idiosyncrasy in prosody

Speaker and speaker group identification in Dutch using melodic and temporal information

Hans Kraayeveld

Idiosyncrasy in Prosody

Speaker and speaker group identification in Dutch
using melodic and temporal information

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Kraayeveld, Johannes

Idiosyncrasy in Prosody: Speaker and speaker group identification in Dutch using melodic and temporal information / Johannes Kraayeveld. - [S.l.: s.n.]. - Ill.
Proefschrift Katholieke Universiteit Nijmegen. - Met literatuuropgave - met samenvatting in het Nederlands.
ISBN 90-9010846-7
Trefw.: fonetiek, prosodie, sprekerherkenning, sprekerspecificiteit.

This Ph.D. thesis is also available on the world wide web. It can be obtained from http://lands.let.kun.nl.

Idiosyncrasy in Prosody

Speaker and speaker group identification in Dutch using melodic and temporal information

Een wetenschappelijke proeve op het gebied van de Letteren

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen

volgens besluit van het College van Decanen
in het openbaar te verdedigen
op vrijdag 5 september 1997
des namiddags om 1.30 uur precies

door

Johannes Kraayeveld

Geboren op 8 juli 1962 te Dordrecht

FACULTEIT DER LETTEREN

promotor:	prof. dr. W.H. Vieregge
co-promotores:	dr. A.C.M. Rietveld dr. V.J. van Heuven (R.U. Leiden)
manuscriptcommissie:	prof. dr. W.J. Barry dr. R.A.M.G. van Bezooijen prof. dr. L. Boves prof. dr. R. Collier dr. J.M.B. Terken

This research was supported by the Linguistic Research Foundation, which is funded by the Netherlands Organization for Scientific Research, NWO.

Marco Polo descrive un ponte, pietra per pietra.
"Ma qual è la pietra che sostiene il ponte?", chiede Kublai Kan.
"Il ponte non è sostenuto da questa o quella pietra", risponde Marco, "ma dalla linea dell'arco che esse formano."
Kublai Kan rimane silenzioso, riflettendo. Poi soggiunge: "Perché mi parli delle pietre? È solo dell'arco che m'importa."
Polo risponde: "Senza pietre non c'è arco."

Marco Polo describes a bridge, stone by stone.
"But which is the stone that sustains the bridge?", asks Kubilay Khan.
"The bridge is not sustained by this stone or that", replies Marco, "but by the line of the arch that is formed."
Kubilay Khan remains silent, thinking. Next he adds: "Why talk to me about stones? It's only the arch that interests me."
Polo answers: "Without stones there is no arch."

from: Italo Calvino, *Le città invisibili*, cap. 5, p. 89

Voorwoord

Het onderzoek waarover in dit boek verslag wordt gedaan werd gesteund door de Stichting Taalwetenschap (projectnr. 300-173-006), die op haar beurt wordt gesubsidieerd door de Nederlandse organisatie voor Wetenschappelijk Onderzoek (N.W.O.).
Geld alleen maakt nog geen proefschrift mogelijk. Familie en vrienden waren, hoewel zij weinig ophadden met fonetiek en niet veel begrepen van mijn interesse voor toonhoogtebewegingen, perturbaties, enzovoort, een belangrijke steun tijdens de periode van mijn onderzoek. Vooral mijn moeder bedank ik voor de vele jaren steun op afstand en voor de aanmoedigingen om verder te leren.
Het besproken onderzoek was een samenwerkingsproject van de universiteiten van Nijmegen en Leiden, uitgevoerd in Nijmegen. Zonder mijn begeleiders uit de genoemde steden, Toni Rietveld en Vincent van Heuven, was dit proefschrift er niet geweest. Toni Rietveld, mijn Nijmeegse begeleider, zag en sprak ik uiteraard het meest. Hartelijk bedankt, Toni, voor je inspanningen om mij wegwijs te maken in fonetiek in het algemeen, prosodie in het bijzonder, statistiek, schrijfstijl, BMDP, CMS, en nog veel meer. Ook mijn Leidse begeleider, Vincent van Heuven, wil ik bedanken voor de steun die ik steeds kreeg. De wereld wordt steeds kleiner en dat geldt gelukkig ook voor de afstand Nijmegen-Leiden. Vincent, bedankt voor de vele aandacht en inspiratie!
Mijn promotor, Felix Vieregge, placht mij er altijd op te wijzen dat bedankjes voor geleverde kritiek niet nodig zijn omdat hij slechts zijn werk deed. Laat ik daarom alleen benadrukken dat zijn kritisch leeswerk de uiteindelijke tekst zeer ten goede is gekomen.
Mijn collega's van de vakgroep Taal en Spraak van de Katholieke Universiteit Nijmegen bedank ik voor de steun die ik gedurende mijn verblijf in Nijmegen van hen kreeg. Zij wijdden mij in in allerlei aspecten van spraak en spraaktechnologie, en wij deelden veel plezierige momenten tijdens koffie- en lunchpauzes. Bedankt voor de gezelligheid.
Marina Koopmans was gedurende de meeste jaren mijn kamergenoot. We bespraken velerlei onderwerpen, zowel ons onderzoek als ook de alledaagse problemen waar je als assistent in opleiding mee te maken krijgt. Bedankt voor je bemoedigingen en je opgewektheid, Marina!
Alle in dit proefschrift te vinden fouten en onvolkomenheden komen geheel voor rekening van de schrijver. Waarschijnlijk zouden er echter meer gebreken te vinden zijn geweest zonder mijn collega's Dick van Bergem, Henk van den Heuvel, Helmer Strik en Catia Cucchiarini, die (delen van) de tekst doorlazen en van kritisch commentaar voorzagen. Catia dank ik bovendien voor het op een hoger plan tillen van mijn Engels.
Onderzoek doen kan een eenzame bezigheid zijn zonder een enthousiaste groep toehoorders die graag op de hoogte gehouden wordt over de meest recente resultaten. Ik vond een heel inspirerend publiek in de 'prosodieclub', en wil haar leden daarom hartelijk bedanken; zonder jullie zou dit project wellicht tot een 'downstepping contour' verworden zijn.
Ik werd van een flink deel van het praktische werk, zoals A/D-conversie en segmentatie, verlost door de hulp van mijn enthousiaste studentassistent Geert

Wiegeraad. Bedankt!
Na mijn verhuizing naar Helmond kon ik iedere week een paar lange dagen in Nijmegen maken dankzij André de Vries en Carol van Hulten, en later Leo Bosland. Bedankt voor jullie gastvrijheid.
Een deel van mijn tijd besteedde ik aan het helpen van studenten met allerlei praktische problemen. Zij hebben er geen idee van hoeveel ik heb gehad aan die sessies. Bedankt allemaal, vooral Angelique Hoogstrate.
Tenslotte zou dit werk niet mogelijk zijn geweest zonder de medewerking van mijn sprekers, de heren Van Bakel, Barkema, Beekelaar, Van den Bercken, Van Berkel, Bongaerts, Boonstra, Van de Borgt, De Bot, Broeders, De Bruin, Colard, Ester, Hoogakker, Koreman, Van Leeuwen, Van Meurs, Poiesz, Slis, Strik, Veenman, Verhoeven, Voeten, Wermers en Wesselingh, en de dames Bange, Bekkering, Berends, Bernards, Claassen, Commissaris, Faaber, Foppen, Gaalman, Gerritsen, Heebing, Van der Heijden, Hofhuis, Van Houtum, Jacobs, Janssen, De Jong, Konst, Kostelijk, Kruize, Laheij, Lindenbergh, De Mönnink en De Vries. Voor veel van deze 49 mensen geldt, dat ik hun stemmen beter ken dan henzelf. Bedankt voor het mij lenen van jullie stemmen.
De vijftigste stem is die van Nicole van Dam. Zij leende mij niet alleen haar stem, maar ook haar oor. Bedankt voor je geduld. Het is goed het relatieve van je werk in te zien, maar een fixatie hierop vormt niet bepaald een vruchtbare bodem voor een proefschrift. Bedankt dat je me leerde dat niet alles relatief is!

CONTENTS

List of abbreviations

APQ	maximimum Amplitude Perturbation Quotient
AR	Articulation Rate
AVI	Amplitude Variability Index
AZR	maximimum Amplitude Zero-crossing Ratio
CA	Correct Assignment (identification score over two sessions)
CB	Contour-Bound
CV	identification score in Cross-Validations
CVA	Coefficient of Variation of maximimum Amplitude
CVP	Coefficient of Variation of Period
DPF	Directional Perturbation Factor
DRC	Discriminant Ratio Coefficient
DTW	Dynamic Time Warping
DURFAL	DURation of the FALl*
DURFIL	DURation of the FInal Lowering*
DURRI1	DURation of the first RIse*
DURRI2	DURation of the second RIse*
ERB	Equivalent Rectangular Bandwidth
F_0END	Final F_0 value of an utterance
F_0MEAN	Mean F_0 value
GDI	Grammar of Dutch Intonation
HMM	Hidden Markov Modelling
IL	Intensity Level
IS	Identification Score
JF	Jitter Factor
LDA	Linear Discriminant Analysis
LOWFAL	ST difference between the LOWest F_0 value of the FALl and the final F_0 value of the utterance*
LOWRI1	ST difference between the LOWest F_0 value of the first RIse and the final F_0 value of the utterance*
LOWRI2	ST difference between the LOWest F_0 value of the second RIse and the final F_0 value of the utterance*
LPC	Linear Predictive Coding
PPQ	Period Perturbation Quotient
PZR	Period Zero-crossing Rate
RAP	Relative Average Perturbation
SEMDEC	SEMitone difference in the DEClination*
SEMFAL	SEMitone difference between onset and end of FALl*
SEMRI1	SEMitone difference between onset and end of first RIse*
SEMRI2	SEMitone difference between onset and end of second RIse*
SES	Socio Economic Status

* Cf. Chapter 4.

SESAM	Signal Editing System AMsterdam
SHS	SubHarmonic Summation
SLODEC	SLOpe of the DEClination over the utterance*
SLOFAL	SLOpe of the FALl*
SLOFIL	SLOpe of the FInal Lowering*
SLORI1	SLOpe of the first RIse*
SLORI2	SLOpe of the second RIse*
SNL	Spectral Noise Level
SPL	Sound Pressure Level
SR	Speech Rate
ST	SemiTone
SYNFAL	SYNchronisation, interval between onset of FALl and vowel onset*
SYNFIL	SYNchronisation, interval between onset of FInal Lowering and vowel onset*
SYNRI1	SYNchronisation, interval between onset of first RIse and vowel onset*
SYNRI2	SYNchronisation, interval between onset of second RIse and vowel onset*
TI	Time-Integrated
VDA	Voicing-Determination Algorithm
VOI	percentage of VOIced frames
WS	identification score Within Sessions

1. Introduction

1.1 Object of this study

Voices differ. Although speech is characterized by a large amount of variation, we generally have few problems recognizing the voices of most of the people we are familiar with, even over telephone lines. The number of voices we are familiar with is large, and because of the advent of radio and television it is probably even larger than in earlier days.

One cannot be but impressed by the human capacity of discerning hundreds of people by their voices while there is also enough similarity in the speech signal to understand the message conveyed by all these different voices. In phonetic research, the study of the communicative aspect of speech has prevailed, and attention has mainly been directed at possibilities of isolating invariant properties of linguistic units (e.g. phonemes) by factoring out contextual influences, including speaker dependencies. The problem of speaker-dependent variation was circumvented in various ways, e.g. by using only data from one speaker, or by averaging out the speakers' idiosyncrasies.

One motive for increasing our knowledge of speaker specificity in speech is a linguistic one. Assuming that the main goal of all speakers is to be understood by their audience (the adaptation principle, Nooteboom and Eefting, 1991), a reasonable hypothesis is that most speaker variation will be found where the restraints imposed by the linguistic system are few, as listeners would allow least variation where the linguistic system defines the expected output most clearly. Thus, linguistically meaningful properties might be characterized by little speaker-dependent variation. In other words "Studies of variabilities pave the way for studies of invariance and vice versa" (Fant et al., 1990: 106).

A second incentive for studying idiosyncrasy in speech originates from the increased technical possibilities available for speech research. These have stimulated the development of a number of new applications such as speech synthesis and speech recognition. Both of these fields have now reached a point in their development where information on speaker characteristics is getting more and more important. In speech synthesis, research is directed at increasing the naturalness of synthetic speech, or even at attaining personalized voice synthesis. Information on what speech characteristics define an individual voice appears to be indispensable in this respect (Carlson et al., 1991; Moulines and Sagisaka, 1995). Speech recognition can profit from information concerning idiosyncratic speech properties of the speaker, because knowledge on the sort of individual variation that is to be expected among speakers, possibly supplemented by information concerning the speaker's sex and dialect, can narrow the search space and thus enhance speech recognition (Furui, 1990).

Another reason for doing research on speaker-specific variation can be found in the potentially useful applications of speaker identification (e.g. in forensic research) and verification (e.g. for electronic access systems).

The present study explores the amount of speaker-specific information that is present in various prosodic parameters, but does not aim at constructing readily usable

speaker identification devices. Some general aspects of speaker identification will be presented in section 1.2, followed by a discussion of the literature on speaker identification in section 1.3.

When in the literature attention was devoted at all to speaker specific properties of the speech signal, this attention was mostly focused on the spectral properties of segments. Previous research has shown that much speaker specificity can be found in the segmental domain (e.g. Nolan, 1983). As yet, however, little is known about individual variation in *prosodic* properties, such as fundamental frequency contours and the temporal organisation of the utterance[1]. These prosodic characteristics form the subject of the present investigation. We do not only try to attain a high percentage of speaker identification by combining these prosodic properties, as is done in most studies, but also aim to identify the speaker specificity in the individual parameters. A necessary first step in a study of prosodic characteristics is, of course, to indicate which measures will be regarded as "prosodic". This issue is taken up in the first part of section 1.3.2.

In this study we will *not* examine speaker identification by human listeners[2]. An important finding in this kind of research is that a high speaker identification performance can be reached by presenting human listeners with prosodic information only (e.g. about 90 % in Schmidt-Neilsen and Stern, 1985). From these results we deduce that there must be much speaker-specific information in the prosodic properties of the speech signal. However, because of the aforementioned preoccupation of phonetic research with invariant properties of linguistic units, not much is known about the acoustic features of the speech signal where speaker specificity is to be found.

Some speaker-specific voice settings exercise influence over longer stretches of a person's speech (Laver, 1980). In such cases more certainty about a person's identity can generally be gained by integrating information from the speech signal over some time period. This integration process has two important advantages. First, by enlarging the integration time period, the value obtained will become more stable and second, by integrating over a sufficient period of time, characteristics get less context-dependent. In this book we will refer to parameters that are obtained by integrating their values over some period of time as *time-integrated*, or TI parameters. A time-integrated parameter that has received particular attention in the study of speaker identification is mean fundamental frequency, or mean F_0. It was found that mean F_0 and the standard deviation of F_0 often enable a good discrimination between subjects (e.g. Jassem et al., 1973; Doherty, 1976).

By applying only time-integrated parameters to speaker identification, more detailed knowledge on the actual prosodic events is lost. In the value of mean F_0, for instance, no information can be found on the specific course of individual speakers' F_0 over the utterance. Prosodic events, such as pitch movements, are dependent on various aspects of the utterances concerned. They depend on the lexical, syntactic, and rhythmic structure of the

---

[1] At our department, the Department of Language and Speech of the University of Nijmegen, the speaker specificity of both segmental and prosodic parameters was studied. In the present project, we delimit our research domain to prosodic phenomena, because in a parallel project (van den Heuvel, 1996), speaker specificity in the segmental domain was examined.

[2] For information on speaker identification on the basis of prosodic parameters, as performed by human listeners, the reader is referred to Abberton and Fourcin (1978), Hollien et al. (1982), Schmidt-Neilsen and Stern (1985) and van Dommelen (1987, 1990)

utterances. Information on prosodic events is therefore best obtained in specified utterances. Henceforth we refer to these measures as *contour-bound*, or CB parameters.

To be able to describe the speaker specificity in CB parameters, one needs a descriptive system for the F_0 contours of speakers. Without a categorization criterion for the movements of F_0, it would be impossible to decide whether individual realizations of the same phonological "event" are being compared. Fortunately, an important advantage of Dutch is that consolidated knowledge is available in relation to prosody, particularly in the area of intonation: the Grammar of Dutch Intonation developed by 't Hart et al. (Collier and 't Hart, 1981; 't Hart et al., 1990). Possibly this transcription system will benefit from some of the data that will be presented in this book, as until recently not much was known about the ways in which speakers realize the pitch movements that are laid down in it.

The main innovation of the present study is that both types of parameters, time-integrated *and* contour-bound, will be applied to speaker identification, whereas in the majority of studies in which prosodic parameters were used for speaker identification, only TI parameters were employed.

The characteristics of a voice originate from numerous extralinguistic factors. Some of these factors stem from idiosyncratic properties, such as the unique anatomic and physiological structure of the speaker's speech organs. Idiosyncratic features are often defined as features that cannot be correlated with group factors such as sex, age, regional origin, social status, health, etc. (Brown, 1982; van den Heuvel, 1996). Apart from these group factors, we should also consider the influence of the task at hand (e.g. reading, conversation) on individual speech behaviour. The influence of the most important extralinguistic determinants on the prosodic parameters will be varied systematically in this study, to be able to factor out their influence on speaker identification. In section 1.3.3 a literature survey of extralinguistic influences on voice behaviour will be presented.

In the remainder of this first chapter different aspects of speaker identification and prosody are elucidated, and their bearing on the current study is explained. In section 1.4 the aims and limitations of our study are briefly recapitulated.

To summarize, we will explicitly formulate the research objective as follows:

*In this study we aim to find out to what extent which **prosodic** parameters can be used to identify speakers. These parameters will be of both the time-integrated and the contour-bound type. The influence of some clearly influential factors - sex, age, speech style - will be strictly controlled, assessed and factored out so as to reveal what speaker idiosyncrasies remain.*

1.2 SPEAKER RECOGNITION

Above it was explained that the goal of the present study is to determine the amount of speaker specificity in prosodic parameters and to find out to what extent this information can be used to identify speakers by their voice.

With regard to the process of matching a voice sample to some reference speaker, many terms are used; recognition, identification, discrimination, etc. Nolan (1983) considers "speaker recognition" to be the general term for the matching of voices to speakers. Recognition would include both speaker identification and speaker verification.

In speaker identification an utterance from an unknown speaker has to be related to a speaker of some known population of speakers. In speaker verification, on the other hand, an individual claims a certain identity and the task of the verification procedure is to accept or reject this claim[3].

The most basic difference between procedures of matching voices to speakers is the difference between open and closed tests. A closed test is defined as a situation in which a sample of speech has to be matched with one of a number of possible speakers. If it is not certain whether the person who produced the sample is in the examined population of speakers, the test is called "open". While in the closed test only an error of false identification may occur, in the open tests it is also possible to incorrectly reject all the reference population. The procedure to ascertain whether or not two samples of speech are similar enough to have been produced by the same speaker is called a discrimination test. Thus, an open test is in fact an iterative discrimination test.

As the present study is mainly concerned with the evaluation of the contribution of certain parameters to speaker recognition, we will apply the most simple test, the closed identification test. We will refer to the closed identification procedure as speaker *identification.*

Speaker identification is essentially a two-stage process: parameter extraction and parameter comparison. During the process of parameter extraction the values of the parameters to be applied are obtained for each of the experimental speech units (e.g. utterances or speech fragments). Speakers may be thought of as being points in a "parameter space" of as many dimensions as there are parameters. The points are defined as the mean scores of the speakers on the parameters. Next, a distance measure is used to estimate the distance between the parameter scores in the test samples and the reference scores of the speakers stored in the database. In the case of an open test, where it is unknown beforehand whether the reference database contains parameter scores of the test sample speaker, a nearness threshold is used to decide whether the test sample should be attributed to any speaker in the database at all.

The statistical transformation of the parameters can be an intermediate step of practical value, as such a transformation can be used to reduce the dimensionality of the space within which the distances between arrays of measurements are to be measured. Some statistical transformation algorithms can also be applied as classification algorithms and as a means of finding out which parameters are most important for the identification task concerned. Moreover, percentages of correct identification can be procured, the criterion for the suitability of parameters to our objectives. In the present study linear discriminant analysis (LDA) will be used to this end (see Chapter 3).

Percentage of correct identification is not the only criterion that a speaker identification parameter must fulfil. Nolan (1983) discerned six criteria that such parameters should meet, the first two of which are directly related to the percentage of correct identification obtained in discriminant analysis:

---

[3] Although identification and verification tests are sometimes treated as quite distinct tasks (e.g. Tosi, 1979), we regard this difference as being only due to differences in the problems surrounding these tasks; in the first type of task investigators face the problems of disguise and lack of cooperation, in the second mimicry by impostors There is no fundamental difference in the nature of the decision procedure between speaker identification and verification tasks.

1. High between-speaker variability;
2. Low within-speaker variability;
3. Resistance to attempted disguise or mimicry;
4. Availability;
5. Robustness in transmission;
6. Measurability.

In the following we will discuss these criteria in more detail.

The present study is of an exploratory nature, and we do not aim to construct a readily usable speaker identification device. Instead, we try to find out which prosodic parameters are most speaker-specific. The criteria that determine the speaker specificity of a parameter, between and within-speaker variability, are therefore considered to be the most important of the above-mentioned criteria. Between and within-speaker variability have been of common concern in the field of speaker identification and are often combined into an *F*-ratio that is calculated as the ratio of the between-speaker variance and the within-speaker variance (Wolf, 1972; Bonastre et al., 1991; Hiraoka et al., 1984; van den Heuvel, 1996).

The third criterion for a speaker-identifying parameter, resistance to the special communicative intents of disguise or mimicry, could in speaker identification systems lead to the inclusion of parameters which by normal standards are inefficient. Doherty and Hollien (1978), for instance, found that although a parameter based on the proportion of voiced speech and articulation rate yielded a poor identification rate in optimum conditions, its performance sank only little in disguise. For the present study we consider the influence of disguise and mimicry on speaker identification as beyond our scope.

Criterion 4 is the availability of a certain parameter in a speech sample. It is of little use basing speaker identification on a parameter which occurs only seldom in speech and therefore necessitates large amounts of data in both test and reference corpora. The availability problem can be circumvented by using only time-integrated measurements. As such measurements do not require the occurrence of specific speech events, they only require a speech sample of sufficient duration. However, the availability of samples of sufficient duration can be problematic as well. For the time-integrated parameter mean F_0, for instance, it was found that rather long speech samples are needed to attain a stable value. Barry et al. (1991) found that between different speech samples of no less than 2 minutes, a within-speaker variation could be found of as much as 15 %. In subsequent chapters it will be explained that, in order to obtain useful values for TI parameters, it is important to control the speaking style of the speakers, while for CB parameters the utterances themselves must be carefully controlled.

A fifth aspect of importance is robustness in transmission. An important advantage of prosodic parameters over e.g. spectral studies is that the former appear to be unaffected by telephone or coding distortions (McGonegal et al., 1979). Nevertheless, judging from the number of publications, most research effort seems to have been invested in spectral information studies. In the next section, we will briefly discuss part of the numerous speaker identification studies in which spectral information was used.

The extent to which the measurement of certain parameters is problematic depends on the type of parameters used and on the kind of identification system which is envisaged. The difficulties inherent in the automatic location of particular phonetic events make it difficult to measure features derived from such events in a fully automatic system. The

reliability of measurements in a semi-automatic system with a human operator will perhaps always be higher. In Chapter 2 the measurability aspect will be discussed with regard to the parameters to be used in the present study.

Wolf (1972) used a list of criteria for speaker identification parameters that to a large extent overlaps with Nolan's. However, he does not mention the important criterion of robustness in transmission, while he introduces another criterion, which comes close to Nolan's "low within-speaker variability": "stability over time". If the speaker specificity of parameters is to be of use in real-life applications, it should not vary too much over time. Therefore, we will perform cross-validation in this study; discriminant functions will be derived from the parameters of one recording session and used to classify speech material from another session. Parameters that change little over time are particularly important for cross-validation. Many researchers (Wolf being one of them) fail to really give credit to the stability criterion and only use data from one single recording session. Such studies fail to recognize that a parameter which is efficient in the short run, may fail in the longer term due to purely physiological variation, such as suffering from a head cold (Sambur, 1975), psychological stress (Doherty and Hollien, 1978), etc.

1.3 Review of the literature

1.3.1 Speaker identification using non-prosodic measures

The origin of a large part of speaker identification research lies in speech recognition research. For a long time, many more speech recognition studies were carried out, because more was understood about segmental phonetics than about the aspects of the speech signal that identify the speaker.

It is often assumed that the information used for speaker recognition is coded in a way that is fundamentally different from the way in which information used for understanding the transmitted message is coded (O'Shaughnessy, 1987). While in speech recognition studies one can utilize the correlation that exists between phonemes and spectral resonances, there seem to be no acoustic cues that deal specifically or even exclusively with speaker identity. Most of the parameters used in speech analysis contain two types of information: information useful for the identification of the message and information useful for identifying extralinguistic factors, such as the speaker.

Speech recognition decisions have to be made from phoneme to phoneme and from word to word, while in speaker identification only one decision must be taken: "who is speaking?" To be able to take this decision, knowledge of the phonemes themselves is not always essential, as some speaker recognizers utilize long-term statistics that are averaged over entire utterances or speech fragments.

In the previous section it was explained that speaker identification is mostly based on a process of parameter extraction and parameter comparison. The parameters most frequently extracted in speech recognition, and therefore in speaker identification too, are spectral and cepstral representations, LPC (Linear Predictive Coding) parameters and transformations of LPC parameters (such as orthogonal LPC parameters)[4]. Hollien and

[4] LPC primarily provides a small set of speech parameters that offer a precise representation of the speech spectral magnitude (Markel and Gray, 1976).

Majewski (1977) reached high recognition accuracy by applying long-term spectra. In another study (Markel and Davis, 1979) material from 17 speakers was used to obtain 22-dimensional vectors for the speakers. Each vector contained the means and standard deviations of F_0 and the reflection coefficients in a tenth-order LPC analysis. Averaging over 1000 speech frames (about 39 seconds per speaker), a percentage of correct identification of 98 was obtained. A disadvantage of long-term spectra is the fact that they are sensitive to speaker effort and to variations in transmission channels such as the telephone (Doddington, 1985).

In speech recognition studies it was found that it is easier to attain a high percentage of correct speech identification for utterances of one single speaker, than for several speakers (e.g. O'Shaughnessy, 1987). In speaker recognition studies it was found that better speaker identification was attained in speech samples of equal lexical content (e.g. O'Shaughnessy, 1987). Therefore, an important distinction that can be drawn is that between text-dependent and text-independent speaker identification methods (e.g. Furui, 1990). The former requires a speaker to produce a predetermined utterance, while the latter does not rely on a specific text being spoken.

In speech samples of equal lexical content statistical averaging can be applied, but it is also possible to directly compare the (spectral) characteristics of phonemes of different speakers. The important advantage of using utterances with the same lexical content is that the phonemes come from the same words and occupy the same positions in the words. The utterance parts to be used in speaker comparison can be made even more comparable by first performing a time normalisation technique, such as dynamic time warping (DTW). The DTW technique aligns the corresponding parts of different utterances. Furui (1981), for example, extracted cepstral coefficients from a short sentence and applied DTW. By comparing the distance between the coefficients of the test and reference material, he was able to attain a very high percentage of correct speaker identification (97 %).

Speaker comparison within the same speech parts can also be accomplished by Hidden Markov Modelling (HMM). With HMM one can compare features of the same sub-word units over different utterances. Rosenberg et al. (1990) showed that this technique yields very high speaker verification results: the equal-error rate for seven-digit test utterances was 1 % or less[5].

Comparisons of spectra from comparable test and reference phones can be made in speech samples that do not have equal lexical content. In this approach the test and reference phones must first be selected on the basis of techniques that originate from speech recognition studies, such as vector quantization (e.g. Burton et al., 1985; Soong et al., 1985), Hidden Markov Modelling (e.g. Savic and Gupta, 1990) and neural networks-based methods (e.g. Oglesby and Mason, 1990).

---

[5] In speaker verification, the percentages of incorrectly rejected and of incorrectly accepted identity claims both depend on the distance threshold for accepting these claims. Lowering the threshold raises the percentage of incorrectly rejected identity claims at the expense of the percentage of incorrectly accepted claims. The equal-error rate is the percentage false acceptance or rejection at the threshold level at which these two errors are equal. It is often used to characterize the performance of a speaker verification system. Most studies apply either the equal-error rate or the percentage of correct identification/verification as a measure of identification/verification performance.

1.3.2 Speaker identification and prosody

The term "prosody" refers to "variations in pitch, loudness, tempo and rhythm" (Crystal, 1985: 249)[6]. Lehiste (1970: 1) observes that "The study of prosody is perhaps one of the oldest branches of the scientific study of language... Yet a certain degree of vagueness seems to characterize most discussions of prosodic features. They seem more elusive than segmental features, and their incorporation into a linguistic system sometimes seems to strain the limits of an otherwise coherent framework." Probably as a result of the vagueness mentioned by Lehiste, many definitions of "prosody" have been proposed. An example is the definition of the suprasegmental strand offered by Nolan (1983: 32): "the suprasegmental strand comprises phonetic systems whose contrastive patterns occupy a linear domain greater than the extent of a segment; the norm is for suprasegmental contrasts to be realised over units of the extent of a syllable up to the tone unit". However, it is not that clear what the upper limit of the suprasegmental strand really is, as phonetic systems have been identified with contrastive patterns even above the sentence level (Sluijter and Terken, 1993)[7].

Probably the most practical way of defining prosody is not in terms of the level of the units (such as exceeding the segment), but as those phenomena that are *not* part of segmental phonetics. In such a negative definition prosody is conceived of as a term covering all aspects of speech that cannot be attributed to the individual speech sounds (de Rooij, 1979; van Heuven, 1994a; van Heuven and Sluijter, 1996).

A major functional difference exists between segmental and prosodic variation. Segmental variation is mainly concerned with the production of contrastive features, phonemes, that have a direct and identifying relationship to the words produced. Prosodic variation cannot only be distinctive in a structural linguistic sense; it can also add expressive power to the message conveyed by the segments. By appropriately handling prosodic parameters emotions can be conveyed (van Bezooijen, 1984; Murray and Arnott, 1993), sentences can be subdivided into phrases (Cooper and Paccia-Cooper, 1980; Price et al., 1991), and prominence can be given to parts of an utterance (Bolinger, 1958; Terken, 1984, 1991).

In section 1.3.1 it was pointed out that most of the studies on speaker identification have some combination of spectral parameters as their input variables. Some studies, however, have included different types of prosodic parameters in an attempt to find out which are most useful for speaker identification. The prosodic parameters most often included in such studies were time-integrated measures: mean F_0, intensity level and speech rate.

A typical example of a study applying time integration is Doherty (1976). Doherty combined mean F_0 and the standard deviation of F_0 for 50 male speakers into a vector and

---

[6] The term "prosodic" is often used interchangeably with the term "suprasegmental". Which of these terms is preferred largely depends on the background of the researcher: the term "suprasegmental" is mainly used in the American Structuralist tradition, and often has the added connotation that the study of non-segmental features is only a "secondary" level of analysis. "Prosody", on the other hand, is often used in the British tradition of intonation study (Couper-Kuhlen, 1986).

[7] As for the smallest prosodic domain, some doubt is thrown on this issue by Van Heuven (1994b). He found that at least some speakers are capable of expressing narrow focus on a linguistic unit below the level of the syllable, by purely prosodic means, viz. by changing the shape and location of a pitch movement.

found that the vector enabled correct speaker identification at 30 %. Various durational measures were combined into a "speaking time vector", that enabled a correct identification of 12 %. A combination of the two vectors yielded 56 % correct identifications[8].

In the early 1970s a number of text-dependent studies were performed. Wolf (1972) used prosodic parameters that were obtained at the location of particular phonetic events, e.g. the fundamental frequencies in the mid-section of some vowels, and at the peak of pitch accents in the contour. Furthermore, parameters such as consonant spectra estimations, word durations and voice onset times were used. The local F_0 measures were the most speaker-specific parameters in terms of the ratio of the between to the within-speaker variability. Examples of such local F_0 measures are the F_0 value at the peak and in the middle of the word "few" in "A few boys bought them". Other useful prosodic measures were the word durations. Using only nine fairly unrelated parameters, Wolf attained a correct identification of 99 % in 210 utterances that were realized by 21 male speakers.

Most of the text-dependent studies were of a rather "holistic" nature (e.g. Das and Mohn, 1971; Doddington, 1971; Lummis, 1973; Rosenberg and Sambur, 1975). Distance measures were often obtained for utterances "as a whole". The first step in obtaining these measures was usually nonlinear time alignment, which is necessary because of the differences in speech rate between speakers. After alignment the utterances were divided into a number of fragments of equal duration. The distance measures for the parameters that were used were the total of the differences between the parameter scores on all of the fragments. Comparisons were made between test and reference templates on many different speech parameters, such as pitch, intensity, filter bank output levels, formants and LPC coefficients.

A holistic approach towards F_0 contours was also taken by Atal (1972). The importance of his approach is that he showed that speaker identification on the basis of F_0 contours is possible and can yield a high identification percentage (97 %). Atal's approach was to simply divide up a sentence into 40 segments of equal duration and to use the mean F_0 value of each of the segments as the input to data reduction and analysis techniques.

Concerning Atal's study, an important point must be made. Atal mixed up two entirely different matters: the speaker-specific realization of pitch movements and the choice of the pitch movements that the speakers realized. On the basis of the examples presented by Atal, we speculated (Kraayeveld et al., 1991) that the contour differences that he found were, at least partly, the result of the fact that the speakers were realizing different types of pitch movements. We compared Atal's holistic approach to a method of our own, in which measurements are only taken at the "pivot points" in the contour: starting points and end points of specific pitch movements.

As there is a difference between the speaker-specificity in the *choice* of pitch movements and in the *realization* of the pitch movements, two approaches to the study of speaker specificity in pitch movements are possible, the "qualitative approach", i.e., comparing the differences in the speakers' preferences of pitch movements, and the

---

[8] Adding a long term spectrum vector consisting of no less than 23 values raised this percentage to 100 %. Doherty and Hollien (1978) later showed that this speaker identification method was quite sensitive to conditions of psychological stress and disguise.

"quantitative approach", i.e., comparing the differences in the speakers' realization of similar pitch movements.

A qualitative approach is often used in studies in the segmental domain. The main segmental category is, of course, the phoneme realization. Utterances can be transcribed in terms of a limited set of phonetic symbols, such as those of the international phonetic alphabet (International Phonetic Association, 1949). A transcription of this kind can be used as a tool for speaker identification. Wells (1982) described four possible types of phonetic differences between speakers: systemic (differences in the inventory of allophones used), phonotactic (differences in the environments in which a phoneme realization can occur), incidental (differences in the use of allophones within a certain word) and realizational differences (differences in the phonetic realization of the phoneme).

Similar studies for prosodic features are rare, because it is unclear to what extent the discreteness of form found in the phonemes in the segmental area is parallelled in the suprasegmental strand. However, some of the suprasegmental systems, such as the Grammar of Dutch Intonation, henceforth referred to as *GDI*, do involve discrete primes. For instance, by means of GDI it should be possible, in principle, to study speaker-specific prosodic styles. Perhaps speaker identification could be helped by establishing differences in the frequency with which different speakers realize different pitch contours.

An example of a study in which the choice of pitch patterns is applied is a study by Woods (1992). She wanted to find out whether the choice for certain non-segmental features in utterances conveys information about various aspects of the social identities of the speakers and of the social situation in which speakers converse. Following Crystal (1969) and Cruttenden (1986), Woods took the phonetic features of pause, anacrusis (the occurrence of a sequence of unstressed syllables articulated at a very fast rate), and changes in pitch on unstressed syllables as the identifying features of tone-unit boundaries. Using these boundaries, she was able to examine characteristics of tone units, such as the tone-unit length (number of words contained in tone group), the rate of articulation (tone units per minute) and the structure of tone-units (number of nuclear tones within the tone-unit). Woods's results were promising: in her transcriptions she found sex differences (e.g. the percentage of fall-rise tones for women was 11, for men 5), age differences (e.g. adults produced a higher percentage of low fall tones: 19 vs. 15), and speech style-related differences (e.g. more low fall tones in informal speech: 23 % vs. 15 %).

Research on the choice of pitch movements by different speakers should perhaps precede a study of speaker differences in the realization of the pitch movements chosen. The present study does not concern this qualitative approach, but in an earlier study (Kraayeveld, 1995) we tried to find out whether the choice of pitch movements is speaker specific. Five speakers read out 29 sentences on two occasions. The frequencies of occurrence of the GDI pitch movements were used as input to a logit-analysis (Rietveld and van Hout, 1993). It was found that there was no interaction of the factors Speaker and Pitch movement. In other words, the frequencies of occurrence of the pitch movements were more or less evenly distributed over the readings of the five speakers. Thus, speaker identification on the basis of specificity in the choice of pitch movements does not seem very promising. Although supplementary research might produce better results for different stimulus material (e.g. for spontaneous speech), in this study we will assess the speaker specificity in the *realization* of pitch movements. To ensure that the pitch movements of the speakers are compared in similar settings, it is vital to control the utterances on the level of the pitch contour. Therefore, somewhat analogous to the text-dependent and text-

independent distinction mentioned above, we distinguish between time-integrated and contour-bound parameters in the prosodic realm.

1.3.3 Extralinguistic influences

In this study we will investigate speaker specificity in selected prosodic characteristics. Speaker specificity is a language-external factor that is related to physiological, social and psychological factors. This very relatedness poses an important problem for studies on speaker specificity. The characteristics of a speaker's voice are partly determined by his sex, age, social status, etc. However, trying to disentangle speaker specificity and the influence of factors such as these is as peeling away the layers of an onion; ever more factors can be found that are related to speaker characteristics, but how useful is it to continue peeling the onion? Will we ever find a "stone" in the onion that could be considered "pure" speaker specificity?

From the literature it is clear that some factors are more related to speaker specificity than others. As will be discussed below, three of these speech-affecting factors, i.e., sex, age and speech style, will be systematically varied in order to factor out their influence on the prosodic speech characteristics under study[9]. The influence of other factors will be kept constant (see Chapter 3). Below we will discuss literature that demonstrates the large influence of the factors sex, age and speech style. These illustrations will be centred around mean F_0, since this is the prosodic parameter that has been studied most extensively in the past.

Sex: A wealth of empirical work exists on male-female speech differences at the phonetic level. Many of the differences between male and female speech are probably related to the physiological differences between men and women. The most striking difference between the speech of men and women, the much higher level of F_0 found for women (e.g. Tielen, 1992), for instance, is mainly a consequence of differences in anatomical constitution. Male speakers have thicker, longer and slacker vocal folds. The range of the lengths of vocal folds for adult males is 17-24 mm and for females it is 12.5-17 mm (Zemlin, 1981).

However, we do not consider the biological distinction to be the sole cause of the sex differences in prosodic behaviour. It appears that the range of possible mean F_0's for men and women is physically bound to be in different, though partly overlapping, parts of the frequency continuum, but the average F_0 level actually found for a speaker might depend on social and cultural/economic factors as well. Zimmerman and West (1975), for example, claim that differences between men's and women's speech must also take into account patterns of male "dominance". Woods (1992), discussing her finding that women realize lower maximum fundamental frequencies in formal speech, thereby making their speech more like men's speech, suggests that women behave in this way because men's speech would be the norm for discussing serious topics.

---

[9] Many other factors can influence prosodic parameters. Braun (1992) gave an elaborate overview of the many factors that influence mean F_0. To mention just a few: emotions (Williams and Stevens, 1972), psychological stress (Scherer, 1977) and time of day (Garrett and Healey, 1987).

Age: The development in life of mean F_0 shows that changes in prosodic characteristics occur over the entire life-span period. Between 1 and 3 years average F_0 decreases rapidly, while in the middle childhood years, a more gradual drop is observed until adolescence (e.g. between 7 and 11 years: Bennett, 1983). The onset of puberty shows a more rapid and dramatic drop of F_0, particularly in males, although in females a decrease in mean F_0 is found during adolescence as well.

During adulthood a relative stabilization of F_0 takes place (Hollien and Paul, 1969). However, for men patterns of gradual decrease were found until around 40-50 years of age, followed by a reverse trend of increase in average F_0 (Hollien and Shipp, 1972). For women a gradual decrease is found until some time after menopause (Stoicheff, 1981a, 1981b; De Pinto and Hollien, 1982; Pegaro Krook, 1988; Higgins and Saxman, 1991). Once the post-menopausal voice has been achieved, F_0 appears to remain more or less stable, as was shown by Russell et al. (1995) in one of the scarce longitudinal studies of voice change. For a group of 15 Australian women they found a decrease by 48 Hz from the ages of about 18 in 1945 to 66 in 1993. For six of the speakers recordings had also been made in 1981 (mean age: 55 yrs), and between then and 1993 a decrease by only 3 Hz was found[10].

As in sex-related differences, age-related prosodic differences are attributable not only to the developments in the physique of speakers, such as changes in central and peripheral neurological functions and to the developments in the articulatory, resonatory, phonatory and respiratory systems (Meyerson, 1976). They also depend on social factors associated with age. Speech may be influenced by expectations of "how people of a certain age speak", and there may also be habitual speaking differences between speakers of different generations.

In the present study we are not primarily interested in developmental changes in prosodic parameters and we will therefore only consider age groups in the relatively stable period of adulthood, after puberty and before the onset of senescence. In this period, between the ages of 18 and 65, we do not expect great changes in speaking behaviour, but some change still appears to occur, such as the pattern found for average F_0 by Hollien and Shipp (1972): a gradual decrease until around 40-50 years followed by an increase after that age period.

Speech style: One of the factors influencing the prosody of speakers is the specific speech style employed by a speaker. "Styles" can be defined as collections of situationally distinctive varieties of language that (partly) account for specific linguistic choices made by an individual or a social group. Sometimes, especially in sociolinguistic studies, "style" refers to the relations among the participants in a language activity, primarily the level of formality they adopt (Robins, 1980).

One division between speech styles that has often been studied is the spontaneous vs. read difference. Phoneticians often need to record specific speech samples, and an easy way to elicit the same samples from different speakers is to have them read aloud textual material. Naturally, this practice invites the question whether findings in read speech can

[10] Further evidence for age-specific information in the speech of post-adolescent speakers comes from the fact that the speech of elderly speakers sounds different from that of younger speakers (Ryan and Burk, 1974; Amerman and Parnell, 1990).

be generalized to or compared with those pertaining to spontaneous speech.

The relevance of stylistic differences for prosodic linguistic features is clearly suggested by the fact that specific discourse modes or speech genres can be identified in the absence of information about segmental, lexical or grammatical aspects of language (e.g. Blaauw, 1991). Fònagy's (1978) results showed that the prosodic pattern of a verbal genre is even sufficiently characteristic to enable correct identification of speech style on the basis of laryngographic recordings alone.

Read speech is often found to have a higher fundamental frequency than spontaneous speech (e.g. Koopmans-van Beinum, 1991; Ramig and Ringel, 1983; Hollien and Jackson, 1973). Daly and Zue (1992) studied F_0 differences between speech styles in human-machine problem-solving dialogues. In a corpus of 4000 utterances they found results that were in conflict with the general tendency: mean F_0 was significantly higher for *spontaneous* speech than for read speech. This reversed pattern is probably also a situational effect: spontaneous speech directed at a human listener is probably different from the speech produced in the rather special situation of human-machine problem-solving interaction.

Although the spontaneous-read difference is a very salient characteristic of speech, it is not the only distinguishable speech style characterization. Patterns of language use cannot be adequately explained without reference to the social situation in which they occur. This claim, first proposed by the social anthropologist Malinowski (1935; in Woods, 1992), has several consequences for linguistic study today. It is generally recognized that in eliciting all forms and types of speech data, it is necessary to account and control for the effects of social situation. An example of another stylistic distinction is the difference between formal and informal speech. In informal speech, Woods (1992) found higher maximum F_0 values, but she did not find a significant influence of social situation on the use of mean F_0. Graddol (1986) found that, in read speech too, subjects employed a higher F_0 range in informal speech modes.

The number of possible social contexts of speech production is probably unlimited, and many of these situations exercise influence on prosodic parameters. Crystal and Davy (1969) list many different styles, such as: conversation, unscripted commentary, language of religion (praying, preaching), broadcast talks and news, public speaking, television advertisement, etc. However, listing instances of speech genres does not lead to a taxonomy of types of speech activity. It is not clear whether such a taxonomy is possible. Johns-Lewis (1986) mentions some dimensions: degree of preparedness, the public-private dimension, and the status of expertise of a speaker as compared with his audience.

The conclusion of this overview is that within the multitude of possible speech genres, conversation and oral reading are just two separate parts of a speaker's linguistic repertoire, not two extremes of a scale. Therefore, results of studies comparing read and spontaneous speech cannot be generalized to all kinds of read and spontaneous speech; this can only be done within the specific genre studied.

1.4 Scope of this study

The central theme of this book is the speaker specificity in prosodic parameters, both of the time-integrated and of the contour-bound type. This theme is elaborated into four specific questions that we try to answer. These questions are:

1) To what extent is speaker identification possible on the basis of prosodic parameters alone?
2) Which parameters are most important for speaker identification?
3) How stable is prosodic speaker identification, i.e., how dependent is it on speech style and date of recording (are the speaker characterizations at time T1 equally valid at time T2)?
4) To what extent does analysing data within sex and age groups affect speaker identification?

We sought to answer questions that were similar to those that were indicated above for "speaker" for the extralinguistic factors mentioned in question 3 and 4, i.e., speech style, date of recording, sex and age. In the same way that we assessed the possibility of speaker identification, we also tried to find out whether it is possible to discriminate, for instance, the speaker's sex. For the identification of the sex of the speakers the influence of the overall F_0 level was expected to be at least as important as for speaker identification, and therefore, as in speaker identification, we determined the influence of the F_0-related parameters on sex identification.

Until now we only briefly touched on the experimental approach that is taken in the present study. We now discuss our approach more thoroughly.

First of all, we study the extent to which *prosodic* parameters can be used to identify speakers. In practical terms, we will take closely to heart the list-like definition of "prosody" that was offered by Crystal (1969: 128): "prosodic features may be defined as vocal effects constituted by variations along the parameters of pitch, loudness, duration and silence... This then excludes vocal effects which are primarily the result of physiological mechanisms other than the vocal cords, such as the direct result of the workings of the pharyngeal, oral or nasal cavities: these are referred to as paralinguistic features". Spectral data, such as long-term average spectra thus fall under the heading "paralinguistic", not under "prosodic".

Crystal's proposal to consider as prosodic features the parameters of pitch, loudness, duration and silence presupposes a perceptual approach. In this study only production data will be considered, not perception data. We therefore translate the perceptual terms in Crystal's proposal into the acoustic domain and read "fundamental frequency" for "pitch". The term "loudness" must be replaced by one or more acoustic terms such as amplitude, spectral tilt, etc.

Although we emphasized the possible uses of speaker identification the present study does *not* aim at constructing readily usable speaker identification devices, but instead explores the amount of speaker-specific information that is present in various prosodic parameters. Therefore, we do not only try to attain a high percentage of speaker identification by combining these prosodic properties, as is done in most studies, but also aim to identify the speaker specificity in the individual parameters.

The exploratory nature of this study also has implications for the importance placed on some of the criteria for speaker specificity, as mentioned in section 1.2. The criteria that determine the percentage of correct speaker identification are considered to be most important: a high between-speaker variability and a low within-speaker variability are good indicators of the speaker-identifying properties of a parameter. In the current study these features play a key role in the statistical analyses that are used to establish the importance of the parameters: analyses of variance and linear discriminant analyses.

The main innovation of the present study is that two different types of parameters, time-integrated and contour-bound measures, are *both* applied to speaker identification, while in most studies only TI parameters are employed.

Time-integration can be applied to attain a certain independence of texts and contours. In this study time-integrated parameters were obtained in non-controlled utterances of considerable length, which appear to be the most appropriate stimulus material.

By applying time-integration one loses the more detailed information about the specific course of individual speakers' F_0 over an utterance. The realization of pitch contours depends on the lexical and phonological-prosodic content of the utterance concerned, and speaker parameters that are related to the pitch contours should therefore only be obtained in strictly controlled utterances. Accordingly, such parameters are referred to as contour-bound parameters.

The characteristics of a voice do not originate from idiosyncratic properties alone, but also from other extralinguistic factors. The influence of some of the speaker-related factors (sex and age) and of some of the task attributes (speech style, recording session, speech fragment, utterance) on the prosodic parameters were controlled, assessed and factored out so as to rule out their influence on speaker identification. Apart from using these extralinguistic factors as control factors in speaker identification, we also tried to find out whether prosodic parameters can be used to identify subgroups that are defined by the levels of these these factors, i.e., the sex and age bracket of the speakers, and the speech style, recording session, etc. of the fragments.

