondant à une même émotion ont été pourvues d'une même configuration intonative, en l'occurrence celle qui avait été produite dans la phrase naturelle correspondant à l'émotion considérée. Lors de ces expériences de perception, les sujets devaient sélectionner et ordonner les phrases qui, selon eux, exprimaient le mieux une émotion donnée. Sur la base des résultats obtenus, on a pu déterminer des valeurs optimales pour la hauteur mélodique, la dynamique fréquentielle et le débit de parole. Afin de tester si ces valeurs permettent effectivement de générer, à partir de phrases neutres, de la parole perçue comme porteuse d'émotion, une nouvelle série d'expériences de perception a été effectuée, au cours de laquelle les sujets devaient signaler, parmi les sept émotions proposées, celle

qu'ils pensaient être exprimée dans la phrase écoutée. Dans ces tests, de la parole manipulée par re-synthèse a d'abord été utilisée, puis de la parole de synthèse. L'utilisation, en parole de synthèse, des valeurs trouvées optimales, a permis d'obtenir 63% d'identification correcte de l'émotion voulue. Bien que certaines émotions aient été moins bien identifiées que d'autres, l'ensemble des résultats est assez encourageant. Ces résultats montrent bien que les variations en fréquence fondamentale et en débit de parole sont des éléments puissants pour exprimer l'émotion dans la parole.

Dans le chapitre III, une étude extensive a été consacrée aux fluctuations en fréquence fondamentale (F_0) produites lors de l'expression de l'émotion, et à la pertinence des variations mélodiques perçues pour l'identification des émotions dans la parole. Une estimation de la hauteur mélodique et de la dynamique fréquentielle, basée sur des mesures de F_0 moyenne et d'écart type de F_0 a été effectuée dans les 315 phrases. On a constaté qu'après normalisation concernant les locuteurs, les valeurs obtenues à l'issue de l'analyse de production correspondent étroitement aux valeurs optimales obtenues grâce aux tests de perception du chapitre II. Les courbes mélodiques de chacune des phrases ont été décrites dans le cadre du modèle d'intonation de 't Hart, Collier et Cohen (1990). Ce modèle décrit la courbe mélodique comme la combinaison d'une composante décroissant lentement - la ligne de déclinaison - et de mouvements mélodiques relativement rapides surimposés sur cette ligne de déclinaison. Dans ce modèle, la fin de la ligne de déclinaison représente la hauteur mélodique, alors que la dimension de l'excursion des mouvements mélodiques représente la dynamique fréquentielle. En principe, la dimension des mouvements mélodiques est considérée comme constante dans toute la phrase. Les courbes mélodiques peuvent donc être décrites en utilisant une ligne de déclinaison inférieure - 'baseline' -, et une ligne de déclinaison supérieure - 'topline' - , entre lesquelles les mouvements mélodiques sont réalisés. La dimension de l'excursion des mouvements mélodiques sur l'ensemble de la phrase correspond alors à l'écart entre la ligne de déclinaison inférieure et la ligne de déclinaison supérieure. Dans le chapitre III, une évaluation de deux estimations de la hauteur mélodique et de la dynamique fréquentielle a été effectuée. Dans l'estimation, basée sur le modèle décrit ci-dessus, on utilise respectivement la fin de la ligne de déclinaison inférieure et la différence entre cette ligne et la ligne de déclinaison supérieure. L'autre estimation, basée sur la moyenne de F_0 dans les phrases naturelles et sur l'écart type de F_0 dans ces phrases, est plus strictement orientée sur les données de production.