1.5 Outline of the book

The prosodic parameters used in this study, both those of the time-integrated (TI) and of the contour-bound (CB) type, are introduced and discussed in the first part of *Chapter 2*. The TI parameters can be subdivided into three groups: measures of fundamental frequency, amplitude and duration. We discuss four fundamental frequency measures: the mean, the coefficient of variation and two measures of perturbation. In the amplitude domain the same measures were used with the exception of the mean. The durational measures studied were pause time, articulation rate and a measure of voicedness.

In the realm of the contour-bound parameters some complementary measures were determined: the measures taken at specific pivot points in the pitch contour, the alignment of these pivot points with the segmental structure, and measures of declination.

The second part of Chapter 2 deals with the collection of the data. First we describe the selection of the speaker group and the criteria according to which the selection was carried out. Next, we turn to the speech material and the tasks that were used to elicit the speech samples. For the measurement of time-integrated parameters, a

reading task and an interview were carried out. For the contour-bound parameters the speakers were asked to read out sentences for which we expected to find uniform prosodic behaviour in all our speakers. The selection of our experimental material from the sentences that were actually read by all speakers is described. Next, the exact procedure of the recordings and a description of the recording equipment are provided.

The speaker-specifying properties of the time-integrated parameters are tested in *Chapter 3*. First the interrelatedness of the parameters was established to find out to what extent the parameters are independent of each other. If parameters correlate too much, they measure the same phenomenon and they should not both be selected for further analysis. Next, by means of analyses of variance the strength of association and the significance of a number of factors, viz. speaker, sex, age, speech style, fragment and session, were tested for each of the parameters separately. Finally, the combined speaker-identifying properties of the total group of TI parameters was assessed by means of discriminant analyses. These were performed both over the total material, and in so-called cross-validation analyses. In such analyses, discriminant functions were determined using material from one of the recording sessions. On the basis of these functions speech material from the other session was assigned to the speakers. From an application point of view these analyses are essential. If speakers attain highly variable parameter values in different sessions, the parameters are of little use in real-life applications. Discriminant analyses are also applied to the characterization of sex, age, speech style, fragment and session. Chapter 3 is concluded by an assessment of the influence of the duration of the fragments (15 seconds) by combining the material of different fragments into larger ones.

In *Chapter 4* the speaker-identifying properties of a number of CB parameters, as well as TI parameters, are established in a set of utterances in which the speakers' pitch contours were controlled as much as possible. The structure of Chapter 4 is comparable to that of Chapter 3; again the interrelatedness of the variables is determined, analyses of variance are performed, as well as LDA's. Separate analyses of the TI and the CB parameters were performed to enable a comparison between the two types. From analyses in which both the TI and the CB parameters were used we can learn whether combining the two types raises speaker identification performance. Furthermore, such analyses reveal which of the parameters are most speaker-specific overall. Apart from the speaker-identifying properties of CB parameters, Chapter 4 also allows for a comparison of the performance of TI measures in 15-second speech fragments and in much shorter sentences, in which the prosodic structure realized by the different speakers is comparable. Attention is also devoted to the influences of sex, age, sentence and session on the parameters.

In *Chapter 5* the main findings are integrated in a general discussion. The results are discussed in the same order as in the data presentation: we start with the presentation of the knowledge gained from the 15-second fragments as presented in Chapter 3, and then proceed to the findings from specific utterances, which are presented in Chapter 4. In the discussion we relate our results to the findings in the literature.

In the course of this project, a large database of speech material was obtained. We were forced to make choices regarding what data to process and analyse. Inevitably, we had to discard possibly interesting options, and we are convinced that our data can answer more questions than were posed here. However, we hope this study sheds some light on the influence of the speaker variable on prosodic parameters, and is helpful to and stimulating for future research in the study of both prosody and speaker identification.

2. Method

2.1 Introduction

In this chapter we will present the research method of the present study. The first part of the chapter deals with the definition of the acoustic prosodic parameters to be used. In the second part of this chapter the choice of our speakers and material will be described, and an account will be given of the recording procedures.

The prosodic parameters used in this study can be classified into two categories. The first category contains the parameters that are obtained by averaging over a certain period of time. These we call *time-integrated (TI)* parameters. The second category comprises the *contour-bound (CB)* parameters, i.e., measurements taken at specific pivot points in the F_0 contours of specific utterances. The main innovation of the present study is that parameters from both categories are applied to speaker identification and characterization. In the majority of studies in which prosodic parameters were used for speaker identification, only TI parameters were used.

We will start the description of the parameters to be applied in the subsequent chapters with the introduction of the TI parameters. The vocal effects to be studied are selected on the basis of Crystal's list-like definition of *prosodic* parameters: "the vocal effects constituted by variations along the parameters of pitch, loudness, duration and silence" (Crystal, 1969: 128). The acoustic equivalents of these parameters are F_0-related, amplitude-related and temporal TI measures. As most of the F_0-related parameters have an amplitude-related counterpart, the F_0 and amplitude-related measures will be described concurrently, in section 2.2.

Measures involving F_0 and amplitude fall into three groups. First we present measures that contain information on *central tendencies* within an utterance or a fragment of speech. Next measures for the *variability* of F_0 and amplitude are dealt with. These parameters represent the dispersion of values around a central tendency. Finally, we discuss *perturbation* measures. These reflect the amount of short-term variability, more specifically cycle-to-cycle variability, in the signal. Measures of both F_0 and amplitude perturbation will be discussed.

In the past, the different kinds of measures mentioned above have been used for different aims. The use of F_0 was above all descriptive, its average value and variability were often used to differentiate between male and female voices and between the voices of speakers of different ages. An overview of the literature was given in Chapter 1.

Perturbation measures have received much attention in the context of research on speech pathology. The reason for this emphasis is that perturbation measures quantify short-term instability of the vocal signal. High perturbation values in the domains of F_0 and amplitude are supposed to be related to malfunctions of the vocal apparatus.

In addition to F_0 and amplitude measures, we will also use three temporal measures. In section 2.3 articulation rate, pausing, and the voicedness in the speakers' utterances will be discussed.

After having concluded the description of the TI parameters, in section 2.4 we will turn to the description of the CB parameters. Section 2.4 starts with a description of the Grammar of Dutch Intonation, GDI (Collier and 't Hart, 1981; 't Hart et al., 1990). In the present study GDI is used as a descriptive tool that enables the parameterization of F_0 movements and, subsequently, the comparison of the prosodic behaviour of different speakers.

A well-known prosodic feature that can be observed in many utterances is declination, which can be described as the slow and gradual decrease of F_0 from the start to the end of the utterance (e.g. Cohen et al., 1982, Ladd, 1984). In section 2.4.4 a number of measures are introduced to describe the declination line. One of these measures, the F_0 measurement at the end of the sentence, is a potentially very promising parameter, as it has been reported that it is quite speaker-specific (Liberman and Pierrehumbert, 1984).

After having presented the acoustic prosodic parameters used in this study, we go on to describe the collection of our experimental material in the second part of this chapter.

In section 2.5 the selection of the speakers will be accounted for. Next, in section 2.6 we will discuss the choice of the speech material. The material consists of two types. In the first place we used rather large fragments of speech, of which the F_0 contours were not strictly controlled. The other part of the material consisted of isolated sentences with controlled contours. The larger speech fragments comprised both read and spontaneous speech. A description of the reading task and of the procedure for the elicitation of the spontaneous fragments, will be given. Furthermore, the criteria for the selection of the isolated sentences are explained, as well as the procedure by which these sentences were obtained. Finally, in section 2.7, details of the recording procedures are considered.

Part I: Definition of acoustic prosodic parameters

2.2 F_0-RELATED AND AMPLITUDE-RELATED TI PARAMETERS

2.2.1 Fundamental frequency

A central notion in the description of speech production data is *fundamental frequency* or F_0. When a speaker produces voiced speech his vocal folds periodically interrupt the airflow, thus producing a (quasi-)periodic sound. The main feature of a periodic signal is that it repeats itself every "period" of T seconds. The fundamental frequency or repetiton rate of a periodic sound is $F = 1/T$.

The first problem in using the term "fundamental frequency" is the terminological confusion that surrounds it. The terms "fundamental frequency" and "pitch" are often mixed up. In the present study we will use "fundamental frequency" for the acoustic measurement of the periodicity of the speech signal in terms of the number of cycles per second. In other words, it will only be used in reference to the acoustic speech signal. The term "pitch", on the other hand, will be used when referring to the perceptual impression that is caused by (among other factors) the fundamental frequency.

Another terminological problem is related to the name of the period T. Some

authors (e.g. O'Shaughnessy, 1987) refer to this period as "fundamental period". Although the term is not in common use, we will sometimes use it as well, as its meaning is clearer than the more general term "period".

In speech processing, periodicity estimation is an important topic. In the past, many algorithms have been proposed for this purpose (see e.g. Hess, 1983). The main difference between the algorithms lies in the domain in which F_0 is measured. All F_0 detection algorithms either work in the *time domain* or in the *frequency domain.*

The analyses working in the time domain often obtain fundamental period estimates by low-pass filtering the speech signal. The time domain algorithms are not very demanding and perform reasonably well, as long as the recordings to which they are applied are of high quality. An advantage of these algorithms is that they can be used to determine the moment of glottal closure (and zero crossings). The possibility of determining the moment of glottal closure is important to the goals of this study, as it enables a rather precise measurement of the period duration, which is necessary for calculating perturbation measures (see sections 2.2.7 to 2.2.9). The F_0-filtering program used, which was developed by van Bergem (1990), will be discussed in section 2.2.2.

An alternative to determining fundamental frequency in the time domain is offered by frequency domain algorithms. These algorithms estimate F_0 values for a short fragment of speech, a *window*. This window is moved through the material in steps of fixed duration. For each step a value of F_0 is obtained, which is assigned to the temporal midpoint of the window. The fragments of speech that thus obtain an F_0 value are called *frames*. Their duration is equal to the step size with which the analysis window is moved through the the signal. In frequency domain algorithms, spectral information relating to the harmonics within the window is used to measure F_0. An important advantage of these algorithms is that they are robust under circumstances with background noise, and that they enable a fundamental frequency estimate even if the fundamental is absent from the speech signal. An example of a speech signal that does not contain a first harmonic is the telephone speech signal. The fundamental frequency itself is (mostly) not present in such a signal, but a sufficient number of higher harmonics in the band-filtered signal may still enable F_0 measurement.

One of the oldest frequency domain algorithms was based on the cepstrum, the inverse Fourier transform of the logarithm of the amplitude spectrum. In a cepstrum, it is often easy to find the fundamental frequency, which is represented by a sharp peak (Noll, 1967). A newer generation of F_0 determination algorithms aims at applying knowledge bearing on the way in which the human auditory system deduces the fundamental frequency of a signal from the available harmonics. Examples of frequency domain algorithms are the harmonic sieve method (Duifhuis et al., 1982), and Hermes's method of subharmonic summation, SHS (Hermes, 1988). The output of this latter program will be used to determine some of the parameters of this study (see section 2.2.2).

2.2.2 F_0 algorithms used in this study

In order to calculate perturbation measures, information on the onsets of the fundamental periods is needed. In the present study this information is obtained from the algorithm that was developed by van Bergem (1990). This algorithm is based on a two-stage process and operates in the time domain. First, on the basis of a simple autocorrelation algorithm (Rabiner et al., 1976), a rough estimate of the fundamental frequency contour is made. An analysis window of 30 ms and a step size of 10 ms are applied. Unvoiced frames are

supplied with the pitch value of their voiced neighbour, or in the absence of such a neighbour, with the overall median fundamental frequency. Next, the speech signal is bandpass filtered around this measured contour with a narrow passband. The filter used is a Kaiser window (Crochiere and Rabiner, 1983) with a centre frequency that is equal to the fundamental frequency of the current 100 ms segment. Pass band and transition band are both equal to half this centre frequency, and the attenuation in the stopband is 40 dB[1].

The result of the filtering operation is an almost sinusoidal signal. In a (quasi) sinusoidal signal it is easy to obtain estimates of individual period durations; pitch markers are placed at minima in the filtered signal (see Figure 2.1 in section 2.2.4). This position is preferred over other possible positions because the position of these markers has been found to correspond closely to the closure moment of the vocal cords, as measured by a laryngograph (Dologlou and Carayannis, 1989). Van Bergem's method of measuring F_0 is not the only F_0 determination algorithm that produces fundamental period estimates. When compared to e.g. Reetz (1989), however, it is better applicable to perturbation measurement because it provides exact measurements of distances between pitch markers, whereas Reetz's algorithm gives local pitch estimates through an averaging mechanism.

In what follows van Bergem's algorithm will be used for pitch period estimation and perturbation calculation. However, the CB parameters used in this study were measured in F_0 contours obtained from Hermes's subharmonic summation algorithm. In the SHS algorithm every 10 ms an F_0 value is determined for a 40-ms segment of speech. The algorithm consists of a number of steps. First, an amplitude spectrum is determined using a linear scale. This spectrum is transformed to a logarithmic scale, and then for each of the following 14 harmonics the spectrum is shifted along the logarithmic frequency abscissa until the harmonic involved coincides with the peak of the first harmonic in the unshifted version of the spectrum. All shifted subharmonic spectra are added to obtain the so-called subharmonic sum spectrum. The maximum value in the sum spectrum is the estimation of the pitch for the frame concerned. Finally, a process of dynamic programming determines the optimal path through the candidate peaks of the subharmonic sum-spectra of the frames.

2.2.3 Scalings of F_0

Fundamental frequency measures are used in a number of different disciplines, such as in physics, in music, in hearing research and in phonetics. In physics the most widely used unit is the hertz (Hz), which is defined as the number of cycles per second ($1/T$).

A measure that is more related to the perception of musical tone intervals is the semitone, one twelfth of an octave, or a 5.95 % increment or decrement. The semitone is a relative measure, as it expresses the relative distance between two tones in a musical interval. To enable the use of semitones in a more absolute way, the Acoustical Society of America proposed a baseline tone: the musical tone C_0, which corresponds to 16.35 Hz (Fletcher, 1934)[2]. By means of this standard, fundamental frequency values can be

---

[1] This method complies with the demands formulated by Titze et al. (1987) in relation to jitter measurements. They found that filtering with a bandwidth of less than 20 % of the fundamental frequency began to deteriorate the measurement.

[2] In practical applications, C_0 is often set at 50 Hz. For our study the value of C_0 is irrelevant, as we only use ST to measure pitch *differences*.

expressed in semitones:

$$F_0 = 12 \times^2 \log\left(\frac{F_0^*}{C_0}\right) [ST], \quad (1)$$

where F_0 is the fundamental frequency in [ST= semitones], F_0^* the fundamental frequency in [Hz], and C_0 the reference tone in [Hz].

In this study we will generally measure absolute fundamental frequency in Hz. However, relative differences in fundamental frequency, such as the F_0 difference between the onset and the end of a pitch movement, will not be measured in Hz. In higher F_0 regions, higher relative F_0 differences are found. If these differences were expressed in Hz, they would become spurious measures, because speakers with higher F_0's, for instance female speakers, would always show larger F_0 differences, even if these differences were perceptually equal. Therefore, F_0 differences will be expressed in semitones (ST).

The semitone measure has long been the most important rival of the Hz measure, because of its relationship to musical scales, and assuming that speech melody is perceptually evaluated as a sung melody. However, we are aware of the advancements in psychoacoustics, where attempts were made to develop scales that are more accurately related to the perceived pitch, such as the Mel scale (Stevens et al., 1937), the Bark scale (Fletcher, 1940), and the Equivalent Rectangular Bandwidth (ERB) scale (Patterson, 1976)[3].

2.2.4 Acoustic correlates of loudness: intensity, sound pressure, amplitude

A characteristic readily attributed to sounds is *loudness*. As early as 1934, Fletcher showed that the psychological scaling of loudness is quite complex, and not independent of the fundamental frequency and spectral and durational properties of a sound. However, the measurement of some of the important physical correlates of loudness, i.e., intensity and sound pressure, is not very difficult.

The universally accepted scaling system for intensity level is the logarithmic *decibel* measure. Intensity level (*IL*), in bels, is equal to the logarithm of the ratio of an intensity (i.e., sound power per unit area) to a reference intensity. By convention, the reference intensity used is the threshold of normal hearing, 10^{-12} W/m^2. Expressed in decibels [dB], IL must be multiplied by 10. A sound of, say, 10^{-9} W/m^2 is associated with an IL of:

$$IL = \log\left(\frac{10^{-9}}{10^{-12}}\right) = \log 10^3 = 3\ [bel] = 30\ [dB]. \quad (2)$$

The logarithmic nature of this measure corresponds rather closely to the perception of intensity.

---

[3] The ERB scale, like the Bark scale, is derived from measurements of the frequency selectivity of the human auditory system. At the outset of this study the semitone scale was more widely used than the above-mentioned psychoacoustic measures. However, the ERB scale has become more widely used since the publication of Hermes and Van Gestel (1991), who found that the perception of pitch movements is best fitted to the ERB scale.

In practice, sound intensities are mostly measured in terms of sound pressures as transduced by a microphone. Sound intensity is directly proportional to the square of the sound pressure. Therefore, the Sound Pressure Level (*SPL*), i.e., the sound pressure equivalent of Intensity Level, is defined by the formula:

$$SPL = 20 \times \log(\frac{P}{P_0}) \ [dB], \tag{3}$$

which means that the sound pressure level is the logarithmic transform of the ratio of a sound pressure P and a reference pressure P_0.

Not the actual sound pressure signal, but its electrical potential equivalent is stored in the computer. The electrical potential representation of the wave form is treated as an equivalent of the original sound pressure pattern. After digitization of a speech recording, the amplitude of the wave form is expressed in sample values, the maximum and minimum of which depend on the number of quantization bits. After digitization with a bit rate of 12 bits per sample, for instance, the maximum possible amplitude value is 2047, and the minimal value is −2048.

It is not immediately clear how the sound pressures from which SPL is obtained should be derived from the sample values. The sample values exhibit a constantly changing pattern, and some sort of integration process appears to be necessary to express the magnitude of the sound pressures over a period of time. Perhaps the most common specification of the "mean" pressure in an alternating current signal is the RMS, the root-mean-square (e.g. Baken, 1987). It is the amplitude that a direct current signal would need in order to deliver the same power as the alternating current signal. For any period of time, the RMS can be determined by taking the square of the amplitude of the speech wave at every point in time, finding the mean of these squared values for the measurement period, and taking the square root of this mean. The period of time for which an RMS is determined is called the integration time.

A much simpler measure that is related to loudness is the absolute (in the mathematical sense) peak amplitude of a cycle. This measure is different from the measures that are often used as a basis for SPL (Baken, 1987), and the two must not be confused. In a pilot study with a hundred speech fragments, each consisting of 15 seconds of spontaneous speech, obtained from 50 speakers[4], we found a high correlation between our peak amplitude-based parameters and the intensity-based ones. In (especially the older) literature, measurements of peak amplitude are often used as a correlate for loudness, for reasons of computational simplicity. We will measure peak amplitude on a period-to-period basis. In Figure 2.1 it is shown how van Bergem's pitch algorithm determines boundary lines between pitch periods. The peak amplitude value of a period is the maximum value between these boundaries.

[4] These speech fragments are a part of the material that will be discussed in subsequent chapters.

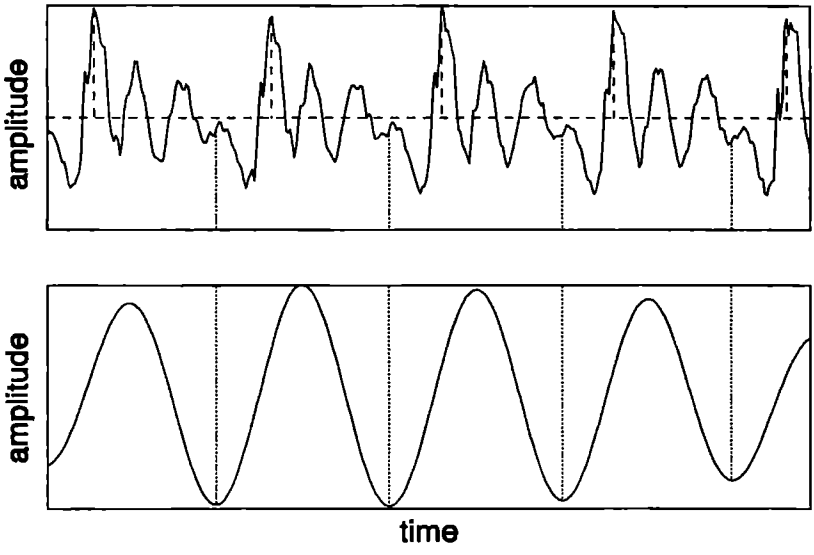

Figure 2.1 Illustration of measurement of period duration and peak amplitude in the periods. In the upper part of the figure the original speech waveform is displayed. In the lower part the filtered waveform is shown, with markers at minima in the signal. The peak amplitude values are determined between the markers *in the original waveform.*

2.2.5 Measures of central tendency

The mean pitch period duration and the mean fundamental frequency are measures that have often been found to be powerful speaker identifying measures (see Chapter 1). Determining a measure of central tendency for the maximum amplitudes in the cycles requires rather severe restrictions. It is vital that a constant mouth-to-microphone distance is maintained, and at an exact angle. To reach this goal, "either a bulky apparatus must be worn on the head or the talker's head position must be severely stricted" (Hollien, 1990). These measures were considered to frustrate the naturalness of the speech. Therefore it was decided not to use any central tendency measure of an acoustic correlate of loudness.

Before discussing mean F_0 itself, we will briefly focus on different measures of central tendency as such. Instead of the arithmetic mean (from now on simply referred to as "mean"), one can also use other measures of central tendency, such as the median (the value with an equal number of values above and below it) or the modal value (the value with the highest frequency of occurrence).

The relevance of discussing the modal and median values beside the mean depends on the *shape* of the sampling distribution. This shape can be described by the amount of *skewness*, the degree to which a distribution is symmetrical, and the amount of *kurtosis*, the degree of "peakedness". The more skewed a distribution is, the larger will be the distance between the mean, the median, and the modal value of the distribution. This difference will be larger if the distribution has a *platykurtic* (flat-topped) rather than a *leptokurtic* (sharp-topped) shape.

In speech samples, the mean, median and modal values of F_0 are probably different from each other. For two samples of 10 speakers of Polish, Jassem et al. (1973) found that most of the distributions of the F_0 values of different frames were positively skewed, which means that the median value was lower than the mean value. Tielen (1992), however, found no large differences between the mean, the median and the modal values of the fundamental frequency of 10 male and 10 female speakers of Dutch.

It is important to know whether mean, median and modal F_0 values measure the same phenomenon and share a large portion of their variance (i.e., have a high correlation). Boves (1984), in an analysis of data obtained from six male speakers, found a correlation of .892 between linear F_0 mean and F_0 mode, and of .994 between the mean F_0 and the median F_0 value. Horii (1975) found a correlation of .995 between mean F_0 and median F_0. In Chapter 4 we shall report on data analyses of 1000 speech fragments with a duration of 15 seconds each, obtained from 50 speakers. For all 1000 fragments, we found a correlation between mean and median F_0 of .999. On the basis of these high correlations we decided to use only one of these measures in the present study, the mean F_0, in Hz.

To establish mean F_0 we will use the period durations obtained by van Bergem's algorithm (discussed in section 2.2.2). All periods with a duration exceeding 20 ms (i.e., a frequency lower than 50 Hz) will be discarded as being either unvoiced speech parts or the result of measurement errors. Next, all periods with a duration that is removed more than 2.5 times the standard deviation from the mean cycle duration of the speaker are discarded. Finally, the mean of the remaining period durations will be determined and expressed as a frequency value. The inverse of the mean period duration corresponds to the number of fundamental periods per second. The measure will be referred to as **F_0MEAN**.

2.2.6 Distributional measures

Variability of F_0: People who speak with equal mean F_0 can differ substantially in their F_0 variability: they can speak monotonously, staying close to the mean F_0 value, or they can have a vivid intonation, in which case they will show greater variability of F_0.

Variability of F_0 is often expressed as the standard deviation of F_0 or as the *range* of F_0: the difference between the highest and the lowest values found in the sample. A disadvantage of the range, however, is that it is derived from only two measurements, which need not be representative of the distribution. As an alternative to this "100 % F_0 range", some authors leave the highest and the lowest 5 % of the data out of consideration, thus using 90 % of the total F_0 distribution (e.g. McGlone and Hollien, 1963), thus reducing the influence of measurement errors. Van Bezooijen (1984) reported a high correlation between an F_0 range based on 80 % of the F_0 distribution (expressed in ST) and the coefficient of variation of F_0: it was .95. For six male and four female speakers, Horii (1975) found that an F_0 range based on 90 % of the F_0 distribution was highly correlated with the standard deviation (r = .982, n= 65), but that the former showed less constancy between different sentences. Thus, it seems that the standard deviation and the range measures are strongly related to each other.

A disadvantage of both the standard deviation and the range of F_0 is the fact that high mean values of F_0 are often accompanied by high values of these variability measures. It is possible to control for this effect by using the *coefficient of variation* (Cramer, 1946) instead of the raw standard deviation. Deal and Emanuel (1978) first used

the coefficient of variation for F_0 and referred to this measure as the variability index[5,6]. The Coefficient of Variation of Period (CVP) in a signal is defined as the standard deviation of the period durations (σ) divided by their mean (M), times 1000:

$$CVP = \frac{\sigma}{M} \times 1000. \quad (4)$$

The coefficient of variation of the period durations will be used in the present study as a measure of F_0 dispersion It will be referred to as CVP.

Variability of peak amplitude In section 2.2.5 it was explained that the mean peak amplitude is considered to be a measure that is influenced too much by uncontrolled recording conditions, such as the mouth-microphone distance, to be of any use. As long as it is independent of this mean amplitude, however, the *dispersion measure* of the peak amplitudes *can* be used Deal and Emanuel (1978) defined the amplitude variability index (AVI) as

$$AVI = \log(\frac{\sigma}{M} \times 1000), \quad (5)$$

where M is the mean peak amplitude and σ the standard deviation of the peak amplitudes. An important reason for applying a logarithmic transformation is that Deal and Emanuel found a more linear relationship between the log-transformed AVI on the one hand and median roughness ratings and Spectral Noise Levels (SNL's) on the other Previous investigations had shown that a high correlation exists between roughness ratings and SNL measures Another reason is the general finding that logarithms of intensity levels are the measurements proportional to human loudness judgements We will apply the same measure in this study and, although it is not actually a coefficient of variation, will refer to it as CVA, for Coefficient of Variation of the maximum Amplitude

2 2 7 Short-term voice instability

Periodicity in sounds is a feature that primarily depends on the regular repetition of sound pressure patterns in time. There are, however, limits to this regularity Simon (1927) already noted that neither in vocal, nor in instrumental sounds, are there tones of constant pitch The seemingly random cycle-to-cycle variations in fundamental period duration are referred to as fundamental frequency perturbation, when discussed in the context of production data, or as jitter when discussed in terms of the perceptual correlate of this phenomenon The cycle-to-cycle variations in amplitude are likewise referred to as

---

[5] The definition of the coefficient of variation used by Deal and Emanuel (1978) does not agree with the generally used statistical index, because they divided the *variance* by the *squared* mean instead of the standard deviation by the mean

[6] Deal and Emanuel used the coefficient of variation of F_0 as an index of frequency perturbation (perturbation phenomena are discussed in section 2 5) This index cannot be regarded as a proper perturbation index, because it is not based on cycle-to cycle variation In the present study this measure will be used as a general measure of variation

amplitude perturbation (in the production realm) or shimmer (for the perceptual correlate).

The differences in the duration or amplitude of adjacent fundamental periods can have two quite different causes. They can be due to more or less gradual changes in pitch or amplitude level, which together form the pitch or amplitude *contour*, or they can be the result of more random vibratory qualities of the voice source.

The first of these causes, changes in the pitch or amplitude contour, should not be studied on the level of individual periods. It is therefore important to control for these gradual changes either by only measuring sustained voice sounds, or by compensating for these long-term changes, using a technique we will discuss in section 2.2.8.

Titze et al. (1987) mention four physiological sources of perturbations: *neurological* (randomness in the action potentials of laryngeal muscles, creating fluctuations in the muscle forces and the configuration of the larynx), *biomechanic* (randomness in the distribution of mucus on the folds and asymmetries in vocal fold structure), *aerodynamic* (randomness in the flow emerging from the glottis), and *acoustic* ones (irregularity in source-vocal tract interactions that stem from nonstationary articulatory configurations). We will not go into details here. It suffices to note that extreme jitter and shimmer measures might be indicative of some malfunction of the vocal apparatus.

Because of this hypothesized relationship between perturbation measures and the physiological condition of the vocal apparatus, perturbation measures have received much attention in speech pathology research. One of the first to observe a tendency for jitter to increase in the presence of vocal pathology was Lieberman (1963). His observation stimulated many researchers of vocal pathology to use acoustic measures in their study of the speech signal. It was assumed that the use of these measures could lead to simple and reliable methods for the detection and diagnosis of speech disorders, and for monitoring the treatment process. If this were true, it would make perturbation measures particularly amenable to speech pathology practice because of their nonvasive character. Until the advent of acoustic measurement methods, speech pathology diagnosis was traditionally based on auditory judgement by the speech therapist. Therefore, research directed at using acoustic measures was at first mainly concerned with defining the relationship between these measures and perceived qualities. Indeed, the perceived roughness in speech increased with increasing perturbation of amplitude and pitch (Lieberman, 1963; Wendahl, 1966a,b; Deal and Emanuel, 1978). In general, when laryngeal disorders are present, the magnitudes of jitter and shimmer measures have been found to increase considerably (Koike et al., 1977). An example of the use of perturbation measures in speech pathology research is the work of Murry and Doherty (1980). On the basis of perturbation measures they were able to discriminate between normal speakers and speakers with laryngeal cancer.

Some amount of jitter and shimmer will always be present in the speech of non-pathological speakers too, since it is impossible for the human voice to produce tones of exactly constant F_0. It is not quite clear how differences in perturbation measures should be interpreted for normal speakers. Eskenazi et al. (1990) investigated the relationship between various voice quality ratings and several acoustic measures obtained from a vowel phonated by normal and pathological speakers. They could establish a clear relationship between perturbation measures and voice quality ratings of pathological, but not of normal voices. The minor role of perturbation in normal voice ratings can result from the fact that a normal amount of perturbation is not a salient speech feature for listeners.

Although perturbation seems to play a minor role in overall voice quality ratings for normal speakers, it is not the result of a measurement artefact, such as sampling noise. One example of a systematic perturbation effect among normal speakers is the study of Ramig and Ringel (1983). They studied the effects of physical condition and aging on some acoustical characteristics of the voice. It was found that subjects in good physical condition produced vowels of maximum duration with significantly less F_0 and amplitude perturbation than subjects of similar chronological ages who were in poor physical condition. Another example of a (fundamental frequency) perturbation effect in non-pathological speech is the change of perturbation across the menstrual cycle in female speakers (Higgins and Saxman, 1989b).

2.2.8 F_0 perturbation measurement

Measures of F_0 and amplitude perturbation are analogous in that both serve to quantify short-term instability of the vocal signal. In this section the measurement of F_0 perturbation will be discussed. By and large, what will be said about F_0 perturbation bears on amplitude perturbation as well, since the ways in which these perturbation types are quantified are quite similar (Baken, 1987). Data bearing only on amplitude perturbation will be presented later.

Perturbation measures represent involuntary, short-term variations. Baken (1987: 166) describes jitter as "a measurement of how much a given period differs from the period immediately following it, and not how much it differs from a cycle at the other end of the utterance. Jitter, then, is a measure of the frequency variability not accounted for by voluntary changes in F_0".

Unfortunately, there is little agreement about the *way* of measuring F_0 perturbation. Over the years, many different measures have come into use, some of which are very much alike. Pinto and Titze (1990: 1279) comment on this state of affairs "many of the perturbation measures in use today were defined in a rather *ad hoc* fashion. Little attention, if any, was paid to the task of comparatively assessing a proposed measure against measures already in existence". In their article, Pinto and Titze show that many of these perturbation measures are related to each other.

A basic division of the measures is yielded by the difference between measures of perturbation *extent* and of perturbation *rate*. In the first type the absolute magnitude of the durational differences between adjacent periods is expressed, whereas the second type only expresses the relative number of sign changes, regardless of the values themselves. A measure of perturbation rate represents the number of times that the durational difference between adjacent period pairs changes from increasing to decreasing and vice versa. A measure of perturbation extent can also show durational period differences within increasing or decreasing period sequences.

The measures of perturbation extent will be addressed first. To illustrate the abundancy of available perturbation measures, we will briefly discuss some of the measures that were proposed in the literature.

Jitter Ratio: Among other things, such as type of phonatory initiation and termination of speech sounds, perturbation measures are dependent on the overall level of F_0. The magnitude of the durational differences between adjacent periods is in general larger if the mean period duration is larger (see e.g. Lieberman, 1963; Koike, 1973; Horii, 1980). To separate the effects of the overall level of the period durations from the perturbation

per se, Hollien and Jackson (1973) constructed the Jitter Ratio: i.e., the mean of the duration differences for all adjacent pairs of two periods divided by the mean period duration, times 1000.

Horii (1979) examined the relationship between the Jitter Ratio and "mean jitter", i.e., the mean of the duration differences for all adjacent pairs of two periods, in milliseconds. Six male adults were asked to phonate /i/ at F_0's ranging from 98 to 298 Hz. The middle segments of the vowels were examined for period-to-period perturbation. Results showed that the Jitter Ratio is indeed relatively constant for mean F_0 values between 98 Hz and 210 Hz. Above about 210 Hz, however, "mean jitter" remains relatively constant and, consequently, the jitter ratio increases as mean F_0 increases. This shows that dividing the frequency variation by the mean F_0 only separates the perturbation effect from the overall mean F_0 in the lower frequency ranges. Orlikoff and Baken (1990) found that the relation between the extent of perturbation and mean F_0 is neither simple nor linear. Proportional measures obscure the perturbation values of women and should only be used for the lower male fundamental frequencies. Since there is no better way to compensate for the mean F_0 effect, however, most perturbation measures still employ the normalization originally applied in the Jitter Ratio.

Relative Average Perturbation: As was mentioned earlier, changes of fundamental frequency and amplitude can be of two types; relatively slow and steady changes, due to the intonation contour, and rapid, quasi-random changes. Changes of the first type belong to the contour-bound parameters and will be discussed in section 2.4.

If the material to be investigated does not consist of sustained vowels, but of utterances, it is necessary to control for the intonational changes. In order to compensate for the effect of slow F_0 movements on the measurement of F_0 perturbation, Koike (1973) introduced the concept of Relative Average Perturbation (RAP). In this parameter the effects of the slow movements are factored out. For all trios of successive cycles, the mean is determined and compared with the duration of the middle cycle. In a slow, gradual pitch movement, these two values will be very close to each other. As a result of such a movement, the individual period perturbation will be low. The mean of all these durational differences is divided by the mean period duration to compensate for the effect of mean F_0 on the perturbation measure.

The process of comparing the mean of a number of consecutive elements in a numerical sequence is often referred to as determining the *moving average* of that sequence and the number of periods considered (in RAP three) is called the *window*. Of course, instead of using a window of three adjacent period durations, one could use other window sizes, e.g. five, seven, or nine. The optimal number of periods to be used in the calculation appears to depend on the material to be analysed. In stretches of connected speech, pitch movements resulting from the intonation contour will influence the perturbation function. These movements belong to the domain of intonation phenomena, and not to perturbation. By applying higher-order perturbation functions, i.e., perturbation measures with larger moving average windows, one can filter out most of the influence of these intonational pitch movements.

Period Perturbation Quotient: Davis (1976) systematically investigated the benefit of changing this window size, and found that five-point averaging windows produced the best differentiation between normal and pathological speakers, both for pitch *and* amplitude perturbation. Davis calls the former measure the Pitch period Perturbation Quotient, **PPQ**:

$$PPQ = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}\left|\frac{1}{5}\sum_{j=-2}^{2}P_{i+j}-P_i\right|}{\frac{1}{N}\sum_{i=1}^{N}P_i}, \qquad (6)$$

where P_i denotes the duration of the ith period and N the number of periods in the speech sample. The five-point averaging window can only be used between the third period and period (N–2), since for the first two periods in the sample no period durations are available to determine P_{i-2} and/or P_{i-1}, while for the last two periods in the sample there are no values for P_{i+1} and/or P_{i+2}. The fact that four periods of the sample are not used in the perturbation measurement explains why, instead of dividing the total of the perturbations by N, we must divide by N–4.[7]

Since Davis's study, the five-point window-size has been applied in many studies (e.g. Kasuya and Kobayashi, 1983; Higgins and Saxman, 1989a; Schoentgen, 1989). Keeping in mind the above-mentioned assumption of Pinto and Titze (1990) that higher-order perturbation functions emphasize quicker variations or shorter-term phenomena, in the present study a five-point window will be applied as well, since a smaller window-size (i.e., a function of a lower order) might amplify the effects of the slower period variations that are related to the pitch contour.

The basis for all measures of perturbation extent is a moving average score of period durations. To illustrate measures of the extent type, an example will be used. Later, this example will also be applied to a measure of perturbation rate. Suppose we want to assess a short fragment of speech, with the following series of period durations, in [ms]:

{10.2 9.5 9.7 9.9 9.2 9.7 9.0}

The mean of this sequence is 9.6 ms. In measures of perturbation extent a window (consisting of e.g. three or five periods) is moved through the entire sequence of periods in the speech sample. In each step, the difference between the value of the middle period in the window is compared with the mean value of all the periods in the window. If a speech sample consists of segments of smoothly increasing, decreasing, or stable F_0 values, the differences between individual period durations and the moving average will be small. If we establish the perturbation extent in the example by comparing within windows of five periods, we find perturbation values of:

[7] Actually, four perturbation values are lost for each of the voiced speech stretches in a speech sample. If a speech sample contains three stretches of voiced speech, separated by two intervals of unvoiced speech, the number of periods used in the calculation is N–12, instead of N–4.

	values	mean value	middle value	abs. diff.
1st group	{10.2 9.5 9.7 9.9 9.2}	9.7	9.7	0.0
2nd group	{9.5 9.7 9.9 9.2 9.7}	9.6	9.9	0.3
3rd group	{9.7 9.9 9.2 9.7 9.0}	9.5	9.2	0.3
mean				0.2

Strictly speaking, all the differences between the middle values and the moving average values are called perturbations, and these values are summarized by their mean[8], in the present example 0.2 ms. In the PPQ measure, this mean perturbation is divided by the mean cycle duration, 9.6, yielding about 0.02. Normally, however, the term perturbation is used to refer only to the mean perturbation. Since we will primarily use the mean perturbation, we will also simply refer to this mean perturbation as perturbation.

Titze et al. (1987) investigated the technical limitations that need to be considered in voice perturbation measurements. A major problem in jitter and shimmer measurement is the fact that all digital recording systems have some amount of quantization noise. Quantization noise is the result of the finite precision of all measurement systems: minute differences between two measurements can be amplified by rounding. By comparing the theoretical quantization noise with normal vocal shimmer and jitter in vowels, Titze et al. found that 500 samples per cycle are needed to minimize the influence of sampling noise. For a speech sample with a period durations of 5 ms (corresponding to F_0 = 200 Hz), this recommendation would necessitate a sampling frequency of 100 kHz. However, Schoentgen (1989) found that PPQ values measured in voiced portions of connected speech were roughly 10 times higher than those measured in sustained vowels produced by the same speaker. Therefore, a sampling frequency of 10 kHz appears to meet the requirements for connected speech.

So far we have discussed parameters that Pinto and Titze (1990) classified as measures of perturbation extent. The second kind of measures they distinguish are measures of perturbation rate. These are again determined by comparing the period durations of adjacent periods. The number of sign changes is counted from period to period, irrespective of the values themselves, and is divided by the maximum number of sign changes possible within the utterance. As an example, the perturbation rate of the period durations presented earlier to elucidate measures of perturbation extent, will now be demonstrated. The durations of the periods in the example were:

{10.2 9.5 9.7 9.9 9.2 9.7 9.0}

From 10.2 to 9.5, the period duration decreases. Here a trend change is impossible, since the difference between the first two periods determines the direction of the change:

[8] Apart from the mean, sample characteristics of the perturbations, such as their standard deviation are sometimes used (Askenfelt and Hammarberg, 1986). Horii (1976) showed that the standard deviation of period perturbation values correlates highly with mean F_0 perturbation.

decreasing. Between 9.5 and 9.7, a change of direction occurs: from decreasing to increasing. A change does not occur between the next two periods, as the duration increases from 9.7 to 9.9. In the next two steps changes of sign take place; first the duration decreases, from 9.9 to 9.2, then it increases, from 9.2 to 9.7. Finally, in the last step, the duration decreases from 9.7 to 9.0. In total, five sign changes could have occurred, and four actually occurred. Thus, the perturbation rate is 80 %.

Hecker and Kreul (1971) first proposed this measure under the name *Directional Perturbation Factor* (DPF). Hecker and Kreul (and Murry and Doherty, 1980) found that their DPF, contrary to a measure of perturbation extent, could differentiate normal from pathologic speakers.

Higgins and Saxman (1989a) compared intrasubject variation across sessions of three F_0 perturbation measures in steady-state productions: the Jitter Factor (JF, strongly related to the jitter ratio), the pitch perturbation quotient (PPQ), and the directional perturbation factor (DPF). The Jitter Factor and PPQ turned out to be highly correlated, while DPF apparently measured a different perturbation phenomenon. Since JF and PPQ do not seem to measure different aspects of vocal behaviour, in this study only PPQ will be considered. Another difference reported by Higgins and Saxman is that JF and PPQ varied considerably within individuals across sessions while DPF was the most temporally stable measure. From the Higgins and Saxman study we conclude that the difference between measures of perturbation extent and of perturbation rate is large enough to justify the inclusion of both types in the present study.

In signal processing terminology, perturbation rate is called *zero-crossing rate*, because it expresses the percentage of times that the derivative of the period duration function changes sign, i.e., crosses the zero-line. We will adhere to this terminology and will refer to the perturbation rate of the periods in a signal as **PZR**, Pitch period Zero-crossing Rate. The formula for its calculation is:

$$PZR = \frac{1}{N-2} \sum_{i=1}^{N-2} \frac{1}{2} |sign(P_{i+2}-P_{i+1}) - sign(P_{i+1}-P_i)|, \qquad (7)$$

where P_i denotes the period duration of the ith period, N the number of periods in the speech sample, and *sign* a function producing an output of "1" in the event of a positive difference between the first and the second of its terms and "−1" in case of a negative difference. In a series of N periods, the number of possible sign changes is $N-2$.

2.2.9 Peak amplitude perturbation measures

As explained above, the magnitude of the durational differences between adjacent periods is in general larger if the mean period duration is larger. For peak amplitude perturbation no dependence between the values of adjacent periods (as expressed on the decibel scale) and the overall peak amplitude level has been reported[9]. An example of a very straightforward shimmer measure is the measure used by Horii (1980: 205):

---

[9] Very low amplitude levels affect shimmer measures, because of the increased influence of system noise (Horii, 1980).

$$shimmer = \frac{20}{N-1} \sum_{i=1}^{N-1} \left| \log \frac{A_i}{A_{i+1}} \right| \ [dB], \tag{8}$$

where A_i denotes the maximum amplitude in the *i*th period and N the number of periods in the speech sample.

Following the advice of Davis (1976), we incorporated the use of a five-point averaging window into our measure of amplitude perturbation by taking the ratio of the mean of the amplitudes of contiguous period quintuplets, and the amplitude of the middle period of the quintuplet. Henceforth, we will call this measure the Amplitude Perturbation Quotient, **APQ**:

$$APQ = \frac{20}{N-4} \sum_{i=3}^{N-2} \left| \log \frac{\frac{1}{5} \sum_{j=-2}^{2} A_{i+j}}{A_i} \right| \ [dB], \tag{9}$$

where A_i denotes the maximum amplitude in the *i*th period and N the number of periods in the sample.

APQ will be used as a measure of perturbation extent for amplitude in this study. Since it is only possible to determine period demarcations in voiced periods, only data from voiced speech parts will be used; the amplitude values in voiceless parts of the speech signal will not be analysed.

As for F_0 perturbation rate, it is also possible to establish amplitude perturbation rate. We will refer to this amplitude perturbation measure as **AZR**, the Amplitude Zero-crossing Rate. The formula for AZR is:

$$AZR = \frac{1}{N-2} \sum_{i=1}^{N-2} \frac{1}{2} \left| sign(A_{i+2} - A_{i+1}) - sign(A_{i+1} - A_i) \right|, \tag{10}$$

where A_i denotes the maximum amplitude of the *i*th period, N the number of periods, and *sign* a function producing an output of "1" in case of a positive difference between the first and the second term and "−1" in case of a negative difference.

2.3 Temporal TI parameters

2.3.1 Pause time

In the preceding sections we discussed parameters that were *time-integrated.* By integrating period durations over a certain stretch of time, and summarizing them only in terms of distributional F_0 characteristics, one loses information concerning the amount of time the speakers really spend on talking.

Goldman-Eisler (1951), originally interested in time sequences and turn-taking in interview situations, studied the relation between the time speakers actually spent on articulating and the time during which they paused. She came to the conclusion that the duration of the intervals of inactivity between stretches of speech were more speaker-

specific than the measures which are concerned with the speakers' active behaviour. In other words, "tendencies for maintaining long periods of silence or holding up action at one extreme, or incapacity to do so and precipitate action at the other, were found to constitute a relatively permanent feature of individuals' conversational behaviour" (Goldman-Eisler, 1968: 4). In the framework of the present study such a conclusion makes pausing time a promising parameter for speaker identification. Therefore, we introduce the parameter pause time, or **PAUSE**, which is defined as the percentage of the speech frames that do not reach a certain threshold intensity[10].

2.3.2 Articulation rate

An innovation in the work of Goldman-Eisler (1968) is the distinction she makes between speech rate (*SR*), i.e., the number of syllables divided by the duration of the whole utterance, and articulation rate (*AR*), i.e., the number of syllables divided by the duration of the utterance *minus the pause time*.

In material that was obtained from three different speech types, highly significant individual differences in articulation rate were found, which did not depend on the speech style condition (Goldman-Eisler, 1961b). Although there appeared to be differences in speech tempo between the different speech types, these proved to be a direct consequence of differences in pausing time, not in articulation time (Goldman-Eisler, 1961a, 1961b; Miller and Grosjean, 1981). What appeared to be increases in speech tempo was merely a closing of gaps Grosjean and Deschamps (1975) replicated the finding that changes in speech rate were to a large extent attributable to changes in pausing. The differences found in articulation rate were, however, not negligible In a reanalysis of the data from the study of Grosjean and Deschamps (1975), Miller et al. (1984) found that the variation of the articulation rate was substantial, also within speakers.

The reason for applying only articulation rate in the present study is that in speech rate the pauses are included. Since PAUSE is among the TI parameters to be used, the variable speech rate becomes redundant in the presence of articulation rate[11].

To determine the parameter **RATE**, first the number of syllables is counted. In the reading text the syllables can simply be counted from the text (while correcting for mispronunciations), and for a stretch of spontaneous speech the syllables must be counted from the transcripts of the utterances. Next, the number of syllables is divided by the time spent in vocal activity: speaking time minus PAUSE.

2.3.3 Measures of voicedness

Some evidence that the percentage of voicedness can be applied to speaker identification comes from a study by Johnson et al. (1984), who applied a large number of durational parameters to speaker identification. They combined two parameters, "total duration of phonation" and "duration of voiced activity", to form the so-called "voiced/voiceless speech time vector". This vector was found to enable some degree of speaker identifica-

---

[10] An undesirable side-effect of this procedure is that silent intervals in plosives (occlusions) will also be considered "pause time"

[11] Furthermore, we compared the speaker specificity of articulation rate and of speech rate for the data presented in Chapter 3 The speaker specificity, as expressed by the ratios of the between to the within speaker variability, was about equal it was 3 76 for articulation rate and 3 49 for speech rate

tion: 65 %. Thus it appears that the relation between the number of voiced and unvoiced speech samples is a possible speaker-identifying index. In the present study this index will be used as well.

An automatic voiced-unvoiced classification algorithm is an important part of any pitch determination algorithm, as it specifies to which part of the signal no F_0 values should be assigned (Atal and Rabiner, 1976; Siegel and Bessey, 1982). A dividing line between voiced and voiceless speech fragments cannot be defined in absolute terms and the criteria for attributing the feature "voiced" depend, at least partly, on the application they are used in.

In the present study the voicedness of the signal will be determined by the voicing-determination algorithm (VDA) that is included in the subharmonic summation pitch determination software of Hermes (1988). After the actual process of subharmonic summation, a process of dynamic programming determines the optimal path through the candidate peaks of the subharmonic sum spectra in the frames. In such spectra a maximum value can always be found, even when the signal consists of noise. To decide whether the maximum found in a spectrum is really the maximum of a voiced speech fragment, the correlation coefficient is determined between the samples in two adjacent signal intervals of the duration of the estimated pitch period T (i.e., $1/F_0$). The first of these periods starts at T (the duration of the period corresponding to the F_0 value established by SHS) seconds before the middle of the frame, the second starts at the middle of the frame. If the fragment is really voiced, there must be a sizable correlation between the two periods. Frames with correlation coefficients smaller than .52 are provisionally classified as "unvoiced". Since the correlations sometimes fluctuate even within vowels, it is necessary to ensure that such voiced segments are not classified as "unvoiced". Therefore, the sequence of provisional voiced/unvoiced judgements is re-evaluated by means of a five-frame window. Within each window the middle frame is classified as "voiced" if three or more of the provisional judgements in the window were "voiced".

The quality of the voiced/voiceless decision taken by the SHS algorithm was tested by Hermes (1988, 1993). In the first of these studies, Hermes compared the performance of his voicing-determination algorithm with the decisions of the VDA used in the harmonic sieve method (Duifhuis et al., 1982) and with a visual inspection of the wave form. In the second study the parallel-processor algorithm of Gold and Rabiner (1969) was applied as well. The voicedness attribution performance of the SHS algorithm was clearly better than that of the algorithm by Duifhuis et al. and that by Gold and Rabiner.

In this study the speech fragments will be analysed by means of Hermes's VDA, as described above. The number of voiced frames is expressed as a percentage of the total number of frames with speech activity (frames for which the intensity exceeds the threshold set for determining PAUSE). This measure will be called **VOI**.

2.4 CONTOUR-BOUND PARAMETERS

2.4.1 GDI as descriptive model for pitch movements

So far we have discussed only time-integrated parameters. These parameters are an abstraction from reality, since they do not give any information on the actual course of events in the speech signal; from mean F_0, for instance, we cannot deduce what the fundamental frequency at a certain point in time is. Yet, this sort of information, from

measurements at specific points of the contour itself, might well be applicable to speaker identification. Therefore, apart from comparing speakers on time-integrated parameters, we will also study the speaker-identifying properties of contour-bound parameters, i.e., of measurements related to specific pitch movements.

In Chapter 1 we explained that, in order to compare the realizations of pitch movements by different speakers, a descriptive intonation model is first needed to be able to determine whether pitch movements are linguistically comparable.

The first problem in describing the intonation grammar of a language is, of course, to find out on the basis of what information a typification should be formulated. No general agreement exists on the level at which measurements should be taken, and on what constitutes the basic unit of intonation. In the history of intonation research, for a long time the dominant approach has been the "levels" approach, which is characterized by the idea that the speaker primarily tries to hit a certain pitch level, and that the resulting pitch movement is nothing more than the physiological realization of the transition from one level to the other (Pike, 1945). This approach was abandoned after Bolinger (1951) and Lieberman (1965) showed that pitch changes of more than a quarter of a speaker's range can be linguistically irrelevant, while quite small changes can be relevant.

An opposing viewpoint is the "movements" approach, which gives priority to the movements themselves as the most basic units of intonation. A clear choice for the movements approach was made by researchers of the Institute for Perception Research (IPO) in Eindhoven, The Netherlands. Their studies have resulted in a rather detailed description of the perceptually relevant pitch movements of Dutch ('t Hart et al., 1990) and other languages (British English: de Pijper, 1983, Willems et al., 1988; German: Adriaens, 1991; Russian: Odé, 1989; French: Beaugendre et al., 1992; Indonesian: Ebing, 1994, 1997). The IPO model of Dutch intonation is often referred to as the Grammar of Dutch Intonation (GDI).

A third approach to intonation is the "targets" approach, in some sense a synthesis of the above-mentioned viewpoints. According to the adherents of the targets approach, speakers aim to reach certain intonational targets, and the intonation contour can be regarded as the result of an interpolation between these targets (Pierrehumbert, 1980; Ladd, 1983; Gussenhoven, 1984; Liberman and Pierrehumbert, 1984; Pierrehumbert and Beckman, 1988). Although the target approach might have advantages in the sense of the phonological interpretation of pitch contours, it is not very useful to our present purposes due to its abstractness.

GDI appears to be well fit for our purposes, as it offers a useful classification scheme for pitch movements. GDI will be described in this section and used throughout this book.

The approach by which the IPO researchers characterized the ten pitch movements of GDI was based on perception. On the relation between perception and production, 't Hart et al. (1990: 70) remark that "only those F_0 changes are relevant for perception that have been voluntarily produced by the speaker as physical properties that are cues to the intonation pattern that he wants to produce". 't Hart et al. acknowledge that there are systematically occurring F_0 changes that are not voluntarily produced[12], but these move-

[12] An example of an involuntary F_0 movement is the fall-rise that is often found at intervocalic consonants. Such a movement is related rather specifically to the segmental layer and therefore not truly "prosodic".

ments are usually perceptually irrelevant. Although the present study is concerned with production data, the use of perception-based transcription criteria does not appear to be unfit for our purposes, as long as these criteria could justifiably be called "prosodic". As this is true for GDI's pitch movements, we consider GDI a useful categorization scheme for our purposes.

GDI distinguishes ten pitch movements, five rises and five falls. The identity of each pitch movement and the extent to which it differs from all the others can be expressed with a notation in terms of binary distinctive features. It appeared to be necessary to distinguish five features for the specification of the pitch movements in Dutch:

RISE indicating that the movement is a rise [+rise] or a fall [−rise],

EARLY indicating that the offset of the movement is located near the beginning of the voiced part of the syllable [+early] or not [−early],

LATE indicating that the offset of the movement is located near the end of the voiced part of the syllable [+late] or not [−late]. A movement can be both [−early] and [−late], in which case it is located near the middle of the syllable,

SPREAD indicating that the movement is associated with two or more successive syllables [+spread] or confined to one syllable [−spread],

FULL indicating that the movement covers the full distance between the upper and lower declination lines (see section 2.4.4) [+full] or that it is smaller than the standard size [−full].

To summarize the description of the IPO pitch movements, in Table 2.1 we reproduce table 6.1, as presented in 't Hart et al. (1990), in which the 10 pitch movements of Dutch are defined:

Table 2.1
The pitch movements of Dutch, taken from 't Hart et al., 1990 (table 6.1)

	/1/	/2/	/3/	/4/	/5/	/A/	/B/	/C/	/D/	/E/
RISE	+	+	+	+	+	−	−	−	−	−
EARLY	+	−	−	−	+	−	+	−	−	+
LATE	−	+	−	+	−	−	−	+	+	−
SPREAD	−	−	−	+	−	−	−	−	+	−
FULL	+	+	+	+	−	+	+	+	+	−

Having defined the pitch movements of GDI, we now go on to find out how these basic prosodic units can be combined to form intonationally correct Dutch utterances. 't Hart et al. showed that pitch movements enter into a limited number of sequences, called *configurations*. Fall "C", for instance, always has to be preceded by rise "3", forming the configuration "3C", while the converse is not true. Combinations of configurations build up complete intonation contours, extending over a domain that more or less corresponds to

a syntactic clause. Much more can be said about the ways in which pitch movements are combined into configurations, and configurations into contours. For our present purposes, however, it is only important to be familiar with the pitch movements per se, as in this study these are considered to be the basic units of intonation.

2.4.2 Measures taken at specific pitch movements

In this study we describe pitch movements by the characteristics of their start and end points. All intonation approaches discussed above (levels, movements and targets) stress the importance of the start and end points, or *pivot points*, for the description of pitch movements. Other reasons for applying pivot points is the increased measurement simplicity and the fact that at the outset of this study little was known about the most appropriate mathematical functions by which the movements would have to be characterized (for recent progress in this domain cf. Fujisaki and Ohno, 1995).

We shall parameterize pitch movements by using the F_0 at the start and end of the movements, the time interval in which they take place and, derived from these two measures, the slope of the movements. Ideally, the contour-bound parameters should be measured in the pitch movements of utterances with identical pitch contours, since only in such utterances can the differences in the performances of the speakers be attributed to the speakers, and not to contour differences. Therefore we will try to find sentences that, when read aloud, elicit such utterances from our speakers. In Chapter 3 it is described how test sentences were selected from a large corpus of readings of sentences.

The pivot point parameters will be denoted by six-character acronyms of which the first three letters represent the actual measure (e.g. SLO for slope, DUR for duration). The last three letters are reserved for denoting the pitch movement to which the parameter relates. The precise naming of the parameters used in the present study is postponed until Chapter 5.

Finding the exact location of the pivot points in an F_0 contour is difficult. The process would be facilitated considerably if the contour could first be stylized into a set of straight lines. Hermes (Hermes, p.c.) developed a computer program that simulates so-called "close-copy" stylizations (see 't Hart et al., 1990), and produces perceptually equivalent F_0 contours with straight lines. However, we found that, although the program probably performs well at stylizing intonation contours, it sometimes results in pivot points that fall outside the original F_0 trace, even when an unambiguous pivot point is present in the contour.

We decided to demarcate the pivot points by visual inspection of the F_0 traces, as obtained from Hermes's (1988) SHS algorithm, combined with audio-feedback. For the majority of the pitch movements studied, it was possible to find local minima and maxima in the F_0 trace that appeared appropriate as pivot points for the movement under consideration. In some cases problems arose, however, and a decision procedure was formalized in a protocol, which is presented in Appendix A. In Figure 2.2 some of the problems encountered in manually measuring pivot points are shown.

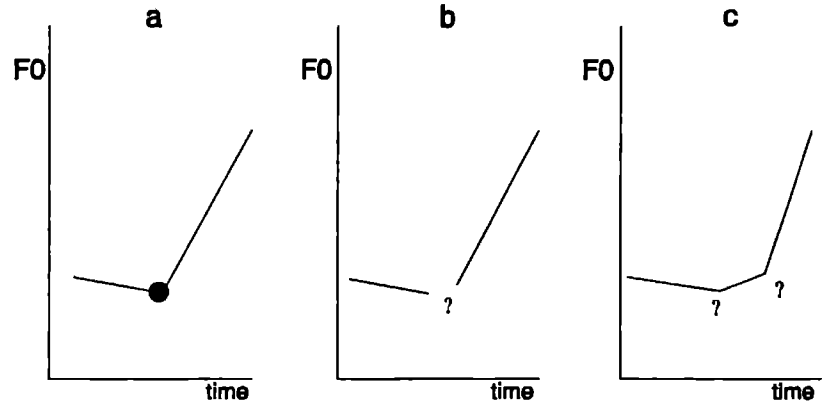

Figure 2.2 Examples of the problems encountered in measuring pivot points: (a) represents the ideal situation, since an unambiguous pivot point is available; in example (b) the start of the rise cannot be determined directly, as the pivot point is situated in an unvoiced part of the utterance; in (c) a point of inflection is found between the lower and upper pivot point. In such a case it is hard to decide where the pitch movement starts.

2.4.3 Synchronization: alignment of pivot points with segmental structure

Above it was explained that one of the distinguishing features in the description of the GDI pitch movements is the position of the movement within the syllable. The peak of a pitch rise of type "1" is aligned with the segmental structure at 40 ms (Collier and Terken, 1987) or 50 ms ('t Hart et al., 1990) after the onset of the vocalic nucleus of the relevant syllable, and the start of the pitch fall of type "A" is located near the middle of the syllable. At least for this fall "A" the timing of the peak appears to be related to the configuration that it belongs to. In a synthesis application of GDI, a value of 20 ms before the vowel onset was used in a "flat hat" configuration, and a value of 80 ms after the vowel onset when the fall was part of a "pointed hat" configuration (Collier, 1991).

The fact that timing is one of the features by which pitch movements are distinguished from each other suggests that something like a fixed timing point does exist for the accent-lending pitch rise and fall. Caspers (1994) compared the start and end points of the pitch movements "1" and "A" and found clear differences between these two movements. To begin with the rise: Caspers found that (1) the onset of the pitch rise, rather than the offset, shows a more or less fixed position relative to the segmental structure; (2) time intervals between the onset of the movements and either of three candidate anchor points (viz. syllable onset, voice onset and vowel onset) were dependent on speaker, time pressure and syllable structure; and (3) the syllable onset appears to be somewhat superior to the other candidate anchor points, since an alignment with the syllable onset showed the least influence of time pressure and syllable structure. The timing of fall "A" was found to be less rigid. Instead, the shape of the fall is relatively invariant within speakers.

At the outset of the present study we decided to use the vowel onset as our reference point, which is in line with GDI. As Caspers found a relatively small advantage of the syllable onset over the vowel onset, we consider the vowel onset to be an efficient

anchor point for the timing of the pitch movements, too.

As speaker specificity is not to be expected where linguistic requirements must be met, some expectations for the present study can be derived from Caspers's work. She found that the timing of the start of the rise "1" relative to the syllable onset and the shape of the fall "A" were rather stable over conditions; for these features little speaker specificity is to be expected.

The demarcation of the vowel onsets was performed by hand, using the high-resolution wave editor SESAM (Broeder, 1990). The segmentation was based on visual wave form information following standard criteria (see e.g. van Zanten et al., 1991 or Rietveld and van Heuven, 1997). We will refer to these time intervals as *synchronization* times. The first three letters of the six-character acronym that is used to represent these parameters are SYN. The last three letters are reserved for denoting the pitch movement to which the parameters relate.

2.4.4 Declination measures

Pitch contours are made up of local phenomena, the pitch movements, as well as of a more global one. As early as 1945, Pike reported a general tendency for F_0 to drop over the duration of a sentence. Cohen and 't Hart (1965) experimentally demonstrated the perceptual relevance of this phenomenon, and named it declination. The downtrend in F_0 appears to be related to the downtrend in subglottal pressure (P_s), which is probably under the active control of the respiratory and/or laryngeal muscles (Strik, 1994)[13].

One parameter that is related to declination has often been found to be speaker-specific: the utterance-final F_0 value. With regard to these values Liberman and Pierrehumbert (1984: 180) state that: "for a given speaker, the utterance-final low-point values in the first experiment are essentially the same as the utterance-final low-point values in the second experiment ... It appears that this final low value is a relatively invariant characteristic of a speaker's voice." For Dutch, Sluijter and Terken (1993) found that speakers have a relatively constant final F_0 value, while sentence-initial F_0 values depend on the duration of the sentence (the longer the sentence, the higher the initial F_0) and the position in the paragraph (the later the sentence appears in the paragraph, the lower the initial F_0. In many other studies the invariance of the "utterance-final low-point value", henceforward F_0END, has been attested as well (e.g. Maeda, 1976; Cooper and Sorenson, 1981; Kutik et al., 1983).

According to GDI, pitch movements occur between two imaginary lines, the upper and lower declination line[14]. It is difficult to establish the exact location of the upper declination line because the number of peak values available is often small (or even absent). Furthermore, the F_0 heights of the peaks possibly reflect differences in the strength of the accents (Rietveld and Gussenhoven, 1985). Because of these problems connected with the upper declination line, we will only take measurements at the lower declination line, which we will simply refer to as "declination".

---

[13] In some studies, however (e.g. Gelfer et al., 1985), it is suggested that the correlation between the decline of F_0 and P_s does not result from of cognitively generated planning processes, but from the intrinsic properties of underlying physiological mechanisms, possibly in the respiratory system.

[14] Since half rises and half falls are possible pitch movements as well, an intermediate trend line can be postulated between these two declination lines.

We consider declination a contour-bound parameter because it is related to contextual factors. It has been found that declination is somehow related to the length of the utterance; in long utterances, higher initial frequencies and more gentle slopes have been found than in shorter ones ('t Hart et al., 1990) and declination also depends on the position of a sentence in the paragraph (Sluijter and Terken, 1993). As a result of our way of parameterizing declination, i.e., as the difference between the F_0 at the beginning and at the end of the utterance, our declination measures do not only depend on the contour of the utterance, but also on its segmental content. Since the present study is concerned with speaker differences in prosodic parameters, and not with segmental factors, such as intrinsic pitch, the F_0 values will be obtained from the same utterances for all speakers.

The best way of representing the amount of declination in an utterance appears to be the use of trend lines that are established according to formal procedures. However, the existence of a number of essentially different definitions of declination (see e.g. Maeda, 1976; Cooper and Sorenson, 1981) complicates the design of such procedures. It is unclear, for instance, whether the initial F_0 rise and the final F_0 lowering, which are often observed in speech (Collier, 1975; Maeda, 1976), should influence the F_0 declination values, or whether they belong to a more local domain, such as pich movements. Furthermore, it is difficult to determine the margins of local F_0 events, such as pitch movements.

Because of this problems, and for reasons of measurement simplicity, we decided to determine as declination measurements the same data as for the pitch movements (see section 2.4.2): the pitch at the start and end of the sentence and the slope of the declination line.

As in the case of the pitch movement measurements, we decided to assess these points simply by visual inspection of the F_0 traces, as obtained from Hermes's SHS algorithm. For most of the utterances that were selected as our experimental material (see Chapter 4), it was mostly not difficult to locate the first and last voiced frame. Some problems did arise, however, and a protocol was designed to allow as systematic a choice as possible. The protocol is presented in Appendix A. In Figure 2.3 some of the problems encountered in manually measuring the end point of an utterance are shown.

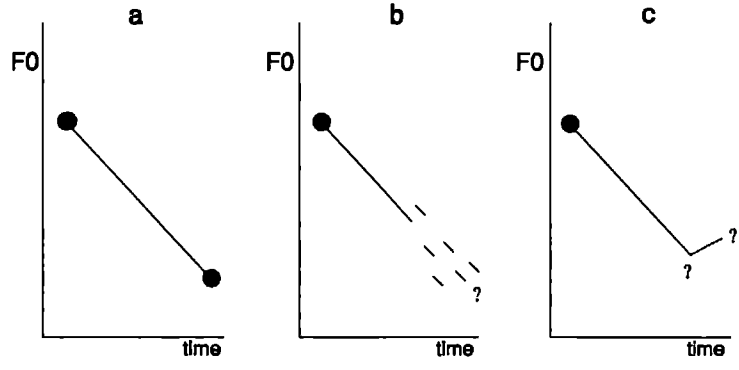

Figure 2.3 Declination determination: examples of the problems encountered in measuring F_0END: (a) ideal situation; (b) it is difficult to determine F_0END, as the utterance ends in creak; (c) F_0END is masked by a rising movement at the end of the utterance (a continuation rise).

Part II: Speech material

2 5 SPEAKERS

In this study speech characteristics are evaluated in terms of speaker-specifying properties To this end, it is of crucial importance to select a well-described group of speakers The classical procedures of statistical inference rest on a sampling scheme in which each population element sampled is independent of all other elements, and is equally likely to be included in any sample (Hays, 1973) The method used to obtain such a sample, *simple random sampling*, allows us to draw inferences about a population of speakers that is larger than the 50 speakers that were investigated in the present study The first problem we face is to define for *what* population we want to draw inferences

In Chapter 1 it was explained that our primary aim in this study is to determine idiosyncratic speaker characteristics It is a well-known fact that factors such as sex and age influence the values found for most parameters to a large extent Furthermore, the effects of sex and age on prosodic parameters other than mean F_0 have not been studied systematically in the context of Dutch For these two reasons we decided to factor out the influences of sex and age as determinants of prosodic behaviour by selecting our speakers so as to obtain a speaker group that was stratified with regard to these factors

Fifty persons, 25 males and 25 females, participated as speakers in this study For each sex, five age groups containing five speakers each were selected The groups consisted of speakers whose ages were 18-25, 26-35, 36-45, 46-55 and 56-65 The composition of the speaker groups is illustrated in Table 2 2

Table 2.2
Stratified sampling of speakers by sex and age The rows represent the two sex groups and the columns the five age groups

	18 25 yr	26 35 yr	36-45 yr	46 55 yr	56 65 yr	total
♀	5	5	5	5	5	25
♂	5	5	5	5	5	25
total	10	10	10	10	10	50

The stratification applied involves somewhat higher requirements than does simple random sampling For *stratified sampling* one has to select the members of the different sex/age groups according to the same criteria, to be able to compare these groups

The factors sex and age were systematically varied, enabling us both to factor out their influence as determinants of prosodic behaviour and to assess their contribution to speaker identification Some of the other (possibly) influential factors were merely controlled by keeping them constant By selecting appropriate speakers we wanted to exclude the influences of speech deficiencies, socioeconomic background, and dialectal colourings

In the author's opinion, none of the speakers showed any speech deficiencies. As an extra check the speakers were asked whether they had ever had any voice problems. Some had, but none of these problems would occur during the sort of tasks they were presently required to perform[15].

Socioeconomic background can be operationalized in terms of an index of Socio Economic Status (SES), in which the level of education of the subjects or a rank order that is related to the subjects' occupations are important criteria (Hollingshead and Redlich, 1958). The speakers in the present study were recruited from students and staff of the Faculty of Arts of the University of Nijmegen, and most definitions would therefore classify our speakers as having high SES[16].

Since this study is not directed at finding dialect or accent-related differences, the author selected speakers who, in his view, spoke Standard Dutch, or a variety very close to it. Further evidence for the presumed low degree of dialectal colouring in the speech of these subjects was provided by a study carried out by van Rie and van Bezooijen (1995), in which the 50 speakers involved in our study were judged together with 64 other speakers on their degree of accentedness. The 114 speakers were judged by a panel of five experts in Dutch phonetics, who were instructed to rate the speakers along an accentedness scale from 1 to 10. On this scale "1" meant that the fragment was spoken with very heavy accent and "10" meant that the fragment only consisted of Standard Dutch speech.

The speakers were judged on three scales of accentedness: general impression, segmental aspects and suprasegmental aspects. For the total data set the inter-rater reliability, Cronbach's α, was high for two of the three scales; it was .92, .90, and .75 for the respective scales[17]. The mean scores over all the 50 speakers involved in our study were 8.28, 8.06, and 8.93, with standard deviations of 1.04, 1.13, and 1.31 on the scales of general impression, segmental aspects and suprasegmental aspects, respectively. The lowest mean scores for general impression and segmental aspects came from female speaker nr. 2 from age group 5; it was 6.3 for both scales[18]. The lowest mean score for the scale of suprasegmental aspects of accentedness, which appears to be most important to our purposes, was even higher: 7.5 (for speaker F51). We conclude that our speakers indeed spoke Standard Dutch[19].

Of course, there are more potentially relevant speaker characteristics than can

---

[15] In fact, the few speakers who reported voice problems experienced these only during prolonged and strenuous vocal effort as in classroom teaching or singing.

[16] By and large the members of the sex/age groups belonged to the same SES group, although there were some secretaries in the oldest group of female speakers, who perhaps should not be assumed to belong to the same socioeconomic status group as the other speakers.

[17] The variability within the speaker group of the present study was small, and therefore the inter-rater reliability α within our set was lower; .80, .74, and .57 for general impression, segmental aspects and suprasegmental aspects respectively.

[18] To denote specific speakers we use a descriptive code in this book. The first character denotes the speaker's sex, the second his age group, and the third indicates one of the five speakers in the sex/age group; female speaker nr. 2 from age group 5 will thus be referred to as F52.

[19] Thanks are due to van Rie and van Bezooijen for making their results available to us.

practically be controlled. One of these characteristics was smoking behaviour[20]. If we define smokers as people who smoke more than five cigarettes a day, or who did so until recently (less than five years ago), 12 of our 50 subjects were smokers; five women (F13, F35, F41, F43, and F45) and seven men (M11, M15, M32, M34, M42, M52, and M53).

Some other conditions that were not strictly controlled are: the time of day of the recordings and the physical and emotional situation of the speakers. The time of day at which the recordings took place was during office hours, with the exception of the earlier hours: from 11.00 a.m. to 5.00 p.m. The precise times at which the recordings took place, as well as most of the speaker characteristics discussed above, are reported in Appendix B.

We neither inquired about nor assessed the general physical condition of our speakers or their emotional situation. To make the speakers feel more at ease they were globally informed about the object of the study and were given coffee or tea during a short break in the first session.

2.6 Material

2.6.1 Introduction

As was explained earlier, two types of prosodic parameters were included in this study, time-integrated (TI) and contour-bound (CB) parameters. The shorter the fragment of speech is, the more TI parameters depend on short-term phenomena. In the first place, the influence of segmental features, such as intrinsic pitch, is larger in short speech fragments. Furthermore, the role of local prosodic phenomena, such as the presence or absence of specific pitch movements is larger. Over longer stretches of speech we assume that the influence of both kinds of short-term phenomena is averaged out, and that the TI measures will become more stable. The time-integrated variables are most commonly used in speech samples for which no attempts were made to control the pitch movements produced by the speakers.

With respect to the necessary duration of the speech fragment, Barry et al. (1991: 38) assert that, at least for average F_0 "little has been done to clarify the general question of individual stability". They themselves found considerable variation in fragments of two minutes' duration. However, the amount of speaker-specificity in acoustic parameters depends on the ratio of the between-speaker variation and the within-speaker variation. A large amount of within-speaker variation can be compensated for by an even larger amount of between-speaker variation.

For most of the possible applications of speaker characterization (e.g. forensic applications), even two minutes of speech is a very long period of time. Therefore, we studied contiguous 15-second fragments, as well as 75-second fragments, obtained by combining five adjacent 15-second fragments into one 75-second fragment. As we announced in Chapter 1, where the influence of different speech styles was explained, the fragments were of two speech styles: read and spontaneous speech. In section 2.6.2 it shall be explained how the speech fragments were obtained.

Apart from the speaker specificity of time-integrated parameters, we also used

---

[20] A difference between the speech of female smokers vs. non-smokers is observed by Gilbert and Weismer (1974). Smokers have significantly lower mean F_0 values than have non-smokers.

contour-bound parameters, i.e., measurements at specific points in the F_0 contours of speakers. To this end we intended to compose sentences that, when read by different speakers, would evoke uniform prosodic behaviour, in terms of the pitch movements as described in the Grammar of Dutch Intonation ('t Hart et al., 1990). In section 2.6.3 the results of our search for such sentences shall be discussed.

The experimental material of this study was obtained from two recording sessions, separated by a time interval of about seven months. In Chapter 3 the speaker-identifying properties of TI parameters are assessed in a study with 15-second fragments that were obtained from an interview situation (spontaneous speech) and from the reading of a newspaper-like text (read speech). The speech material used in Chapter 3 is described in section 2.6.2. In Chapter 4 the CB parameters are examined in sentence material derived from a large corpus of sentences that were read immediately after the newspaper-like text. The speech material of Chapter 4 is described in section 2.6.3.

In addition to the spontaneous speech in the interview, unpremeditated speech utterances were obtained as well, but these were not used in the present study. Thus, the experimental material will be described as follows:

1. Interview (§ 2.6.2)
2. Reading task 1: newspaper-like text (§ 2.6.2)
3. Reading task 2: isolated sentences and short stories (§ 2.6.3)

2.6.2 Speech fragments

In this study the speaker-identifying properties of a number of time-integrated parameters are determined in speech fragments with a duration of 15 seconds. To enable the evaluation of the stability of the fragments, for each of the two different speech styles (read vs. spontaneous) five fragments were obtained, on two different occasions. Thus we had $2 \times 5 \times 2 \times 15 = 300$ seconds, or five minutes of speech per speaker, at our disposal. The total number of fragments was $2 \times 5 \times 2 \times 50$ speakers = 1000.

The fragments of *spontaneous speech* were gathered from two interviews, held at the beginning of each of the two recording sessions. The questions in the interview had to do with the speakers' food preferences and eating habits (session I) and holidays and travelling experiences (session II). The questions asked by the interviewer and their translation into English are given in Appendix C.

The fragments of *read speech* were obtained from the speakers' readings of a newspaper-like text. This text was not composed in such a way that it would evoke any specific prosodic behaviour. The text was composed to contain a high proportion of voiced sounds, so as to increase the number of F_0 measurements in the fragments. The original text in Dutch and its translation into English are shown in Appendix C.

In the spontaneous part of the data we have fragments of different content for all combinations of the factors Speaker and Session. The read fragments, however, were more or less the same for all speakers, as all speakers read the same text on both occasions.

2.6.3 Sentences

At the outset of this study we intended to compose sentences that, when read by different speakers, would evoke uniform prosodic behaviour, in terms of GDI. We even hoped to be able to elicit all or most of the GDI pitch movements, so as to determine which of them

would be most speaker specific.

What was obviously needed was a technique to elicit specified pitch movements from the speakers. We discarded the possibility of asking speakers to imitate pitch movements because it is quite difficult to get naive subjects to reproduce pitch movements (Boves et al., 1984). Furthermore, it is not clear to what extent speakers would imitate only the pitch patterns, and none of the other characteristics of the speaker modelling the required utterance.

We hoped to be able to compose sentences that "by their very nature" would evoke fixed pitch contours. In a number of pilot studies (e.g. Kraayeveld et al., 1990) we tried to find such sentences. To this end we had to rely largely on intuition, since at present there exist no useful formalized insights into the relationships between the choice of pitch movements and the communicative intentions conveyed by them. We tried using proverbs, such as *Eigen haard is goud waard* "There is no place like home", we asked speakers to describe configurations of toy blocks, we used sentences that were embedded in a certain context (forcing speakers to a certain interpretation of the sentence), and we tried the opposite, presenting sentences in isolation. The results of all these attempts were meagre; for most of the pitch movements (pitch rises 3 to 5, and falls C to E) we did not succeed *at all* in eliciting them in the utterances of the majority of the speakers.

Our intuition that there are sentences that elicit uniform pitch behaviour, in terms of GDI transcriptions, was not completely wrong, however. In the pilot studies, we did find some sentences in which certain pitch movements were produced by all speakers. By composing related sentences we enlarged this group of suitable sentences. For example: the sentence *De Denen wonnen van de Noren met één-nul* "The Danes beat the Norwegians by 1-0", in which we found a pitch movement of type "1" on *De*nen, *No*ren, and *één*, was supplemented by *De Ieren wonnen van de Denen met drie-één* "The Irish beat the Danes by 3-1" and *De Noren wonnen van de Roemenen met drie-nul* "The Norwegians beat the Rumanians by 3-0". These three related sentences constitute a sentence type that we describe as the "sports sentences". By supplementing sentences that appeared useful in pilot experiments by related sentences, eight groups of sentences (sentence types) came into being with 30 sentences in all (cf. Appendix D)[21].

As was mentioned above, unpremeditated speech utterances were obtained as well, but these were not used in the present study. Per speaker 18 of these utterances and 30 read-out sentences were recorded, amounting to 48 utterances per speaker. As the 50 speakers realized all 48 utterances on two occasions, the corpus contained 4800 utterances. For practical reasons we decided to study only a subset of this corpus. Of course, the main selection criterion for the experimental subset had to be uniformity in prosodic behaviour (in terms of GDI transcriptions) of the speakers for each of the sentences. During the recording of our material we noticed that there were still many sentences that did not evoke uniform prosodic behaviour. Therefore, to be able to select sentences that were realised with uniform contours, (part of) the material had to be transcribed in terms of GDI.

---

[21] In an attempt to acquire strictly controlled utterances that were not read out, a simple question-answer scheme was designed, in which the speakers had to respond to a question with an accompanying picture by producing an utterance with a narrow focus word accent. As the 18 utterances obtained from this procedure were not be used in this study, the stimuli used will not be presented either.

A large part of the speech material that was obtained in the first recording session was transcribed at IPO by five master transcribers. After completion of the individual transcriptions the transcribers together drew up a consensus transcription of all utterances for which the individual transcriptions diverged[22].

At the time of the transcriptions, only part of the corpus was available. For each of the sentences transcriptions were made of the realisations of between 10 and 18 speakers. From the total 632 utterances transcribed (13 % of the corpus) we selected utterances in such a way that we had a design at our disposal in which for all sentences there were transcriptions from 10 different speakers (10 % of the corpus). For each sentence the pitch movements that were found in at least eight out of ten transcriptions are shown in Appendix D[23].

For our experimental corpus we did not only want to select sentences that would constitute a largely homogeneous group, but also sentences that would allow for a fixed data structure. Because of this second requirement, only sentences with the same structure (the same number of rises and falls at comparable positions in the sentence) could be used. In practice this implied that we could only use sentences from one sentence type.

As can be seen in Appendix D, on the basis of the pitch movements found in the sentence types, various sentence types were candidates for selection We chose to select the sports sentences as our experimental sentences, because in these sentences three pitch movements were present in all transcriptions: two rises of type "1", on the first stressed syllable of the first nationality and of the first number in the result, and one type "A" fall, on the stressed syllable of the second number in the result. The fact that several different pitch movements could be studied in this sentence type allowed us to study the relation between the CB measures in different movements. As the speakers read out three sports sentences on two occasions, there were six replications in all.

2.7 Recordings

Recordings were made in two recording sessions that were separated by a time interval of about seven months[24]. The total recording procedure took about 25 minutes per speaker. It consisted of:

---

[22] The transcriptions were used as test material in an IPO research project that aimed to design a system for the automatic transcription of pitch movements (ten Bosch, 1995)

[23] We consider composite GDI movements such as "1&A" as primarily consisting of the constituting movements "1" and "A" This is disputed by some authors Caspers and van Heuven (1993), for instance, claim that production data for a rise "1" in isolation are not at all like data for a rise "1" in a pointed hat pattern (i e "1&A") Furthermore, Collier (1991), in his description of a Dutch intonation synthesis implementation, prescribes a different alignment for a fall "A" in a pointed hat than for an "A" in a flat hat configuration

[24] In fact, three recordings were made. Following the first recording there was a short break, after which the reading tasks were repeated. Due to time limitations, the material from the second recording of the first meeting was not used.

1. Unpremeditated speech task 1: interview
2. Reading task 1: newspaper-like text
3. Reading task 2: isolated sentences and small stories
4. Unpremeditated speech task 2 (not used in the present study)

During the recording sessions, the speakers were seated in a sound-treated room and received instructions about the interview. The speakers were asked not to make any unnecessary sounds, such as tilting the chair or rustling the pages of the reading-booklet, and to avoid talking while the interviewer was speaking.

At the end of the interview the speakers were asked to read the first page of a booklet with reading-texts. The first page of this booklet contained the instructions to the first and second reading tasks. The instructions and their translation into English are given in Appendix C. The speakers were asked to study the texts carefully before reading them out.

In the first reading task, the speakers had to read out the newspaper-like text while in the second one they had to read out the isolated sentences and the short stories that were supposed to lead to uniform pitch contours.

Following this second reading task the speakers performed a task which was designed to elicit unpremeditated speech. The recordings from this task were not used in the present study.

The audio-recordings were made in a sound-treated room with a Studer 089 tape recorder and a Sennheiser MKH416T microphone. The microphone was located at approximately 30 cm from the mouth of the speaker. Tape speed was 19 cm per second. Both before the interview recordings and before the recordings of the reading session, the recording level was adapted to the speaker's intensity level in such a way that the amplifier was kept functioning within its linear region.

The utterances were digitized with a 10 kHz sampling frequency, at 12 bits resolution and low-pass filtered at 4.7 kHz through an eighth order Butterworth filter onto a DEC μVAX II computer.

Using a multi-channel speech editing system, the questions and interferences of the interviewer were removed from the original recordings of the interviews. Thus we obtained speech files with several minutes of "monologue". From the end of the interview backwards, five adjacent fragments with a duration of 15 seconds each were segmented into separate files, the spontaneous fragments. In order to exclude as much as possible any habituation effects, the experimental material was extracted from the latter part of the interview, rather than from the initial part.

From the reading text, which consisted of five paragraphs, five fragments of 15 seconds taken from the beginning of each paragraph onwards were segmented into separate files. Thus, these files contained more or less the same textual material for all speakers, depending on their speech rate.

Finally, the recordings of the sentence material were segmented into separate files containing the individual sentences only, from the first to the last observable diversion from the zero-level amplitude.

2.8 SUMMARY

In this chapter the method of the present study was presented. Because much attention was devoted to the description of the prosodic parameters to be used in this study, the chapter was divided into two parts; the first part contains a description of the acoustic prosodic parameters while the second part describes the speakers, speech material and recording procedure of the study.

In part I a survey of the literature on acoustic prosodic parameters was presented, as well as information on the speaker-identifying properties of these measures. Throughout this book a classification of parameters into time-integrated (TI) and contour-bound (CB) parameters will be used. We started this chapter with a description of the TI parameters.

There are various methods of measuring F_0 and amplitude. In the present study, two different algorithms will be used to determine F_0; Hermes's subharmonic summation algorithm (Hermes, 1988) and van Bergem's algorithm (van Bergem, 1990) of filtering out the fundamental frequency. As an acoustical correlate of "loudness" we chose the absolute peak amplitude of each cycle.

Measures involving F_0 and amplitude fall into three groups: (1) central tendency measures, (2) variability measures, and (3) perturbation measures. As a measure of central tendency for F_0 the mean of the period durations, as determined by van Bergem's algorithm for filtering out the fundamental frequency was selected. No measure of central tendency will be applied for amplitude.

As a measure of F_0 dispersion the coefficient of variation of the period durations will be used in this book. An adapted version of the coefficient of variation will be used for the amplitude measures.

In sections 2.2.8 and 2.2.9 different measures of voice instability, or perturbation, were introduced. The parameters selected for the present study were two measures of perturbation extent, the Period Perturbation Quotient (PPQ) and the Amplitude Perturbation Quotient (APQ), and two measures of perturbation rate, the Period Zero-crossing Rate (PZR) and the Amplitude Zero-crossing Rate (AZR).

To conclude the description of our TI parameters, three temporal phenomena were discussed in section 2.3: pausing, speaking rate and voicedness. For each of these phenomena a parameter was defined to be used in this study.

In section 2.4 we discussed the contour-bound parameters. In a group of experimental sentences measurements will be taken at pivot points, the start and end F_0 of the pitch rises and falls. The temporal relation between the pitch movements and the vowel onsets in the relevant syllables, will be determined as well.

Finally, declination was discussed. We decided to study baseline declination measures only. We will measure F_0 at the beginning and end of the test utterances, and the slope of the F_0 declination.

In part II of this chapter the remaining aspects of the methodology were described. First the selection criteria for the speakers to be used were given. Next we described the selection of the speech material to be used in this study, and finally the recording procedure was delineated.

3. Time-integrated parameters

3 1 Introduction

In this chapter[1] we will study the relation between a number of extralinguistic factors, particularly the speaker factor, and ten time-integrated parameters in the speech material that was described in section 2 6 2 (1000 fragments with a duration of 15 seconds) The TI parameters that were introduced in Chapter 2 are

1	F_0MEAN[2]	mean F_0 [Hz]
2	CVP	coefficient of variation of period
3	PPQ	period perturbation quotient
4	PZR	period zero-crossing rate
5	CVA	coefficient of variation of maximum amplitude per cycle
6	APQ	amplitude perturbation quotient
7	AZR	amplitude zero-crossing rate
8	PAUSE	silence as a proportion of the speaking time [%]
9	RATE	articulation rate [syllables/s]
10	VOI	voiced speech as a proportion of the speaking time [%]

The factor of greatest importance in this study is Speaker[3] Most of the other extralinguistic factors that are examined in this chapter were introduced in Chapter 1 Speech style, Sex and Age group and Session The sixth factor, Fragment, indicates the influence of the five read and the five spontaneous 15-second fragments per speaker per session

Before studying the relationship between the TI parameters and the extralinguistic factors, we must first assess the interrelatedness of the TI parameters If there are (groups of) parameters that covary to a large extent, data reduction should be employed, the TI parameters should then be transformed into a smaller set of more independent factors In order to evaluate the parameters' interrelatedness, we shall first present a correlation matrix and a factor analysis of the TI parameters

The relationship between the time-integrated variables and the grouping variables can be determined in two different ways

---

[1] Some of the results reported in this chapter were presented earlier in the proceedings of the 1993 ESCA Workshop on Prosody, September 27 29, Lund, Sweden (Kraayeveld et al , 1993)

[2] The names of dependent variables will be printed in small capitals

[3] The names of extralinguistic factors will be printed with capitalized initial letters

(1) Analyses of variance can be performed for each of the predictor variables. Thus, one can find out more about the ways in which the individual predictor variables are related to specific grouping variables. The advantage of this approach is that the effects of the interactions of the grouping variables on the predictor variable can be determined.

(2) The scores of the speakers on the predictor variables in the 15-second fragments can be used to perform linear discriminant analyses (*LDA's*). From these analyses we can learn to what extent it is possible to assign the cases (the individual speech fragments) to the grouping variables on the basis of the scores on the predictor variables.

Analysis of variance is used to determine how the dependent variables, in our case the ten time-integrated variables, vary as a function of the independent variables, here the grouping variables. In discriminant analysis the dependent variables are used to distinguish levels of the independent ones. There are some important differences between analysis of variance on the one hand and discriminant analysis on the other: the former has the advantage of showing interactions between the independent variables (the extralinguistic factors), while LDA's show to what extent the dependent variables (the prosodic parameters) can be used to assign cases to the levels of the independent ones. Because of the respective advantages of the two analysis techniques, both will be applied here.

The influence of some of the extralinguistic factors on the factor analyses, the analyses of variance and the discriminant analyses will be clarified by performing separate analyses for both the total material and for different subsets of it. For instance, separate analyses will be performed for the spontaneous fragments and the read fragments.

The results of the LDA's in terms of the identification performance for the various extralinguistic factors will set a base-line against which, in the next chapter, we will set off the results of the contour-bound parameters.

3.2 Preliminary considerations: the interrelatedness of the variables

For this study we tried to select parameters that cover different aspects of the speakers' prosodic behaviour. In order to find out whether, despite these exertions, there are high correlations between parameters, or underlying factors (components) that can summarize the values of different predictor variables, we determined the correlation matrix and performed a Principal Component factor analysis (PC-analysis) for the ten TI parameters in the 15-second fragments that were described in Chapter 2.

The Pearson Product-Moment correlation coefficients of the predictor variables are presented in Table 3.1.

Table 3.1
Correlation matrix of ten predictor variables in the 1000 fragments (critical values of r | 06 | (p < 5 %) and | 08 | (p < 1 %), two tailed)

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAUSE	RATE
CVP	31								
PPQ	05	34							
PZR	62	- 07	37						
CVA	03	- 11	21	34					
APQ	- 45	- 09	35	- 13	32				
AZR	21	- 09	36	59	42	01			
PAUSE	- 29	- 07	- 13	- 34	- 03	01	- 31		
RATE	- 01	13	- 03	- 30	- 24	18	- 44	24	
VOI	14	09	42	- 34	- 51	29	- 58	26	56

The variables with the highest correlation are mean F_0 and PZR. The correlation of these two variables, 62, indicates that they share only 38 % of their variance Therefore we conclude that even the strongest relationship between any two predictor variables is weak, which means that all ten predictors provide potentially independent information.

Next, a Principal Component factor analysis was carried out This analysis resulted in four factors with eigenvalues higher than one[4].

To find out how well these four factors cover the variation in the variables, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, "an index that expresses the appropriateness of the factor-analytical model for our variables" (Rietveld and van Hout, 1993 277) was consulted. The value of this measure, .57, can be evaluated as poor (see Rietveld and van Hout). Although for most applications of factor analysis a poor fit to the data is undesirable, for the present study it has a favourable consequence: we can conclude that all variables measure more or less independent aspects of the speakers' behaviour and that all ten variables should be retained in the subsequent analyses.

The low sampling adequacy may be the result of large differences between the levels of the extra-linguistic grouping variables A higher KMO sampling adequacy and, consequently, a motive for data reduction, could perhaps be found in PC analyses that were performed on subsets of the data, for instance, in the male speakers' data Table 3 2 shows the KMO sampling adequacies of analyses in which the speech styles, the sexes and the sessions are analysed separately

[4] This criterium for retaining components is known as the Guttman-Kaiser rule (Rietveld and van Hout, 1993)

Table 3.2
Sampling adequacy as a function of considering only parts of the data

variable	levels	nr of speakers	nr of fragments	KMO sampling adequacy
overall		50	1000	575
speech style	read	50	500	551
	spontaneous	50	500	541
sex	females	25	500	599
	males	25	500	678
replications	1	50	500	581
	2	50	500	573

Only the fit of the model for the male speakers is somewhat better than that of the overall analysis, with a value of .68 it can be classified as "mediocre", and data reduction by means of a PC factor analysis could be applied to the male speakers However, to remain in congruence with the rest of the data, and because a mediocre fit is still a somewhat questionable basis for data reduction, we will retain all ten predictor variables in the analyses to come[5]

To summarize our assumption that the ten time-integrated variables are not very dependent on each other was confirmed in the PC analyses Apparently our reasons for choosing these parameters were solid each parameter covers a rather unique aspect of speech

3 3 Analyses of variance

3 3 1 Introduction

To be able to see how each of the time-integrated variables alone depends on the extra-linguistic factors Speaker, Speech style, Sex, Age, Session, Fragment/Paragraph (see below), and the interactions of these grouping variables, analyses of variance were carried out for all time-integrated variables separately

The model used in these analyses is a mixed model, with both random and fixed effects The decision of assigning fixedness or randomness to the factors is important because it influences the way in which pairs of mean squares are combined to form *F*-ratios Rietveld and van Hout call a factor

- *fixed*, "when all possible levels of that factor are included in the experiment",
- *random*, "when the number of possible levels of that factor greatly exceeds that of

[5] According to some authors (e g Gorsuch, 1990) a proper Factor Analysis should be preferred to the Principal Component Analysis used here, because the former tends to be more conservative For the analyses discussed this would make the results even less impressive

> the number of levels included in the experiment; furthermore, the levels included should have been selected at random" (Rietveld and van Hout, 1993: 31).

We discuss analyses of all our 1000 fragments, and we study the differences between the separate analyses of the two speech styles, read and spontaneous speech.

In the analyses of variance, we defined the factors Speaker and Fragment as random and Sex, Age, Session, and Speech style as fixed. For obvious reasons the factor Sex is fixed. The factors Age and Speech style are considered to be fixed, too, since the levels of the factors were not selected at random: a specified number of speakers were assigned to specified age groups, and the "read" and "spontaneous" speech styles are two rather basic, not randomly chosen, levels. The results of two recording sessions are included in the design. There was a specific time interval between the two sessions and Session is therefore defined as a fixed factor.

The factor Speaker is clearly a random factor. Although we did not make a random selection from all possible Dutch speakers, the number of speakers we could have selected on the basis of the same criteria far exceeds the number we actually used. Furthermore we did not use specified selection criteria when assigning the subjects to the different groups.

Speaker is nested under Sex and Age in our design since there were different speakers for each combination of Sex and Age. In that way it was possible to determine the influence of Sex and Age. The consequence of this nesting is that the influences of Sex and Age have been removed from the Speaker variable, and a significant speaker effect must be interpreted as a difference in the means of speakers *within* sex/age groups. Thus, the factor Speaker is defined in a narrower sense here than in the sections on discriminant analyses.

For the factor Fragment the choice between random and fixed is complicated by the fact that the conditions for the spontaneous speech fragments and the read fragments are different. The fragments of spontaneous speech can be considered to have been selected more or less at random from the population of all possible 15-second fragments. This is not entirely true for the read fragments, which were obtained from five paragraphs of a text. Although it is evident that the paragraphs in a text are semantically and lexically related, we have no idea whether this also implies that they are related in a prosodic sense. To stay in line with the spontaneous fragments we will assume this is not the case. The number of possible fragments greatly exceeds the number of fragments used, and we will consider Fragment to be a random factor.

In our design Fragment is nested under Sex, Age, Session, Speech style, and Speaker, since in the spontaneous part of the data we have different fragments for all combinations of the predictor variables.

The read fragments are more or less the same for all speakers, however. The recording of each fragment started at the beginning of a paragraph and lasted 15 seconds. Therefore it depends on the individuals' speaking rate how much alike the fragments from different speakers are. Separate analyses were performed for the read and the spontaneous part of the data. A separate analysis of the read fragments has the advantage that Fragment is *not* a nested factor; in such a design one can estimate how dependent the time-integrated variables are on the lexical content of the material. To distinguish the fragment factor in the read speech condition from that in the overall and in the spontaneous condition, the former will henceforth be referred to as *Paragraph*.

A disadvantage of this design is that for the main effects Sex, Age, and Session,

and for the interaction terms of these effects there is no appropriate error term available. However, it was possible to obtain *F'*, the quasi *F*-ratio, for these effects[6]. *F'*-ratios are known to be rather conservative (Forster and Dickinson, 1976), and the effects of Sex, Age, Session, and the interactions of these effects might therefore be underestimated to some extent.

Some preliminary considerations regarding the analyses of variance are:

(1) The factor *Speaker* is of capital interest for this study. If this factor does not reach significance for a parameter, we must reject the idea that speakers differ on this parameter; consequently, it will be of no relevance with regard to speaker identification.

(2) For many variables differences have been found for *Speech style* (see section 1.3.3) It is therefore to be expected, that some of these differences will be replicated here. It is important to find out whether an interaction exists between Speech style and Speaker[7].

(3) For some of the prosodic parameters we expect to find significant effects for *Sex* and *Age* (see section 1.3.3).

(4) A significant result for *Fragment/Paragraph* would mean that the linguistic content of the fragments strongly influences the values measured, and that 15 seconds of speech would not be enough to factor out these effects.