Par ailleurs, hauteur mélodique et dynamique fréquentielle sont des propriétés définies au niveau de la phrase entière. Dans les courbes mélodiques d'une phrase produite par un locuteur, de nombreux détails peuvent être discernés qui ne peuvent pas être représentés dans un tel modèle d'intonation. Afin d'étudier les fluctuations de F_0 à l'intérieur des phrases, F_0 a été mesurée à un certain nombre de points fixes dans les phrases. Ces mesures ont été effectuées dans la première partie voisée de la phrase, dans la voyelle du premier accent mélodique, dans une voyelle suivant ce premier accent, dans une voyelle précédent l'accent mélodique final, dans la voyelle de ce dernier accent et dans la dernière partie voisée des segments de la phrase. Il s'est avéré que les phrases produites pendant l'expression de différentes émotions peuvent présenter des différences considérables en ce qui concerne la hauteur relative des pics mélodiques et l'importance de la baisse mélodique se manifestant en fin de phrase. Par exemple, F_0 mesurée dans le pic du dernier accent avait souvent une valeur supérieure à celle trouvée dans le pic du premier accent, ce qui ne s'explique pas simplement sur la base du phénomène de déclinaison. Pour certaines émotions en particulier, F_0 mesurée dans la partie finale de la phrase avait une valeur inférieure à celle escomptée sur la base des mesures effectuées dans les parties précédentes de la phrase étant supposées être représentatives de la 'baseline'. La pertinence de ces différences a été testée au cours d'une étude de perception. Bien que certains effets se soient avérés significatifs - par exemple, la réalisation d'une baisse mélodique en fin de phrase a provoqué une augmentation du nombre de réponses des sujets dans la catégorie 'indignation' -, les effets trouvés étaient d'ordre relativement limité.

Les 315 phrases sélectionnées comme matériel ont toutes été étiquetées en termes de configurations intonatives et la répartition des configurations de mouvements mélodiques entre les différentes catégories correspondant aux émotions a été étudiée pour chacun des locuteurs. Les résultats sont présentés dans le chapitre IV. Il est apparu que les configurations ne sont pas réparties de façon égale entre les sept émotions. La configuration '1&A', une montée et une descente mélodiques réalisées sur une seule syllabe et la rendant proéminente, a été la plus fréquemment utilisée. Elle revient régulièrement dans les phrases exprimant les différentes émotions. L'hypothèse a donc été émise que, lorsqu'on souhaite exclure la variabilité introduite par la réalisation de différentes configurations intonatives, la configuration '1&A' pourrait être utilisée dans l'expression de toutes les émotions. Toutefois, l'étude de production montre aussi que de

nombreuses phrases ont été produites avec d'autres configurations intonatives que '1&A' et que certaines configurations intonatives semblent être plus caractéristiques de certaines émotions que d'autres. On a remarqué, en particulier, que les configurations '12' (une montée mélodique suivie d'une montée mélodique très tardive) et '3C' (une montée mélodique tardive et une descente mélodique très tardive) ne sont jamais apparues en position finale dans les phrases exprimant la neutralité. Une seconde hypothèse s'est imposée, à savoir que ces deux configurations pourraient signaler l'émotion dans la parole. Une expérience de perception a été conduite pour étudier la pertinence des configurations intonatives pour l'identification des émotions dans la parole. Cette expérience a confirmé que des configurations intonatives spécifiques contribuent à la perception de certaines des émotions étudiées. Pour finir, des groupes de configurations intonatives ont été formés sur la base des résultats de l'expérience de perception. La partie finale de la configuration s'est avérée particulièrement pertinente. Les groupes reflètent les caractéristiques perceptives permettant de distinguer les différentes configurations intonatives.

Le chapitre V porte sur les variations temporelles véhiculant l'émotion dans la parole. Une analyse du débit de parole a d'abord été effectuée au niveau de la phrase entière. Cette analyse globale a consisté à mesurer la durée totale des 315 phrases sélectionnées comme matériel. La moyenne des résultats et l'écart type correspondant ont été calculés par émotion pour chaque locuteur. Ensuite, afin de déterminer si une approche linéaire consistant simplement à étirer ou à comprimer toute la phrase de façon linéaire - c'est à dire une approche globale du débit de parole moyen considéré sur l'ensemble de la phrase - suffirait pour l'expression de l'émotion dans la parole, ou si une approche plus détaillée serait nécessaire, une analyse a été effectuée à un niveau plus local, à l'intérieur des phrases. Cette analyse a consisté à mesurer la durée des segments de parole (syllabes ou groupes de syllabes) en position accentuée, indépendemment de celle des segments en position non-accentuée. Une description de la structure temporelle interne des phrases émotionnelles a ainsi été obtenue en considérant la durée relative des segments accentués et non-accentués composant la phrase. Bien que les différences constatées soient faibles et que l'étude de production ne soit pas concluante quant à l'usage systématique de variations dans la structure temporelle des phrases exprimant une émotion, certaines de ces informations à un niveau détaillé ne pouvaient pas s'inscrire dans un modèle linéaire ne permettant que de comprimer ou d'étirer la phrase. Il importait alors de savoir s'il est ou