The absence of a significant Speaker × Fragment interaction would imply that the inter-speaker differences are stable between fragments. If a Speaker × Fragment interaction were to occur, it still would not be problematic provided that the *strength* of the Speaker effect (of which significance is *not* a direct reflection) would amply surpass the strength of the Speaker × Fragment interaction. An index for the strength of association is ω^2 (Hays, 1973); it is the estimate of the proportion of the total variance that is associated with the factor or interaction term at issue. The ω^2 index is defined as:

$$\omega_x^2 = \frac{\sigma_x^2}{\sum_{i=1}^{N} \sigma_i^2}, \qquad (11)$$

where ω_x^2 denotes the strength of association of factor x, σ_x^2 the variance associated with the factor x, and the denominator of the right term the total variance.

If the ω^2 value of the Speaker × Fragment interaction is of the same order as the ω^2 value of Speaker, it means that the scores of the speakers largely depend on the fragment in which they are measured. This could be the consequence of the

---

[6] The quasi *F* ratio consists of a numerator and a denominator in which the mean squares of a number of effects are combined in such a way that the expected value for the numerator only exceeds the expected value of the denominator by the variance of the factor tested (Winer, 1971; Clark, 1973).

[7] If a Speaker x Speech style interaction were to be found, it could have consequences for the applicability in speaker identification of the variable concerned, although a speaker can probably be trained to speak in a rather "formal" way.

fragments being too short in duration to reach really stable values. If a speech fragment of about 15 seconds is not long enough to attain stable values on a variable, it would mean that, for many applications, the variable is not usable.

(5) Significant differences for the grouping variable *Session* could perhaps be associated with certain emotions, like nervousness in the first session or boredom in the second. Emotions can influence voice characteristics, as was shown by van Bezooijen (1984). In the present, emotionally rather neutral task, we do not expect to find significant differences between the sessions.
A significant Speaker × Session interaction would constitute a problem. It would imply that the speaker differences are not stable from session to session.

3.3.2 Results

In this section we discuss the effects of the grouping variables on each of the ten predictor variables as observed in ten separate analyses of variance. We start the discussion of the analyses of variance with a summary of the results in terms of the presence and absence of significant effects, and of the strength of effects. Summary tables of the overall analyses (pooled over both the read and the spontaneous fragments), and of separate analyses of the read and the spontaneous data are presented (Tables 3.3 through 3.5). In the next sections the results will be discussed variable by variable.

To facilitate the discussion of the analyses of variance, the data are not presented in full; ω^2 values are not inserted in the tables if the effect concerned is not significant and if ω^2 does not exceed .05.

In the analyses of variance of the read data, we defined Fragment as not being nested under Sex, Age, Session, Speech style, and Speaker. Roughly speaking, the fragments correspond with paragraphs of the text, and we therefore refer to the read fragments as Paragraphs. In this design, values for Paragraph and its interaction terms are obtained, which enables the assessment of the influence of the texts that were read on the values of the time-integrated variables. The results of these analyses are presented in Table 3.4.

Table 3.3
Summary of analyses of variance of all data, for ten TI parameters (in the columns). For significant effects the cells are shaded and the ω^2 values are presented; if effects are not significant, ω^2 values are presented if they are at least .05; S= Speaker, G= Gender (sex), A= Age, R= Replication (session), T= speech Type (style), F= Fragment.

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAUSE	RATE	VOI
S(GA)	.06	.19	.32	.12	.08	.20	.13	.25	.15	.16
G	.88	.10		.51	.03	.21	.11	.14	.03	
A		.06		.04						
R			.03	.01						.01
T	.03	.21	.02	.15	.53	.06	.41		.35	.52
GT									.02	
RT									.01	
SR(GA)	.01	.06	.12	.02	.02	.03	.04	.15	.05	.06
ST(GA)	.01	.15	.11	.02	.07	.07	.10	.15	.09	.05
SRT(GA)	.00	.02	.03	.03		.07	.06	.12	.05	.05
F(GARTS)[8]		.20	.28	.11	.26	.38	.15	.25	.24	.13

Table 3.4
Summary of analyses of variance of the read data, for ten TI parameters (in the columns). For significant effects the cells are shaded and the ω^2 values are presented; if effects are not significant, ω^2 values are presented if they are at least .05; S= Speaker, G= Gender (sex), A= Age, R= Replication (session), P= Paragraph (read fragments).

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAUSE	RATE	VOI
S(GA)	.07	.46	.29	.12	.22	.18	.29	.34	.45	.34
G	.90			.73	.05	.30	.24	.22	.16	
A	.01	.10		.04						.06
R					.01	.01	.02			
P	.00	.01	.02	.01	.48	.31	.03	.09	.08	.22
GA		.06	.11				.10			.06
SR(GA)	.01	.06	.13	.02	.06	.06	.07	.16	.09	.14
SP(GA)[8]		.11	.15		.07	.06	.10	.09	.08	.06
SRP(GA)[8]		.13	.20	.05	.09	.09	.16	.14	.11	.10

[8] For this effect no significance could be established because there was no appropriate error term with which to construct an F-ratio.

To be able to compare the outcomes for the two speech styles, analyses of the spontaneous data were carried out as well. The relevant ω^2 values are reported in Table 3.5.

Table 3.5
Summary of analyses of variance of the spontaneous data, for ten TI parameters (in the columns). For significant effects the cells are shaded and the ω^2 values are presented; if effects are not significant, ω^2 values are presented if they are at least .05; S= speaker, G= gender (sex), A= age, R= replication (session), F= fragment.

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAUSE	RATE	VOI
S(GA)	.06	.27	.45	.19	.33	.34	.37	.37	.24	.47
G	.92	.23		.50	.05	.15	.17	.10		
A		.06		.07						
R			.04	.02						.02
SR(GA)	.01	.13	.16	.08	.19	.15	.17	.29	.16	.27
F(GARS)[8]		.34	.32	.19	.49	.39	.32	.29	.53	.28

We do not want to discuss all the data reported in these tables, but only effects of substantial significance, i.e., effects for which a considerable strength of association was found. We need an ω^2 threshold to determine whether the strength of association of an effect is large enough to discuss the result. Although each possible ω^2 threshold is to some extent arbitrary, we found that most of the *in*significant effects have an ω^2 value that is lower than .05. The threshold ω^2= .05 therefore prevents us from having to discuss many insignificant effects.

Although we are most interested in the Speaker effect, we first turn our attention to the interaction effects involving Speaker. The reason for this order is that, as was mentioned in section 3.3.1, interactions between the factor Speaker on the one hand and the factors Session and Fragment on the other would point to instability of the parameter estimates, which reduces the applicability of the results to some extent. After discussing these interactions, the interactions of the other main effects are presented for all ten parameters. Next we turn to the Speaker effects. Finally, the other main effects are discussed.

Interactions involving Speaker
In Table 3.3, the summary of the analyses for the total set of fragments, the interactions of Speaker × Speech style and Speaker × Session were both significant for all ten parameters and the three-way interaction Speaker × Speech style × Session was significant in all but one: CVA. In all the separate analyses of the read and the spontaneous fragments, Speaker × Session was significant, too.

To find out whether these findings pose problems, the proportion of the total variance that is associated with each of the effects was determined, and the ω^2 values of the main effects and of the interactions were compared. In Table 3.3 it can be seen that in all cases the ω^2 values of the factor Speaker were larger than those of the interaction terms, and mostly they were of quite a different order. For F_0MEAN, for instance, ω^2(Speaker) was .06 and ω^2(Speaker × Session) was .01. However, as is shown in Table

3.6, there are cases where the differences were much smaller.

Table 3.6
Parameters for which the ω^2 value of the factor Speaker was only little higher than that of the interaction term Speaker × Speech style.

	ω^2(Speaker)	ω^2(Speaker × Speech style)
CVP	.19	.15
CVA	.08	.07
AZR	.13	.10

For these parameters it is difficult to assert that Speaker was the most important factor, as ω^2(Speaker) did not clearly exceed the ω^2 values associated with these interaction terms. The high value of ω^2(Speaker × Speech style) for CVP, for instance, implies that the differences in the scores of the speakers on the two speech styles are important, and CVP can best be used for speaker identification within the context of one speech style.

In the analysis of the read fragments the significance of the interaction terms involving Speaker and Paragraph can be determined. These interaction terms, Speaker × Paragraph and Speaker × Session × Paragraph, are of the same undesirable kind as the above-mentioned interactions of Speaker on the one hand and Session and Speech style on the other. Interactions of Speaker and Paragraph would indicate that the value found for a speaker reading a passage of text depends on the content of that text, which would reduce the applicability of the results somewhat. Unfortunately, the significance of both Speaker × Paragraph and of Speaker × Session × Paragraph cannot be determined in the design employed here. We can only compare their ω^2 values to ω^2(Speaker). For most of the parameters the interaction effects had low ω^2 values, signifying that the variance components associated with the effects were not important.

In section 3.4, LDA's will be applied to assign the cases (i.e., the individual speech fragments) to the speakers. If the speaker specificity of prosodic parameters is to be of use in real-life applications, it should not vary too much from session to session. Therefore, cross-validation LDA's will be performed; discriminant functions will be derived from one session and used to classify speech fragments from the other session. If the interaction of Speaker and Session is an important variance component in the analysis of variance of a parameter, that parameter will not contribute to, and might even weaken the identification performance of the cross-validation LDA's. It is important to remember, however, that in the present design Speaker is nested under Sex and Age. In a design in which the effects of Sex, Age, and Speaker are not separated, Speaker will be a more important factor.

Other interaction effects

As we explained above for Speaker, prior to determining the significance of a main effect, we must check the interaction terms in which such an effect is involved. A significant Sex effect, for instance, can be subordinate to a Sex × Session effect, and such interaction effects should therefore be discussed before turning our attention to the main effects of Sex, Age, Fragment and Session.

There are not many interaction effects of substantial importance (in terms of our ω^2

threshold) among the interaction effects discussed here. We only found significant Sex × Age group interaction effects in the analyses of the read fragments. For the read data, ω^2(Sex × Age group) of PPQ and AZR was .11 and .10, respectively. Figures 3.1 and 3.2 show how Sex and Age interact for these parameters. For PPQ the significant interaction was caused to a large extent by the high score of the male speakers in Age group 3 (especially by the unusually high PPQ score of speaker M35). For the significant Sex × Age group interaction of AZR the strength of association, ω^2(Sex × Age), was clearly lower than ω^2(Sex). From Figure 3.2 it becomes clear why the strength of association of the factor Sex surpasses that of the Sex × Age group interaction. The female speakers had a higher AZR in all age groups except in age group 4. The main effect is therefore clearly more representative of the values of AZR actually found than the interaction effect.

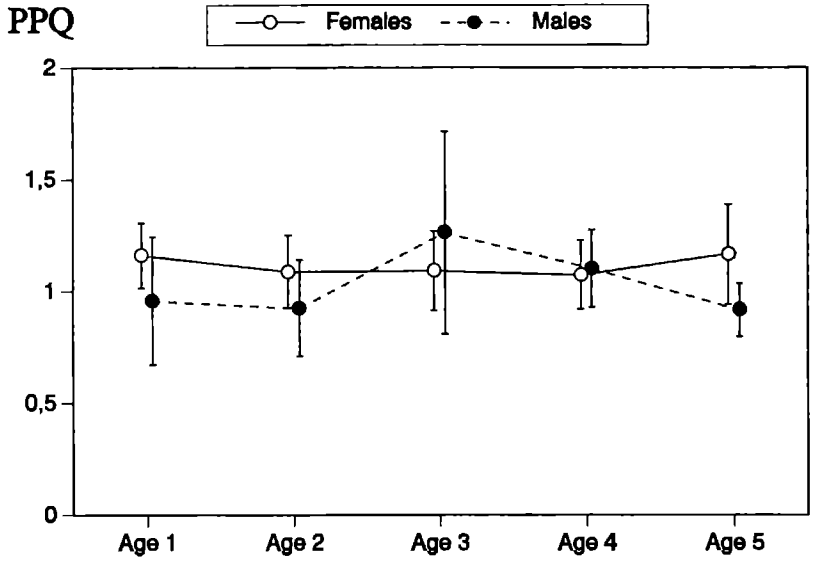

Figure 3.1 Pitch Perturbation Quotient (PPQ) in the read fragments, as a function of age group, plotted for male and female speakers.

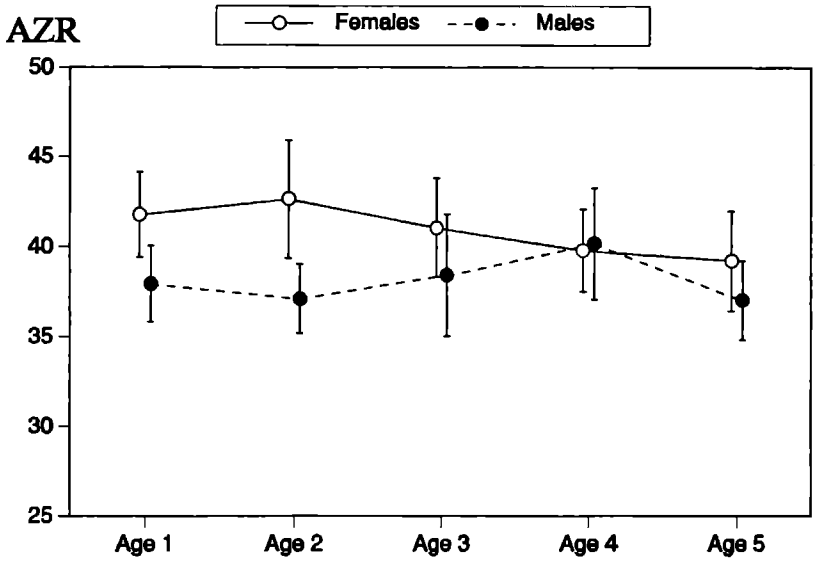

Figure 3.2 Amplitude Zero-crossing Rate (AZR) in the read fragments, as a function of age group, plotted for male and female speakers.

Speaker effect

The main effect we are most interested in is, of course, Speaker. For all ten TI parameters this factor reached significance, in the overall analyses as well as in the separate analyses of the read and the spontaneous fragments. Apparently, for all variables there were significant speaker differences that were not related to the fact that the speakers belonged to different age groups and different sexes.

For each of the parameters the speaker differences can be illustrated in a figure such as Figure 3.3, in which the means and standard deviations of all individual speakers on the parameter F_0MEAN are shown. Since these figures take up a lot of space, we only present the speaker differences on the parameter F_0MEAN which, as we will come to see, is an important speaker-identifying parameter. For an impression of the speaker scores on the other parameters, the reader is referred to Appendix E, in which for all parameters the speakers' mean values are provided.

Figure 3.3 demonstrates how a large effect of one of the factors (Sex), can mask the effects of other factors, such as Speaker and Speech style. Although ω^2(Speaker) was only .06, F_0MEAN is quite speaker-specific within sex groups.

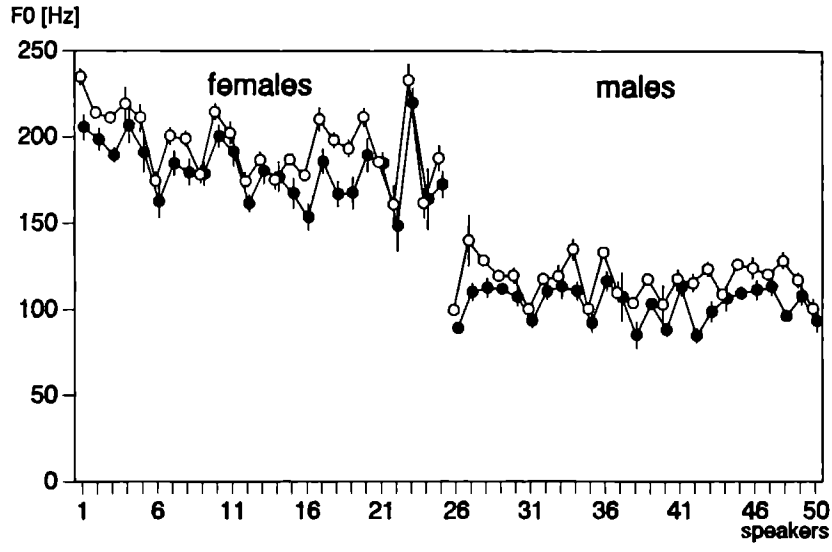

Figure 3.3 Means and standard deviations on the most speaker-specific parameter, F_0MEAN, for all 50 speakers The values of the read (O) and the spontaneous fragments (●) are given In the left half of the figure, the values of the female speakers are given, in the right those of the male speakers Within each of the sex groups, the speakers are arranged according to their age groups Thus, the data of the first five speakers belong to the youngest group of female speakers

Speech style effect

The main effect of Speech style is significant and of substantial importance (i.e., $\omega^2 > .05$) for all predictor variables except F_0MEAN, PPQ and PAUSE. The first of these negative results is somewhat surprising, as F_0 differences between speech styles have often been reported (e.g. Hollien and Jackson, 1973; Ramig and Ringel, 1983; Koopmans-van Beinum, 1991). In fact, we did find a significant F_0 difference between read and spontaneous speech (156 6 Hz vs. 142 3 Hz, respectively), but the value of ω^2(Speech style) was quite low: .03. In section 3.5 it will be shown that F_0MEAN is not among the most important variables for discriminating the two speech styles.

For all main effects except Speaker (due to space limitations), the significant differences will be presented in *profiles*, figures with the means and standard deviations of the levels expressed in Z-scores relative to the mean and standard deviation of the total material. The raw data (means and standard deviations before Z-transformation) are presented in Appendix E

In Figure 3 4, the Speech style Profile, it is shown that in spontaneous speech the parameters PZR, CVA, APQ and AZR have higher values, while CVP, RATE and VOI have lower values than in read speech.

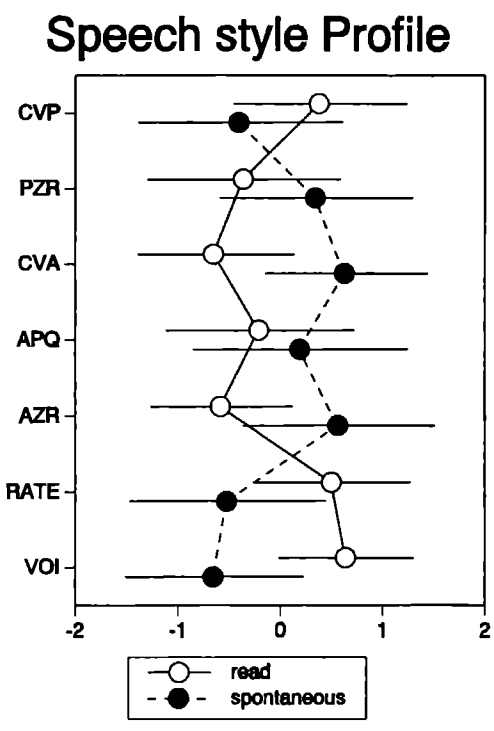

Figure 3.4 Speech style Profile, means and standard deviations of the parameters for which a substantially significant Speech style effect was found. Means and standard deviations are expressed in Z-scores relative to the overall mean and standard deviation of each measure

For some of these results it is easy to find an explanation. The fact that VOI was higher for read speech, meaning that the signal was evaluated as "voiced" in a higher proportion of time, can at least partly be explained by the fact that the reading text was deliberately composed in such a way that it contained many voiced sounds.

The rate of read speech was higher than in spontaneous speech, which is not surprising either, since in the latter type of speech people need time to think about what they are going to say. This difference is an indication that the comparable amount of silence in the two speech styles must result from more filled pauses or lengthening of segments in spontaneous speech. This lengthening may be a side effect of the larger number of short sentences that is found in spontaneous speech (Haselager et al., 1991). More short sentences result in more final lengthenings.

Sex effect

From everyday experience and from a large body of phonetic literature (e.g. Tielen, 1992) it is a well-known fact that large differences in F_0 exist between the speech of men and women. Indeed we found that the factor Sex was the most important variance component

for F_0MEAN: ω^2(Sex) was .88.

As can be seen in Figure 3.5, the Sex profile, differences were not only found for mean F_0, but also for five other time-integrated variables: CVP, PZR, APQ, AZR and PAUSE. For CVA and RATE the difference between the two sexes was significant, but the strength of association was low: ω^2(Sex) was .03 for both parameters. In the analyses with the read data, the sex difference of CVP was not significant, while in the analyses of the spontaneous data we found significant Sex differences for CVP and PPQ.

Figure 3.5 shows that for F_0MEAN, CVP, PZR and AZR higher values were found for female speakers, while for APQ and PAUSE higher values were found for males.

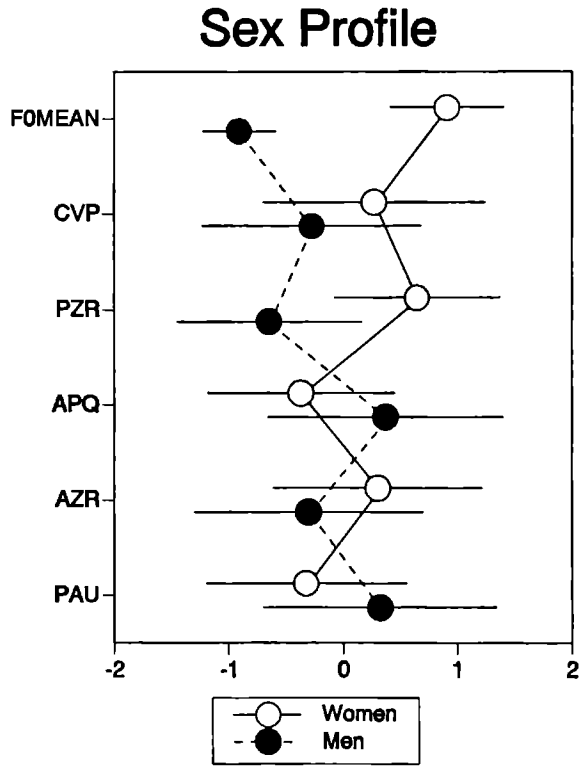

Figure 3.5 Sex Profile, means and standard deviations of the parameters for which a significant Sex effect was found Means and standard deviations are expressed in Z-scores relative to the overall mean and standard deviation of each measure

In section 3 6 we will show that in a discriminant analysis in which the sexes are to be predicted, a successful identification *without* F_0MEAN is possible. This proves once more that other parameters than F_0MEAN are sex-specific as well.

The sex-specificity of PZR at first sight appears to be related to the moderately high correlation of this variable with F_0MEAN (r= 62, n= 1000). However, it would be too easy

an explanation to simply assert that PZR and F_0MEAN measure the same thing, as a large part of the correlation between these measures directly results from the fact that for both F_0MEAN and PZR higher values were found for women. Within the sex groups the correlation of PZR with F_0MEAN was much smaller (for women: r= 38, n= 500; for men: r= −29, n= 500). It is important to note that the measurement accuracy for the male and female speakers' data was different in our measurement method. The sampling frequency for the fragments of male and female speakers was equal, 10 kHz, and the smallest difference in pitch periods that can be measured is therefore 0.1 ms. The pitch periods of women are shorter than those of men (about 5 ms vs. about 10 ms) and consequently the measuring accuracy is higher for men, as (relatively) smaller differences are discernible It is unclear, however, to what extent the PZR difference between the sexes is due to this measuring difference. In stretches of speech of constant F_0, a higher measurement accuracy should lead to a higher number of dissimilar period durations and a higher PZR. As we found a *lower* PZR for men, we do not attribute the sex difference to a measuring artefact.

For the coefficient of variation of the period durations (CVP), a significant Sex difference was found as well This parameter was selected as a measure to express the dispersion of F_0 values (instead of, for instance, the standard deviation), because in CVP the influence of mean F_0 is controlled for (by dividing the standard deviation of F_0 by its mean). The low correlation of F_0MEAN and CVP (overall r= .31, for males r= .23 and for females r= .11) is an indication that the influence of F_0MEAN was indeed removed and that women's F_0 values vary more than men's.

Age group effect

In the overall analyses of variance, we only found a substantially (i.e., ω^2> .05) significant Age group effect for CVP; ω^2(Age group) was .06 for this parameter. Figure 3.6 shows that CVP increased somewhat over the age groups. The correlation between age group and CVP was rather low, however (r= .20, n= 1000).

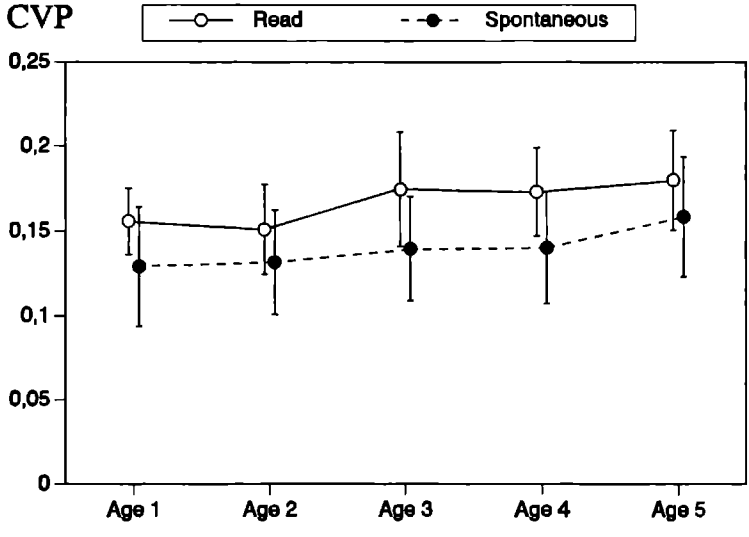

Figure 3.6 Coefficient of Variation of the Pitch period (CVP) as a function of age group

In separate analyses of the two speech styles, two other substantially significant effects were found beside CVP· in the analyses of the read fragments, a significant Age group effect was found for VOI, and in the analyses of the spontaneous speech fragments, PZR turned out to be significant.

Fragment/Paragraph effect

It was not possible to determine the significance of the factor Fragment in the analyses of variance of the total material, as there was no appropriate error term available to construct an F-ratio. From the ω^2(Fragment) values in Table 3.3 we can see, however, that the fragment-to-fragment variability must be high, since the Fragment effect explains a large part of the variance The same is true for the analyses of the spontaneous speech data. There, the ω^2(Fragment) values are even higher than in the overall data analyses, which is not surprising, as in the spontaneous speech condition *all* fragments had a different lexical content, while in the read speech condition the variation was much smaller[9].

In the overall and in the spontaneous analyses we considered Fragment to be nested under Speech style, Sex, Age group, Session, and Speaker. In the analyses of the read speech, however, we did not treat Fragment as a nested variable because of the fact that the speakers read out the same paragraphs. In this somewhat different design it *was* possible to determine the significance of Paragraph. In the remainder of this section we will discuss only the Paragraph effect, the effect of the fragments *in the reading-style analyses only.*

There were five parameters for which we found a substantially ($\omega^2 > .05$) significant effect: CVA, APQ, PAUSE, RATE and VOI. The Z-scores of the different paragraphs on these parameters are shown in Figure 3.7.

When the values of a parameter differ significantly from paragraph to paragraph, this is most probably the result of a dependence of the parameter on the linguistic content of the paragraph in which it is measured[10].

Session effect

For some parameters a significant effect of Session was found, but for none of them did the strength of association of this effect exceed .05. Still, as will be shown in section 3.9, to some (small) extent it is possible to correctly assign the cases to the sessions. The most important parameter in these discriminant analyses is PPQ. The value of ω^2(Session) found for PPQ was .03.

---

[9] In fact, the material read by the speakers was (about) the same, the first fragment of each speaker contained the material from the first paragraph of the reading-text, the second fragment the material of the second paragraph, etc

[10] It might also be the result of some sort of "text-prosody", but it is practically impossible to differentiate these two causes

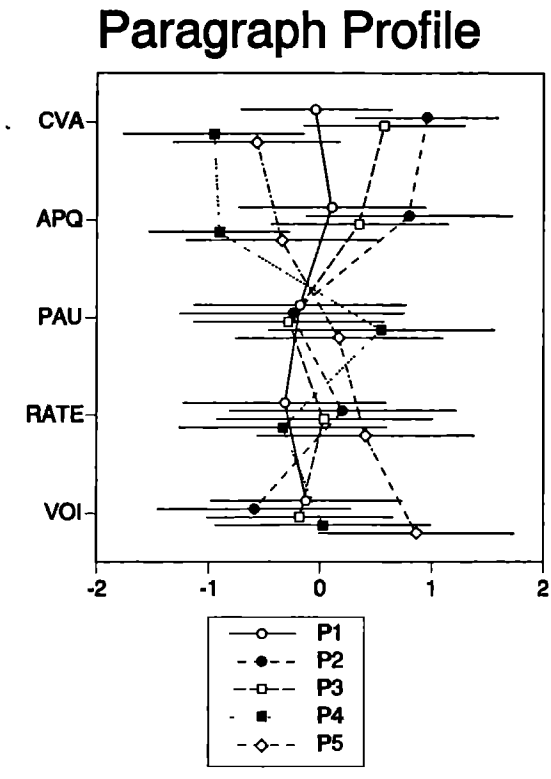

Figure 3.7 Paragraph Profile, means and standard deviations of the parameters for which a significant Paragraph effect was found in the read material Means and standard deviations are expressed in Z-scores relative to the overall mean and standard deviation of each measure. The codes "P1" to "P5" stand for the first to the fifth paragraph

3.4 Speaker identification by LDA

In the next sections, we shall describe the results of the discriminant analyses that were applied to assess how well the TI parameters can be used to assign cases to extralinguistic factors. The assignment of the cases to the speakers forms the crucial part of this study.

Discriminant analysis is a statistical technique that is to some extent comparable to the more widely used technique of factor analysis. Factor analysis aims at finding *factors*, linear combinations of the scores on the input variables weighed by factor loadings. The weights are determined in such a way as to maximize the amount of explained variance in the factors. The *functions* in discriminant analysis and the factors in factor analysis are somewhat alike, the difference being that the discriminant functions maximize the discrimination of the groups (in this study e.g. Speakers), instead of the amount of explained

variance. The functions, again analogous to the factors of factor analysis, span a multi-dimensional space (hyperspace) in which the variables can be marked[11].

Figure 3.8 shows a fictitious data set. The scores of two groups (e.g. data from men and women) on two variables are plotted in a two-dimensional space. Dimension 1 maximizes the amount of explained variance and is the best one-dimensional representation of the total data set. Dimension 1 corresponds to the first factor found in a factor analysis for these data. Dimension 2 maximally separates the two groups, and therefore Dimension 2 corresponds to the first function in a discriminant analysis in which the two groups are to be discerned.

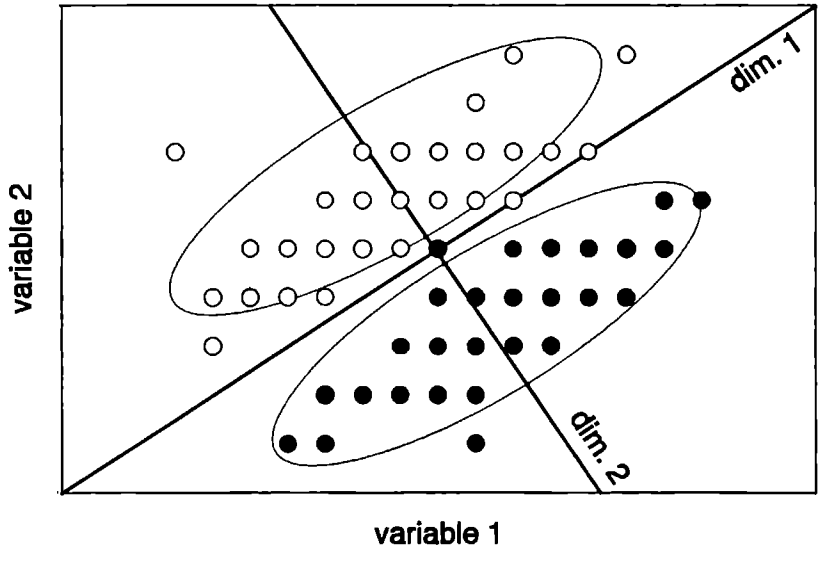

Figure 3.8 Illustration of the difference between factor and discriminant analysis by means of a fictitious data set: Dimension 1 corresponds to the first factor in a factor analysis, Dimension 2 to the first function in a discriminant analysis.

Once discriminant functions have been constructed, hyperplanes mark off sub-spaces in the hyperspace where cases are assigned to specified groups. By assigning the cases according to these criteria and then comparing the assignments to the groups from which the cases really originated, the discriminating qualities of the variables can be determined.

The different degrees of identification performance that are obtained in different LDA's cannot always be directly compared to each other. Equal speaker identification accuracy is a better achievement if more groups are to be discriminated. In the analyses with the entire group of 50 speakers as discriminant groups it is more difficult to obtain good discrimination than in the LDA's with either male or female speakers, where there are only 25 groups to be discerned. To make the outcomes of the analyses more comparable, we will not report the identification performance per se, but the proportion identifica-

[11] For an introduction to discriminant analysis techniques, see e.g. Klecka (1980) or Stevens (1986).

tion exceeding chance (Klecka's tau, see Klecka, 1980). To determine this score, which we refer to as Identification Score, or *IS*, one needs the raw proportion of correct identification, P_{raw}, and the proportion correct at chance level, P_{ch}. The formula used is:

$$IS = \frac{P_{raw} - P_{ch}}{100 - P_{ch}} \times 100 \ [\%]. \qquad (12)$$

It is important to keep in mind that the identification scores mentioned in this study are lower than the raw proportions of correct identification actually found. In analyses where the percentage correct at chance level is low, such as in analyses with the speakers as the discriminant groups (with 50 speakers P_{ch} is only 2 %), the difference between the reported IS and P_{raw} is only marginal. However, in analyses in which for instance the speech styles are to be discerned (with two speech styles P_{ch} is 50 %) the difference is more important. The raw percentages of correct identification are presented in Appendix F.

To be useful in any application, the variables should not only enable discriminant analyses with a high percentage of correct group assignment, it is also necessary that discriminant functions can be applied to correctly assign new cases to the groups. As was explained in section 1.2, to this end we collected speech material from two sessions, with a time interval of about seven months. The influence of the course of time on the discriminating power of the discriminant functions will be tested by means of *cross-validation* analyses; first discriminant functions are constructed on the basis of one session and then these functions are used to assign the cases from the other session to the groups. In this study we performed both of the possible cross-validations, i.e., assigning the material of the first session on the basis of functions derived from the material in the second session, and vice versa. These two cross-validations lead to somewhat different identification scores, but since they both reflect the same attribute (usefulness over different sessions) we only present the average value of the two analyses. In Table 3.7 and all following tables with discriminant analyses, the identification score in cross-validation will be denoted as "c.v.".

The identification scores for LDA's that are performed with data from one of the sessions are indicative of the session-to-session variation as well. The identification score *within sessions* will be indicated by "w.s.". The reported percentage of correct identification within sessions is the average of the identification scores found in the analyses of the two sessions separately.

To determine the importance of the predictor variables for the assignment of the cases to the groups, we need a measure of the association of discriminant functions and predictor variables. Some authors recommend the use of discriminant function coefficients for this, while others favour the use of structure coefficients, the within-groups correlations of the response variables with each discriminant function (Huberty and Morris, 1989). Recently, Thomas (1992) proposed a new index, the *Discriminant Ratio Coefficient* (DRC). This is the product of the standardized discriminant function coefficients and the corresponding structure coefficients. Thomas showed that the discriminant ratio coefficient provides a measure of relative importance with important advantages over its composite parts, the standardized discriminant function coefficients and the structure coefficients.

In order to enhance the interpretability of the multidimensional space, the DRC values are determined after performing a Varimax rotation. This rotation spins the axes of

the hyperspace around the origin in such a way that the *vari*ance of the loadings of the predictor variables is *max*imized (hence its name), while orthogonality is preserved. Using the Varimax procedure it was possible to find a clearly related parameter for most of the rotated functions (i.e., one for which DRC exceeds .50).

In the present study discriminant analyses are carried out to determine how well the ten time-integrated variables can be applied to the task of identifying the speakers and the levels of other extra-linguistic factors. The results of the LDA's will be presented in tables with a fixed format. As we will present many such tables in the remainder of this book, we shall first describe their format in some detail. The first of these tables, Table 3.7, is used as an illustration in this description.

In Table 3.7 discriminant analyses are summarized in which the 50 speakers functioned as discriminant groups. As in the factor analyses discussed in section 3.2, discriminant analyses were performed over subsets of the data to assess the extent to which the grouping variables influence the outcome of the analysis; apart from analyses with the total data set, analyses were performed with only the read half of the fragments, only the spontaneous half, only the data from female speakers and only the male speakers' data. The results of these five analyses are reported in the columns of the table.

As a measure of the importance of the predictor variables, the parameters with DRC's exceeding .50 are reported for most of the functions. For reasons of conciseness, we decided to report only the most related variables for as many functions as are necessary to explain more than 85 % of the variance. This variance criterion was chosen because we found that in most cases a relatively small number of functions can explain up to 85 % of the variance, while many more are needed to enhance this percentage any further. The DRC's themselves are not reported; the Varimax rotation brings about a strong relatedness of the functions and the variables, which leads to DRC's that are mostly about equal to the maximum value of 1.

In the same columns as the parameter names, the percentage of explained variance relating to the *unrotated* functions is reported. The relation between the parameter name and the percentage should not be misinterpreted. From the percentage one can infer how many functions are necessary to describe the between-groups variation to a reasonable extent, while the order of the percentage names gives an indication of the order of importance of the parameters.

The identification score of the cases to the speakers is specified below the functions. It is denoted as c.a., the percentage of *correct assignment* exceeding chance.

Now we go on to the discussion of the results that are reported in Table 3.7. In the overall analysis the 50 speakers functioned as discriminant groups, with 20 data points (2 sessions × 2 speech styles × 5 fragments) per group. In the analyses of speech fragments of one speech style there were 50 speakers with 10 data points per group, while in analyses with speakers of the same sex there were 25 speakers with 20 data points per speaker.

All discriminant analyses resulted in 10-dimensional representations. This is in fact the maximum number of functions because the number of functions in a discriminant analysis cannot exceed the number of groups minus one or the number of predictor variables, whichever is the smallest (Stevens, 1986: 234).

Table 3.7
Summary of discriminant analyses with the speakers as groups (speaker identification) over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males. The most related TI parameters and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are given, followed by the IS within sessions and the cross-validation IS.

	total		read		spontaneous		females		males	
f. 1	F_0MEAN	86.56	F_0MEAN	84.00	F_0MEAN	84.86	F_0MEAN	58.62	F_0MEAN	42.62
f. 2			PZR	4.42	PZR	3.62	PZR	13.60	PPQ	15.19
f. 3							CVP	7.85	CVP	12.30
f. 4							VOI	6.45	PZR	10.87
f. 5									PAUSE	7.52
c.a.	60.1		88.0		70.2		65.0		63.1	
w.s.	71.5		96.5		89.4		75.6		71.0	
c.v.	33.0		53.9		29.4		33.1		30.2	

When discriminant functions were determined in relation to Speakers on the basis of the total data set, the assignment of the fragments to the speakers was successful in 60 % of the cases. This rather low percentage was caused by the heterogeneity in the data; as we saw in the previous section, large differences exist between the two speech styles and the sexes. The variability in the prosodic parameters is further increased by combining data from two different sessions. This becomes clear in the analysis of the data from separate sessions; averaged over both sessions, the identification accuracy within sessions (w.s.) was 72 %, which is about 10 % higher than in the analysis with data from the two sessions combined (c.a.).

From the decrease in the IS that was caused by the combination of data from the two sessions one can predict that cross-validation (the assignment of fragments from one session on the basis of discriminant functions that were derived in the other session) cannot be very successful. Indeed it was not; the IS in cross-validation was very low: 33 %.

In the analysis of all 1000 fragments, the most important predictor variable, the one most related with the most important discriminant function, was F_0MEAN (as can be seen in the first data column of Table 3.7).

Separate analyses were carried out for read and spontaneous speech data. In both analyses the assignment of the cases to the speakers was clearly better than in the overall analysis: IS was 88 % for the analysis of read speech and 70 % of the cases was assigned correctly in the analysis of the spontaneous speech fragments. We assume that such large differences exist between read and spontaneous speech, that this extra variance blurs the speaker differences to some extent. Compared to the cross-validation results in the overall analyses, cross-validation was more than 20 points higher in the analysis of the read fragments, while in the analysis with the spontaneous speech fragments it was somewhat lower. Apparently the properties of read speech are more constant over different recording sessions.

Differences between the speech styles do not influence the relative importance of the parameters. In both speech styles F_0MEAN was highly related with the first (rotated) function and PZR with the second.

In the same way that we performed separate analyses for the two speech styles, we also studied the two sexes separately. A somewhat higher identification performance was found in these latter analyses than in the overall analysis (60 %). In the analysis of the female speakers we found 65 % correct assignment and for the male ones the IS was 63 %. In the separate analyses of the two sexes we need more than two functions to explain 85 % of the variance. The first few functions were not as speaker-specific in these analyses as they were in the former ones. Speaker differences in the parameters that were related to these functions (F_0MEAN and PZR) were probably smaller in the present analyses. When we compare the analyses of the male and female speakers, some striking differences can be observed in the roles played by the different TI parameters. The most important difference is found in the role of PPQ. In the analysis of the male data this variable is related to the second (rotated) discriminant function, while in the analysis of the females it is not related to any of the most important functions. PPQ therefore appears to be much more speaker-specific for men than for women.

Mean F_0 played a major role in all analyses so far. The importance of this variable for speaker identification was attested before (e.g. Sambur, 1975), and it is therefore important to test whether the other nine time-integrated parameters contribute anything to the already well-known effect of mean F_0. Therefore, in Table 3.8 the outcomes of two kinds of analyses are presented: discriminant analyses with all time-integrated variables *except* mean F_0, and analyses with mean F_0 *only*.

Table 3.8

Summary of discriminant analyses with all TI parameters *except* mean F_0, and with mean F_0 *only*, LDA's were carried out with the speakers as groups over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males The most related parameters and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance Below the functions the percentages of correct identification are given

	total		read		spontaneous		females		males	
f 1	PZR	47 61	PZR	48 67	PZR	32 75	PZR	39 11	PPQ	35 03
f 2	CVP	14 43	CVP	14 85	CVP	17 52	VOI	19 87	PZR	18.04
f 3	APQ	9 60	RATE	10 41	VOI	12 55	RATE	13 07	PAUSE	16 95
f 4	PAUSE	8 26	APQ	7 83	PPQ	11.52	CVP	8 04	CVP	10 85
f 5	PPQ	7 64	PAUSE	6 73	APQ	7 59	PAUSE	6 83	AZR	8 38
f 6					PAUSE	6 80				
c a without F_0	47 9		79 2		56 1		51 9		54 2	
c a only F_0	8 3		18 6		15 7		10 0		4 6	

Removing F_0MEAN from the analysis clearly reduced the percentage of correct identification. In the overall analysis the percentage fell from 60 % to 48 %, which shows the importance of F_0MEAN as a predictor variable. Although this decrease in IS is substantial, it is clear that speaker discrimination does not exclusively rely on F_0MEAN, as only a relatively low IS was found in the analyses in which F_0MEAN was the only predictor variable. The other parameters must have been important for speaker identification as well.

Both in the analyses with only F_0MEAN and in those with all parameters except F_0MEAN, the best results were found for the read speech data. In general, read speech seems to be more speaker-specific than spontaneous speech.

After the removal of F_0MEAN, considerably more functions were needed to explain more than 85 % of the variance. In all analyses the parameter that was formerly related to the second-most important function was now related to the first discriminant function. In all LDA's except that of the male data, PZR was related to the first function and in the analysis of the male data PPQ was associated with that function.

In Table 3.7 we reported the outcomes of discriminant analyses of the total material and subsets of it. Even smaller partial analyses are now performed by considering the spontaneous and the read speech data separately for male and female speakers. The results of these four analyses are summarized in Table 3.9.

Table 3.9
Summary of discriminant analyses with speakers as groups per sex and per speech style, over (1) read material, female speakers, (2) spontaneous material, female speakers, (3) read material, male speakers and (4) spontaneous material, male speakers. The most related TI parameters and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are given, followed by the IS within sessions and the cross-validation IS.

	female speakers				male speakers			
	read		spontaneous		read		spontaneous	
f. 1	F_0MEAN	55.98	F_0MEAN	49.38	F_0MEAN	34.43	F_0MEAN	46.06
f. 2	CVP	15.02	VOI	14 38	CVP	20.54	CVP	15.71
f. 3	PZR	10.10	PZR	13.00	RATE	16.66	PPQ	11.11
f. 4	RATE	5.68	CVP	8.76	PAUSE	8.84	AZR	9.15
f. 5					AZR	7.85	PAUSE	5.50
c.a.	90.0		68.3		87.9		75.4	
w.s.	97.9		89.2		95.0		94.6	
c.v.	55.4		25.4		49 6		26.3	

As compared with the identification performance in the overall analysis (60 % of the cases), these partial analyses led to a much higher IS. Apart from the large difference between the two speech styles, which was shown earlier in Table 3.7 and 3.8, the most striking fact is that in the analyses with the read fragments, comparable IS's were found for the male and female speakers (90 % and 88 %, respectively), while in the spontaneous material the percentage was lower for females (75 % vs. 68 %). Perhaps the low IS for

women's spontaneous speech is caused by the less important role of CVP in that condition; in female spontaneous speech CVP was related to the fourth discriminant function while in the other LDA's it was related to the second function.

As in the earlier analyses, F_0MEAN turned out to be of major importance; in all four analyses it was highly related to the first function.

In the LDA's reported in Table 3.7 we found that PZR was an important speaker-identifying parameter, except perhaps in the analysis of the male speakers' data, where PPQ was more important. In the analyses reported here, PZR was again more important for the identification of the female speakers; apparently PZR is less suitable for male speaker identification.

As in the analyses reported in Table 3.7, comparing only material from one session boosts the amount of correct identification considerably, while in cross-validation a sharp decrease in identification accuracy is found.

Here we conclude the discussion of the application of the prosodic parameters to speaker identification. In the remainder of this chapter we will present the outcomes of LDA's in which the speech styles, the sexes, the age groups, the fragments, and the sessions serve as discriminant groups.

3.5 Identification of speech style by LDA

In the previous section, the influence of the two different speech styles on speaker identification was demonstrated. Read speech fragments clearly differ from those obtained from spontaneous speech. Many (sometimes conflicting) observations on the differences between read and spontaneous speech have been reported in the literature (see the summary given in section 1.3.3), and in Table 3.3 it was shown that for seven TI parameters substantially significant differences were found between the speech styles.

By performing LDA's in which the two speech styles function as discriminant groups, we determine how well the styles can be discriminated and which of the predictor variables really matter in characterizing them.

As in the analyses with the speakers as discriminant groups, we also performed LDA's over subsets of the data. This was done to determine the influence of another important extralinguistic factor, Sex, on the outcome of the analysis.

In the overall analysis with the two Speech styles as discriminant groups there were 500 data points (2 sessions × 50 speakers × 5 fragments) per group. Thus, in separate analyses with the female and the male speakers there were 250 data points per group (2 sessions × 25 speakers × 5 fragments). As the grouping variable (i.e., Speech style) has only two levels, only one discriminant function can be determined. The outcomes of the analyses are reported in Table 3.10.

Table 3.10
Summary of discriminant analyses with the speech styles as groups, over (1) the total material (2) the females and (3) the males Below the function the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS

	total	female speakers	male speakers
f 1	– 100 00	– 100 00	– 100 00
c a.	87 0	87 6	91 2
w s	88 0	87 6	92 4
c v	86 6	83 6	90 8

The assignment of the fragments to the speech styles was successful for a large number of cases: 87 % Using material from female speakers only did not substantially change this result, but somewhat better Speech style identification was possible using male speech For analyses with material from only one of the sessions, as well as for cross-validation analyses, we found essentially the same results as for the analysis of the total data set. Apparently the parameter values of the speech styles were stable over the recording sessions.

Discriminating read and spontaneous speech was easy in all three analyses. All of the parameters for which substantially significant Speech style differences were found in the analyses of variance contributed in some degree to the discriminant function Consequently, this function was not related to just one predictor variable and none of the variables had a DRC that exceeded 50 (which is indicated in Table 3 10 by the character "–").

It is important to stress that we do not make any claims as to the validity of the differences that we found between read and spontaneous speech *in general*, as we argued earlier (see section 1 3) that these are very broad categories from which subtypes had to be selected. The differences found were at least partly the result of our choice of these subsets VOI, for instance, is higher in the read speech fragments, but this was caused by our deliberately including many voiced phonemes in the reading-text[12] We speculate, however, that the newspaper-like character of the reading-text and the casualness of the interview lead to speech styles that are relatively far apart on the read-spontaneous continuum.

3.6 Identification of speaker characteristics: sex and age

In this section the possibility of assigning utterances to speaker sex and age is assessed In section 3.4 we stressed that our primary interest is speaker identification The stratification of the speaker group for sex and age allows us to test whether and how well prosodic parameters can be applied to assign utterances to the sex and age groups First we present

[12] Removing the parameter VOI from the LDA's does not lead to a large decrement in the percentage of correct identification In the overall analysis, this percentage even *increases* from 93 6 % with VOI to 94 3 % without it

the results concerning sex identification, next we turn to the identification of the speakers' age groups.

3.6.1 Sex identification by LDA

In the previous section we tried to find out how well our predictor variables could distinguish between the two speech styles. To this end we performed LDA's with the two speech styles as the discriminant groups. Apart from an overall analysis, we also performed analyses for the material from the male and female speakers. In this section we will present the outcomes LDA's in which the two sex groups function as discriminant groups. Furthermore, we will discuss separate analyses for the two speech styles.

As in the analyses of Speech style, in the overall analysis each of the discriminant groups (i.e., the sex groups) had 500 data points (2 sessions × 2 speech styles × 25 speakers × 5 fragments). In the analyses with material of only one speech style, there were 250 data points per sex group (2 sessions × 25 speakers × 5 fragments). Again, with two discriminant groups, only one function can be found. In Table 3.11 the outcomes of the analyses are presented.

Table 3.11

Summary of discriminant analyses with the sexes as groups, over (1) the total material (2) the read material and (3) the spontaneous material. The most related TI parameter and the percentage of explained variance are reported for the discriminant function. Below the function the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS.

	total		read		spontaneous	
f. 1	F_0MEAN	100.00	F_0MEAN	100.00	F_0MEAN	100.00
c.a.	97.0		97.6		98.0	
w.s.	97.4		98.0		98.0	
c.v.	97.0		97.6		98.0	

The success of assigning the cases to the sexes even surpassed that of the Speech style identification in the previous section: 97 % of the cases was correctly assigned. With such high percentages there is not much room for improvement in separate analyses of read and spontaneous speech (ceiling effect).

As was to be expected, this success is mainly based on the all-pervasive but already well-known difference in mean F_0 between the sexes. Removing mean F_0 from the analyses allows us to find out how well the remaining predictor variables can discriminate between the sexes. The outcomes of these analyses, plus the results of discriminant analyses in which *only* F_0MEAN was applied as a predictor variable, are reported in Table 3.12.

Table 3.12
Summary of discriminant analyses with all TI parameters *except* F_0MEAN and with F_0MEAN *only*; with the sexes as groups, over (1) the total material (2) the read material and (3) the spontaneous material. The most related TI parameter and the percentage of explained variance are reported for the discriminant function.

	total	read	spontaneous
f. 1	PZR 100.00	PZR 100.00	PZR 100.00
c.a. without F_0MEAN	80.0	90.0	72.0
c.a. only F_0MEAN	97.2	96.8	98.4

An analysis with only F_0MEAN as a predictor variable led to about the same IS as the analysis with all ten predictor variables did. However, it was also quite well possible to predict sex on the basis of the nine other predictor variables. PZR, which is moderately correlated with mean F_0 (r= .62, n= 1000), was now the variable most closely related to the discriminant function.

In Table 3.3 it can be seen that there are substantially significant differences for Sex on four parameters besides F_0MEAN and PZR. For the parameters F_0MEAN, CVP, PZR and AZR higher values were found for female speakers, APQ and PAUSE were higher for male ones.

3.6.2 Age group identification by LDA

The age groups in this study cover a period of life in which no large mutations take place; all speakers have passed the age of puberty and have not yet reached old age. We therefore do not expect to find clear Age group identification. If there were differences between the groups that would be large enough to permit some degree of Age group assignment, the best results would probably be obtained in separate analyses for the different sexes and the different speech styles because the large variation due to Speech style and Sex might conceal the much smaller age group differences.

In the overall analysis, the five Age groups functioned as discriminant groups, with 200 data points (2 speech styles × 10 speakers × 5 fragments × 2 sessions) per group. Consequently, 100 data points per group were available for the analyses with material from either one sex group or one speech style.

The highest possible number of functions in the analyses is four. In all LDA's we indeed found four functions. The outcomes of the analyses are reported in Table 3.13.

Table 3.13
Summary of LDA's with the age groups as groups, over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males The most TI parameter and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance Below the functions the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS

	total		read		spontaneous		females		males	
f 1	RATE	72 93	RATE	65 15	PPQ	71 59	F_0MEAN	69 81	–	59 34
f 2	PZR	13 09	–	17 64	–	13 27	PPQ	20 91	PPQ	23 47
f 3			VOI	11 35	RATF	8 86			–	13 16
c a	26 3		30 8		30 5		35 8		37 8	
w s	28 9		34 0		33 0		39 3		38 8	
c v	21 9		24 0		19 0		30 8		28 5	

Using the total material to determine discriminant functions, the assignment was successful above chance level in 26 % of the cases

The relations between the TI parameters and the discriminant functions do not show a coherent pattern for the five analyses However, some variables repeatedly turned up and seem to be somewhat more important for the attribution of fragments to age groups than others RATE appears to be of importance, since it was related to the first function in the overall analysis In Table 3 3 it can be seen that the factor Age was significant in an analysis of variance with this variable

A rather striking result is the importance of mean F_0 for the Age group identification of female speakers, which is not found for male speakers The youngest groups of male and female speakers had higher mean F_0 values than the other groups, but for female speakers the difference was particularly striking mean F_0 of age group 1 was 21 Hz higher than that of age group 2 Regarding the identification accuracy there appear to be no large differences between the analyses of the two sexes

Using material from only one of the two sessions does not reduce the variation in the data very much Within sessions the IS is not much higher than in the overall analysis, and cross-validation does not lead to a much lower identification performance Thus we can conclude that the differences in prosodic behaviour between the age groups were small, but consistent from session to session

3 7 Identification of task characteristics paragraph and session

In the preceding sections we discussed the possibilities of applying prosodic parameters to the identification of the extralinguistic factors Speaker, Speech style, and Sex and Age groups The remaining two factors were related to the tasks Fragment/Paragraph and Session As explained earlier, the reason for making recordings on two different occasions and for dividing the material into different fragments was that good speaker identification is too easy if there is no variation in the linguistic material uttered by the speakers and if all recordings are made on one single occasion We consider discerning the levels of these

factors to be much less interesting.

However, although LDA's with fragments and sessions as discriminant groups are not of primary importance, they are not useless. From LDA's with the paragraphs as discriminant groups we can learn in how far the TI parameters depend on the (lexical) content of the utterances in which they are measured. Furthermore, a high level of paragraph identification accuracy would be an indication that an integration time of 15 seconds is not enough to attain stable TI measures. From LDA's for the two sessions we can learn whether the recording circumstances were stable enough to draw any conclusions from our cross-validations.

In Table 3.8 it can be seen that, in the read fragments, the effect of Paragraph was substantially ($\omega^2 > .05$) significant for five parameters: CVA, APQ, PAUSE, RATE and VOI. For none of the parameters was the effect of Session substantially significant. In this last part of the presentation of our results we try to identify the Paragraphs and the Sessions from which we obtained the cases.

3.7.1 Paragraph identification by LDA

The LDA's with the different paragraphs as discriminant groups were performed expecting *not* to find any differences. If we found clear discrimination of the fragments, it would mean that the texts read by the speakers influence the values of the predictor variables. We chose an integration time of 15 seconds as we expected this interval to be long enough to remove the possible effects of the linguistic structure of what is being said.

From the analyses of variance reported in Table 3.4 we know that for some of the predictor variables an interaction exists between Paragraph and Sex. Separate analyses were therefore carried out for each of the sex groups, to see whether higher IS's could be attained.

In the overall analysis the five discriminant groups contained 100 data points each (2 sessions × 50 speakers × 1 speech style). In the analysis of data from only one of the sexes we had 50 data points per group at our disposal. The outcomes of the analyses are reported in Table 3.14.

Table 3.14

Summary of discriminant analyses with the paragraphs as groups, over (1) the total material (2) the females and (3) the males. The most related TI parameter and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS.

	total		females		males	
f. 1	CVA	80.03	APQ	81.38	CVA	74.96
f. 2	APQ	14.47	CVA	12.59	APQ	20.09
c.a.	51.3		57.0		56.5	
w.s.	53.8		54.5		59.5	
c.v.	45.3		43.0		49.5	

The assignment of the cases to the paragraphs was reasonably successful; the identification score was 51 % (i.e., 51 % of the 80 % that can be gained, as the chance level of success

is 20 %). Two discriminant functions were needed to explain more than 85 % of the variance. The parameters that were related to these functions were CVA and APQ. In the overall analysis and within the male speakers' data the first function was related to CVA and the second to APQ. For the female speakers the order was reversed and APQ was related to the most important function.

The IS in the separate sessions is about the same as that in the combined analysis, and the identification accuracy in cross-validation analyses was only slightly lower.

3.7.2 Session identification by LDA

The main reason for including speech material from different recording sessions is to assess the stability of the identification performance of the LDA's. In order to be able to draw valid conclusions from the cross-validations, we tried to keep the recording circumstances during the sessions as constant as possible. Therefore, we do not hope to find large differences between the two sessions, as such a finding could result from differences between the recording circumstances. Differences between the sessions may also be the result of attitude changes of (some of) the speakers in between sessions.

In the LDA's the two sessions functioned as discriminant groups, with 500 data points (2 speech styles × 50 speakers × 5 fragments) per group. The outcomes of the analyses are reported in Table 3.15.

Table 3.15

Summary of discriminant analyses with the two sessions as groups, over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males. The most related TI parameter and the percentage of explained variance are reported for the discriminant function. Below the function the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS.

	total	read	spontaneous	female	male
f. 1	– 100.0	– 100.0	– 100.0	PPQ 100.0	– 100.0
c.a.	15.2	22.8	17.2	15.6	20.0

Using the total material to determine discriminant functions in relation to the sessions, the number of cases that was assigned successfully was 15 % above chance level. Since the amount of success was so low, we determined Cohen's kappa, K. This is a measure of the chance-corrected percentage of agreement between actual and predicted group memberships and its standard error can be used to set confidence limits for the accuracy of the discriminant prediction (Wiedemann and Fenster, 1978). For the outcome of the overall analysis we found a significant Z-value (Z_K= 4.81, $p<.01$), which means that the K-value was so large that the probability that it would have occurred by random sampling from a population with K= 0 is extremely low.

In analyses of the separate speech styles and sexes, we always found IS's slightly above chance level. For these partial analyses we found Z_K values of 5.10 (read), 3.85 (spontaneous), 3.49 (females), and 4.47 (males). All these Z-values were significant

($p<.01$)[13].

3.8 Longer stretches of speech

In section 2.6.1 we referred to Barry et al. (1991), who found that speech fragments of a duration of 15 seconds are too short to attain stable mean F_0 scores. The results reported in this chapter show that the ten time-integrated parameters together enabled a substantial attribution of the cases to the speakers, and that the contribution of the parameter mean F_0 to the identification of the speakers was substantial. Still, with a longer integration time a higher stability of the parameters might be reached, resulting in a higher identification accuracy.

To find out to what extent longer integration time leads to higher IS's, we repeat the discriminant analyses, this time with fragments of a duration of 75 seconds. The values of the time-integrated parameters were determined for the entire 75 seconds of speech that were available per speaker per speech style per session. In the overall analysis the number of fragments was 200 (50 speakers × 2 speech styles × 2 sessions).

The results of the LDA's with the 75-second fragments and the outcomes of the analyses with fragments of 15 seconds are both summarized in Table 3.16, to enable a comparison of the results for fragments of different duration. A problem with comparing the outcomes of analyses of the 75-second and the 15-second fragments is that the number of observations in the analyses was not equal. In the analyses of the 15-second fragments there were five fragments per speaker per speech style per session, 1000 fragments in all, while the number of cases in the 75-second fragments was only 200. To allow a fair comparison, this difference should be taken into account. In the last line of Table 3.16 the Z-score of a proportional difference in two samples of different size is given. The formula for this Z-score is:

$$Z_{diff} = \frac{p_1 - p_2}{\sqrt{p^* \times (1 - p^*) \times (\frac{1}{n_1} + \frac{1}{n_2})}}, \qquad (13)$$

where p_1 and p_2 are the proportions correct in the two samples, n_1 and n_2 are the sizes of the samples, and p^* is defined as:

$$p^* = \frac{k_1 + k_2}{n_1 + n_2}, \qquad (14)$$

in which k_1 and k_2 are the numbers of correct identification in the two samples. All Z_{diff} values in Table 3.16 exceed 2.57, which means that the differences are significant ($p<.01$).

---

[13] In a two-tailed approach a Z-score that exceeds 1.96 signifies that $p < .05$, while a Z-score exceeding 2.57 signifies that $p < .01$.

Table 3.16
Summary of discriminant analyses with the speakers as groups and 200 fragments of 75 seconds as cases, over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males The most related TI parameter and the percentage of explained variance are reported for as many factors as are necessary to explain 85 % of the variance Below the functions the percentages of correct identification are presented, followed by the percentages for LDA's with 1000 fragments of 15 seconds

	total	read	spontaneous	females	males
f 1	F_0MEAN 80 93	F_0MEAN 72 11	F_0MEAN 83 84	F_0MEAN 50 13	PZR 40 26
f 2	PZR 5 52	PZR 8 30	PZR 6 10	PZR 15 96	PPQ 16 60
f 3		VOI 6 42		CVP 9 50	F_0MEAN 14 13
f 4				VOI 7 58	VOI 10 10
f 5				A7R 5 94	CVP 7 59
c a 75 s	86 7	99 0	87 8	91 7	81 3
c a 15 s	60 1	88 0	70 2	65 0	63 1
Z_{diff}	7 18	3 32	3 63	5 29	3 51

Overall, longer integration times led to a clearly better identification performance. In the analysis of the 15-second fragments, an IS of 60 % was reached, while for the longer fragments 87 % of the material was identified correctly In the separate analyses of the read and the spontaneous fragments we found important differences as well In the read part of the material an almost perfect speaker identification was attained in the 75-second fragments, while a lower IS was found in the LDA of the spontaneous part of the data The lower IS in the spontaneous condition is probably caused by the fact that in the read fragments all speakers read out the same text, thus reducing the variability within the data.

In separate LDA's of the data from male and female speakers we also found a clearly better speaker identification in the 75-second fragments, especially for the female speakers Apparently the increased integration time had a more beneficial effect on the identification of female speakers than on the identification of men.

In the 75-second fragments' analyses the most important discriminant function explained a smaller part of the variance In four out of five analyses this first function was still related mainly to F_0MEAN Possibly the other TI parameters contributed more to speaker identification for the 75-second fragments Especially PZR, which was related to the relatively more important second function, benefited from the increased integration time. The order in which the parameters were related to the discriminant functions in the 75-second fragments' analyses was about equal to that of the 15-second fragments' LDA's, with the exception of the analyses of the male speakers There we found that the variable most related with the first discriminant function was PZR, not F_0MEAN. The percentage of explained variance of this first function was 40 %. F_0MEAN was not even related to the second-most important function, which was related to PPQ Apparently the period perturbation parameters gained importance by the increased integration time

We conclude that integrating over a much larger stretch of time enables clearly better speaker identification, probably because more stable values can be obtained for some of the TI parameters

3.9 SUMMARY OF RESULTS

In this chapter we investigated the possible application of ten time-integrated parameters to the identification of the extralinguistic factors Speaker, Speech style, Sex, Age, Fragment/Paragraph and Session. The values of the time-integrated (TI) parameters were determined for 1000 fragments of 15 seconds.

In section 3.2 we established, on the basis of both the correlation matrix of the parameters and of the outcome of a Principal Component analysis, that the parameters measure relatively independent speech characteristics. Therefore they were all applied to the identification of the levels of the extralinguistic factors.

In section 3.3 analyses of variance were performed for each individual TI parameter, over the total material and over the data from the read and spontaneous fragments separately. Significant interactions between the factor Speaker and other extralinguistic factors (Speech style, Paragraph and Session) could hamper speaker identification. Even more important than the significance of the interaction effects is their strength of association ω^2, the part of the total variance that is associated with an effect. In all analyses a larger ω^2 was found for Speaker than for the above-mentioned interaction terms. However, for some significant interaction effects the differences were not very large. In these cases the interactions did reduce the applicability of the parameter involved.

Furthermore, for all extralinguistic variables the group differences on the parameters for which a substantially ($\omega^2 > .05$) significant effect was found were presented in profiles. For the sexes, for example, the profile in Figure 3.5 shows that for F_0MEAN, CVP, PZR and AZR[14] higher values were found for female speakers, while for APQ and PAUSE higher values were found for males. Such outcomes of the analyses of variance were later used to explain the outcomes of the discriminant analyses that were subsequently performed.

In section 3.4 the time-integrated parameters were used as predictor variables in discriminant analyses that were performed with the 50 speakers as the discriminant groups. In a linear discriminant analysis (LDA) over all fragments we found 60 % correct identification. Although we expected better speaker identification in an LDA with speakers of both sexes, the identification performance could actually be increased somewhat by partitioning the data set in subsets of male and female speech samples. A larger increase in the IS could be realized by analysing only fragments from one of the speech styles (especially the read speech fragments were speaker-specific).

To be of use in practical applications, it should be possible to correctly assign new cases to the speakers on the basis of discriminant functions that were determined at a different point of time. We found that cross-validation (the assignment of fragments from one of the sessions on the basis of discriminant functions derived from the other session), resulted in a low identification score (IS); overall it was 33 %. In the analyses of read fragments, the best cross-validation results were found: 54 %.

The most important time-integrated variable for speaker identification was mean F_0. In an analysis with mean F_0 alone, however, only a low IS was reached. Therefore we must conclude that the other time-integrated variables were necessary for speaker identifi-

[14] The TI parameters were listed in full in section 3.1.

cation as well.

The IS of the speakers increased considerably if fragments from only one of the two speech styles, read or spontaneous speech, were used. Indeed, the differences between the speech styles were so large that, in section 3.5, we had little problem identifying the speech styles by means of discriminant analyses; the IS was 87 %. In the analyses of variance we had found that substantially significant differences between the speech styles were found in seven of the parameters: CVP, RATE, and VOI were higher for read speech, while PZR, CVA, APQ and AZR had higher values for spontaneous speech.

In section 3.6 we applied the ten TI parameters to the identification of the sex and age groups that the speakers belong to. Many studies have focused on sex differences in voices. A well-known difference is of course the higher mean F_0 of women. In LDA's with the sexes as the discriminant groups we found that a very high percentage of correct sex identification is still possible if mean F_0 is excluded from the analysis. We are not surprised at this finding since, as was noticed above, for five of the other variables we had found a substantially significant difference between the sexes as well.

For this study, we selected speakers from five different age groups; 18-25, 26-35, 36-45, 46-55, and 56-65 years old. In this period of life, after adolescence and before old age, one would not expect to find large differences between the age groups. To some extent, however, it was possible to identify the age groups. In an LDA with the age groups as discriminant groups, an IS of 26 % was reached.

In section 3.7 we tried to identify the paragraphs (the fragments in the read part of the data) and the sessions from which the cases originated. For both of these task characteristics parameter differences are somewhat undesirable. If the parameters differed from one paragraph to the other, this would imply that they were dependent on the (lexical) content of the fragment. An integration time of 15 seconds would not have been enough to get rid of this dependence. Differences between the two recording sessions are undesirable because they might result from differences in the recording procedures.

It was possible, above chance level, to assign cases to paragraphs; the IS was 51 %. The variables that were related to the discriminant functions, CVA and APQ, depended to a large extent on the (lexical?) content of the fragments. In the analyses of variance of these variables the ω^2 values for Paragraph were indeed quite high. It was hardly possible to assign fragments to sessions: IS was only 15 %.

Above we discussed that high values of ω^2 for interaction terms with Speaker on the one hand, and Fragment/Paragraph and Session on the other, are unfavourable. Such high values are due to instability of the parameter estimates, which reduces the applicability of the results to some extent. Although the strength of association for the factor Speaker was mostly larger than for these interaction terms, even better speaker identification might be possible after determining values for the ten parameters over longer stretches of time. Therefore, five 15-second fragments per speaker, per speech style, and per session were combined into one 75-second fragment. In discriminant analyses over these longer stretches it was found that a much higher identification accuracy was reached. This higher accuracy appears to be related to a larger role for other TI parameters than F_0MEAN.

The ten time-integrated variables that were studied in this chapter were to a large extent independent of each other. They could be used to identify the speakers far above chance level, but not to an extent that is encouraging with regard to practical applications. The TI parameters were related to the sex and age of the speaker, to the speech style, to the content of the utterance, and even to some extent to the recording session in which

they were obtained.

In the next chapter we will add to the time-integrated variables information from pivot points in the F_0 contours of specific utterances to find out if this leads to an increase in the accuracy of speaker identification.

4. Contour-bound parameters

4.1 INTRODUCTION

In the previous chapter we studied the relationship between ten time-integrated (TI) prosodic parameters and six extralinguistic factors: Speaker, Sex and Age of the speaker, Speech style, the content of the speech material, and the recording Session from which the material was obtained. The primary aim was to determine to what extent the TI parameters were related to the factor Speaker. The results showed that the time-integrated parameters could be used to identify speakers reasonably well. In the present chapter we try to find out to what extent contour-bound (CB) parameters (measurements taken at pivot points of the F_0 contour in specific utterances) can be applied to the identification of the speakers. We also employ the CB parameters for the identification of the levels of other extra-linguistic factors. Finally, we test whether the combined use of both types of parameters improves identification performance.

Contour-bound parameters require utterances for which the F_0 contours are realized in a comparable way, in terms of the component pitch movements, as defined in the Grammar of Dutch Intonation, GDI ('t Hart et al., 1990). In Chapter 2 we described the selection of the sentence material to be used in the present chapter. Although we found it difficult to direct the speakers' choice of pitch contours, we did find some promising stimulus sentences in a pilot experiment. In the actual recording sessions we compiled a corpus of 4800 utterances (see section 2.6.3): 48 different utterances, produced by 50 speakers on two occasions.

For practical reasons we decided to study only a subset of this corpus. Part of the material was transcribed in terms of GDI pitch movements. On the basis of the uniformity in the speakers' choices of pitch movements we selected three sentences, the "sports sentences", as the experimental material for the analyses reported in this chapter. The sports sentences set consisted of:

De Denen wonnen van de Noren met één-nul
De Ieren wonnen van de Denen met drie-één
De Noren wonnen van de Roemenen met drie-nul

or in English: "The Danes beat the Norwegians by 1-0", "The Irish beat the Danes by 3-1", and "The Norwegians beat the Rumanians by 3-0". The intonation contours in the transcriptions of the sports sentences contained three pitch movements that were realized by all speakers: a pitch movement of type "1" on the stressed syllable of the first nationality and on the first number of the score, and a fall of type "A" on the second number of the score.

The TI variables were identical to the ones applied to the 15-second speech fragments in the previous chapter: F_0MEAN, CVP, PPQ, PZR, CVA, APQ, AZR, PAUSE, RATE and VOI.

In Chapter 2 it was explained that a number of properties of pitch movements can be measured: F_0 at the start and end of the movement, the difference between these two, and the time interval in which the movement takes place. From the last two measurements the slope of the movement can be derived. For the sports sentences, Figure 4.1 shows the pivot points of the pitch movements that were used.

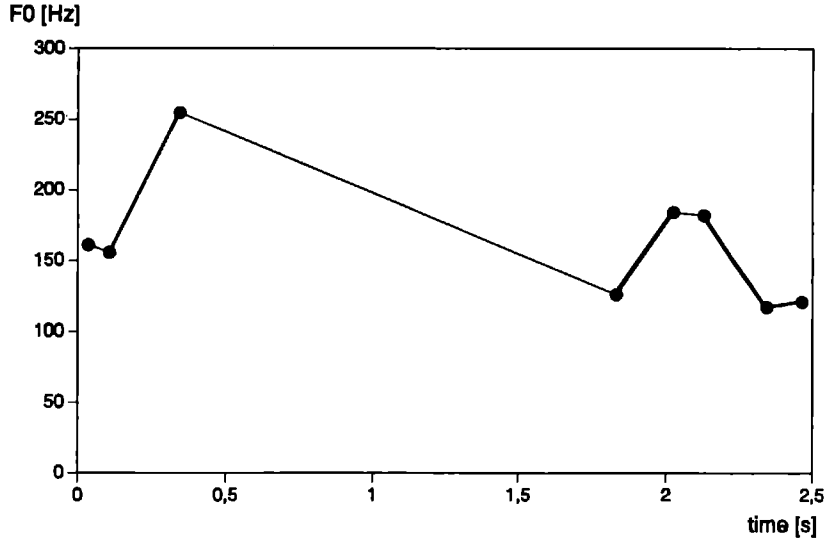

Figure 4.1 F_0 contour: pivot points in the "sports sentences". The coordinates of the pivot points are the actual mean values found for the present speaker group.

F_0 was measured at eight pivot points: the start and end of the three pitch movements and the start and end of the utterance as a whole. It is to be expected that the F_0 values of the pivot points and the F_0MEAN, which is an indicator of the general F_0 level, are closely related. Indeed all intercorrelations exceeded .70. Therefore we determined which of the pivot points was most speaker-specific, in order to maintain only that point as an absolute measure (i.e., in Hz) in the subsequent analyses. The other F_0 values are to be expressed as intervals (in ST) relative to that most speaker-specific pivot point.

As a measure of speaker specificity we used *F*, the ratio of the mean squares between speakers to the mean squares within speakers. The speaker specificity of F_0 at the different pivot points turned out to be more or less equal, with the exception of the F_0 at the start of the first rise, where speaker specificity was low[1], and of the final F_0 value, where speaker specificity was high (as could be expected on the basis of Liberman and Pierrehumbert, 1984). Because of the latter finding, the special importance of final F_0 for

[1] The low speaker specificity of F_0 at the beginning of the first rise appears to originate almost exclusively from one of the sentences: *De leren...* etc. It is probably caused by the difficulty of measuring the beginning of the rise in the vowel */i/* that, at the onset of a word, often has irregular periodicity due to a laryngealized onset (Jongenburger and van Heuven, 1991).

speaker identification, we maintain final F_0 in the analysis as a measure that is expressed in Hz. This measure will probably play a role among the CB parameters that is equivalent to that of F_0MEAN among the TI parameters.

As was announced above, the pitch values at the other pivot points were expressed in terms of the semitone distance to the final F_0 value. We are not only interested in the pitch difference between any pivot point and the final F_0 value, but also in the pitch differences within pitch movements, the semitone distance between the beginning and the end of the movements. To avoid redundancy in the data we do not use the semitone distance between the highest pivot point of a pitch movement and the final F_0 value; this value can directly be obtained by adding the semitone distance of the lowest point of the pitch movement (relative to the final F_0 value) and the semitone distance realized in the pitch movement.

The CB parameters that were used in the present chapter are:

F_0END	Final F_0 of utterance [Hz]
SEMDEC[2]	pitch difference between the first F_0 value in the utterance and the final F_0 value [ST]
SEMRI1	pitch difference between onset and end of first rise [ST]
SEMRI2	pitch difference between onset and end of second rise [ST]
SEMFAL	pitch difference between onset and end of fall [ST]
LOWRI1	pitch difference between F_0 value at the start of the first rise and the final F_0 value [ST]
LOWRI2	pitch difference between F_0 value at the start of the second rise and the final F_0 value [ST]
LOWFAL	pitch difference between F_0 value at the end of the fall and the final F_0 value [ST]
DURRI1	duration (time interval between onset and end) of first rise [ms]
DURRI2	duration of second rise [ms]
DURFAL	duration of fall [ms]
DURFIL	duration of final lowering [ms]
SLODEC	slope of declination over the utterance [ST/ms]
SLORI1	slope of first rise [ST/ms]
SLORI2	slope of second rise [ST/ms]
SLOFAL	slope of fall [ST/ms]
SLOFIL	slope of final lowering [ST/ms]
SYNRI1	synchronisation interval, time interval between onset of first rise and vowel onset of the syllable in which it takes place [ms]
SYNRI2	time interval between onset of second rise and vowel onset of syllable in which it takes place [ms]
SYNFAL	time interval between onset of fall and vowel onset of syllable in which it takes place [ms]
SYNFIL	time interval between onset of final lowering and vowel onset of syllable in which it takes place [ms]

In section 4.2, we will first look at the interrelatedness found in the CB parameters, in order to reduce the initial set of parameters to a subset of parameters that are not too related to each other. In section 4.3 we will give an overview of the outcomes of analyses of variance for the parameters used in this chapter (both TI and CB). Finally, the values of the parameters will be combined by means of linear discriminant analyses to assign the speech material to five grouping variables: Speaker, Sex and Age group, Sentence and Session. LDA's will be performed on both the total material and different subsets of it: the utterances that were produced by female or male speakers only.

---

[2] This variable was not included in any further analyses, due to its high correlation with SLODEC (r= .92, n= 300), see text

Table 4.1
Correlation matrix of 21 CB parameters in the 300 utterances (critical values of r: |.114| ($p < 5$ %) and |.149| ($p < 1$ %), two-tailed)

	F_0 END	SEM DEC	SEM RI1	SEM RI2	SEM FAL	LOW RI1	LOW RI2	LOW FAL	DUR RI1	DUR RI2	DUR FAL	DUR FIL	SLO DEC	SLO RI1	SLO RI2	SLO FAL	SLO FIL	SYN RI1	SYN RI2	SYN FAL
SEMDEC	-.44																			
SEMRI1	-.15	.10																		
SEMRI2	.00	.02	.20																	
SEMFAL	-.13	.16	.32	.58																
LOWRI1	-.37	.63	-.30	.19	.23															
LOWRI2	-.41	.52	.14	-.34	.17	.37														
LOWFAL	-.27	.37	-.07	.00	-.38	.34	.45													
DURRI1	.15	-.06	.25	.01	.02	-.13	-.03	-.04												
DURRI2	.10	-.02	.17	.31	.16	.00	-.17	-.06	.15											
DURFAL	-.03	.04	.01	.23	.36	.07	.07	-.04	.13	.30										
DURFIL	.31	-.14	-.23	.04	-.24	.01	-.12	.27	-.04	.00	-.14									
SLODEC	-.40	.92	.02	-.09	.06	.56	.48	.34	-.15	-.14	-.07	-.18								
SLORI1	-.19	.08	.56	.06	.18	-.22	.15	-.04	-.52	-.03	-.06	-.09	.10							
SLORI2	-.07	.02	.01	.46	.28	.14	-.12	.04	-.12	-.56	-.11	.02	.03	.05						
SLOFAL	.06	.01	.20	.22	.39	.02	-.04	-.36	-.08	-.10	-.43	-.09	.03	.17	.32					
SLOFIL	-.14	.24	-.01	-.01	-.28	.21	.32	.65	.02	.02	-.05	.15	.22	-.03	-.06	-.35				
SYNRI1	-.13	.05	.38	.21	.28	-.08	.15	.03	.41	.01	.09	-.02	-.06	-.05	.11	.05	.01			
SYNRI2	.00	.01	.11	.24	.07	.07	-.27	-.08	.01	.68	.15	.02	-.08	.08	-.31	-.01	-.02	.00		
SYNFAL	.14	-.08	-.04	.25	.17	.01	-.07	.09	.07	.35	.69	.25	-.16	-.09	-.10	-.30	.04	.04	.21	
SYNFIL	.21	-.15	-.06	-.01	-.27	-.08	-.17	.16	-.09	.01	-.52	.47	-.09	-.03	.03	.21	.11	-.08	.04	.27

4.2 INTERRELATEDNESS OF THE VARIABLES

As in the previous chapter, we start the presentation of our results by determining the interrelatedness of the parameters in two ways. Firstly, the Pearson Product-Moment correlation coefficients of the parameters are presented in Table 4.1. As this chapter primarily deals with CB parameters, only the correlations of these parameters are shown. The correlations of the TI parameters within the sports sentences are presented in Appendix G.

In the previous chapter we found that the highest correlation between any two predictor variables was poor, meaning that all predictors potentially provided independent information. Among the CB parameters in the Table 4.1 there is one correlation that clearly exceeds the highest correlation found in the previous chapter (.62): the correlation of .92 between SEMDEC and SLODEC[3]. This high correlation indicates that these two parameters are practically interchangeable, which is not surprising since there was not much variation in the duration of the utterances. If all utterances have more or less the same duration, utterances with similar F_0 declination must also have a similar F_0 slope. We decided to omit SEMDEC from all subsequent analyses, because this parameter has a higher correlation with other declination-related parameters, such as LOWRI1.

The highest correlation found among the remaining parameters, .69 for the correlation of DURFAL and SYNFAL, is not much higher than the highest correlation found in the previous chapter. Therefore, apart from the aforementioned parameter SEMDEC, no parameters need to be omitted on the basis of these correlations.

The highest correlation between a TI and a CB parameter (see Appendix G) was found for F_0MEAN and F_0END: .91. This correlation is not surprising. F_0 at specific points in the contour and the general F_0 level must be related, especially since data from male and female speakers are pooled in the correlation. The correlations of F_0MEAN and F_0END within the male and female speakers were clearly lower: .47 and .69 for men and women, respectively. As in the previous chapter, we found a high correlation between F_0MEAN and PZR. In the sports sentences this correlation was even higher than in the 15-second fragments; .75 in the sentences vs. .62 in the fragments. As could be expected from the high correlation between F_0MEAN and F_0END, we also found a high correlation between F_0END and PZR: .72.

The Principal Component factor analysis that was carried out on the 20 remaining CB parameters resulted in eight factors with eigenvalues higher than one. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was consulted to determine the appropriateness of the factor-analytical model. The value of this measure, .50, can be evaluated as poor (see Rietveld and van Hout, 1993). Apparently the variables measure rather unrelated speech properties.

To find out whether considering only data from one of the levels of the grouping variables increases the sampling adequacy, the 20 remaining CB variables were used in PC analyses that were performed on subsets of the data. Table 4.2 shows the KMO sampling adequacies of analyses in which the sexes, the sentences and the sessions were analysed separately.

---

[3] The same correlation was found for data from male and female speakers separately.

Table 4.2
Sampling adequacy in subsets of the data

variable	levels	nr of speakers	nr of utterances	KMO in CB	KMO, in CB and TI
overall		50	300	498	546
sex	females	25	150	488	525
	males	25	150	530	535
sentence	1	50	100	472	505
	2	50	100	538	527
	3	50	100	518	565
session	1	50	150	508	565
	2	50	150	484	528

For none of the analyses was the fit of the model to the data substantially better than that of the overall analysis Consequently, we retained all 20 CB parameters in the subsequent analyses By adding the ten TI parameters that were introduced in the previous chapter, 30 parameters are used in the present chapter

4 3 Analyses of variance

To be able to see how each of the parameters depends on the extralinguistic factors Speaker, Sex, Age, Sentence and Session and on the interactions of these factors, an overview of the analyses of variance carried out for all of the 10 TI and 20 CB parameters is presented in this section Although the TI parameters were extensively analysed in Chapter 3, new analyses of variance were required because the present material is essentially different (e g shorter) from that used in the previous chapter

Another reason for analysing the TI parameters again is that the values of some of the TI measures were strongly influenced by differences in the experimental material of the present and the previous chapter VOI, the percentage voiced sounds, for example, is one of the measures for which the values of the two data sets cannot be compared VOI is very high in the sports sentences, because these were composed in such a way that they would consist almost exclusively of voiced segments We cannot foretell if and how an overall increase in the values of VOI influences its sensitivity as a measurement instrument The same is true for PAUSE, the percentage of time a speaker does not speak The sentences are short enough to be pronounced without the necessity to pause and inter-sentence pauses are of course absent from isolated sentence material Presumably, the proportion of time the speakers pause is very small, and the parameter might become irrelevant in the sports sentences material

The very fact that the pitch contours of the utterances are more controlled than the contours in the 15-second fragments influences the parameter values The parameter CVP,

for instance, is undoubtedly influenced by the *number* of pitch movements in an utterance. If speakers produce the sentences with a strictly controlled number of rises and falls, the value of CVP will be more related to the *way* in which pitch movements are realized.

The model used in the analyses of variance is a mixed model, with both random and fixed effects. Analyses of variance were performed for all 300 utterances. As was explained in section 3.3.1, the factor Speaker (nested under Sex and Age) was considered to be random and Sex, Age and Session were defined as fixed. The sentences used were of a very specific type; they had a very specific syntactic, rhythmic and theme-rheme structure. Because the sentences were not at all randomly chosen, Sentence appears to be a fixed factor as well. A disadvantage of this design was that for all effects that include Speaker, both for the main effect and for all interactions with Speaker, no appropriate error term was available. It was not possible to construct an F'-ratio either. A more desirable outcome of the analyses could be obtained by defining Sentence as a random variable. However, one should be aware of the fact that it is only random in the sense that many other sports sentences could be generated. The factor Sentence is random, but results can only be generalized to other sports sentences.

This model, with Sentence as a random factor, has the disadvantage that no appropriate error terms are present for the factors Sex, Age, Session, Sex × Age, Sex × Session, and Age × Session. However, for these factors it was possible to generate F'-ratios.

As in Chapter 3, we do not only look for significances, but we keep an eye on the strength of the effects as well. Hays (1973: 415) explains that "virtually any study can be made to show significant results if one uses enough subjects, regardless of how nonsensical the content may be". He adds to this statement that "we want to find and refine relationships that 'pay off', that actually increase our ability to predict behaviour. When the results of an experiment suggest that the strength of an association is very low, then perhaps the experimenter should ask himself whether this matter is worth pursuing after all" (Hays, 1973: 419). Thus, a significant Speaker effect can be of marginal importance if ω^2(Speaker) is low.

Another possible reason why a significant Speaker effect might not be worthwhile is a strong interaction of Speaker with Session. If ω^2(Speaker) is amply surpassed by ω^2(Speaker × Session) or if these ω^2 values are of about the same order, it could be that the scores of the speakers largely depend on the session in which they are measured, which renders the variable concerned of little practical use.

We start the discussion of the 30 analyses with a summary of the results. In Table 4.3 the levels of significance and the degrees of strengths of the effects are presented for the 10 TI parameters, and in Table 4.4 for the 20 CB parameters.

Table 4.3
Summary of analyses of variance for 10 TI parameters (in the columns) For significant effects the cells are shaded and the ω^2 values are presented, if effects are not significant, ω^2 values are presented if they are at least 05, S= speaker, G= gender (sex), A= age, U= utterance (sentence), R= replication (session)

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAUSE	RATE	VOI
S(GA)	.08	.40	.13	.09	.28	.12	.19	.37	.53	.37
G	.89		.10	.63		.10	.05			
A									05	
U	.00	.15	.14	.02	.23	.04			.04	.23
R										
GA								07	07	
SU(GA)[4]		13	22	10	20	18	19	14	08	12
SR(GA)	.01				.17	.08	.15	22	.13	.08
GAUR						.07				
SUR(GA)[4]		30	33	15	24	31	34	34	12	17

Table 4.4
Summary of analyses of variance for 20 CB variables (in the columns) For significant effects the cells are shaded and the ω^2 values are presented, if effects are not significant, ω^2 values are presented if they are at least 05, S= Speaker, G= Gender (sex), A= Age, U= Utterance (sentence), R= Replication (session)

	F_0END	SEMRI1	SEMRI2	SEMFAL	LOWRI1	LOWRI2	LOWFAL	DURRI1	DURRI2	DURFAL
S(GA)	.09	.26	.29	.28	.19	.10	.11	.06	28	.18
G	.81					.08		.06		
A	04			16						
U		.05	.11	.02	.09	.03			.03	.02
R										
GA								.11		
SU(GA)[4]		30	17	20	25	34	27	35	28	29
SR(GA)			.10						07	
GAUR					10					
SUR(GA)[4]		46	29	39	42	42	68	41	40	46

[4] For this effect no significance could be determined because there was no appropriate error term with which an F-ratio could be constructed

	DURFIL	SLODEC	SLORI1	SLORI2	SLOFAL	SLOFIL	SYNRI1	SYNRI2	SYNFAL	SYNFIL
S(GA)	06	.18	.10	.11	.10	.05		.21	.23	
G	.05	.06								.07
A							.07			
U	.05	.00		.15			.04		.00	.03
R							.04			
GA			.06							.08
SU(GA)[4]	34	29	.33	.34	.24	.34	.35	.33	32	.35
SR(GA)		06			06			.12		
GAUR							06			
SUR(GA)[4]	53	.41	63	.49	44	.78	.37	42	.39	.45

We will now discuss the analyses of variance of the 30 parameters in some more detail. As in Chapter 3, we do not discuss all significant effects. Instead, we use a threshold of $\omega^2 = .05$ to determine whether the strength of association of an effect is large enough to make a discussion of the effect worthwhile. Significant effects with a strength of association remaining below this threshold will not be discussed.

Interactions involving Speaker

Interactions between the factors Speaker and Session and between Speaker and Sentence reduce the applicability of a parameter for speaker identification. Therefore, before discussing the main effect Speaker, we turn to the interaction effects with Speaker.

The sports sentences were read out, and it seems most appropriate to compare the results of these sentences with the effects found in the read fragments of Chapter 3. There, Speaker × Session was substantially significant for eight of the TI parameters. In the present analyses for six of the ten TI parameters this interaction term was significant with $\omega^2 > .05$. The picture is dramatically different for the CB parameters, where we found a significant Speaker × Session effect for only two parameters (SEMRI2 and SYNRI2). As was explained above, the relevance of factors in analyses of variance can be determined by comparing the ω^2 values of the effects. For all parameters ω^2(Speaker) was larger than ω^2(Speaker × Session), which implies that the stability from session to session is high. Stable session-to-session speaker differences are very important for the use of prosodic parameters in real-life applications.

For the interaction terms Speaker × Sentence and Speaker × Sentence × Session no significance could be determined, since no appropriate error term was available. We can therefore only compare the ω^2 values of these interaction effects to those of the Speaker effect. In Chapter 3, we found that for all parameters ω^2(Speaker) was indeed higher than the ω^2 values of the interaction effects. In the sports sentence analyses, however, for most of the parameters this was not the case; for four of the TI parameters (the perturbation measures PPQ, PZR, APQ and AZR) and for all CB parameters except F_0END we found ω^2 values that exceeded ω^2(Speaker) for at least one of the interaction terms. This finding is not very promising for the speaker-identifying potential of these parameters. For the slope

of the first rise, SLORI1, for instance, we found an ω^2(Speaker) of .10, an ω^2(Speaker × Sentence) of .33, and an ω^2(Speaker × Sentence × Session) of .63. An important interaction effect of Speaker and Sentence shows that the values found for the speakers depend on the exact (lexical) content of the utterance, which reduces the applicability of parameters such as SLORI1. If the slope of the first rise varies considerably for each combination of Sentence and Session it will not be of much use in speaker identification either.

Other interaction effects

Before investigating the significance of extralinguistic factors other than Speaker, one has to check the interaction terms in which they are involved. The significance of a main effect such as Sex can be completely subordinate to an interaction effect such as Sex × Sentence. The relative importance of the effects can again be determined from their ω^2 values.

In Chapter 3, we did not find many interaction effects of substantial importance. We only found significant Sex × Age group interaction effects in the analyses of the read fragments; for PPQ and AZR, the values found for ω^2(Sex × Age group) were .11 and .10, respectively. In the analyses of the sports sentences we found a substantially significant interaction effect for only one of the TI parameters; for APQ the Sex × Age group × Sentence × Session interaction was significant.

Among the CB parameters there were two substantially significant two-way interactions. The ω^2(Sex × Age group) values found for DURRI1 and SYNFIL were .10 and .08, respectively. Figures 4.2 and 4.3 show how Sex and Age group are related for these parameters.

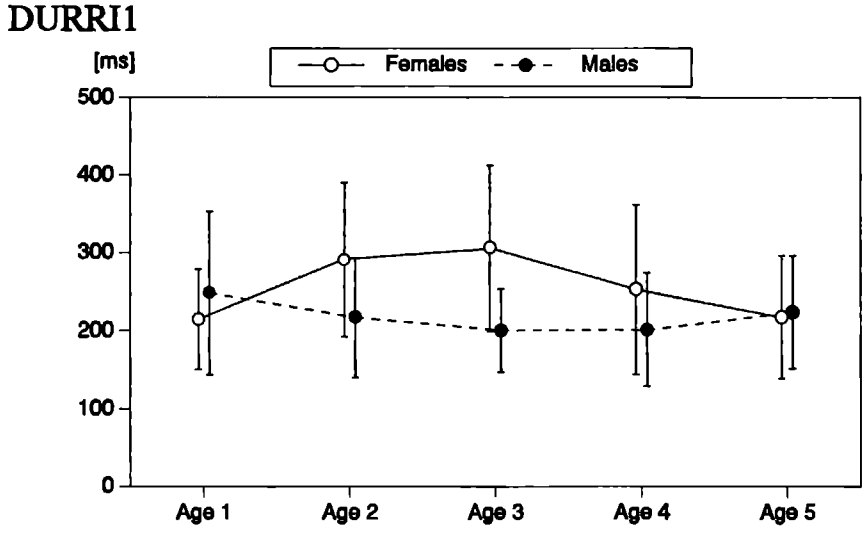

Figure 4.2 Duration of the first rise (DURRI1) as a function of age group, plotted for male (●) and female (○) speakers.

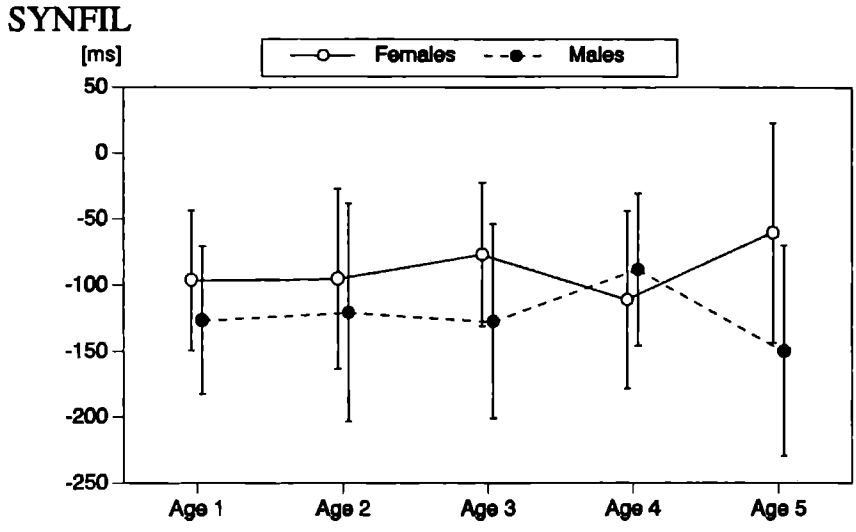

Figure 4.3 Synchronization, time between onset of final lowering and vowel onset of syllable (SYNFIL) as a function of age group, plotted for male (●) and female (○) speakers.

Speaker effect

The crucial main effect for our study is Speaker. For all TI parameters this factor reached significance. Among the 20 CB parameters there were four parameters for which no significant Speaker effect was found: DURFIL, SLOFIL, SYNRI1 and SYNFIL. For all other variables there were significant differences between the speakers. We could illustrate the speaker differences on these parameters by means of figures in which the means and standard deviations of all individual speakers are shown. However, such figures would occupy much room. Therefore we only present the speaker differences on the parameters F_0MEAN and F_0END in Figure 4.4. As we will come to see later, the former is the most important speaker-identifying TI parameter and the latter the most important CB measure. Tables with the exact values of all parameters are provided in Appendix E.

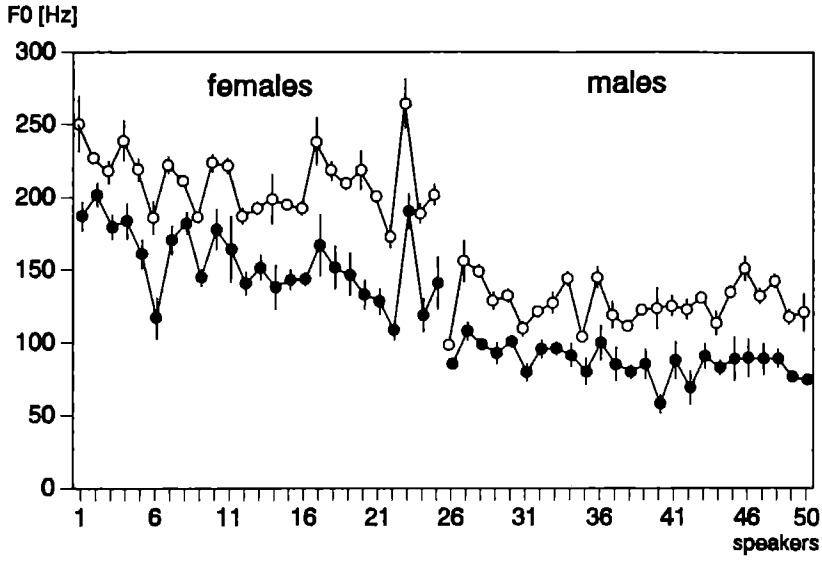

Figure 4.4 Means and standard deviations on the most speaker-specific parameters, F_0MEAN (○) and F_0END (●), for all 50 speakers. The values of the female speakers are given in the left half of the figure, the values of the male speakers in the right. Within each of the sex groups, the speakers are arranged in their age groups. Thus, the data of the first five speakers belong to the youngest group of female speakers.