non pertinent à la perception de l'émotion dans la parole de considérer séparément les segments accentués et les segments non-accentués. La déviation du modèle linéaire pouvait être inhérente à l'expression de l'émotion dans la parole, mais pouvait aussi être liée à la modification du débit de parole sur l'ensemble de la phrase et n'être ainsi qu'indirectement liée à l'expression de l'émotion (auquel cas la déviation serait due non pas à l'expression de l'émotion en soi, mais au changement de débit de parole qu'elle entraîne). Afin d'obtenir la référence qui nous permettrait de trancher entre ces deux interprétations, une même analyse de la durée relative des segments accentués et des segments non-accentués a été effectuée sur de la parole neutre produite à divers débits de parole par l'un des locuteurs masculins. La comparaison des résultats concernant la parole émotionnelle et ceux concernant la parole neutre a révélé que, dans la production, la structure temporelle change de façon non linéaire et qu'elle varie en fonction de certaines émotions. La pertinence de ces variations a alors été testée lors d'une expérience de perception. Des manipulations ont eu lieu afin de générer de la parole émotionnelle, soit par simple manipulation linéaire étirant ou comprimant la totalité de la phrase, soit par manipulation proportionnelle variant la durée des segments accentués indépendemment de celle des segments non-accentués. Les valeurs de durée relative des deux types de segments qui ont été utilisées au cours de cette manipulation ont été inspirées des données de production, et respectivement utilisées telles quelles, augmentées ou diminuées. Les différences de durée proportionnelle des segments accentués et des segments non accentués qui étaient *associées aux changements du débit de parole* se sont avérées non pertinentes pour la perception de l'émotion. Par contre, les différences de durée relative des segments accentués et des segments non accentués *associées à l'expression de l'émotion* se sont avérées tout à fait pertinentes pour l'expression de la neutralité et de l'indignation.

Finalement, dans le chapitre VI, la délimitation du domaine de recherche de la présente étude est rappelée et les résultats sont récapitulés. Il est conclu qu'une interaction de paramètres prosodiques permet l'expression orale de l'émotion et que la plupart des émotions peuvent être véhiculées dans la parole de synthèse grâce aux paramètres ici étudiés. Pour certaines émotions, les résultats restent moins convaincants. Pour ces émotions-là, il se peut que d'autres facteurs tels que la qualité de la voix, son intensité ou d'autres propriétés de l'intonation soient essentiels. Les résultats spécifiques à l'expression de l'émotion dans la parole qui ont été obtenus dans la présente étude sont récapitulés sous

Relations établies entre émotions et paramètres sur la base des études de production et/ou de perception

Paramètres	Emotions						
	neutralité	joie	ennui	colère	tristesse	peur	indignation
hauteur mélodique	65 Hz	155 Hz	65 Hz	110 Hz	102 Hz	200 Hz	170 Hz
dynamique fréquentielle	5 s.t.	10 s.t.	4 s.t.	10 s.t.	7 s.t.	8 s.t.	10 s.t.
baisse mélodique finale	-	-	non	-	oui	oui	oui
hauteur relative des pics	-	-	-	-	oui	oui	-
configuration(s) à préférer en position finale	1&A	1&A et 5&A	3C	5&A, A et EA	3C	12 et 3C	surtout 12, mais aussi 3C
configuration(s) à éviter en position finale	12 et 3C	A, EA, et 12	5&A et 12	1&A et 3C	5&A	A et EA	1&A
durée par rapport à la neutralité	100%	83%	150%	79%	129%	89%	117%
durée proportionnelle des segments acc. et non-acc.	ne pas dévier d'une manipulation linéaire	-	-	-	-	-	allonger les segments acc. de 40% de plus que les segments non-acc.

la forme d'une série de règles permettant de générer de la parole véhiculant chacune des émotions étudiées. Ces règles correspondent aux relations entre émotions et paramètres présentées dans la table ci-dessus, dans laquelle les valeurs optimales sont mentionnées pour chaque émotion, ainsi que les configurations à préférer ou à éviter pour générer de la parole porteuse de ces émotions. Il y est également stipulé, pour la baisse mélodique en fin de phrase et pour la hauteur relative des pics mélodiques, si leur modélisation paraît oui on non pertinente.