Sex effect

For the main effects Sex, Age group, Sentence and Session, the origin of the significant effects is again shown in profiles, i.e., figures with the means and standard deviations of the levels of the main effects expressed in Z-scores relative to the mean and standard deviation of the total material.

In Chapter 3, we found sex effects of substantial importance in six parameters of the overall analyses of the fragments: F_0MEAN, CVP, PZR, APQ, AZR and PAUSE. All of these parameters, except CVP, also had a significant effect in the analyses of the read fragments. In the material studied here, sex differences were significant for five time-integrated variables: F_0MEAN, PPQ, PZR, APQ and AZR. Compared with the outcomes in the 15-second fragments' analyses, PPQ was new among the significant values, and PAUSE was no longer sex-specific. For five of the CB parameters significant sex differences were found: F_0END, LOWRI2, DURRI1, DURFIL and SLODEC.

In Figure 4.5, the Sex Profile, it can be seen that mean F_0 is higher for women, and that the difference between men and women is large compared to that found in other measures. The raw data (means and standard deviations before Z-transformation) are presented in Appendix E.

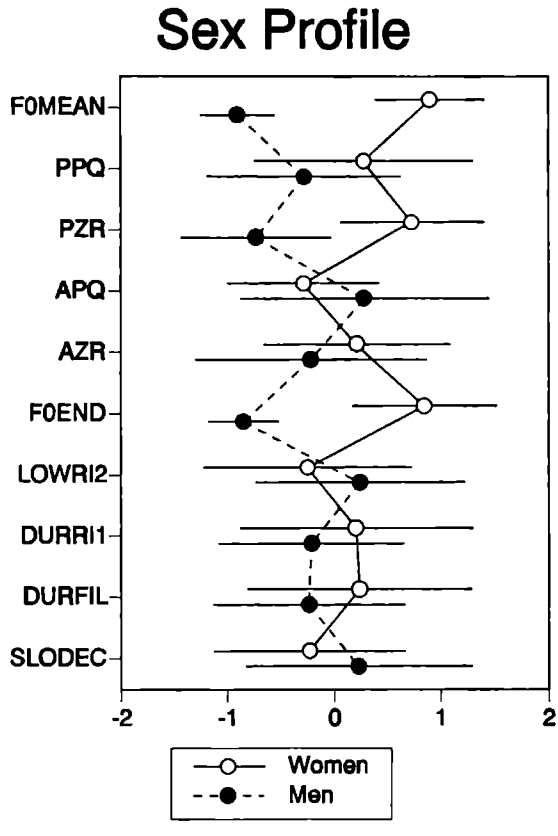

Figure 4.5 Sex Profile: means and standard deviations of the parameters for which a significant Sex effect was found. Means and standard deviations are expressed in Z-scores relative to the overall mean and standard deviation of each measure.

Age group effect

There were two age-specific parameters that were substantially significant in the analyses of the read 15-second fragments: CVP and VOI. Apparently, for these two TI parameters the findings for the reading fragments and for the sports sentences were rather different, since none of the Age group effects were significant in the present analyses.

Among the CB parameters we found two measures on which age group had a substantially significant effect: SEMFAL and SYNRI1 with ω^2(Age group) values of .16 and .07 respectively. Figure 4.6 shows the directions of the differences found:

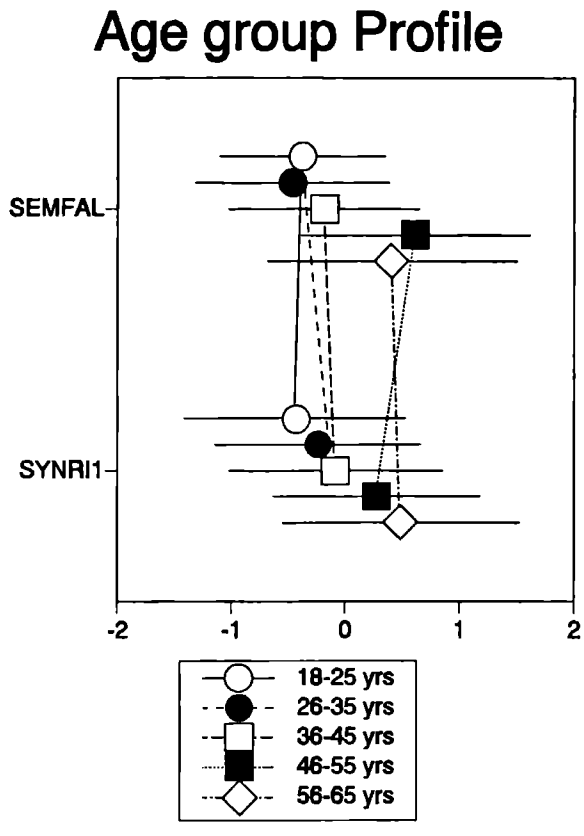

Figure 4.6 Age group Profile: means and standard deviations of the parameters for which a significant Age group effect was found. Means and standard deviations are expressed in Z-scores relative to the overall mean and standard deviation of each measure.

The effect of Age group on the parameter SYNRI1 was systematic: the older the speaker, the longer the time interval between the onset of the rise and the vowel onset. This phenomenon might be related to the changing anatomy of the aging voice apparatus. The data of the second rise and the fall showed a less straightforward trend, but again the two older groups had higher values than the groups with younger speakers.

Sentence effect

The three sports sentences only differed in part of their lexical content: a few words were different. However, as they were prosodically alike, we did not expect to find large differences between the sentences, at least not for the TI parameters. For four measures we did find substantially significant differences, however: for CVP, PPQ, CVA and VOI. These findings cannot be directly compared with the results reported in the previous chapter,

since the sentences were more alike, and shorter, than the fragments.

While we did not expect large differences in the TI parameters, some of the CB parameters are measured in the very words on which the sentences differ. The measures that are related to the first rise, for instance, are measured in different words in each sentence: in *Denen*, *Ieren* and *Noren*. Five parameters were significant and had ω^2 values that exceeded .05: SEMRI1, SEMRI2, LOWRI1, DURFIL and SLORI2. Figure 4.7 shows the directions of the differences found. The sentences were given a numerical code in the following way:

1 = *De Denen wonnen van de Noren met één-nul.*
2 = *De Ieren wonnen van de Denen met drie-één.*
3 = *De Noren wonnen van de Roemenen met drie-nul.*

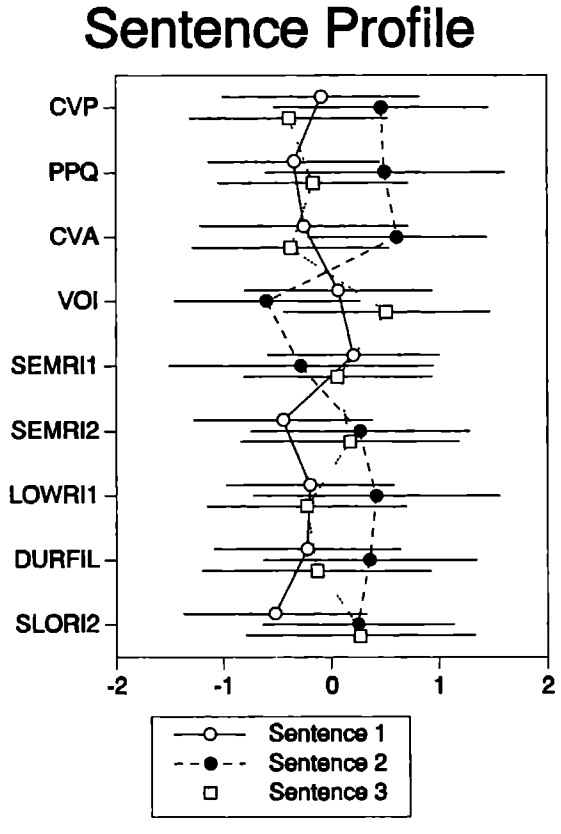

Figure 4.7 Sentence Profile means and standard deviations of the parameters for which a significant Age group effect was found Means and standard deviations are expressed in Z-scores relative to the overall mean and standard deviation of each measure

The differences between the levels of the significant CB parameters can probably be explained by the phonetic consequences of the lexical content of the target syllables. The duration of the time interval between the end of the fall and the end of the utterance was longest in the second sentence. This sentence was the only one that ended in the word *één*, which contains the long vowel /e/, instead of the short vowel /Y/ in the word *nul* in the other two sentences.

The F_0 excursion of the first rise was somewhat smaller in the second sentence and the standard deviation was much larger. The first rise in sentence 2 was on the vowel /i/ of the word *Ieren*. At the onset of a word, the vowel /i/ often has irregular periodicity due to a laryngealized onset (Jongenburger and van Heuven, 1991). The laryngealized onsets might have made it difficult to determine the exact point where the rise began, which could explain the smaller rise and the large standard deviation.

In the second pitch rise, we found a relatively small semitone difference and slope for the first sentence. That sentence is the only one with the word *één* instead of *drie* in the relevant position. Just as *Ieren* in the first rise, *één* starts with a laryngealized onset, but this time the standard deviation was not very large. The smaller pitch rise in *één* is probably the result of another segmental characteristic. In a stressed syllable, F_0 tends to jump to a higher level if the syllable has an initial unvoiced consonant than if it starts with a vowel or voiced consonant. The high F_0 at the initiation of voicing is likely to be related to the adducting motion of the vocal folds, which causes the initial periods to be short. Voiced consonants tend to show a small dip in F_0, presumably resulting from a decreased glottal pressure drop due to the consonantal constriction of the vocal tract (Ohde, 1984). The F_0 level at the onset of the rise was indeed higher (although not significantly) in the first sentence than in the other two; LOWRI2 was 1.72 Hz in the first sentence, 0.72 Hz in the second and 0.39 Hz in the third sentence.

Summarizing, we found that the influence of the exact words in which the measurements were taken was large enough to lead to significant sentence differences for some of the CB parameters.

Session effect

As in the previous chapter, no significant session effects were found for which the strength of association exceeded .05. The fact that no session effects were found must not be misinterpreted: the main effect of Session was not significant, but especially among the TI parameters Session was often involved in a significant interaction with Speaker.

4.4 Speaker identification by LDA

In this section we report the speaker identification results of three types of LDA's: (1) those based on the TI parameters, (2) those based on the CB parameters, and (3) those based on a combination of these two types of parameters.

In these analyses, the 50 speakers functioned as discriminant groups. There were six data points (2 sessions × 3 sentences) per group. Separate LDA's were performed over subsets of the data to find out to what extent the sex of the speakers influences the outcome of the analyses.

LDA's with TI parameters

In Chapter 3 LDA's were presented that were based on fragments with a duration of 15 and 45 seconds, respectively. We performed new LDA's for the sports sentences in order to compare the performance of the TI and the CB parameters

As in the previous chapter, the predictor variables that were most related (DRC exceeding 50) to the most important discriminant functions (the functions that together explain more than 85 % of the variance) are reported in tables of the format introduced in section 3 4 A summary of the analyses involving TI parameters is given in Table 4.5:

Table 4.5

Summary of LDA's with the TI parameters as predictors and with the speakers as groups, over (1) the total material (2) the females and (3) the males The most related predictor and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance Below the functions, the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS

	total		females		males	
f 1	F_0MEAN	78 74	F_0MEAN	44 66	F_0MEAN	35 72
f 2	RATE	7 82	RATE	19 04	RATE	23 78
f 3			VOI	12 44	CVP	15 86
f 4			CVP	7 61	PAUSE	8 33
f 5			AZR	5 36	VOI	5 50
c a	81 3		83 3		86 1	
w s	91 8		88 9		93 1	
c v	41 5		43 8		38 2	

The data in Table 4.5 can be compared to the results for the 15-second fragments that were reported in Table 3 7, preferably with the analysis of the read speech fragments, as the sports sentences were also read The assignment of the utterances to the speakers was successful in 81 % of the cases in the sports sentences, which is somewhat less than the 88 % correct identification found in the analysis of the read fragments. The duration of the fragments was about six times the duration of the sports sentences, however. Even without the CB parameters in the analyses, relatively short, controlled utterances enable about the same quality of speaker identification as much longer, uncontrolled speech fragments.

Cross-validation, the assignment of utterances from one session by means of discriminant functions that are based on material from the other session, also resulted in a poorer IS than was found in the analysis of the fragments, 42 % was correctly identified here, as opposed to 54 % in the cross-validations of the read 15-second fragments.

In addition to discriminant analyses for the total material, analyses were also performed for subsets of the material In the analyses in which either the male or the female speakers' data were included, we found identification scores that were somewhat higher than those in the overall analysis. 83 % for females and 86 % for males vs. 81 % for the combined material

In the LDA's of the sports sentences, F_0MEAN was again related to the most important (Varimax-rotated) discriminant function, both in the overall analysis and in the

separate analyses of male and female data. Of course F_0MEAN is less speaker-specific in the analyses of the two sexes separately, as the large between-sex difference in F_0 is no longer present. As a result, more than two functions were needed to explain 85 % of the variance. Compared to the analyses of the 15-second fragments, the most important difference appears to be in the role of RATE. Perhaps the larger (prosodic and lexical) control in the sentences reduced the influence of some sources of irrelevant variance for RATE, thereby allowing it to play a more important role in speaker identification.

The analyses of the TI parameters again demonstrated the importance of mean F_0. Since the speaker specificity of this parameter is a well-known phenomenon, it is important to test the performance of discriminant analyses with all time-integrated variables *except* mean F_0, and of analyses with *only* mean F_0. The summary of these LDA's is presented in Table 4.6:

Table 4.6
Summary of LDA's with all TI parameters *except* F_0MEAN, and with F_0MEAN *only*; analyses were carried out with the speakers as groups over (1) the total material (2) the females and (3) the males. The most related predictor variables and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are presented.

	total		females		males	
f. 1	RATE	32.54	RATE	36.72	RATE	39.66
f. 2	PZR	24.14	VOI	22.90	CVP	23.03
f. 3	CVP	13.05	CVP	11.75	PAUSE	12.21
f. 4	VOI	8.72	AZR	9.10	VOI	7.99
f. 5	PAUSE	7.28	PZR	7.22	CVA	6.47
c.a. without F_0MEAN	68.1		73.6		75.0	
c.a. only F_0MEAN	18.7		20.8		14.6	

Removing F_0MEAN from the analysis clearly reduced the identification accuracy. The largest decrease was in the overall analysis, where the percentage fell from 81 % to 68 %. This decrease by 13 % is equal to that found in the 15-second fragments. Although this finding confirms the importance of F_0MEAN as a predictor variable, the high IS found after removing F_0MEAN, as well as the rather low IS (19 %) found in an analysis with F_0MEAN as the only predictor variable, shows that speaker identification does not rely exclusively on F_0MEAN.

The results found for the ten TI parameters were not essentially different from the findings in Chapter 3. Identification accuracy was higher, probably because of the increased homogeneity in the material, and somewhat better identification results were obtained when the two sex groups were analysed separately. Again F_0MEAN was the most important, but by no means all-important, speaker-specific parameter.

LDA's with CB parameters
The overall analysis and the separate analyses of male and female data are reported in Table 4.7.

Table 4.7
Summary of LDA's with the CB parameters as predictors and with the speakers as groups over (1) the total material (2) the females and (3) the males. The most related predictor and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS.

	total		females		males	
f. 1	F_0END	80.57	F_0END	53.64	F_0END	34.83
f. 2	LOWRI1	6.16	LOWFAL	12.79	LOWRI1	27.44
f. 3			LOWRI1	10.43	SEMRI2	8.60
f. 4			SLODEC	8.09	LOWRI2	7.62
f. 5			DURRI2	4.45	SLODEC	5.69
f. 6					SLOFAL	4.89
c.a.	86.4		91.7		83.3	
w.s.	86.4		90.3		68.1	
c.v.	41.8		43.1		16.7	

The most striking conclusion that can be drawn on the basis of these analyses, is that the CB parameters enable a level of speaker identification accuracy that is slightly higher than that found in the LDA's of the TI parameters. Cross-validation performance was more or less equal to that found in the LDA's with TI parameters as well. Using only measurements taken at a limited number of carefully selected 10-ms frames, the same performance was reached as with parameters that were obtained by averaging over the total duration of the utterance (mean duration of the sports sentences was about 2.5 s).

In the overall analysis of all 300 utterances, as well as in the separate analyses of the male and female data, the prominent role of F_0MEAN in the analyses of the TI parameters was parallelled by F_0END. The percentages of explained variance of F_0END were about equal to the percentages of explained variance in the LDA's with TI parameters. The second-most important function was related to LOWRI1, the semitone-difference between the beginning of the first rise and the end of the utterance.

In the LDA based on the male speakers' data, IS was somewhat lower (83 %) than in the overall analysis (86 %), while in the LDA of female data it was higher (92 %). In the TI analyses we found that F_0MEAN was less speaker-specific in the separate analyses of the two sexes than in the overall analysis. Again like F_0MEAN in the analyses of the TI parameters, F_0END was still related to the most important function in the analyses for one of the sex groups only, but that function explained a much smaller part of the variance.

The importance of F_0END does not surprise us given its high correlation with F_0MEAN, which already proved to be important in the TI analyses. The importance of F_0END has been attested before (e.g. Liberman and Pierrehumbert, 1984) and it is therefore interesting to find out how important the other CB parameters were for speaker identification. Therefore, the performance of discriminant analyses with all CB variables *except* F_0END, and that of analyses with F_0END *only* was tested. The results of these analyses are reported in Table 4.8.

Table 4.8
Summary of LDA's with all CB parameters as predictors *except* F_0END, and with F_0END *only*; analyses were carried out with the speakers as groups over (1) the total material (2) the females and (3) the males. The most related predictors and the percentages of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are presented.

	total		females		males	
f. 1	LOWRI1	32.19	LOWRI1	29.33	LOWRI1	42.64
f. 2	SEMFAL	15.14	SEMFAL LOWFAL	20.74	SEMFAL	14.79
f. 3	SEMRI2	11.87	DURRI2	16.28	SEMRI2	11.02
f. 4	DURRI1	10.19	DURFAL	9.20	SLODEC	9.31
f. 5	DURFAL	7.70	SEMRI2	8.46	DURFIL	7.67
f. 6	SLODEC	5.37	DURRI1	5.84		
f. 7	DURRI2	4.12				
c.a. without F_0END	66.7		77.1		68.1	
c.a. only F_0END	12.9		16.7		9.7	

The conclusion to be drawn from these analyses is that the effect of removing F_0END from the CB analyses is somewhat larger than that of removing F_0MEAN from the TI analyses: the identification score fell by 19 %, proving the importance of F_0END as a predictor variable. At the same time, however, the high IS found after removing F_0END, and the low IS found in an analysis with only F_0END, show that F_0END was not all-important for speaker identification.

The most important parameters in the overall LDA without F_0END were LOWRI1, SEMFAL and SEMRI2, in that order.

LDA's with TI and CB parameters combined

The behaviour of F_0MEAN among the TI parameters and that of F_0END among the CB parameters were very similar. Regarding the outcome of analyses over both TI and CB parameters, it is to be expected that, since these two F_0 parameters are related (r= .91, n= 300), both of them will be related to the first discriminant function. Which of the other parameters is related to the second-most important function is not clear, as both for the TI and for the CB parameters a reasonably high level speaker identification without the two F_0 parameters was possible.

Discriminant analyses were performed with both the TI and the CB parameters as the predictor variables. The overall LDA and the analyses for male and female speakers separately are reported in Table 4.9.

Table 4.9
Summary of LDA's with both the TI and the CB parameters as predictors and with the speakers as groups over (1) the total material (2) the females and (3) the males The most related predictors and the percentages of explained variance are reported for as many functions as are necessary to explain 85 % of the variance Below the functions the percentages of correct identification are presented, followed by the IS within sessions and the cross-validation IS

	total		females		males	
f 1	F_0MEAN	74 23	F_0MEAN	44 81	F_0MEAN	26 86
f 2	F_0END	5 89	F_0END	13 81	RATE	20 00
f 3	LOWFAL	4 01	RATE	9 14	LOWRI1	12 81
f 4	RATE	2 63	LOWFAL	8 38	CVP	11 22
f 5			VOI	5 45	F_0END	7 30
f 6			AZR	3 91	SEMFAL	4 71
f 7					PAUSE	3 78
c a	96 9		97 9		95 8	
w s	99 3		97 9		95 1	
c v	51 4		50 0		43 1	

In the overall analysis and in the separate analyses of the male and female data, F_0MEAN was related to the most important function In all three analyses the important role of F_0MEAN did not exclude an independent role for the strongly correlated variable F_0END[5] Even though F_0MEAN was related to the first function in the overall analysis, F_0END was related to the second function In the analysis of the female speakers F_0END was also related to the second discriminant function, but in the analysis of the male speakers' utterances F_0END was less important for speaker identification, as it was related to the fifth discriminant function The functions resulting from an LDA are orthogonal, which means that they are unrelated to each other Apparently, even though their correlation is very high, F_0MEAN and F_0END can play an independent, and important, role in speaker identification We speculate that this role is more prominent for speaker identification with material from the female speakers because of F_0 determination problems in the creaky final part of the male speakers' utterances

Thus far we have shown the importance of F_0MEAN and F_0END for speaker identification We also showed that this importance is not all-pervasive, and that speaker identification on the basis of all TI parameters except F_0MEAN and of all CB parameters except F_0END is quite well possible For the sake of completeness we conclude this presentation of speaker identification LDA's with analyses of both the TI and the CB parameters, without F_0MEAN and F_0END The results of these analyses are reported in Table 4 10.

[5] Remember that the high correlation found for F_0MEAN and F_0END is for an important part the result of the large F_0 difference between male and female speakers Within the sex groups, the correlations are clearly lower (for female speakers r= 69 and for male speakers r= 47)

Table 4.10
Summary of LDA's with all TI and CB parameters *except* F_0MEAN and F_0END, and with F_0MEAN and F_0END *only*; analyses were carried out with the speakers as groups over (1) the total material (2) the females and (3) the males. The most related parameters and the percentage of explained variance are reported for as many functions as are necessary to explain 85 % of the variance. Below the functions the percentages of correct identification are presented.

	total		females		males	
f. 1	RATE	20.46	RATE	25.43	RATE	29.85
f. 2	PZR	18.23	VOI	17.85	LOWRI1	18.79
f. 3	LOWRI1	11.42	AZR	13.09	CVP	12.34
f. 4	SEMFAL	8.98	SEMFAL LOWFAL	8.96	PAUSE	7.51
f. 5	VOI	7.20	SEMRI1	8.26	SEMFAL	6.21
f. 6	CVP	5.47	DURRI2	5.95	VOI	5.54
f. 7	PAUSE	4.69	PZR	4.82	CVA	4.66
f. 8	AZR	4.41	CVP	3.80	SEMRI2	3.56
f. 9	DURFAL	3.31				
f.10	PPQ	2.72				
c.a. without F_0MEAN or F_0END	91.2		91.7		91.0	
c.a. F_0MEAN and F_0END only	33.7		37.5		28.5	

Speaker identification performance on the basis of all parameters except the two F_0 related ones was quite successful, and the results of LDA's with only F_0MEAN and F_0END were much poorer, which confirms the conclusions drawn from the separate LDA's of TI and CB parameters. The F_0-related parameters were important, but not all-important for speaker identification.

For the LDA's of the total material and of the material of female and male speakers only, the parameter that was most related to the first discriminant function was RATE. The parameters that were related to other functions, however, were different for each analysis. This implies that the parameters were not of the same importance for the identification of male or female speakers. LOWRI1, for instance, is related to the second-most important function in the LDA of the male speakers' data, but does not play any role in the analysis of the female speakers' data.

To facilitate a direct comparison between the analyses in which TI and CB parameters were combined and those in which they were separated, the identification scores of most of the analyses presented thus far are summarized in Table 4.11. This table is different from the LDA summaries presented earlier. In each of the cells the identification score is presented. In parentheses the IS within sessions and the cross-validation IS are given. This table enables us to focus on the identification performance only.

Table 4.11
Summary of discriminant analyses with the speakers as groups over (1) the total material (2) the females and (3) the males The percentages of correct identification of LDA's with the TI parameters, the CB parameters and a combination of the two types are shown in the columns two, three and four, respectively IS within sessions and cross validation IS are presented in parentheses

	time-integrated	contour-bound	both types
all material	81 3 (91 8, 41 5)	86 4 (86 4, 41 8)	96 9 (99 3, 51 4)
females	83 3 (88 9, 43 8)	91 7 (90 3, 43 1)	97 9 (97 9, 50 0)
males	86 1 (93 1, 38 2)	83 3 (68 1, 16 7)	95 8 (95 1, 43 1)

From the data presented above, we can conclude that (1) CB parameters enable speaker identification to a degree that is comparable to that found for TI parameters, and that (2) TI and CB parameters combined have a clearly higher discriminative power However, a possible complication in the comparison of the different parameter types could be that the number of parameters is not the same in the different analyses. It is probably easier to reach a high IS with 20 or even 30 parameters than with only 10 Therefore the analyses of the CB parameters and of the combination of the two parameter types were repeated, allowing only 10 parameters to enter into the LDA The results of these analyses are summarized in Table 4 12

Table 4.12
Summary of discriminant analyses into which only ten parameters were entered The speakers were the discriminant groups for analyses over (1) the total material (2) the females and (3) the males The percentages of correct identification of LDA's with the TI parameters, the CB parameters and a combination of the two types are shown in the columns two, three and four, respectively IS within sessions and cross-validation IS are presented in parentheses

	time-integrated	contour bound	both types
all material	81 3 (91 8, 41 5)	76 9 (86 4, 41 8)	84 4 (92 9, 43 5)
females	83 3 (88 9, 43 8)	86 1 (90 3, 43 1)	84 0 (98 6, 45 8)
males	86 1 (93 1, 38 2)	75 7 (68 1, 16 7)	87 5 (95 1, 43 1)

In these analyses the identification accuracy for the CB parameters was slightly lower than that found for the TI parameters 77 % Apparently, measurements taken at only a few pivot points can lead to almost the same degree of accuracy as do measurements resulting from integration over stretches of 2 5 seconds

Having determined that the discrimination performance of the CB parameters was almost equal to that of the TI parameters, we should find out whether combining the two types of measurements raises speaker identification For the same number of parameters applied, the identification score (84 %) was only three points higher than in the analysis of the TI parameters and seven points higher than in the CB measures analysis The predictive power of the combined data, however, was barely higher, the cross-validation performance was only two percentage points higher than in the LDA with only TI parameters

4.5 Identification of speaker characteristics: sex and age

In this section we assess the possibility of identifying the sex and the age of the speakers. Although this is not the primary aim of this study, it is interesting to determine to what extent this can be done on the basis of prosodic parameters. Below we first present the results concerning sex identification and subsequently those pertaining to age group identification.

4.5.1 Sex identification by LDA

In section 3.6.1 the possibility of assigning the 15-second fragments to the sex groups on the basis of TI parameters was discussed. We found that sex identification was almost perfect. At first sight this appeared to be caused by a rather trivial finding: the large F_0 differences between men and women. However, we found that it is still possible to correctly assign a large number of cases to the correct sex group if F_0MEAN is removed from the analysis.

We repeated these analyses for the sports sentences, once for the TI and the CB parameters separately, and once for the two parameter types combined. In the previous chapter we found that F_0MEAN is the most powerful sex-specific TI parameter. Since F_0MEAN and F_0END are strongly related, we expect the role of F_0END among the CB parameters to be prominent with regard to sex identification too. To further clarify the role of F_0MEAN and F_0END in sex identification, we also present the outcomes of analyses from which F_0MEAN and F_0END were removed, as well as those of analyses in which F_0MEAN, F_0END, and the combination of these two parameters were included as predictor parameters.

We start with the summary of LDA's of all parameters. Each of the discriminant groups in the overall analysis had 150 data points (2 sessions × 3 sentences × 25 speakers). Table 4.13 shows the sex identification accuracy with the TI parameters, the CB parameters, and with the two parameter types combined. The results of LDA's with F_0MEAN only are presented in the column with the results of the TI parameters, The results for F_0END only analyses are shown in the column with the CB parameters' results, and the results of the combination of F_0MEAN and F_0END are presented in the last column.

Table 4.13
Summary of discriminant analyses of the sports sentences with the sexes as groups over the total material, with (1) all predictors (2) all predictors except F_0MEAN and F_0END, and (3) only F_0MEAN and F_0END. The percentages of correct identification of LDA's with the TI parameters, the CB parameters and a combination of the two types are shown in the columns two, three and four, respectively. IS between sessions and cross-validation IS are presented in parentheses.

	time-integrated	contour-bound	both TI and CB
all parameters	97.4 (97.4, 97.4)	98.0 (96.6, 92.0)	99.4 (99.4, 98.6)
all par., except F_0MEAN and F_0END	77.4 (77.4, 73.4)	39.4 (29.4, 22.0)	78.0 (78.6, 69.4)
only F_0MEAN and/or F_0END	97.4 (97.4, 97.4)	95.4 (86.6, 86.0)	96.6 (97.4, 97.4)

Sex group identification was so successful that possible differences between analyses based on TI, CB and the two parameter types combined could hardly emerge (ceiling effect) The assignment of the utterances to the sex groups on the basis of both parameter types was successful for all cases but one, the sentence of a female speaker (F52) was assigned to the male group The effect of session is apparently low, since cross-validation analyses were very successful

As we expected, the success of the analysis with the TI parameters alone was mainly based on the difference in F_0MEAN between the sex groups and that of the analysis with the CB parameters on the F_0END differences In the analysis with TI and CB parameters combined, F_0MEAN was the parameter most related to the discriminant function

To find out whether the other parameters could also discriminate between the sex groups, F_0MEAN and F_0END were removed from the analyses The outcomes of these analyses are reported in the second data row of Table 4 13 The results of the analyses with the TI parameters were reasonably good, due to the presence of PZR, a parameter that is closely related to mean F_0 (r= 75, n= 300) In the analysis with the CB parameters, however, there was apparently no variable that could replace F_0END as a sex-specific measure and IS dropped considerably, to 39 % None of the parameters was particularly related to the discriminant function

The results of discriminant analyses in which F_0MEAN, F_0END or both were allowed as predictor variables are reported in the third data row of Table 4 13 From this data row we conclude that analyses with only F_0MEAN and F_0END as predictor variables yielded about the same results as did analyses in which the other prosodic parameters were applied as well

4 5 2 Age group identification by LDA

In Chapter 3 we already mentioned that we did not expect to find clear results for the Age groups, because our age groups cover a period of life where no large voice mutations take place Although we did find significant differences on some of the TI parameters in the fragments (CVP and VOI), their ω^2 values were low and Age group identification was just barely possible The accuracy of age group identification (26 %) was significant, but low

In the overall analysis of the sports sentences the five Age groups were the discriminant groups, with 60 data points (2 sessions × 10 speakers × 3 sentences) per group Consequently, the analyses based on one sex group contained 30 data points The outcomes of the analyses are reported in Table 4 14

Table 4.14

Summary of discriminant analyses with the age groups as groups over (1) the total material (2) the females and (3) the males The percentages of correct identification of LDA's with the TI parameters, the CB parameters and a combination of the two types are shown in the columns two, three and four, respectively IS within sessions and cross validation IS are presented in parentheses

	time integrated	contour bound	both types
all material	22 5 (16 6, 8 8)	25 4 (22 5, 15 0)	33 4 (22 5, 16 6)
females	40 0 (29 1, 24 1)	56 6 (55 0, 39 1)	66 6 (63 4, 40 9)
males	25 0 (17 5, 3 4)	31 6 (25 9, 10 0)	35 9 (26 6, 9 1)

In the analysis based on the TI parameters alone the identification score for the age groups was 23 %. This is somewhat less than the 26 % that was found in the analysis of the 15-second fragments, but is still clearly above chance level.

The low IS was to be expected because of the low age specificity in the TI parameters in the sports sentences. There was not a single parameter for which the age group differences were substantially significant. None of the parameters was clearly related to the first function; APQ was related to the second discriminant function and PAUSE to the third.

In the analyses of the CB parameters, the IS was somewhat higher (25 %). The parameter that was most related to the first discriminant function was SEMFAL. In other words, the amount of pitch change in the fall was important for age group identification. Indeed SEMFAL was one of the two CB parameters for which a significant age group effect (ω^2= .16) was found. As shown in Figure 4.6, the speakers in the two oldest age groups had a larger pitch fall. The second discriminant function was most related with SYNRI1. The higher the age of the speakers, the longer the time interval between the start of the first rise and the onset of the vowel in the syllable in which the rise took place.

As can be seen in Table 4.14, adding TI and CB parameters together raised the identification performance somewhat above the level of the CB parameters. A larger increase in IS could be attained by analysing the data of male and female speakers separately. In an LDA with both types of parameters the IS for the male speakers was only some points higher, but for the female speakers the accuracy of age group identification was 33 % higher than in the overall analysis. The origin of the performance in the combined material is unclear because several different parameters were related to the first functions. In the TI analysis of the female data, RATE was related to the most important discriminant function and in the CB analysis DURRI1 and F_0END were important parameters. RATE and F_0END decreased with age, while for DURRI1 no straightforward trend was found.

Using only material from one of the two sessions did not reduce the variation in the data very much, which becomes clear in analyses with data from one of the sessions only. In these analyses the IS was not noticeably higher than in the overall analysis. Likewise, cross-validation analyses did not result in a much lower degree of accuracy of age group assignment. We conclude that the differences in prosodic behaviour between the age groups did not change much from session to session.

When the outcomes of the present analyses are compared to the results of the LDA's for the read 15-second fragments in the previous chapter, similar levels of accuracy were found. The age group identification performance in the overall analysis of the read 15-second fragments was 26 % and that of the sports sentences was 23 %.

4.6 Identification of task characteristics: sentence and session

The main objective of our study is to find out how well prosodic parameters can be applied to speaker identification. The group of speakers studied was stratified for sex and age in order to control for the influence of these speaker characteristics. The influence of the exact sentence and session from which the measurements were obtained is not in itself interesting. As explained earlier, speaker identification is relatively easy if there is no variation in the linguistic material uttered by the speakers and if all recordings are made

on one single occasion The reason for our not finding more perfect speaker identification lies in the sentence-to-sentence and session-to-session variation in the behaviour of the speakers

Important interactions of the factors Sentence and Session on the one hand and Speaker on the other were to be expected Substantially significant effects of the main factors Sentence and Session themselves are less important In the analyses of variance reported in section 4 4 a substantially significant Sentence effect was found for four TI and five CB parameters For none of the parameters was the effect of Session substantially significant In this last part of the presentation of the results we try to apply the prosodic parameters to relate the cases to the Sentences and the Sessions

4 6 1 Sentence identification by LDA

Above we noticed that there were nine parameters for which the effect of Sentence was significant and of substantial importance Therefore, it might well be possible to assign the cases to the sentences

In the overall analysis the three discriminant groups (i e , the sentences) contained 100 data points each (2 sessions × 50 speakers) The outcomes of the analyses are reported in Table 4 15

Table 4 15

Summary of discriminant analyses with the sentences as groups over (1) the total material (2) the females and (3) the males The percentages of correct identification of LDA's with the TI parameters, the CB parameters and a combination of the two types are shown in the columns two, three and four, respectively IS within sessions and cross validation IS are presented in parentheses

	time integrated	contour bound	both types
all material	34 0 (32 5, 28 5)	51 0 (43 0 48 6)	67 5 (64 0, 56 1)
females	29 1 (28 0, 19 0)	47 1 (36 0, 20 1)	65 1 (42 0 29 1)
males	44 1 (37 0 32 1)	51 0 (43 0, 39 0)	72 0 (55 0, 48 0)

As expected, TI and CB measures can both be used to discriminate among the three sentences to some degree and, again as expected, using CB parameters better sentence identification was accomplished than using TI measures

In none of the analyses did we find parameters that were strongly related to any of the discriminant functions Apparently, even though none of the relationships was very strong, the combination of the many parameters for which Sentence was significant led to some degree of sentence identification

4 6 2 Session identification

We did not expect to find session identification above chance level because no substantially significant session effect was found for any of the parameters in the analyses of variance reported in section 4 4

In the overall analysis of the sports sentences the two sessions were the discriminant groups, with 150 data points each (50 speakers × 3 sentences) The outcomes of the analyses are reported in Table 4 16

Table 4.16
Summary of discriminant analyses with the sessions as groups over (1) the total material (2) the females and (3) the males. The percentages of correct identification of LDA's with the TI parameters, the CB parameters and a combination of the two types are shown in the columns two, three and four, respectively.

	time-integrated	contour-bound	both types
all material	13.4	25.4	25.4
females	12.0	12.0	12.0
males	16.0	26.6	26.6

In the analysis of the total material, assignment of the utterances to the sessions was successful in 13 % of the cases above chance. To be able to set confidence limits for the accuracy of the discriminant prediction, Cohen's kappa, κ, was determined. For all discriminant predictions significant Z_κ-values (i.e., Z_κ exceeding 1.96) were found. This means that the κ-values were so large, that the probability that they would have occurred by random sampling from a population with $\kappa = 0$ is very low; Session identification was possible above chance level.

4.7 SUMMARY OF RESULTS

In the present chapter we studied contour-bound (CB) parameters, i.e., measurements made at specific points in an utterance. Our aim was to determine whether these, alone or in combination with TI parameters, would lead to higher percentages of speaker identification. The speech material used consisted of readings of three sentences of a specific type: the "sports sentences". These sentences were of the form *De Ieren wonnen van de Denen met drie-één* ("The Irish beat the Danes by 3-1"). Measurements were taken at the first rise (on the first syllable of the first nationality), the last rise (on the first number of the score), and on the last fall (on the second number of the score). In most cases, the final fall did not proceed up to the end of the utterance. The stretch of pitch between the end of the fall and the end of the utterance was regarded as a kind of pitch movement as well.

At the outset of this chapter we studied the interrelatedness of 21 CB parameters: the final F_0 value, the semitone difference between this final F_0 value and the lowest point of the three pitch movements, the semitone difference between the final F_0 value and the onset of the utterance, the semitone difference between the highest and the lowest pitch values in the pitch movements, the duration of the three pitch movements and the final lowering (the time interval from the end of the fall to the end of the utterance), the slopes of the declination, the pitch movements and the final lowering, and the synchronization intervals (the time intervals between the onset of a pitch movement and the vowel onset of the syllable in which it takes place) of the three pitch movements and the final lowering.

The correlation between SLODEC (the slope of the declination) and SEMDEC (the semitone difference between the onset and the end of the utterance) was so high that only SLODEC was maintained in subsequent analyses. A factor analysis was performed to determine whether further data reduction was necessary among the remaining CB parameters. This analysis resulted in an eight-factor solution with a poor fit to the data.

Because of this poor fit we retained all parameters except SEMDEC in the subsequent analyses.

For all 10 TI parameters and 20 CB parameters, analyses of variance were performed. Interaction effects between the factor Speaker on the one hand and the factors Sentence and Session on the other, could render significant speaker differences irrelevant for speaker identification. In seven of the TI parameters and two of the CB parameters, the Speaker × Session interaction was substantially significant. In none of these analyses, however, did the ω^2 values of the interaction effect exceed ω^2(Speaker). Interactions between the factors Speaker and Session are therefore not expected to make speaker identification impossible.

The significance of the Speaker × Sentence effect could not be determined, because in the design used no appropriate error term was available to calculate an *F*-ratio. Comparing the ω^2 values of the Speaker × Sentence interaction and the ω^2 values of the Speaker factor, we found higher ω^2 values for the interaction effect in three of the TI parameters and in 16 of the CB parameters. Apparently, many CB parameters are particularly sensitive to the exact syllables in which they are measured. It is to be expected that speaker identification using CB parameters leads to much better results if only measurements within one utterance are used.

The main effect of speaker, which is of primary importance in this study, was significant and of substantial importance for all parameters but four: DURFIL[6], SLOFIL, SYNRI1 and SYNFIL. A substantially significant effect of Sex was found for five TI parameters and five CB measures. An age effect was found only for SEMFAL and SYNRI1. Sentence was significant for four TI and five CB parameters, while no effect was found for Session.

From the LDA's with the speakers as the discriminant groups, an important first conclusion is that, in the total material, the CB parameters led to about the same quality of speaker identification as did the TI parameters. Measurements taken at a few frames only apparently led to almost the same results as measurements that result from integration over a few seconds of time. The results of a separate LDA with the CB measurements for men lag behind somewhat.

About as important a finding is that combining the two types of measurements raises speaker identification: from 81 % for the TI measures and 86 % for the CB measures, to 97 % correct identification for the combination of the two types.

Analysing the different sessions either separately or in combination led to slight differences in degree speaker identification accuracy. Cross-validation (the assignment of cases from one session on the basis of discriminant functions obtained for material in the other session) was not very successful, though. The cross-validation IS for the total material was 51 % for the combination of the parameter types and 42 % for both the CB parameters and the TI parameters separately.

In the LDA with the TI parameters as well as in the LDA with TI and CB parameters combined, the variable related to the most important function was F_0MEAN. In the analyses of the CB parameters, F_0END played a comparable role. Without these two measures speaker identification performance decreased considerably in LDA's with TI and CB parameters separately. Smaller decrease was observed in an LDA with both parameter

---

[6] The CB parameters were listed in full in section 4.1.

types. In the previous chapter we found that, although F_0MEAN is an important TI parameter, speaker identification does not exclusively rely on it. To some extent this conclusion also applies to the role of F_0END among the CB parameters: it is important, but omitting it from the analysis does not make speaker identification impossible.

F_0MEAN and F_0END were related to a considerable extent (r= .91, n= 300). Nevertheless, in the combined analyses of TI and CB parameters, it was found that both parameters were related to separate discriminant functions; F_0MEAN was related to the first function and F_0END to the second. Since discriminant functions are mutually independent, the small difference between the two parameters must have some relevance for speaker identification. We speculate that F_0END indeed plays an independent role which is more prominent for speaker identification with material from the female speakers, because of measuring problems in the very low final part of the male speakers' utterances.

With regard to the distinctiveness of the sex groups, we confirmed the finding in the previous chapter that almost perfect sex identification is possible on the basis of F_0MEAN alone. A very high Sex IS was again possible when mean F_0 was excluded from the analysis. In LDA's with the CB parameters alone, a high sex identification performance was possible because of F_0END. Analyses with F_0END only were much less successful: in an LDA with all CB parameters except F_0END IS was 39 %.

For this study speakers were selected from age groups between which no large differences were expected. Indeed, in the previous chapter we found an age group identification performance that was not impressively high, but above chance level. The IS's found for the sports sentences were similar to those of the 15-second fragments. The CB parameters could distinguish between the age groups slightly better than the TI parameters (25 % vs. 23 %), and the combination of the parameter types resulted in still better results (33 %).

Differences between the three sentences and between the two recording sessions are not of primary interest to this study. Different sentences and different sessions were mainly included in the design of the study to create a more realistic setting: better speaker identification results will be found for linguistically identical material that is obtained from recordings that were made on one single occasion. It was possible to identify the utterances read. Especially for the CB parameters this was not surprising, as it is to be expected that the pitch movements depend on the segments on which they are measured. The identification of recording sessions was barely possible; the identification scores were just above chance level.

The 20 contour-bound variables that were studied in this chapter were to a large extent independent of each other. Some of them were related to the speaker's sex and age, and to the content of the utterances. Still, they could be used to identify speakers reasonably well. The most important speaker-identifying CB parameter was F_0END. In the LDA with all CB parameters except F_0END, the importance of other CB parameters became clear. The most important parameter was LOWRI1, the semitone distance between the pitch at the start of the first rise and at the end of the utterance. Not all parameters seem to be important for the identification of speakers or sex and age groups. The synchronization intervals, the slopes and the durations of the pitch movements were not very important.

LDA's with the TI parameters were about as able to discriminate speakers as were the CB parameters. When the two types were combined, F_0MEAN appeared to be the most important speaker-identifying parameter.

5. Conclusions

5.1 INTRODUCTION

The aim of this study, as formulated in Chapter 1, was to find out to what extent *prosodic* parameters, both of the time-integrated and of the contour-bound type, can be used to identify speakers. The effects of some influential extralinguistic factors (Sex, Age, Speech style) were strictly controlled, assessed, and factored out so as to reveal what speaker idiosyncrasies remained. This general goal was elaborated into four specific questions, the answering of which will guide us through a discussion of our results with regard to speaker specificity:

1) To what extent is speaker identification possible on the basis of prosodic parameters alone?
2) Which parameters are most important for speaker identification?
3) How stable is prosodic speaker identification, i.e., how dependent is it on speech style and date of recording (are the speaker characterizations at time T1 equally valid at time T2)?
4) To what extent does analysing data within sex and age groups affect speaker identification?

The answering of each of these four questions covers a subsection of section 5.2. In section 5.3, similar questions will be answered for the extralinguistic factors mentioned in question 3 and 4, i.e., speech style, sex and age, date of recording and paragraph and sentence. All these factors are discussed together in one section to underline the importance of the Speaker factor in this study.

The two types of prosodic parameters that were applied to speaker identification in this book, time-integrated (TI) and contour-bound (CB) parameters, were measured in two different sets of experimental material. Many prosodic parameters depend heavily on segmental features when they are measured over short fragments of time. Assuming that the influence of such short-term phenomena is averaged out in longer stretches of speech (and that the measures thus become more stable), the most appropriate domain in which to measure time-integrated variables appears to be a relatively long stretch of speech. The speech material used in Chapter 3 consisted of a thousand 15-second fragments of read and spontaneous speech, in which no attempts were made to control the pitch movements produced by the speakers. The ten TI parameters that were applied to speaker identification were: mean F_0 (or F_0MEAN), the coefficient of variation of the pitch period durations (CVP), the pitch perturbation quotient (PPQ) and the pitch period zero-crossing rate (PZR), the coefficient of variation of the maximum amplitude per cycle (CVA), the amplitude perturbation quotient (APQ) and the amplitude zero-crossing rate (AZR), the articulation rate (RATE), silence as a percentage of the speaking time (PAUSE), and voiced speech as a percentage of the speaking time (VOI). Both analyses of variance and discriminant analyses

were performed on the experimental data set. Chapter 3 was concluded by discriminant analyses with a somewhat reorganized data set, in which groups of five contiguous 15-second fragments were joined into single 75-second fragments.

For the study of contour-bound parameters quite a different sort of experimental material is required than for TI measures. Many parameters, e.g. those that are related to specific F_0 rises or falls, must be measured within comparable linguistic contexts. For such parameters, sentences with fixed intonation contours are the most natural domain in which to make measurements. In Chapter 4 contour-bound parameters were measured in sentences that elicited a number of identical pitch movements in all 50 speakers. These "sports sentences" were of the form *De Ieren wonnen van de Denen met drie-één* ("The Irish beat the Danes by 3-1"). Measurements were taken at the first rise (on the first syllable of the first nationality), the last rise (on the first number of the score), and on the last fall (on the second number of the score). In most cases, the fall did not extend until the end of the utterance. The stretch of F_0 contour between the end of the fall and the end of the utterance was regarded as a pitch movement as well, and was labelled "final lowering".

Altogether we used 20 CB parameters in Chapter 4: the final F_0 value (in [Hz]) the pitch difference between this final F_0 value and the lowest point of the two rises and the fall (in [ST]), the pitch excursion in the two rises and the fall (in [ST]), the duration of the three pitch movement and the "final lowering" (in [ms]), the slopes of the declination line (pitch difference between the onset of the utterance and the final F_0 value), the pitch movements and the final lowering (in [ST/s]), and the synchronization intervals, i.e., the time intervals between the onset of the pitch movements and the vowel onset of the syllables in which these took place (in [ms]).

For the sports sentences the values of the TI parameters were established as well, to be able to compare the speaker-identifying possibilities offered by the two parameter types. We indicated above that TI parameters gain stability when they are measured in longer utterances. Assuming that a higher degree of stability enables better speaker identification, it might appear to be unfair to the TI parameters to compare them with CB parameters in fairly short sentences. However, as will be discussed below, speaker identification on the basis of the TI parameters was actually more successful in the sports sentences than in the 15-second fragments.

In the present chapter we integrate the findings from Chapter 3 and Chapter 4 in the discussion of each of the above-mentioned questions.

5.2 Speaker identification performance

5.2.1 Introduction

To facilitate the discussion of the results of linear discriminant analyses (LDA's) over TI and CB parameters in both the 15-second fragments and in the sports sentences, the identification performance found in the respective LDA's are summarized in Table 5.1, where identification scores[1] are presented for the total material, as well as for males and

[1] The Identification Score, or *IS*, was defined in section 3.4 as the proportion identification exceeding chance (Klecka's tau, see Klecka, 1980).

females and read and spontaneous speech separately. The first two columns specify the identification scores of LDA's with TI parameters, as they were reported in Chapter 3, for fragments of both 15 seconds' duration and of 75 seconds' duration. In the last three columns the identification scores that were found for the sports sentences are listed, first for the TI parameters, next for the CB measures, and finally for both parameter types combined.

Table 5.1
Summary of discriminant analyses concerning speaker identification, over the total material, for females and males, and for read and spontaneous speech separately In the data columns the percentages of correct identification of LDA's with (1) TI parameters, in 15-second fragments (2) TI parameters, in 75-second fragments (3) TI parameters in the sports sentences (4) CB parameters in sports sentences and (5) TI and CB parameters in the sentences In (4) and (5) IS is obtained in LDA's with maximally 10 parameters. In parentheses the IS is given for LDA's with all available parameters

	fragments		sports sentences (n= 300)		
	TI, 15 s (n= 1000)	TI, 75 s (n= 200)	TI, ± 2.5 s	CB	TI + CB
all material	60	87	81	77 (86)	84 (97)
females	65	92	83	86 (92)	84 (98)
males	63	81	86	76 (83)	88 (96)
read	88	99			
spontaneous	70	88			

5.2.2 Extent of speaker identification
First we discuss the identification scores of LDA's that were performed over the pooled fragments (i.e., from both speech styles, both sessions, and all sex and age groups) and over the pooled sports sentences (i.e., from both sessions, all sex and age groups, and all sentences). In these overall analyses we want to find the answer to the question:

1) To what extent is speaker identification possible on the basis of prosodic parameters alone?

The percentage of speaker identification found for the TI measures in the sports sentences was clearly higher than that found in the 15-second fragments (81 % vs. 60 %[2], respectively), even though the time interval in which the parameters were measured was six times shorter in the sports sentences as compared to the 15-second fragments; the sentences lasted only about 2.5 seconds. Even the results of the 75-second fragments in Chapter 3 (87 %) only just exceeded the results of the TI parameters in the sports sentences. This

[2] The percentages reported in this chapter are "chance-corrected", which means that they represent the percentage of identification exceeding chance. The actual percentage of correct identification in the fragments was 60 9 The identification score, the percentage exceeding the chance level of 2 % was 60.1: (60.9 - 2) / 98 In Appendix F the raw identification percentages are presented

finding underlines the favourable influence of an increased amount of (prosodic) control. Apparently this increased control could well compensate for the negative influence of the shorter integration time[3].

For the sports sentences material, a comparison can be made of the speaker identification potential of TI and CB parameters. An LDA with only the CB parameters as predictor variables led to a somewhat better speaker identification than did the TI parameters. For the TI parameters an IS of 81 % was obtained, while for the CB measures IS was 86 %. However, the number of parameters in the analysis of the CB parameters was twice the number used in the analysis of the TI measures. To compensate for this biased situation, an LDA was performed with only the best 10 CB parameters; the percentage of correct identification was now slightly lower than that found for TI parameters, 77 %. Apparently, measurements taken at only a small number of pivot points can give almost the same results as measurements resulting from integration over stretches of 2.5 seconds.

Having established that the CB parameters were about as able to discriminate speakers as the TI parameters were, we should find out whether combining the two types of measurements raises speaker identification. Indeed, speaker identification increased by 16 and 11 percentage points for the TI and the CB parameters, respectively. When the number of parameters was kept under control (i.e., when only ten parameters were allowed to enter into the analysis) the identification score was 84 %, which is only three points higher than in the analysis of the TI parameters alone. Aparently, most of the individual CB parameters add little to the discriminative power of the TI measures, but when many of them are applied, they can raise the speaker identification performance of the TI parameters somewhat.

5.2.3 Importance of parameters for speaker identification

In this section we will discuss the relevance of the individual TI parameters for speaker identification to find the answer to the second question raised in section 5.1:

2) Which parameters are most important for speaker identification?

The relevance of the measures for speaker identification can be deduced from the extent to which they are related to the discriminant functions. In the discriminant analysis of the TI parameters in the 15-second fragments, the most important discriminant function surpassed the other functions by far: it accounted for 87 % of the variance. The parameter that was most closely related to this first function was F_0MEAN.

The importance of F_0MEAN for speaker identification has often been attested in the literature (e.g. Sambur, 1975). To find out how speaker-specific the other parameters were, analyses were carried out from which F_0MEAN was excluded. The identification scores remained rather high and PZR was now related to the first discriminant function. At first sight this appears to be partly caused by the relatively high correlation of this parameter with F_0MEAN. However, the correlation between F_0MEAN and PZR is probably caused by the large PZR difference between the sex groups. Within the sex groups the correlation was

[3] Perhaps it would be better to compare the results of the sports sentences to the read speech fragments. The identification score in the read fragments was 7 % higher than in the sports sentences (which is a significant difference: Z= 374.7).

much smaller, and we assume that the importance of PZR for speaker identification lies in its ability to differentiate speakers within the two sex groups. CVP was related to the second discriminant function.

In the TI analyses of the sports sentences the most important discriminant function was again closely related to F_0MEAN, and again the importance of the first function surpassed that of the other ones by far, accounting for 79 % of the variance. The parameter related to the second-most important function for the sports sentence material was not PZR, but RATE, which was not among the most important parameters in the analysis of the 15-second fragments. Apparently, the increased control over the utterances was favourable for the speaker specificity of the articulation rate. In an LDA from which F_0MEAN was excluded, RATE was the parameter most related to the first discriminant function.

Among the CB parameters one function clearly stood out as well, compared to the others. This first function, accounting for no less than 81 % of the variance, was most related to F_0END, the parameter for which, on the basis of the literature (e.g. Liberman and Pierrehumbert, 1984), a high degree of speaker specificity was expected. The second-most important function was related to LOWRI1, the pitch difference between the start of the first rise and the end of the utterance. LOWRI1 is also the most important parameter in an LDA with all CB parameters except F_0END[4].

Even though F_0MEAN and F_0END are strongly related ($r = .91$, $n = 300$), they were associated with separate discriminant functions in the LDA in which TI and CB parameters were combined. F_0MEAN was related to the first function and F_0END to the second. Since discriminant functions are mutually independent, the small difference between the two parameters must have some relevance for speaker identification[5].

In an LDA with all parameters except F_0MEAN and F_0END the relevance for speaker identification of parameters that were not directly related to the general F_0 level became clear. The most important parameters were TI parameters; RATE was related to the first discriminant function and PZR to the second. In all LDA's from which F_0MEAN and F_0END were excluded, both in Chapter 3 and in Chapter 4, the identification performance decreased markedly. Omitting F_0MEAN and F_0END from the analyses does not make speaker identification impossible, however. From these results we conclude that F_0MEAN and F_0END are important, but not all-important parameters.

In Chapter 1 we noted that parameters for which little speaker specificity would be found were perhaps linguistically relevant because linguistic constraints do not allow for much variation. On the other hand, a lack of speaker specificity could just as well point to a large amount of random variation.

The speaker specificity of the parameters is not only deduced from their relatedness to the discriminant functions. Another important element is the amount of explained variance for the factor Speaker in the analyses of variance reported in Chapters 3 and 4.

---

[4] This finding is the more surprising since we noted earlier, in Chapter 4, that the start of the first rise, when expressed in Hz, is the least speaker-specific of the pivot points used in these utterances.

[5] F_0END is less important for speaker identification in the analysis with male data. At the end of phrases and breath-groups of male speakers, creak is a common phenomenon (Hirson and Duckworth, 1995). Perhaps the creaky final part of some of the male speakers' utterances led to F_0 determination problems and less reliable measurements.

We found that speaker specificity was high for all TI measures and for all of the CB parameters except SLOFIL, SYNRI1 and SYNFIL. For the parameter SLOFIL, for which the factor Speaker was associated with a low percentage of explained variance, we did not find substantially significant results on other factors either. None of the main effects turned out to be significant, and perhaps this parameter was determined to a large extent by random variation.

For the parameters SYNRI1 and SYNFIL, on the other hand, no large Speaker effects were found, while we did find effects for other factors (e.g. Sentence). For these two parameters we therefore conclude that a Speaker effect was not a priori unobtainable. SYNRI1, the time period between the onset of the first rise and the vowel onset, thus appears to be linguistically determined, which is more or less in agreement with the results reported by Caspers (1994), who found that the time interval between the syllable onset and the onset of an accent-lending rise is rather invariant and perceptually relevant. There are two counter-arguments against linguistic determination, however. First, a linguistically relevant feature should not vary too much over age, sentence and session. Second, it is not clear why a similar result was not found for the second rise.

5.2.4 Cross-validation: a prerequisite for application

The finding that a combination of parameters can be used to identify speakers indicates that the variance within the data of individual speakers is smaller than the variance within the entire speaker group. Large speaker specificity is a factor that should be taken into account in studies of prosody. In many studies on prosody (and other phonetic topics, for that matter) data of a limited number of subjects are conceived of as being indicative of a much larger population of subjects. The rather high speaker specificity found in this study should be taken as a warning that what appear to be universal characteristics of speech might in fact be idiosyncrasies.

Parameter values are not only dependent on the factor Speaker, however; they also vary over time. In Chapter 1 it was pointed out that the stability of speaker identification over time must be considered the first prerequisite for any application of parameters in real-life situations. Therefore, we performed cross-validation in this study; the parameter scores of one recording session were used in a discriminant analysis, the functions of which were used to classify speech material from another session. Thus we dealt with the third of the questions that were raised in section 5.1:

3) *How stable is prosodic speaker identification, i.e., how dependent is it on differences in speech style and in time? Are the speaker characterizations at time T1 equally valid at time T2?*

The percentage of speaker identification found is determined partly by the homogeneity of the input speech material. A diversification of the speech material such as including recordings from different sessions can be expected to lead to an increase in the variability in the data and consequently to a decrease in the speaker identification performance that can be attained. In this study data from two recording sessions were applied. This was done to create a more realistic setting, as it is obvious that good speaker identification is relatively easy if all recordings are made on one single occasion. To assess the influence of session-to-session variability, different approaches can be chosen. One is to find out whether speaker identification within data from one session is superior to identification by

means of the pooled data of both sessions. Another approach is cross-validation, i.e., the assignment of fragments from one of the sessions to the speakers, on the basis of discriminant functions that were derived from the other session.

To facilitate the discussion of the data, a summary of the relevant findings is presented in Table 5.2. The reported within-session and cross-validation scores are the average percentages of correct identification over the two within-sessions analyses and the two cross-validation LDA's: one in which fragments from the first session were assigned to the speakers by means of functions determined for the second, and one in which the procedure was reversed.

Table 5.2
Summary of LDA results concerning IS within sessions (before the comma) and cross-validation IS (after the comma), for the total material, as well as for female, male, read and spontaneous data separately. In the data columns are the IS's of LDA's with (1) TI parameters in 15-second fragments (2) TI parameters in sports sentences (3) CB parameters in sports sentences and (4) TI and CB parameters in sports sentences. In (3) and (4) the percentages of correct identification of LDA's with maximally 10 parameters are presented first. The IS's for LDA's with all available parameters are shown in parentheses.

	fragments	sports sentences (n= 300)		
	TI, 15 s (n= 1000)	TI, 2.5 s	CB	TI + CB
all material	72, 33	92, 42	86, 42 (86, 42)	93, 44 (99, 51)
females	76, 33	89, 44	90, 43 (90, 43)	99, 46 (98, 50)
males	71, 30	93, 38	68, 17 (68, 17)	98, 43 (95, 43)
read	97, 54			
spontaneous	89, 29			

The identification performance of LDA's within sessions was higher than that of LDA's in which material from both sessions was used. Both in analyses with the 15-second fragments as the experimental material and in analyses with the sports sentences material, the identification performance for a single session was always higher than for the total data set, for TI and CB parameters as well as for the combination of these parameter types.

To be of any use in practical applications, correct assignment to the speaker groups should be possible for data from new speech fragments. This situation was simulated by means of cross-validation. The results of cross-validation LDA's were disappointing in all analyses. For the 15-second fragments, the percentage of correct identification was 33. For the sentences we found cross-validation percentages of 42 %, 42 % and 44 % for the TI, the CB and the TI + CB analyses, respectively.

The cross-validation IS's that were obtained for the parameters studied do not seem to enable a level of speaker identification that allows application for practical purposes

(e.g. in forensic settings). To explain the poor cross-validation results the results of the analyses of variance that were reported in the Tables 3.3, 3.4, 3.5, 4.3 and 4.4 should be considered. Although the Speaker factor explained a large part of the variance for most parameters, we also found high ω^2 values for the interactions of Speaker × Session and of Speaker × Session × Paragraph/Utterance. Apparently the speaker values of many parameters are different in different recording sessions. To some extent this could be expected beforehand. A Speaker × Paragraph/Utterance interaction could be expected as well as speakers can be supposed to have their own speaking/reading styles. The fact, however, that these speaking/reading styles themselves also vary over recording sessions makes speaker identification extremely difficult. For the time-integrated parameters a solution can perhaps be found in integrating over larger periods of time, while for the contour-bound parameters we must probably resort to using many utterances with many pitch movements. Both these "solutions" are disadvantageous in practical applications.

It is important to note that many speaker identification studies do not attempt any cross-validation at all. Such an approach is often substituted by "split-half" studies, in which part of the material is assigned to the speakers on the basis of the rest. We consider it essential to integrate in identification studies the influence of recording at different moments, as such an approach offers the most realistic results.

5.2.5 Stability of speaker identification over speech styles

Each diversification of the speech material can be expected to lead to an increase in the variability in the data and consequently to a decrease in the speaker identification performance that can be attained. The opposite expectation holds for clear groupings in the data. Between men and women, for instance, large differences are expected on the basis of the vast literature describing these differences. If it is very easy to differentiate between men and women, an LDA including only speakers from one of these sex groups will probably be less effective than one in which both male and female speakers are included. In an analysis of the latter type speakers of opposite sex will probably not get mixed up very often, thus reducing the percentage of incorrect attribution.

Thus, the overall identification scores that were presented in Table 5.1 could probably be enhanced by measuring within more restricted data sets, and were expected to drop if speakers of equal sex and age were compared. We performed analyses over subsets of our material in order to find out whether analysing data within levels of extra-linguistic factors can indeed affect speaker identification, and if so, whether it raises or lowers it. In the previous section we discussed the favourable effects of analysing material within recording sessions. We will now consider the influence of Speech style.

In the 15-second fragments higher identification scores were obtained in LDA's within one speech style. Within the set of spontaneous speech fragments IS was 70 % and within read fragments it was 88 %, while in the overall analysis only 60 % of the speakers was identified correctly. As suggested earlier, the reason for the better identification performance within speech styles probably lies in the higher within-speaker variation in an analysis over both speech styles. In such an analysis, both the effects of Speech style and of the Speaker × Speech style interaction add to the within-speaker variation.

The cause of the clear superiority of speaker identification in read speech is that read speech contains less irrelevant variation. This reduced variability probably has a number of causes. To name a few: the read fragments were more homogeneous since the speakers all read the same sentences; speaking behaviour in read fragments may be more

uniform due to the fact that speakers tend to adhere to a culturally defined reading style ideal; and read and spontaneous speech differ in the "level of preparedness" (Blaauw, 1995), thus reducing the irrelevant variation caused by disfluencies. Summarizing, increased control over the speakers' utterances appears to be favourable for speaker identification.

5.2.6 Influence of sex and age on speaker identification

The last question raised in section 5.1 was related to the influence of the speakers' sex and age on speaker identification:

4) *To what extent does analysing data within sex and age groups affect speaker identification? We want to factor out, as much as possible, the influence of these factors.*

Analysing data within levels of these extra-linguistic factors affects speaker identification for different reasons. In the previous section we showed that applying only measures obtained in one of the sessions or in one of the speech styles raises the identification performance. This is not surprising, as we found in Chapter 3 (see Table 3.3) that the Speaker × Session and the Speaker × Speech style interactions were significant (and substantial) for all parameters. Apparently, speaker differences are not stable from style to style and from session to session.

Because of the (generally) small influence of the Age factor in the analyses of variance reported in Chapter 3 and 4, no separate LDA's were performed within the levels of that factor. We did performed separate analyses for the male and female subsets of the material, however. We thus tried to find out whether analysing data within sex groups affects speaker identification, and if so, in what way. For the 15-second fragments we found that performing LDA's within one of the sex groups did not affect results very much. The identification score within female speakers was 65 % and within males it was 63 %, where a combined analysis resulted in an IS of 60 %. Higher results can be obtained in separate analyses of the sex groups because the discriminant functions can be optimized to the speaker differences within the sex groups. Apparently, the speaker differences on the parameters are distributed differently within the sex groups.

F_0MEAN remained the most important speaker-identifying parameter, even though the inter-speaker differences for this measure were much smaller within sex groups. However, F_0MEAN was less important than in the analyses in which data from both sex groups were included; the first discriminant function (related most to F_0MEAN) was associated with a lower percentage of explained variance. The decrease in speaker-identifying power of F_0MEAN was compensated for, apparently, by some of the parameters that played a less prominent part in the overall analysis. In the LDA of the female speakers PZR was related to the prominent second discriminant function, and in the analysis of the male speakers the second, third and fourth function explained a large part of the variance. The parameters that were most strongly related to these functions were PPQ, CVP and PZR, respectively.

For the TI analyses of the sports sentences in which only data from one of the sex groups were used, we again found somewhat higher identification scores in the one-sex analyses. Among the CB parameters, the identification performance for female speakers was found to be markedly higher than in the overall analysis, and in the TI + CB analyses

the percentages of correct identification were about equal.

Summarizing, the differences between analyses within sex groups and with the two groups combined were rather small. The smaller importance of mean F_0 in separate analyses of the sex groups was apparently compensated for by parameters that were more speaker-specific within just one of the sex groups.

In the analyses of variance the factor Speaker was nested under Sex and Age. It seems to be a sensible approach to perform discriminant analysis in a similar way. If the speakers within each combination of Sex and Age are more alike than in random combinations of speakers, a lower speaker identification performance is expected within Sex-Age groups. In Appendix H such analyses are presented for the sports sentences data. In the LDA's with TI and CB parameters combined, the IS was about the same for the analysis of the total data group as for within-Sex-Age group analyses. In the LDA's with only TI parameters the performance in within-Sex-Age groups analyses was somewhat *better* than in the overall analysis (which runs contrary to the argumentation above), and in analyses of CB parameters only, the overall analysis resulted in superior results.

The implication of these findings for future applications appears to be that in situations in which information on the speaker's Sex-Age group is already available, such as in speaker verification, TI parameters might be more useful than CB measures. It is important to keep in mind, however, that the databases that were applied in the LDA's within Sex-Age groups were quite small.

From the separate analyses for Sex and Age groups and for Sex-Age groups a picture emerges of large sex differences and relatively small age differences. The interactive effect of Sex and Age was marginal.

5.3 Identification of other extra-linguistic factors

5.3.1 Introduction

In the previous sections it was established that many of the TI and CB parameters are speaker-specific. Furthermore, the influence of other extra-linguistic factors (speech style, sex and session) on speaker identification was manifested in the LDA's in which the material was partitioned according to the levels of these factors. An LDA with only data from read speech fragments, for instance, was much more successful than an LDA with the total data set as its input.

In the same way that the speaker specificity in the data was studied by means of LDA's, it is also possible to establish whether the parameters can be used to determine the levels of the other extra-linguistic factors. In the present section, the results of discriminant analyses in which the Speech styles (§ 5.3.2), Sex groups (§ 5.3.3), Age groups (§ 5.3.4), Sessions (§ 5.3.5), and Paragraphs and Sentences (§ 5.3.6) were used as discriminant groups are discussed.

5.3.2 Speech style

With regard to the factor speech style we tried to answer questions that were concerned with the extent of speech style identification, the importance of the parameters for speech style identification, the stability of the identification, and the importance of the factor Sex for the identification.

To what extent is speech style identification on the basis of prosodic parameters possible?

The percentage of correct identification of the speakers increased considerably if fragments from only one of the two speech styles, i.e., read and spontaneous speech, were used. Indeed, the differences between the speech styles were so large, that, in section 3.5, we had no problem characterizing the speech styles in discriminant analyses with the speech styles as discriminant groups. The assignment of the fragments to the groups was successful in many cases, 87 % above chance.

Which parameters are most important for speech style identification?

In the discriminant analysis in which the 15-second data were used for speech style identification, none of the parameters stood out as particularly style-specific. The fact is that for most parameters speech style differences existed. In analyses of variance substantially significant differences between the speech styles were found in all of the variables except F_0MEAN, PPQ and PAUSE. CVP, RATE and VOI were higher in read speech, whereas APQ, PZR, AZR and CVA had higher values in spontaneous speech.

The fact that we did not find a speech style difference for PAUSE is somewhat surprising. Assuming that in spontaneous speech speakers need time to formulate their utterances, one would expect to find more pausing in spontaneous speech. There are indications in the literature on pausing behaviour that the main difference between the two speech styles is in the *number* of pauses; according to Howell and Kadi-Hanafi (1991) a larger number of pauses is found in spontaneous speech. However, Barik (1977), testing different kinds of speech, did find considerable differences within and across speech styles. Barik explains these differences by referring to the suggestion of Goldman-Eisler (1961b) that greater pausing time is related to speech in which speakers are involved in generalizing or abstracting of the meaning of events. Less pausing would be involved in the factual description of events. The relatively low percentage of pausing that we found in the spontaneous speech material (low in the sense that we expected more pausing than in the reading condition) might be related to the fact that we interviewed our speakers on more or less familiar and factual information, such as their preferences for foods, holidays, etc. Our results do not support Goldman-Eisler's claim that pausing varies "with the different degrees of spontaneity" (Goldman-Eisler, 1968: 58).

In studies by Goldman-Eisler (1961a and 1961b) and Grosjean and Deschamps (1980) it was found that the amount of pausing was not related to articulation rate. Indeed, while we found no speech style differences for PAUSE, we did find such differences for RATE; it was higher for read speech. This finding comes as no surprise, since in spontaneous speech speakers need time to think about what they are going to say. This difference is an indication that the equal percentage of silence (PAUSE) in the speech styles reported above might result from more filled pauses. Another possible explanation lies in the lengthening of speech segments in utterance-final position. Perhaps our spontaneous speech material contained more short utterances (as was found in Haselager et al., 1991) and more short sentences result in more final lengthenings.

In the literature, read speech was often found to have a higher fundamental frequency than spontaneous speech (e.g. Koopmans-van Beinum, 1991; Ramig and Ringel, 1983; Hollien and Jackson, 1973). This difference was replicated in the present study.

However, F_0MEAN was not the only or even the most important parameter for Speech Style identification.

For CVP higher values were found in read speech. Assuming that CVP is related to the speakers' vividness of intonation (see section 2.2.6), this shows that the speakers spoke more vividly in the read condition. The opposite result was found for CVA. The variability of the maximum amplitude in the fundamental periods was larger in the spontaneous speech. This agrees well with our intuition, as one would expect that spontaneous speech is characterized by periods of mumbling as well as by periods of forceful argumentation. Read speech appears to lack such clear differences in speaking effort, and seems to be primarily directed at communicating a message.

All perturbation scores were higher for spontaneous than for read speech. In Chapter 2 we stated that it is not quite clear how differences in perturbation measures for normal speakers should be interpreted (Eskenazi et al., 1990). For the zero-crossing rates we speculate that higher values will be found in contours that contain many "level" line segments, while in parts of the contour with a steadily falling F_0 the zero-crossings rate will be low. Thus, the lower zero-crossing rates for pitch in read speech could result from a steeper declination line in read speech. Evidence for a steeper declination in read speech can be found in e.g. Umeda (1982) and Lieberman et al. (1985)[6].

How stable are the prosodic speech style characterizations, i.e., how dependent are they on differences in time of recording?

Cross-validation results were barely lower than the results of LDA's with material from one of the sessions, or with both sessions combined, which proves that the speech style differences are not only large, but also stable over sessions.

To what extent does analysing data within levels of the factor Sex affect speech style identification?

In analyses of the male speakers' data, the performance was better than when data from both sex groups were combined. Speech style identification scores within the female data were comparable with the overall analyses.

5.3.3 Sex

With respect to the factor Sex, the questions we tried to answer are related to the extent of sex group identification, the importance of the parameters for the identification of these groups (especially those other than the parameters that are directly related to the general F_0 level), and the stability of the identification.

To what extent is sex identification on the basis of prosodic parameters possible?

In section 3.6 we applied the ten TI parameters to the identification of the sex of the speakers in the 15-second fragments. Many studies have focused on sex differences in

---

[6] Note, however, that the status of declination as a linguistic universal is a disputed issue (Lieberman et al., 1985).

voices and a well-known difference, of course, is the higher mean F_0 found for women. Indeed, in the 15-second fragments, F_0MEAN was by far the most important sex-identifying parameter. An LDA with all TI parameters resulted in an almost perfect identification of the sexes (97 %). A separate analysis with F_0MEAN as the only predictor resulted in a percentage of correct sex group identification that was even higher than the percentage obtained in an LDA with all parameters except F_0MEAN: 97 % vs. 80 %, respectively.

In Chapter 4 we found that sex identification in the sports sentences LDA's led to essentially perfect results as well, both for TI and CB parameters separately and with both parameter types combined (97 %, 98 % and 99 %, respectively). In analyses with one single parameter, high percentages of correct sex identification were obtained both for F_0MEAN (97 %) and for F_0END (95 %). Analyses with all parameters except the two F_0-related parameters were successful for the TI parameters (77 % correct). In an LDA with all CB parameters except F_0END, however, the identification score was only 39 %.

Which parameters are most important for sex identification, and how important are the parameters not related to the overall F_0 level in this respect?

As was indicated above, the parameters F_0MEAN and, to a lower extent, F_0END were the most important parameters for sex identification. However, in LDA's from which these measures were excluded we still found high percentages of correct identification.

For the 15-second fragments this is not surprising because for five of the other variables substantially significant differences between the sex groups were found: CVP, PZR and AZR had higher values for female speakers, while for APQ and PAUSE higher values were found for males. CVP and PAUSE were not substantially significant in the sports sentences, while for one parameter a significant sex difference was found in the sports sentences where it had not been found in the fragments: the PPQ values were lower for males.

The sex difference found for CVP in the 15-second fragments runs contrary to the data that are reported in studies on pitch *range*[7]. Henton (1989), for instance, reports equal ranges for male and female speakers. She reviewed many studies and claimed that ranges should be measured in ST, and that after a conversion to ST for most studies a larger range is found for male speakers. Other researchers have proposed different measures, such as the standard deviation in Barks, ERB's, etc. We consider it an important feature for each measure of F_0 dispersion that it is independent of the general F_0 level, as it has often been found that the general F_0 level is strongly related to the dispersion measures. An important property of our CVP measure is that within sex groups, we did not find a correlation between CVP and F_0MEAN. This indicates that there appears to be no influence of F_0MEAN on CVP (the amount of variation in the women's F_0 values was larger than that in the men's in the *spontaneous* speech condition only).

In the 15-second fragments a larger amount of pausing time was found for male speakers than for females. Since we found no compensatory sex effect in the articulation rate, we must assume that our male subjects realized speech at a slower *speaking rate* (i.e.,

---

[7] Although CVP is, strictly speaking, not a measure of pitch range, van Bezooijen (1984) reported a high correlation between the mid-80 % range (expressed in ST) and the coefficient of variation of F_0: the correlation was 95. Horii (1975) found that the mid-90 % range was highly correlated with the standard deviation (r = .982)

the number of syllables per second *including pauses*). The PAUSE difference was found both in read and spontaneous fragments, but was absent in the sports sentences. It is important to remember that the level of pausing found in the sports sentences is of an entirely different order. In the present study all speaking time in which no vocalisation occurs is marked pausing time, but in many other studies (e.g. Miller et al., 1984) a minimum duration for a proper pause is defined. The larger amount of pausing for men in the 15-second fragments should therefore be taken more seriously than the negative result in the sentences; in the fragments, all speakers, male and female, realized "real" pauses, to breathe or to reach some communicative effect, while in the sports sentences hardly any pauses occurred at all.

In the sports sentences, women attained higher values for both pitch perturbation measures, PPQ and PZR, while in the 15-second fragments a sex difference was only found for PZR. As for PPQ, Schoentgen (1989) reported higher values for (healthy) female speakers, both in sustained vowels and in isolated sentences. He attributed these higher values to the relative character of PPQ, i.e., to the fact that the absolute jitter is divided by the mean period duration. According to Schoentgen, the absolute jitter decreases more slowly than the mean fundamental period. This could explain the fact that the male-female difference found for PPQ in the 15-second fragments was not significant, while it was significant in the sports sentences: the mean period duration for the female speakers had decreased considerably; mean F_0 for the fragments was 189 Hz and for the sentences it was 211 Hz. Horii (1979) reported a non-linear relationship between absolute perturbation and mean period duration as well. He found a decreasing perturbation above 210 Hz.

In section 5.3.2 we hypothesized an underlying mechanism determining the value of the zero-crossing rates; in contours with a clear declination line the number of zero-crossings will be lower and thus PZR will be lower. However, this does not explain the high PZR value found for the female speakers, since their declination, when measured in Hz., was larger than the men's: 46.6 vs. 33.8 Hz.

For the amplitude perturbation measure AZR higher values were again found for women, while for men higher APQ values were found. Interpretation of the amplitude data is difficult, also because the amplitude and pitch data are perhaps not independent: the much shorter pitch periods of the female speakers bring about smaller time distances between the amplitude measurements of women. The higher time resolution for female speakers perhaps partly explains the sex difference that were found for the amplitude perturbation measures.

In the sports sentence material, for five of the CB parameters (F_0END, LOWRI2, DURRI1, DURFIL, and SLODEC) substantially significant sex differences were found. Before discussing these differences we first show the pitch contours for men and women in Figure 5.1. The contours are depicted on the basis of the mean values of their CB parameters.

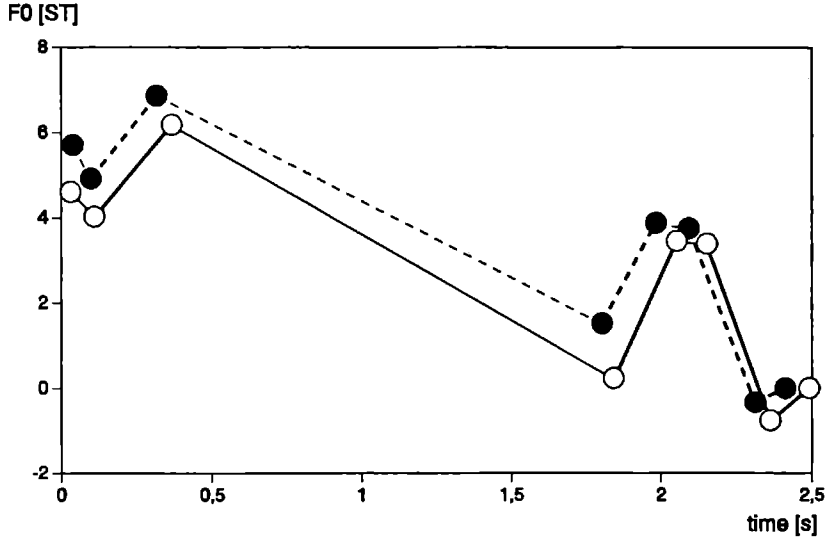

Figure 5.1 F_0 contours of men (●) and women (○), in ST relative to the mean F_0END of the two sex groups.

First of all, it is important to stress that, apart from the F_0END difference, the sex group differences were rather small (see Table 4.4). A higher value of SLODEC was found for the male speakers (2.41 ST/s vs. 1.88 ST/s for the female speakers) and the male speakers' F_0 contour appears to be somewhat higher overall, relative to the final F_0 value. It is hard to give an explanation for this (small) difference, and we can only speculate that it might be caused by an F_0 level-related artefact in the voicing-determination algorithm that, in turn, influenced the measurements of the final F_0 values. Such an artefact would also explain the DURFIL difference between the sexes.

We found two significant sex differences that were related to the F_0 rises in the utterances: male speakers had lower DURRI1 and higher LOWRI2 values. These findings could not be replicated in the other rise, however; the sex differences in DURRI2 and LOWRI1 were not significant.

How stable are the prosodic sex characterizations, i.e., how dependent are they on differences in speech style and time of recording?

Within sessions as well as in cross-validation analyses, a nearly perfect sex group identification was achieved. The sex differences were very stable over sessions. The identification scores in all analyses were so high, that further improvement by applying only fragments from one of the speech styles was not possible. From LDA's without F_0MEAN it is clear that within the read fragments more powerful sex identification was found than in the the spontaneous fragments. Possibly this is the result of the larger amount of variability found in spontaneous speech, as was hypothesized earlier in the

discussion of speaker identification differences between the speech styles.

5.3.4 Age

We tried to assess the degree of success of age group identification, the importance of the different parameters for the identification process, the stability of the identification, and the importance of the factor Sex for the identification.

To what extent is age group identification on the basis of prosodic parameters possible?

For this study, speakers from five different age groups were selected; 18-25, 26-35, 36-45, 46-55, and 56-65 years old. In the period of life that we studied, i.e., after adolescence and before old age, large differences between the groups were not expected. In Chapter 3, however, it proved to be possible to characterize the groups; a level of correct identification of 28 % above chance level was reached.

The percentages of correct identification found for the sports sentences were about equal to the percentages found for the 15-second fragments. The CB parameters could distinguish between the age groups only little better than the TI parameters could, and the combination of the two parameter types resulted in somewhat better results.

Which parameters are most important for age group identification?

In the 15-second fragment material, the parameter that was related to the most important discriminant function was RATE. The youngest group of speakers articulated faster than the other groups did. None of the TI parameters in the sports sentences was related strongly to the first discriminant function in the LDA. APQ was related to the second-most important discriminant function and PAUSE to the third, but in an analysis of variance no significant age group effects were found for any of the TI parameters.

The two parameters that were related to the most important discriminant functions in the age group identification on the basis of CB parameters were the same two measures for which Age group was substantially significant: SEMFAL and SYNRI1. The ST fall at the end of the utterance was larger and the first accent-lending rise started earlier for the younger speakers. These phenomena might be related to the changing anatomy of the aging vocal apparatus, or to language change. Note, however, that the data of the second rise or the fall did not show a straightforward trend.

How stable are the prosodic age group characterizations, i.e., how dependent are they on differences in speech style and in time?

In LDA's with data from only one session, the percentage of correct identification was a few percentage points higher than in an LDA with all material. In cross-validation analyses the identification score was a little lower. Analysing data within one of the speech styles raised the performance a few points as well. There was no difference in age group identification between the speech styles.

To what extent does analysing data within levels of the extra-linguistic factor Sex affect age group identification?

In Chapter 3 a large increase in identification was obtained by analysing only data within sex groups. In the sports sentences material the largest increase was found for the female speakers' data. This suggests that male and female aging patterns are dissimilar.

In a separate analysis of the male speakers' 15-second fragments, we found no TI parameters that were clearly related to the discriminant functions. For the female speakers, however, we found that F_0MEAN was related to the first discriminant function. The youngest group of women had higher mean F_0 values than the older groups[8]. In studies on the relation between mean F_0 and age for female adolescents it was found that mean F_0 decreases during adolescence. The age of the speakers in such studies often ranged from 14 to 18 years (see de Bruin, 1993) and in most of the studies girls' mean F_0 was still decreasing between the two oldest age groups. In the data on Dutch girls' voices by de Bruin (1993), for instance, mean F_0 decreased from 209 Hz for 15-year olds to 188 Hz for 17-year olds. In studies of adults, broad age categories have often been used. Stoicheff (1981a), for instance, used a category of 20-29 years. Although most puberty-related physical changes should have been completed by the age of 18, the decrease of F_0 for women might proceed even after that age. We suggest further research into the development of F_0 for young women. As the sex/age groups consisted of only five speakers, only tentative conclusions are allowed for the present study. The higher F_0MEAN for the youngest female group which was found in the 15-second fragments, was again found in the sports sentences, but the parameter F_0MEAN was not related to the most important discriminant function any more; the parameter related to the most important function in the sentences was RATE.

5.3.5 Other extra-linguistic factors

Differences between the two recording sessions are not of primary interest in this study, nor are the differences between the paragraphs and the sentences. Different sessions and different speech material were included in the design to create a more realistic setting: good speaker identification is more easily accomplished if there is no variation in the linguistic material uttered by the speakers and if all recordings are made on one single occasion. The main importance of these parameters lies in their interaction with the Speaker factor. A high Session × Speaker interaction, for instance, renders the parameter concerned useless for speaker identification. As significant main effects are much less important than the interaction effects, we only briefly consider the percentages of correct identification reached for the main effects of Session, Paragraph and Sentence..

[8] The youngest group of male speakers was found to have a high mean F_0 as well, but the difference with the other age groups was less pronounced.

To what extent is it possible to identify the recording sessions on the basis of prosodic parameters and which parameters are most important for session identification?

For the 15-second fragments it was possible to assign the cases to the sessions above chance level, but only to a marginal degree: IS was 15 %. Session identification within the sports sentences material was possible to some degree as well. None of the parameters was strongly related to the discriminant functions in these analyses.

To what extent is it possible to identify the paragraphs of the reading text on the basis of prosodic parameters and which parameters are most important for paragraph identification?

The read speech fragments used in Chapter 3 were obtained from separate paragraphs of a short newspaper-like story. Paragraph effects on the parameters are somewhat undesirable; if the parameter values differed from one paragraph to the other, this would imply that they are dependent on the (lexical) content of the fragment being read and, consequently, that an integration time of 15 seconds is not long enough to get rid of this dependence. It was possible to assign the cases to the paragraphs; the identification score was 52 %. The variables related to the discriminant functions were CVA, APQ, and VOI. It is not clear whether the significant differences result from lexical or prosodic differences in the content of the paragraphs.

To what extent is it possible to identify each of the three sports sentences on the basis of prosodic parameters and which parameters are most important for speaker identification?

In the sports sentences material it was to some extent possible to identify which of the sentences was read. Especially among the CB parameters this was not surprising, as the pitch movements depend at least partly on the segmental context in which they are realized. The intrinsic pitch of vowels, for instance, influences the F_0 level at the beginning and end of pitch movements.

The three sports sentences differed in the lexical content on a few positions only. Therefore, for the TI parameters, we did not expect large differences between the sentences. Compared to the paragraphs in the read 15-second fragments, the amount of lexical and prosodic variation was clearly lower within the sports sentences. Indeed, the identification performance was poorer for the sentences: 34 % vs. 52 % for the paragraphs. The fact that the sentence differences were not very large supports our decision to combine measurements from these sentences in LDA's.

5.4 Limitations and suggestions for further research

5.4.1 Speaker identification

The aim of this study was to find out whether prosodic parameters can be applied to speaker identification and if so, to what extent. The first part of this question can be answered affirmatively. Prosodic parameters are indeed speaker-specific. By means of ten

time-integrated parameters, obtained in a data set of 1000 fragments of 15 seconds, a speaker identification performance of 60 % was attained. The innovative part of our approach was that we combined time-integrated parameters with contour-bound ones, i.e., measurements at discrete pivot points in the pitch contour. In a set of 300 utterances of a specific type ("sports sentences") we found a speaker identification score of 81 % for the TI parameters, 86 % for the CB parameters, and 97 % for the combination of the two.

The CB measures, up to now only scarcely used in speaker-identification studies, were found to be quite speaker-specific. It remains to be seen, however, how useful these measures are for the implementation in real-life applications, such as in solving the forensic problem of identifying an unknown speaker. Most of the CB measures were dependent on the exact pitch movement on which they were measured and even two pitch movements of the same type (e.g. rise "1") had considerably different attributes. CB parameters must be obtained from sentences that are equal both in lexical and prosodic form, which means facing a major practical problem, for it turned out to be difficult to extract utterances with strictly specified prosodic patterns.

TI measures require less restricted material, but the less restricted, the longer the integration time needed. In this study we suggested that for TI measures too, much can be gained by using prosodically controlled utterances.

The identification scores reached in the discriminant analyses were not very high, and it would seem that prosodic parameters can only play a supportive role in speaker identification. However, a higher percentage can perhaps be reached by means of different analysis techniques. The primary aim of linear discriminant analysis is to determine the dimensionality of the subspace in which groups are optimally separated. Hyperplanes are placed in this multidimensional space in such a way that they optimally divide the groups studied. An important aspect of LDA is that these hyperplanes have a linear relationship with the underlying parameters. If it were possible to use a curved plane, better group separation might be reached. In the present study we chose to use LDA's because the importance of the parameters for speaker identification can be deduced from their relatedness with the dimensions of the space.

The prosodic parameters studied do not seem to enable a level of speaker identification that allows application for practical purposes. This seems to be even more true if we take the low cross-validation IS's into account. The cross-validation identification results were not entirely negative, however, as long as LDA's were performed within speech styles and within sex groups. In the 15-second fragments the identification score in the cross-validation analyses was 54 % within the read speech, which is more promising for application, as it is 21 percentage points higher than in the mixed styles condition. In forensic speaker identification and electronic access systems the 54 % correct cross-validation does not seem to be enough. It is important to note that many speaker identification studies do not attempt any cross-validation. Such an approach is often substituted by "split-half" studies, in which one randomly selected half of the speech material is used as the reference material, while the other half is used as experimental data. The experimental data are assigned to the groups on the basis of analyses of the reference set. In the present study the reference data set and the experimental data set were not selected at random. Instead, we attributed data from one recording session to the reference data set and data from another recording to the experimental set. We consider it essential to integrate the influence of different recording sessions in identification studies. Such an approach offers the most realistic results, as in many practical applications a reference data

set is acquired at time T1, while speaker identification is to be performed with data obtained at time T2.

Two important aspects of speaker identification that were not considered in this study are disguise and mimicry. Lower IS's would probably be found if non-cooperative speakers were involved. The parameters that are particularly speaker-specific, such as mean F_0 and RATE, are plausible candidates for deliberate distortion. Therefore, application of such measures should be restricted to cooperative speaker identification only.

For future research, we recommend the incorporation of some of the prosodic parameters in speaker identification studies with cooperative speakers, as these parameters are generally rather insensitive to data transport and are to some extent speaker-specific. Furthermore, with the use of more powerful identification tools better performance might be possible[9].

5.4.2 Importance of the parameters

In this study we repeatedly reported the relatedness of prosodic parameters to discriminant functions as a measure of the importance of the parameters for the identification of extralinguistic factors, such as speaker. In most cases one parameter clearly stood out as the one most related to the (Varimax-rotated) discriminant function. Sometimes, however, a number of parameters were related to a function, in which case none of them reached the .50 criterion. Although all these parameters were important discriminators, none of them were reported. In Table 3.10, for instance, the results of speech style identification by means of LDA were shown. Earlier, we had found that the main effect of Speech style was significant and of substantial importance for all predictor variables except F_0MEAN, PPQ and PAUSE. Because all of the parameters for which substantially significant Speech style differences were found contributed to the discriminant function, none of them stood out as a clear speech style specific measure.

Summarizing, the relations between parameters and discriminant functions are not always as straightforward as might appear in this study. In analyses where many parameters load on a limited number of functions, such as the analyses in which the speech styles were to be discerned, it seems as if none of the parameters was clearly related to the function, while in fact (almost) all of them were involved in speech style identification. The best way to establish the real importance of a parameter is to perform an LDA from which it has been removed, as we did for F_0MEAN and F_0END.

5.4.3 Speaker and other extra-linguistic factors

In the previous chapters we argued that a diversification of the input speech material leads to an increase in the variability of the speaker scores and consequently to a decrease in the level of speaker identification that can be attained. Indeed, LDA's with input data from one of the speech styles or from one of the sessions resulted in higher percentages of correct identification than an overall analysis did. Especially the analyses with the data

---

[9] We did assess the speaker identification success of another analysis technique, a self-organizing neural network (Kohonen, 1989). We applied Kohonen's self-organizing map program, SOM_PAK (Kohonen, 1990) to the TI measurements that were obtained in the 15-second fragments. The highest percentage of correct speaker identification was obtained with a codebook of 350 vectors and a training length of 15 steps. Speaker identification turned out to be poorer than with LDA; we found 46 % correct identification for SOM_PAK and 60 % for the LDA.

from read fragments resulted in a high speaker identification performance, probably because reading is a more restricted speech mode. In cooperative speaker identification applications it might be useful to use verbal behaviour tasks that restrict the speakers in their freedom to mould their own utterances.

We found that analyses within only one of the sessions were generally more successful than analyses with data from both sessions. Although the differences between the replications themselves were marginal, the interaction of Speaker and Session was important and, as was expected, the speaker identification performance was hampered by including data from different sessions in the analyses. Cross-validation studies resulted in poor speaker identification, as was discussed in the previous section.

For one of the extra-linguistic factors, Sex, there was reason to expect the reverse pattern: perhaps the sex differences would be so large that the number of alternatives in speaker identification was lower than it seemed, since the speaker's sex is obvious anyhow. It was, however, easier to identify speakers within sex groups. LDA's within age groups were not performed as we found in the analyses of variance that the age differences were rather small.

Age is clearly a speaker attribute that is much more difficult to identify than sex, especially since the effects of chronological and biological age seem to be confounded. Ramig and Ringel (1983) found that it is physiological aging, rather than chronological aging, that induces voice changes. They found that subjects in good physical condition produced vowels of maximum duration with significantly less F_0 and amplitude perturbation than subjects of similar chronological ages who were in poor physical condition. Supportive results were obtained by Braun and Rietveld (1995) who found that it was easier to estimate the age of smokers (who can be supposed to be in a non-optimal physical condition) than of non-smokers.

5.4.4 Further research

For this study, which is of an exploratory nature, a large amount of data have been collected. With respect to speaker identification on the basis of prosodic parameters, we showed that both TI and CB parameters can be quite useful.

The results found for the CB parameters should stimulate further research into the production aspects of pitch movements. The rather large differences found in the contributions to speaker identification of comparable parameters measured in different pitch movements raise questions about the correlates of these perceptually-defined movements in the production domain. 't Hart (1976) reviewed some literature on the sensitivity of listeners to differences in size, position and slope of pitch movements and concluded that this sensitivity was low. He speculated that the consequence of the poor perceptibility was that "the limited set of perceptually relevant pitch movements corresponds to the set of mutually distinguishable movements" ('t Hart, 1976: 18). However, even though the perceptual talents of listeners might be limited, still some relation must be found between the production and perception of pitch movements. More conclusive data are necessary for a closer description of this relation. How much, for instance, do different rises of type "1" really have in common? Are differences the result of different positions in the utterance (sentence-initial or sentence-final) or of differences in the configuration that a pitch movement belongs to (pointed hat vs. flat hat pattern)? Our production data do not suggest a very clear demarcation of the movements of rise "1" in the production realm, since the resemblance of different movements of the same type was not very strong. Caspers (1994)

also found limited similarity within the realizations of pitch movements of the same type.

't Hart et al. (1990) underline the fact that their pitch movement types are only perceptually relevant. Therefore, the large production differences within pitch movements of equal type do not appear to be an argument against the GDI model (the difficulty of learning to transcribe pitch movements seems to be a better argument against GDI). Perhaps the shape of the pitch movements is not too important after all, and the main feature of the intonation in Dutch is that it rises and falls, and this to a certain degree. This less explicit description of pitch movements better suits the alternative description of intonation of the "targets" approach (e.g. Pierrehumbert, 1980), in which intonation is described in terms of high and low targets, where the exact F_0 height of the H's and L's is either predictable by rule or regulated by a different system of "prominence".

Since GDI pitch movements do not seem to be clearly defined in the production domain, the next step appears to be the study of the perceptual basis of the GDI pitch movements. Synthetic speech could be used to test whether the major underlying dimensions only allow pitch movements within certain regions of these dimensions, or whether the differences in e.g. timing correspond to emphatic differences signalled by the speaker. A study on the meaning of the melodic elements of Dutch is presently being carried out by Caspers (1996).

This study aimed at finding speaker specificity in production data. Of course the identification of voices by listeners and the relation of the importance of our prosodic parameters with that of their perceptual counterparts are fit subjects for study as well. Some work has been done in this respect, for instance by Abberton and Fourcin (1978), who found that mean F_0 and F_0 contour provide important speaker-identifying information for an age-, sex- and accent-matched group, even in the absence of all supraglottal features. This is of course an interesting finding from our point of view. It might well be that contour-bound parameters, like those studied here, help listeners identify speakers. Our data analyses show on which of the prosodic parameters most speaker specificity is found. It could be expected that the perceptual counterpart of such parameters are also among the features that are used most by listeners. The importance of F_0 in both perception and production studies has been well-described, but the importance of other measures applied in this study, and the possible correspondence between discriminant functions obtained from both types of study could be considered.

As for speaker identification, the perceptual counterpart of speech style, sex and age identification can be related to the findings in the current study as well, although for sex it is already well-known that F_0 related measures are important both for identifying the sex of a speaker from the production data as for determining it impressionistically.

In discriminant functions the prosodic parameters are combined in such a way that they optimally separate the speakers. If such a combination were made by the listener, it could be found that these functions are related to well-known perceptual voice categories, such as the aesthetic appreciation of voices. Voices can be evaluated as being beautiful, melodic, etc. or in terms of underlying emotions (cf. van Bezooijen, 1984). Perceptual voice categories can be related to the prosodic parameters themselves, but also to the discriminant functions as established in this study, by looking for correlations between scores of aesthetic or emotional appreciation and the scores on the discriminant functions.

In this book we showed that prosodic parameters are speaker-specific and can also be applied to the characterisation of other extra-linguistic parameters such as speech style, sex, age group, session and fragment/sentence.