En outre, les résultats concernant l'adéquation des modèles pour traiter les variations extrêmes qui sont réalisées dans la parole émotionnelle sont résumés dans ce dernier chapitre. En conclusion, quelques lignes générales sont suggérées pour de futures recherches concernant l'expression de l'émotion dans la parole.

Index

P

R

S

T

U

V

Curriculum Vitae

Name: Sylvie Jeannette Laure Mozziconacci

Born: 27th October 1961, in Paris (France)

Nationality: French

Diplomas

1979: French "Baccalauréat série D: Mathématiques et Sciences de la Nature"

1982: "Diplôme d'Etat d'Orthophoniste" (Diploma of speech therapist), Université Paris VI, France

1993: Dutch "Doktoraal Alfa-Informatica, Specialisatierichting: Spraak en Informatica" (Specialisation: Speech synthesis and speech recognition), Department of Computer Linguistics and Institute of Phonetic Sciences, University of Amsterdam, the Netherlands

1998: Doctorate (Ph. D.), Eindhoven University of Technology, the Netherlands

Main experience

1982-1983: Speech therapist for the deaf, France

1984-1998: French teacher at the French cultural institute of Amsterdam, Maison Descartes, the Netherlands

1986-1990: Private practice of French speech therapy in the Netherlands

1993-1998: Research concerning the variability in speech, in particular the acoustical and prosodic parameters conveying emotion and attitude in speech, IPO, Eindhoven, the Netherlands

As part of the previous research:

1996-1997: 8 months as guest researcher at KTH (Royal Institute of Technology), Department of Speech, Hearing and Music, Stockholm, Sweden

1998: 9 months as guest researcher at the Intitute of Phonetic Sciences, University of Amsterdam, the Netherlands

Stellingen

behorende bij het proefschrift

Speech Variability and Emotion: Production and Perception

van Sylvie J. L. Mozziconacci

1. Because of the complementarity of the production and the perception processes, which is the basis of spoken communication, the role of speech properties in communication can only be determined by complementing production experiments with perception experiments in which these speech properties are varied independently of each other.

2. The way people interpret a spoken message does not only depend on the semantic content but also on the pitch, duration, and other prosodic properties of the utterance. These properties can be modeled in a quantitative way on the basis of results of both production and perception experiments.

3. A perceptual analysis of pitch curves resulting in a classification into intonation patterns helps in understanding the role of pitch in vocal expression of emotion.

4. In practicing speech research, feelings and emotions are frequently considered as "noise". They are, therefore, disregarded, often unjustly, in the process of identifying objective information.

5. "In onze strijd voor vrijheid is de waarheid ons enige wapen".

   Z. H. de Veertiende Dalai Lama.

6. De stilte waarin een dove leeft is van een heel andere aard dan de stilte waar een horende naar kan verlangen. De stilte is voor een horende aantrekkelijk juist door de zachte geluiden en de geluiden van veraf die men nu wel kan horen; de wereld wordt er wijds door. De "stilte" die de dove kent valt te vergelijken met de situatie in een zeer lawaaiige fabriek; men voelt zich opgesloten door het geluid, de wereld wordt tot benauwende proporties teruggebracht.

7. "... kleur licht op en wil alleen maar schitteren. Als we haar verstandelijk meten en in trillingsgetallen uiteenleggen, is ze weg. Ze toont zich alleen als ze onontborgen en onverklaarbaar blijft."

   Martin Heidegger, De oorsprong van het kunstwerk.

8. Uiteraard dient men de significantie van experimenteel verkregen gegevens statistisch te toetsen. Vaak doet men dit echter verkeerd. Zo niet, dan blijft men vrijwel steeds onbegrepen.

9. Le voyage est une excellente école de la vie.
   La destination importe peu, l'important est de faire la route.

10. Si cette IPO-thèse, qui pendant des années n'a été qu'une hypothèse,
    a fini par être vérifiée c'est bien malgré les IPO-crises et les manIPOlations.