Both time-integrated and contour-bound parameters were found to be useful (but probably not sufficient) speaker-identifying parameters. The problems surrounding time-integrated parameters are well-known: the instability of these measures can only be resolve by using long integration times or, as we suggested in this book, highly controlled speech material (both in a prosodic and a lexical sense). Even then, the stability of the measures can be troublesome, as was shown by Barry et al. (1991), who found that even between speech samples of no less than two minutes, considerable within-speaker variation could be found. It was also exemplified by the rather poor cross-validation results obtained in this study.

The future usefulness of the contour-bound measures hinges on a better understanding of what factors influence the exact positioning of the pivot points in a pitch contour. We hope that this book will stimulate research aimed at obtaining such knowledge.

References

Abberton, E. & Fourcin, A.J. (1978) Intonation and speaker identification, *Language and Speech*, 21, 305-318.

Adriaens, L.M.H. (1991) Ein Modell deutscher Intonation [A model of German intonation]. Dissertation, University of Eindhoven.

Amerman, J.D. & Parnell, M.M. (1990) Clinical impressions of the speech of normal elderly adults, *British Journal of Disorders of Communication*, 25, 34-43.

Askenfelt, A.G. & Hammarberg, B. (1986) Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures, *Journal of Speech and Hearing Research*, 29, 50-64.

Atal, B.S. (1972) Automatic speaker recognition based on pitch contours, *Journal of the Acoustical Society of America*, 52, 1687-1697.

Atal, B.S. & Rabiner, L.R. (1976) A pattern recognition approach to voiced/unvoiced/silence classification with applications to speech recognition, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 24, 201-212.

Baken, R.J. (1987) *Clinical measurement of speech and voice*. Boston: College Hill Press; Little, Brown and Company.

Barik, H.C. (1977) Cross-linguistic study of temporal characteristics of different types of speech material, *Language and Speech*, 20, 116-126.

Barry, W.J., Goldsmith, M., Fourcin, A.J. & Fuller, H. (1991) Stability of voice frequency measures in speech, *Proceedings of the XIIth international congress of phonetic sciences, Aix-en-Provence 1991*, 2, 38-41.

Beaugendre, F., d'Alessandro, C., Lacheret-Dujour & Terken, J. (1992) A perceptual study of French intonation, *Proceedings of the international conference on spoken language processing (ICSLP 92), Banff, 1992*, 1, 739-742.

Bennett, S. (1983) A 3-year longitudinal study of school-aged children's fundamental frequencies, *Journal of Speech and Hearing Research, 26*, 137-142.

Bergem, D.R. van (1990) Pitch period estimation by filtering the fundamental frequency out of the speech waveform, *Proceedings of the Institute of Phonetic Sciences Amsterdam*, 14, 17-26.

Bezooijen, R.A.M.G. van (1984) *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht: Foris Publications.

Blaauw, E. (1991) Phonetic characteristics of spontaneous and read-aloud speech, *Proceedings of the ESCA workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, Barcelona, 12/1-5.

Blaauw, E. (1995) *On the perceptual classification of spontaneous and read speech.* OTS dissertation series, Utrecht: LEd, OTS.

Bolinger, D.L. (1951) Intonation — levels vs. configurations, *Word*, 7, 199-210.

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

Bonastre, J.F., Meloni, H. & Langlais, P. (1991) Analytical strategy for speaker identification, *Proceedings of EUROSPEECH '91*, Genova: ESCA/IIC, 427-430.

Bosch, L. ten (1995) On the representation of acoustic realizations of Dutch pitch movements, *Journal of the Acoustical Society of America* (submitted).

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

Boves, L., Have, B.L. ten & Vieregge, W.H. (1984) Automatic transcriptions of intonation in Dutch. In: *Intonation, accent and rhythm, studies in discourse phonology*, D. Gibbon & H. Richter (eds.), Berlin: Walter de Gruyter, 20-45.

Braun, A. (1992) Zur Bedeutung des Merkmals "Mittlere Sprechstimmlage" [On the meaning of the feature "mean speaking fundamental frequency"]. In: *Phonetik und Dialektologie*. Marburg: Gunter Narr Verlag, 1-26.

Braun, A. & Rietveld, T. (1995) The influence of smoking habits on perceived age, *Proceedings of the XIIIth international congress of phonetic sciences, Stockholm*, 2, 294-297.

Broeder, D. (1990) *Sesam handleiding* [Sesam manual]. Utrecht: Stichting Spraaktechnologie, SPIN/ASSP report no. 24.

Brown, R.S. (1982) What is speaker recognition? *Journal of the IPA*, 12, 13-24.

Bruin, L. de (1993) De ontwikkeling van de stem in de puberteit: een descriptieve en evaluatieve studie bij meisjes [The development of the voice during puberty: a descriptive and evaluative study for girls]. Master's thesis, University of Nijmegen.

Burton, D., Shore, J. & Buck, J. (1985) Isolated-word speech recognition using multisection vector quantization codebooks, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 33, 837-849.

Carlson, R., Granström, B. & Karlsson, I. (1991) Experiments with voice modelling in speech synthesis, *Speech Communication*, 10, 481-489.

Caspers, J. (1994) *Pitch movements under time pressure.* Holland Institute of Generative Linguistics, dissertation series no. 10, The Hague: Holland Academic Graphics.

Caspers, J. (1996) Testing the semantics of single-accent intonation patterns in Dutch, to appear in: *Proceedings workshop "Semantics on the HIL"*, A. Arregui, C. Cremers & J. Doetjes (eds.), Holland Institute of Generative Linguistics, publication series no. 4, The Hague: Holland Academic Graphics.

Caspers, J. & Heuven, V.J. van (1993) Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall, *Phonetica*, 50, 161-171.

Clark, H.H. (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research, *Journal of Verbal Learning and Verbal Behaviour*, 12, 335-359.

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

Cohen, A. & Hart, J. 't. (1965) Perceptual analysis of intonation patterns, *Proceedings of the V^th International congress on acoustics, Liège*, paper A16.

Collier, R. (1975) Physiological correlates of intonation patterns, *Journal of the Acoustical Society of America*, 58, 249-255.

Collier, R. (1991) Multi-language intonation synthesis: the case of Dutch, *Journal of Phonetics*, 19, 61-73.

Collier, R. & Hart, J. 't (1981) *Cursus Nederlandse Intonatie* [Dutch Intonation Course]. Leuven: Acco.

Collier, R. & Terken, J.M.B. (1987) Intonation by rule in text-to-speech applications, *Proceedings of the European conference on speech technology*, J. Laver & M.A. Jack (eds.), Edinburgh, 2, 165-168.

Cooper, W.E. & Paccia-Cooper, J. (1980) *Syntax and Speech*. Cambridge: Harvard University Press.

Cooper, W.E. & Sorenson, J.M. (1981) *Fundamental frequency in sentence production*. New York: Springer-Verlag.

Couper-Kuhlen, E. (1986) *An introduction to English prosody*. Tübingen: Max Niemeyer Verlag.

Cramer, H. (1946) *Mathematical Methods of Statistics*. Princeton (NJ): Princeton University Press.

Crochiere, R.E. & Rabiner, L.R. (1983) *Multirate digital signal processing*. Englewood Cliffs (NJ): Prentice-Hall.

Cruttenden, A. (1986) *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. (1969) *Prosodic systems and intonation in English*. London: Cambridge University Press.

Crystal, D. (1985) *A dictionary of linguistics and phonetics, 2nd ed.* Oxford: Basil Blackwell.

Crystal, D. & Davy, D. (1969) *Investigating English Style*. London: Longman.

Daly, N.A. & Zue, V.W. (1992) Statistical and linguistic analyses of F_0 in read and spontaneous speech, *Proceedings international conference on spoken language processing*, Banff, Canada, 1, 759-762.

Das, S.K. & Mohn, W.S. (1971) A scheme for speech processing in automatic speaker verification, *IEEE Transactions of Audio and Electroacoustics*, 19, 32-34.

Davis, S.B. (1976) Computer evaluation of laryngeal pathology based on inverse filtering of speech, *SCRL Monograph 13*, Santa Barbara: Speech Communications Research Laboratories.

De Pinto, O. & Hollien, H. (1982) Speaking fundamental frequency characteristics of Australian women: Then & now, *Journal of Phonetics*, 10, 367-375.

Deal, R.E. & Emanuel, F.W. (1978) Some waveform and spectral features of vowel roughness, *Journal of Speech and Hearing Research*, 21, 250-264.

Doddington, G.R. (1971) A method for speaker verification, *Journal of the Acoustical Society of America*, 49, 139.

Doddington, G.R. (1985) Speaker recognition – identifying people from their voices, *Proceedings of the IEEE*, 73-11, 1651-1664.

Doherty, E.T. (1976) An evaluation of selected acoustic parameters for use in speaker identification, *Journal of Phonetics*, 4, 321-326.

Doherty, E.T. & Hollien, H. (1978) Multiple-factor speaker identification of normal and distorted speech, *Journal of Phonetics*, 6, 1-8.

Dologlou, I. & Carayannis, G. (1989) Pitch detection based on zero-phase filtering, *Speech Communication*, 8, 309-318.

Dommelen, W.A. van (1987) The contribution of speech rhythm and pitch to speaker recognition, *Language and Speech*, 30, 325-338.

Dommelen, W.A. van (1990) Acoustic parameters in human speaker recognition, *Language and Speech*, 33-3, 259-272.

Duifhuis, H., Willems, L.F. & Sluyter, R.J. (1982) Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception, *Journal of the Acoustical Society of America*, 71, 1568-1580.

Ebing, E.F. (1994) Towards an inventory of perceptually relevant pitch movements for Indonesian. In: *Experimental studies of Indonesian prosody*, C. Odé & V.J. van Heuven (eds.), (Semaian 9), 181-210.

Ebing, E.F. (1997) *Form and function of pitch movements in Indonesian.* Dissertation, Leiden University, CNWS publication series no. 54.

Eskenazi, L., Childers, D.G. & Hicks, D.M. (1990) Acoustic correlates of vocal quality, *Journal of Speech and Hearing Research*, 33, 298-306.

Fant, G., Kruckenberg, A. & Nord, L. (1990) Prosodic and segmental speaker variations, *Proceedings of the tutorial and research workshop on Speaker characterization in speech technology*, J. Laver, M. Jack & A. Gardiner (eds.), Edinburgh, 106-111.

Fletcher, H. (1934) Loudness, pitch, and the timbre of musical tones and their relation to the intensity, *Journal of the Acoustical Society of America*, 6, 59-69.

Fletcher, H. (1940) Auditory patterns, *Review of Modern Physics*, 12, 47-65.

Fònagy, I. (1978) A new method of investigating the perception of phonetic features, *Language and Speech*, 21, 34-49.

Forster, K.I. & Dickinson, R.G. (1976) More on the Language-as-a-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F' and Min F', *Journal of Verbal Learning and Verbal Behavior*, 15, 135-142.

Fujisaki, H. & Ohno, S. (1995) Analysis and modeling of fundamental frequency contours of English utterances, *Proceedings of EUROSPEECH '95 Madrid*, 2, 985-988.

Furui, S. (1981) Cepstral analysis technique for automatic speaker verification, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 29, 257-272.

Furui, S. (1990) Speaker-dependent feature extraction, recognition and processing techniques, *Proceedings of the tutorial and research workshop on Speaker characterization in speech technology*, J. Laver, M. Jack & A. Gardiner (eds.), Edinburgh, 135-139.

Garrett, K.L. & Healey, E.C. (1987) An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day, *Journal of the Acoustical Society of America*, 82, 58-62.

Gelfer, C.E., Harris, K.S., Collier, C. & Baer, T. (1985) Is declination actively controlled? In: *Vocal fold physiology: biomechanics, acoustics and phonatory control*, I. Titze & R. Scherer (eds.), Denver: Denver Center for the Performing Arts, 113-126.

Gilbert, H.R. & Weismer, G.G. (1974) The effects of smoking on the speaking fundamental frequency of adult women, *Journal of Psycholinguistic Research*, 3, 225-231.

Gold, B. & Rabiner, L. (1969) Parallel processing techniques for estimating pitch periods of speech in the time domain, *Journal of the Acoustical Society of America*, 46, 442-448.

Goldman-Eisler, F. (1951) The measurement of time sequences in conversational behaviour, *British Journal of Psychology*, 42, 355-362.

Goldman-Eisler, F. (1961a) The continuity of speech utterance, its determinants and its significance, *Language and Speech*, 4, 220-231.

Goldman-Eisler, F. (1961b) Hesitation and information in speech. In: *Information theory*, C. Cherry (ed.), London: Academic Press, 162-173.

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

Gorsuch, R.L. (1990) Common factor analysis versus component analysis: some well and little known facts, *Multivariate Behavioral Research*, 25, 33-40.

Graddol, D. (1986) Discourse specific pitch behaviour. In: *Intonation in discourse*, C. Johns-Lewis (ed.), London and Sydney: Croom Helm, 221-237.

Grosjean, F. & Deschamps, A. (1975) Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes [Contrastive analysis of temporal variables in English and French: speech rate and component variables], *Phonetica*, 31, 144-184.

Gussenhoven, C. (1984) *On the Grammar and Semantics of Sentence Accents.* Dordrecht: Foris Publications.

Hart, J. 't (1976) Psychoacoustic backgrounds on pitch contour stylisation, *IPO Annual Progress Report*, 11, 11-18.

Hart, J. 't, Collier, R. & Cohen, A. (1990) *A Perceptual Study of Intonation.* Cambridge: Cambridge University Press.

Haselager, G.J.T., Slis, I.H. & Rietveld, A.C.M. (1991) An alternative method of studying the development of speech rate, *Clinical Linguistics & Phonetics*, 5, 53-63.

Hays, W.L. (1973) *Statistics for the social sciences, 2nd ed.* London: Holt, Rinehart and Winston.

Hecker, M. & Kreul, S.J. (1971) Descriptions of the speech of patients with cancer of the vocal folds, Parts 1 and 2, *Journal of the Acoustical Society of America*, 49, 1275-1287.

Henton, C.G. (1989) Fact and fiction in the description of female and male pitch, *Language and Communication*, 9, 299-311.

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

Hermes, D.J. (1993) Pitch analysis. In: *Visual Representations of Speech Signals*, M. Cooke, S. Beet & M. Crawford (eds.), Chichester: John Wiley & Sons, 3-25.

Hermes, D.J. & Gestel, J.C. van (1991) The frequency scale of speech intonation, *Journal of the Acoustical Society of America*, 90-1, 97-102.

Hess, W. (1983) *Pitch determination of speech signals: algorithms and devices.* Berlin: Springer-Verlag.

Heuvel, H. van den (1996) Speaker variability in acoustic properties of Dutch phoneme realisations. Dissertation, University of Nijmegen.

Heuven, V.J. van (1994a) Introducing prosodic phonetics. In: *Experimental studies of Indonesian prosody*, C. Odé & V.J. van Heuven (eds.), (Semaian 9), 1-26.

Heuven, V.J. van (1994b) What is the smallest prosodic domain? In: *Phonological structure and phonetic form, papers in Laboratory Phonology III*, P.A. Keating (ed.), Cambridge: Cambridge University Press, 76-98.

Heuven, V.J. van & Sluijter, A.M.C. (1996) Notes on the phonetics of word prosody. In: *Stress patterns of the world, Part I: Background*, R. Goedemans, H. van der Hulst & E. Visch (eds.), Holland Institute of Generative Linguistics, The Hague: Holland Academic Graphics, 233-269.

Higgins, M.B. & Saxman, J.H. (1989a) A comparison of intrasubject variation across sessions of three vocal frequency perturbation indices, *Journal of the Acoustical Society of America*, 86, 911-916.

Higgins, M.B. & Saxman, J.H. (1989b) Variations in vocal frequency perturbation across the menstrual cycle, *Journal of Voice*, 3, 233-243.

Higgins, M.B. & Saxman, J.H. (1991) A comparison of selected phonatory behaviors of healthy aged and young adults, *Journal of Speech and Hearing Research*, 34, 1000-1010.

Hiraoka, N., Kitazoe, Y., Ueta, H., Tanaka, S. & Tanabe, M. (1984) Harmonic intensity analysis of normal and hoarse voices, *Journal of the Acoustical Society of America*, 76, 1648-1651.

Hirson, A. & Duckworth, M. (1995) Forensic implications of vocal creak as voice disguise. In: *Studies in Forensic Phonetics*, A. Braun & J.-P. Köster (eds.), Trier: Wissenschaftlicher Verlag, 67-76.

Hollien, H. (1990) *The acoustics of crime. The new science of forensic phonetics.* New York, London: Plenum Press.

Hollien, H. & Jackson, B. (1973) Normative data on the speaking fundamental frequency characteristics of young adult males, *Journal of Phonetics*, 1, 117-120.

Hollien, H. & Majewski, W. (1977) Speaker identification by long-term spectra under normal and distorted speech conditions, *Journal of the Acoustical Society of America*, 62, 975-980.

Hollien, H., Majewski, W. & Doherty, E.T. (1982) Perceptual identification of voices under normal, stress and disguise speaking conditions, *Journal of Phonetics*, 10, 139-148.

Hollien, H. & Paul, P. (1969) A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls, *Language and Speech*, 12, 119-124.

Hollien, H. & Shipp, T. (1972) Speaking fundamental frequency and chronologic age in males, *Journal of Speech and Hearing Research*, 15, 155-159.

Hollingshead, A.B. & Redlich, F.C. (1958) *Social class and mental illness.* New York: Wiley.

Horii, Y. (1975) Some statistical characteristics of voice fundamental frequency, *Journal of Speech and Hearing Research*, 18, 192-201.

Horii, Y. (1976) Difference limen for jitter and shimmer. Paper presented at the annual convention of the American Speech and Hearing Association, Houston, Texas.

Horii, Y. (1979) Fundamental frequency perturbation observed in sustained phonation, *Journal of Speech and Hearing Research*, 22, 5-19.

Horii, Y. (1980) Vocal shimmer in sustained phonation, *Journal of Speech and Hearing Research*, 23, 202-209.

Howell, P. & Kadi-Hanifi, K. (1991) Comparison of prosodic properties between read and spontaneous speech material, *Speech Communication*, 10, 163-169.

Huberty, C.J. & Morris, J.D. (1989) Multivariate analysis versus multiple univariate analyses, *Psychological Bulletin*, 105, 302-308.

International Phonetic Association (1949) *The principles of the International Phonetic Association.* London: International Phonetic Association.

Jassem, W., Batóg, M.S. & Czajka, S. (1973) Statistical characteristics of short-term average F_0 distributions as personal voice features. In: *Speech analysis and synthesis*, vol. 3, W. Jassem (ed.), Warsaw: Panstwowe Wydawn. Naukowe, 209-225.

Johns-Lewis, C. (1986) Prosodic differentiation of discourse modes. In: *Intonation in discourse*, C. Johns-Lewis (ed.), London and Sydney: Croom Helm.

Johnson, C.C., Hollien, H. & Hicks, J.W. (1984) Speaker identification utilizing selected temporal speech features, *Journal of Phonetics*, 12, 319-326.

Jongenburger, W. & Heuven, V. van (1991) The distribution of (word initial) glottal stop in Dutch. In: *Linguistics in the Netherlands 1991*, F. Drijkoningen & A. van Kemenade (eds.), Amsterdam, Philadelphia: John Benjamins, 101-110.

Kasuya, K. & Kobayashi (1983) Characteristics of pitch period and amplitude perturbations in pathological voice, *Proceedings International Conference of Acoustics, Speech, Signal Processing*, Boston, 344-353.

Klecka, W.R. (1980) *Discriminant analysis*. Beverly Hills, London: Sage Publications.

Kohonen, T. (1989) *Self-organization and associative memory, 3rd ed.* Berlin: Springer-Verlag.

Kohonen, T. (1990) The self-organizing map, *Proceedings of the IEEE*, 78, 1464-1480.

Koike, Y. (1973) Application of some acoustic measures for the evaluation of laryngeal dysfunction, *Studia Phonologica*, 7, 17-23.

Koike, Y., Takahashi, H. & Calcaterra, T.C (1977) Acoustic measures for detecting laryngeal pathology, *Acta Oto-Laryngologica*, 84, 105-117.

Koopmans-van Beinum, F.J. (1991) The role of focus words in natural and in synthetic continuous speech: Acoustic aspects, *Speech Communication*, 11, 439-452.

Kraayeveld, J. (1995) Speaker specificity in the choice of pitch movements, *Proceedings of the Department of Language and Speech Nijmegen*, 19, 67-75.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J. van (1990) Prosodic speaker characteristics in Dutch. In: *Proceedings of the tutorial and research workshop on Speaker characterization in speech technology*, J. Laver, M. Jack & A. Gardiner (eds.), Edinburgh, 135-139.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J. van (1991) Speaker characterization in Dutch using prosodic parameters, *Proceedings of EUROSPEECH '91*, Genova: ESCA/IIC, 427-430.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J. van (1993) Speaker specificity in prosodic parameters, *Proceedings ESCA Workshop on Prosody, Working Papers Department of Linguistics and Phonetics Lund University*, 264-267.

Kutik, E.J., Cooper, W.E. & Boyce, S. (1983) Declination of fundamental frequency in speakers' production of parenthetical and main clauses, *Journal of the Acoustical Society of America*, 73, 1723-1730.

Ladd, D.R. (1983) Phonological features and intonational peaks, *Language*, 59, 721-759.

Ladd, D.R. (1984) Declination: a review and some hypotheses, *Phonology Yearbook*, 1, 53-74.

Laver, J. (1980) *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

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

Liberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In: *Language sound structure*, M. Aronoff and R. Oerhle (eds.), Cambridge (MA): MIT Press, 157-233.

Lieberman, P. (1963) Some acoustic measures of the fundamental periodicity of normal and pathologic larynges, *Journal of the Acoustical Society of America*, 35, 344-353.

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

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

Lummis, R.C. (1973) Speaker verification by computer using speech intensity for temporal registration, *IEEE Transactions of Audio and Electroacoustics*, 21, 80-89.

Maeda, S. (1976) *A characterization of American English intonation.* Ph.D. thesis, MIT.

Malinowski, B. (1935) *Coral Gardens and their Magic, vol. 2.* London: Allen and Unwin.

Markel, J. & Davis, S. (1979) Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 27, 74-82.

Markel, J. & Gray, A. (1976) *Linear Prediction of Speech.* New York: Springer-Verlag.

McGlone, R.E. & Hollien, H. (1963) Vocal pitch characteristics of aged women, *Journal of Speech and Hearing Research*, 6, 164-170.

McGonegal, C., Rosenberg, A. & Rabiner, L. (1979) The effects of several transmission systems on an automatic speaker verification system, *The Bell System Technical Journal*, 58, 2071-2087.

Meyerson M.D. (1976) The effects of aging on communication, *Journal of Gerontology*, 31, 29-38.

Miller, J.L. & Grosjean, F. (1981) How the components of speaking rate influence perception of phonetic segments, *Journal of Experimental Psychology*, 7, 208-215.

Miller, J.L., Grosjean, F. & Lomanto, C. (1984) Articulation rate and its variability in spontaneous speech: a reanalysis and some implications, *Phonetica*, 41, 215-225.

Moulines, E. & Sagisaka, Y. (eds.) (1995) Voice conversion: state of the art and perspectives, *Speech Communication*, 16.

Murray, I.R. & 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.

Murry, T. & Doherty, E.T. (1980) Selected acoustic characteristics of pathologic and normal speakers, *Journal of Speech and Hearing Research*, 23, 361-369.

Nolan, F. (1983) *The phonetic bases of speaker recognition.* Cambridge, New York, Melbourne: Cambridge University Press.

Noll, A.M. (1967) Cepstrum pitch determination, *Journal of the Acoustical Society of America*, 41, 293-309.

Nooteboom, S.G. & Eefting, W. (1991) To what extent is speech production controlled by speech perception? Some questions and some experimental evidence. In: *Speech perception, production and linguistic structure*, Y. Tokhura, E. Vatikiotis-Bateson & Y. Sagisaki (eds.), Tokyo: Ohmsha and Amsterdam, Oxford, Burke: IOS Press, 439-449.

Odé, C. (1989) *Russian intonation: a perceptual description*. Amsterdam, Atlanta: Rodopi.

Oglesby, J. & Mason, J.S. (1990) Optimization of neural models for speaker identification, *Proceedings International Conference of Acoustics, Speech, Signal Processing*, Albuquerque, S5.1.

Ohde, R. (1984) Fundamental frequency as an acoustic correlate of stop consonant voicing, *Journal of the Acoustical Society of America*, 75, 224-230.

Orlikoff, R.F. & Baken, R.J. (1990) Consideration of the relationship between the fundamental frequency of phonation and vocal jitter, *Folia Phoniatrica*, 42, 31-40.

O'Shaughnessy, D. (1987) *Speech Communication*. Reading (MA): Addison-Wesley.

Patterson, R.D. (1976) Auditory filter shapes derived with noise stimuli, *Journal of the Acoustical Society of America*, 59, 640-654.

Pegaro Krook, M.I. (1988) Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis, *Folia phoniatrica*, 40, 82-90.

Pierrehumbert, J. (1980) The phonology and phonetics of English intonation. Ph.D. thesis, MIT.

Pierrehumbert, J. & Beckman, M. (1988) *Japanese tone structure*. Cambridge (MA): MIT Press.

Pijper, J.R. de (1983) *Modelling British English intonation*. Dordrecht, Cinnaminson: Foris Publications.

Pike, K.L. (1945) *The intonation of American English*. Ann Arbor (MI): University of Michigan Press.

Pinto, N.B. & Titze, I.R. (1990) Unification of perturbation measures in speech signals, *Journal of the Acoustical Society of America*, 87.3, 1278-1289.

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

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. & McGonegal, C.A. (1976) A comparative performance study of several pitch detection algorithms, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 24, 399-418.

Ramig, L.O. & Ringel, R.L. (1983) Effects of physiological aging on selected acoustic characteristics of voice, *Journal of Speech and Hearing Research*, 26, 22-30.

Reetz, H. (1989) A fast expert program for pitch extraction, *Proceedings of EUROSPEECH '89*, 1, Paris, 476-479.

Rie, J. van & Bezooijen, R. van (1995) Perceptual characteristics of voice quality in Dutch males and females from 9 to 85 years, *Proceedings of the XIIIth international congress of phonetic sciences, Stockholm*, 2, 290-293.

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

Rietveld, T. & Heuven, V.J. van (1997) *Algemene Fonetiek* [General phonetics]. Muiderberg: Coutinho.

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

Robins, R.H. (1980) *General linguistics: an introductory survey.* London: Longman.

Rooij, J.J. de (1979) Speech punctuation. An acoustic and perceptual study of some aspects of speech prosody in Dutch. Dissertation, University of Utrecht.

Rosenberg, A.E., Lee, C.-H. & Soong, F.K. (1990) Sub-word unit talker verification using hidden Markov models, *Proceedings International Conference of Acoustics, Speech, Signal Processing*, Albuquerque, 1, 269-272.

Rosenberg, A.E. & Sambur, M. (1975) New techniques for automatic speaker verification, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 23, 169-176.

Russell, A., Penny, L. & Pemberton, C. (1995) Speaking fundamental frequency changes over time in women: a longitudinal study, *Journal of Speech and Hearing Research*, 38, 101-109.

Ryan, W. & Burk, K. (1974) Perceptual and acoustic correlates of aging in the speech of males, *Journal of Communication Disorders*, 7, 181-192.

Sambur, M. (1975) Selection of acoustic features for speaker identification, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 23, 176-182.

Savic, M. & Gupta, S.K. (1990) Variable parameter speaker verification system based on hidden Markov modelling, *Proceedings International Conference of Acoustics, Speech, Signal Processing*, Albuquerque, 1, 281-284.

Scherer, K.R. (1977) The effect of stress on fundamental frequency of the voice, *Journal of the Acoustical Society of America*, 62, 25-26.

Schmidt-Neilsen, A. & Stern, K. (1985) Identification of known voices as a function of familiarity and narrow-band coding, *Journal of the Acoustical Society of America*, 77, 658-663.

Schoentgen, J. (1989) Jitter in sustained vowels and isolated sentences produced by dysphonic speakers, *Speech Communication*, 8, 61-79.

Siegel, L.J. & Bessey, A.C. (1982) Voiced/unvoiced/mixed excitation classification of speech, *IEEE Transactions of the Acoustics, Speech and Signal Processing* ASSP, 30, 451-460.

Simon, C. (1927) The variability of consecutive wavelengths in vocal and instrumental sounds, *Psychological Monographs*, 36.

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

Soong, F.K., Rosenberg, A.E., Rabiner, L.R. & Jang, B.H. (1985) A vector quantization approach to speaker recognition, *Proceedings International Conference of Acoustics, Speech, Signal Processing*, Tampa, 387-390.

Stevens, J. (1986) *Applied multivariate statistics for the social sciences*. Hillsdale (NJ): Lawrence Erlbaum.

Stevens, S.S., Volkman, J., & Newman, E.B. (1937) A scale for the measurement of the psychological magnitude of pitch, *Journal of the Acoustical Society of America*, 8, 185-190.

Stoicheff, M.L. (1981a) Speaking fundamental frequency characteristics of non-smoking female adults, *Journal of Speech and Hearing Research*, 24, 437-441.

Stoicheff, M.L. (1981b) Speaking fundamental frequency of middle-aged females, *Folia Phoniatrica*, 19, 167-172.

Strik, H. (1994) Physiological control and behaviour of the voice source in the production of prosody. Dissertation, University of Nijmegen.

Terken, J.M.B. (1984) The distribution of pitch accents in instructions as a function of discourse structure, *Language and Speech*, 27, 269-289.

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

Thomas, D.R. (1992) Interpreting discriminant functions. A data analytic approach, *Multivariate Behavioral Research*, 27, 335-362.

Tielen, M.T.J. (1992) Male and female speech. Dissertation, University of Amsterdam.

Titze, I.R., Horii, Y. & Scherer, R.C. (1987) Some technical considerations in voice perturbation measurements, *Journal of Speech and Hearing Research*, 30, 252-260.

Tosi, O.I. (1979) *Voice identification: theory and legal applications*. Baltimore: University Park Press.

Umeda, N. (1982) F_0 declination is situation dependent, *Journal of Phonetics*, 10, 279-290.

Wells, J.C. (1982) *Accents of English: an introduction*. Cambridge, New York, Melbourne: Cambridge University Press.

Wendahl, R.W. (1966a) Laryngeal analog synthesis of jitter and shimmer: auditory parameters of harshness, *Folia Phoniatrica*, 18, 98-108.

Wendahl, R.W. (1966b) Some parameters of auditory harshness, *Folia Phoniatrica*, 18, 26-32.

Wiedemann, C.F. & Fenster, C.A. (1978) The use of chance corrected percentage of agreement to interpret the results of a discriminant analysis, *Educational and Psychological Measurement*, 38, 29-35.

Willems, N.J., Collier, R., & Hart, J. 't (1988) A synthesis scheme for British English intonation, *Journal of the Acoustical Society of America*, 84, 1250-1261.

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

Winer, B.J. (1971) *Statistical principles in experimental design, 2nd ed.* New York: McGraw-Hill.

Wolf, J.J. (1972) Efficient acoustic parameters for speaker recognition, *Journal of the Acoustical Society of America*, 51, 2044-2056.

Woods, N.J. (1992) Sociolinguistic patterns in English pitch and intonation. Doctoral dissertation, Linacre College, Oxford.

Zanten, E. van, Damen, L. & van Houten, E. (1991) *The ASSP Speech Database*. Utrecht: Stichting Spraaktechnologie, SPIN/ASSP final report no. 41, 17-24.

Zemlin, W.R. (1981) *Speech and hearing science, anatomy and physiology, 2nd ed.* Englewood Cliffs (NJ): Prentice-Hall.

Zimmerman, D.H. & West, C. (1975) Sex roles, interruptions and silences in conversation. In: *Language and sex: differences and dominance*, B. Thorne & N. Henley (eds.), Rowley (MA): Newbury House, 105-129.

Appendix A

Appendix A: Protocol for CB measurements: pitch movements and declination.

Our starting-point in determining the begin and end frame of the pitch movements and the declination lines was, that we wanted to remain as close as possible to the actually obtained F_0 measurements and to abstain, where possible, from stylization and interpolation. Such an approach necessitates a set of rules to decide on the problematic cases in a verifiable way. As most of the cases did not raise any problems we do not believe the measurement problems encountered influenced our results to a large extent. Below an overview is given of the rules applied. The frames mentioned in the rules had a duration of 10 ms.

First frame of a rise:

- There must be a lowest point, in the sense that from 5 frames before until 5 frames after a possible pivot point there should be no lower point, or there must be a clear breach of tendency in a gradually increasing line. In this latter case there is mostly no earlier point suited as starting-point of the rise.
- In some cases where apparently a rise has started a minor lowering takes place within that rise. If the first part of the rise has lasted at least 5 frames the lowest point of this lowering must not be considered the starting point of the rise if:
  * the direction of the first part of the rise is more or less equal to that of the part of the rise following the minor lowering,
  * the lowering lasts shorter than 5 frames,
  * the lowest point of the minor lowering is higher than the start of the first part of the rise.

  If one of these conditions is not met, the lowering must not be considered a "minor" lowering, and the lowest point of the lowering is taken as the start of the rise.
- If the first point of a rise is preceded by a voiceless part of the utterance, try changing the voiced/unvoiced criterion (in Hermes's pitch editing program PCT two voiced/unvoiced criteria are available). If this does not solve the problem (i.e., does not result in extra voiced frames that comply with the above-mentioned criteria): compare the first point of the voiced part of the rise with the last low-declination point before the voiceless part. If these are not to far apart (i.e., differ less than 15 Hz for women or 10 Hz for men), take the first voiced frame as the starting point of the rise. If this is not the case, try to interpolate the rise until the last voiced low-declination point. (NB: this was never necessary in our study).
- If a voiceless part of the contour is followed by a short (lasting less than 5 frames), small rise (less than 15 Hz for women or 10 Hz for men), which in its turn is followed by a steeper rise, this last rise is considered to be the relevant rise.

Final frame of a rise:

- There must be a highest point, in the sense that from 5 frames before until 5 frames after a possible pivot point there should be no higher point.

- If the last point of a rise is followed by a voiceless part of the utterance, try changing the voiced/unvoiced criterion. If this does not solve the problem compare the last point of the voiced part of the rise with the first high-declination point after the voiceless part. This method is not very reliable, as in some contours the rise is followed by a fall and the high-declination level can only be inferred from a forward-interpolation of the last part of that fall.
- If a steep rise is followed by a short (lasting less than 5 frames), small rise (less than 15 Hz for women or 10 Hz for men), this last rise is not considered to be part of the rise. If the second part of the rise lasts longer or exceeds the F_0 criteria mentioned, the second part of the rise is regarded to be part of the rise.
- If the local maximum is an outlier, consider the datafile with the original waveforms to check whether the high value originates from measurement errors of the subharmonic summation program. If this is the case, see whether the division of the erratic frame value by some integer number corresponds better with the estimated duration of the periods near the top value of the rise.

First frame of a fall:

- There must be a highest point, in the sense that from 5 frames before until 5 frames after a possible pivot point there should be no higher point.
- If the first point of a fall is preceded by a voiceless part of the utterance, try changing the voiced/unvoiced criterion. If this does not solve the problem compare the first point of the voiced part of the fall with the last high-declination point before the voiceless part. This method is not very reliable, as in some contours the fall is preceded by a rise and the high-declination level can only be inferred from an interpolation of the first part of that rise.
- If a steep fall is preceded by a short (lasting less than 5 frames), small (less than 15 Hz for women or 10 Hz for men) fall, this first small fall is not considered to be a relevant part of the fall. If this first part of the fall lasts longer or exceeds the F_0 criteria mentioned, it is indeed regarded to be part of the fall.
- If the local maximum is an outlier, consider the datafile with the original waveforms to check whether the high value originates from measurement errors of the subharmonic summation program. If this is the case, see whether the division of the erratic frame value by some integer number corresponds better with the estimated duration of the periods near the top value of the fall.

Final frame of a fall:

- There must be a lowest point, in the sense that from 5 frames before until 5 frames after a possible pivot point there should be no lower point, or there must be a clear breach of tendency in a rather steeply decreasing line. In this latter case there is mostly no earlier point suited as starting-point of the rise.
- In some cases within a fall a local dip takes place. If after the end of that dip the fall continues for at least 5 frames the lowest point of this dip must not be considered the starting point of the rise if:
  * the direction of the last part of the rise is more or less equal to that of the part of the rise following the dip,
  * the local dip lasts shorter than 5 frames,
  * the lowest point of the dip is higher than the end of the last part of the fall.

If one of these conditions is not met, the dip must not be considered a "minor" lowering, and the lowest point of the dip is taken as the end of the fall.

- If the last point of a fall is followed by a voiceless part of the utterance, try changing the voiced/unvoiced criterion. If this does not solve the problem (i.e., does not result in extra voiced frames that comply with the above-mentioned criteria): compare the last point of the voiced part of the fall with the first low-declination point after the voiceless part. If these are not to far apart (i.e., differ less than 15 Hz for women or 10 Hz for men), take the last voiced frame as the ending point of the rise. If this is not the case, try to interpolate the fall until the last voiced low-declination point. (NB: this was not necessary in our study).
- If a steep fall is followed by a short (lasting less than 5 frames), small fall (less than 15 Hz for women or 10 Hz for men), which in its turn is followed by a voiceless part of the contour, this last fall is not considered to be part of the relevant rise.

First frame of the declination line:

- The first voiced frame should have the highest F_0 value of the first part of the utterance, in the sense that until 5 frames after that first voiced frame there should be no higher point.
- If there are higher points in the vicinity of the first F_0 point, they could be part of a pitch movement, in which case one cannot be sure if the first voiced frame belongs to the low declination line. Also, the first voiced frame is sometimes followed by a clear fall. In that case the first voiced frame is not part of the low declination line. In cases of doubt, look for a part of the contour that reliably forms a part of the low declination line and interpolate back to the first voiced frame. If the F_0 value of that frame is almost equal to the value estimated by interpolating the low declination line, maintain the F_0 value of the first voiced frame. If it is not, replace it by the estimated value.
- If the first voiced frame is an outlier, or if the F_0 values of the first few frames vary considerably (differences exceeding 15 Hz for women or 10 Hz for men), replace the first voiced frame by the first voiced frame that is reliably a part of the low declination line or, when this frame is further than 5 frames away from the first voiced frame, interpolate back to the first voiced frame.

Final frame of the declination line:

- There must be a lowest point, in the sense that from 5 frames before the last F_0 measurement there should be no lower point.
- If there are lower points in the neighbourhood of the last F_0 point, they could be part of a pitch movement, in which case one cannot be sure if the last voiced frame belongs to the low declination line. Also, the last voiced frame is sometimes preceded by a clear rise. In that case the last voiced frame is mostly not part of the low declination line, but of a continuation rise (rise "2" in GDI). In cases of doubt, look for a part of the contour that reliably forms a part of the low declination line and interpolate forward to the last voiced frame. If the F_0 value of that frame is almost equal to the value estimated by interpolating the low declination line, maintain the F_0 value of the last voiced frame. If it is not, replace it by the estimated value.

- If the last voiced frame is an outlier, or if the F_0 values of the last few frames vary considerably (differences exceeding 15 Hz for women or 10 Hz for men), the speaker's voice quality probably degrades to "creaky". Replace the last voiced frame by the last voiced frame that is reliably a part of the low declination line or, when this frame is further than 5 frames away from the last voiced frame, interpolate back to the first voiced frame. Always check the datafile with the original waveforms to find out if the F_0 value used comes close to those corresponding to the "real" period durations. If the difference is substantial try whether the division of the erratic frame value by some integer number corresponds better with the estimated duration of the periods near the top value of the fall.

Appendix B

Appendix B: Summary of extra-linguistic speaker characteristics.

Below an overview is given of some of the extra-linguistic characteristics of the 50 speakers who were recorded for the purposes of this study; the codes in the heading of the table correspond to the following characteristics:

A: Speaker code; the first character stands for female (F) or Male (M), the second for the speaker's age group (1 to 5), and the third for the specific speaker (from 1 to 5 for each Sex-Age group),
B: Age at recording 1,
C: Time interval (in days) between the two recordings,
D: Mean value of overall accentedness as assessed by 5 judges (see Chapter 3),
E: Mean value of accentedness in segmental aspects of speech,
F: Mean value of accentedness in suprasegmental aspects of speech,
G: Any voice complaints? (+ = yes, − = no),
H: Smoking behaviour (+ = yes, − = no),
I: Self-assessed accentedness (+ = with, − = without any accent),
J: Date of birth,
K: Place of residence during largest part of childhood,
L: idem for father,
M: idem for mother.

A	B	C	D	E	F	G	H	I	J	K	L	M
F11	18	220	9 06	8 90	9 60	−	−	−	25-09-70	Dieren	Arnhem	Arnhem
F12	23	358	7 56	7 22	8 50	−	−	~	11-06-67	Tilburg, Nijmegen	Breda	Oosterhout
F13	24	185	7 70	7 20	8 60	−	+	~	14-12-65	Tilburg	Tilburg	Tilburg
F14	23	210	7 50	7 50	7 70	−	−	~	13-03-67	Hengelo	Nederhorst	Broek op L
F15	20	237	7 60	7 40	8 10	−	−	~	14-03-70	Haaksbergen	Haaksbergen	Haaksbergen
F21	26	290	8 50	8 00	9 20	−	−	−	28-07-64	Oegstgeest	Indonesia	Alkmaar
F22	34	196	9 04	8 84	9 60	−	−	−	11-07-56	Ermelo, Amersfoort	Ermelo	Ermelo
F23	30	243	7 70	7 30	8 70	−	−	+	11-11-60	Nijmegen	Hengelo	Delft
F24	28	235	7 70	7 70	8 10	−	−	+	01-01-62	Waalwijk	Waalwijk	Waalwijk
F25	26	252	8 90	8 70	9 60	−	−	−	08-09-64	Bennekom	Utrecht	Utrecht
F31	40	268	9 04	8 94	9 40	−	−	−	20-05-50	Voorburg	misc (west)	misc (west)
F32	37	266	7 94	7 74	8 50	−	−	~	28-03-53	Amsterdam, Nijmegen	Alkmaar	misc. (west)
F33	40	160	7 30	7 00	7 90	−		+	22-01-50	Heerlen	misc. (Lim)	misc (Lim)
F34	41	288	8 14	8 04	9 10	~	−	−	14-08-49	Haarlem	Hilversum	Den Helder
F35	40	287	9 06	9 26	9 00	−	+	−	12-09-50	Nijmegen	Kerkdriel	Nijmegen

F41	51	421	8 20	8 30	8 80	–	+	–	29-08-39	Nijmegen	Nijmegen	Nijmegen
F42	52	219	7 40	7 20	8 60	–	–	~	05-05-38	Vught	StMichielsgst	Roosendaal
F43	48	188	8 00	7 90	9 10	–	+	~	12-09-42	Uden	misc (N-B)	misc (N-B)
F44	51	209	9 14	9 44	8 80	–	–	–	07-06-39	Zwolle	Zwolle	Zwolle
F45	51	225	9 44	9 34	9 50	–	+	–	23-12-39	misc	misc (Z-H)	misc (N-B)
F51	61	170	7 60	7 70	7 50	–	–	–	05-01-30	Nijmegen	Arnhem	Arnhem
F52	63	134	6 30	6 30	7 50	–		–	23-08-28	Weert	misc (Ov)	Düsseldorf
F53	56	183	7 22	6 78	8 40	–	–	+	19-08-35	Beuningen, Nijmegen	Beuningen	Winssen
F54	60	142	8 60	8 50	8 90	–	–	~	08-08-31	Rotterdam	Rotterdam	Rotterdam
F55	64	727	8 60	8 20	8 90	–	–	–	11-07-29	Vlaardingen	Groningen	Groningen
M11	24	243	7 90	7 62	8 40	–	+	~	21-12-66	Groenlo	Arnhem	Groenlo
M12	18	231	8 56	8 26	9 10	–	–	–	04-01-70	Hansweert	Hansweert	Heinkenszand
M13	22	233	7 90	7 40	8 90	–	–	~	16-04-69	Hengelo	Hengelo	Hengelo
M14	21	281	8 70	8 20	9 10	–	–	~	26-07-69	Nijmegen	Elst	Bredevoort
M15	24	307	7 70	7 30	8 50	–	+	+	11-12-66	Hengelo	Hengelo	Hengelo
M21	26	183	8 20	7 90	9 20	–	–	+	23-09-64	Nijmegen	Nuenen	Arnhem
M22	30	211	8 40	7 90	9 20	–	–	~	03-10-60	Groningen	Groningen	misc (Dr)
M23	28	201	8 20	7 70	9 00	–	–	+	06-01-62	Oosterhout	Oosterhout	Terheijden
M24	29	203	8 24	8 04	8 90	–	–	–	02-05-61	Bennebroek	Schiedam	Zoeterwoude
M25	33	257	7 86	7 46	9 40	–		+	18-05-57	Rosmalen	Bergem	Den Bosch
M31	45	245	7 86	7 56	8 80	–	–	–	14-03-45	Horst	Horst	Horst
M32	39	255	9 20	9 00	9 40	–	+	–	16-07-51	Utrecht	misc (Dr)	Brabant
M33	39	260	9 00	8 76	9 60	+	–	–	11-03-59	Rotterdam	Dordrecht	misc
M34	45	216	8 90	8 70	9 70	-*	+	–	19-06-46	Teteringen	Teteringen	Teteringen
M35	45	216	9 06	9 06	9 70	–	–	–	27-07-46	Utrecht	Utrecht	Utrecht
M41	46	216	7 80	7 70	9 00	–	–	~	17-02-44	Roermond	Roermond	Roermond
M42	46	230	8 08	7 58	9 10	–	+	~	07-07-44	Nijmegen	Nijmegen	Venraij
M43	48	228	7 70	7 50	8 60	–		–	22-01-42	Nijmegen	Naaldwijk	Beuningen
M44	53	252	8 26	7 66	9 40	–	–	+	21-04-37	Udenhout	Udenhout	Moergestel
M45	48	223	8 40	8 60	8 80	–	–	–	26 07-42	Rucphen	Rucphen	Rucphen
M51	57	246	9 10	8 70	9 50	–	–	–	21-09-33	Hilversum, Nijmegen	Haarlem	Harnstätte
M52	64	208	8 90	8 70	9 60	–	+	–	14-11-27	Nijmegen	Nijmegen	Grave
M53	57	216	9 00	8 00	9 20	–	+	~	12-03-33	Rotterdam	Rotterdam	Rotterdam
M54	58	419	8 98	8 84	9 50	–	–	–	21-04-32	Amsterdam	Vianen	Indonesia
M55	56	204	8 96	8 66	9 50	–	–	–	26 10-34	Hazerswoude	Hazerswoude	Hazerswoude

* slight inclination towards stuttering

Appendix C

APPENDIX C: Stimuli.

The recording procedure of the two sessions was described in Chapter 2. We started with an interview, to obtain spontaneous speech. The conversation revolved round the following questions:

session 1:

- There are those who say that whatever tastes nice, must be healthy. Others think of "healthy" in terms of "physically healthy". They think of vitamines, minerals, etc. What is your comment to the statement "what tastes well, must be healthy".
- Could you describe your favourite dish?
- In how far do you consider it important whether your meals are healthy?
- Another aspect of food is its social function. How important is this aspect to you, do you attach much importance to being in agreeable company while eating?
- What do you think of people that, on being invited for diner, make demands concerning the composition of the meal?
- In many countries inviting somebody for dinner is a sign of hospitality. What do you do if somebody is paying you a visit around dinner-time? Would you feel obliged to ask him or her to stay for dinner?

session 2:

- I would like to talk with you about the subject of holidays. When did you last go on holiday and where did you go?
- Can you tell me something special about your most recent holidays, such as a nice experience?
- A distinction can be made between the more active type of holidays and the more lazy kind. Which do you prefer, and can you explain why?
- What, in your opinion, is the use of taking a holiday? Or doesn't it have any?
- Could you describe to me a really idyllic place; one that you saw during your holidays?

After the interview, which usually lasted about five minutes, the speakers were to read the first page of their reading-booklet, an instruction to the reading task:

Original instruction

Dit boekje omvat twee taken. Het eerste deel, tot aan de volgende rode bladzijde, is een voorleestaak en in het tweede deel moeten naar aanleiding van plaatjes bepaalde zinnen worden gezegd.

In het eerste deel, dat op de volgende bladzijde begint, ziet u op iedere bladzijde een zin of een stukje tekst. Het is de bedoeling dat u deze zinnen en teksten één voor één voorleest *zonder* te aarzelen of te stotteren. Het is daarom belangrijk dat u, alvorens de tekst op een nieuwe bladzijde voor te lezen, deze eerst heel aandachtig doorneemt.

Let erop, dat u niet op rare plaatsen nadruk legt. In sommige gevallen zijn woorden in een zin schuin gedrukt, bijv.:
Het is hier niet zomaar warm; het is hier *heel* warm.

Het schuin gedrukte woord (hier "heel") moet dan beklemtoond worden.

English translation

This booklet comprises two tasks. The first part, up to the next red page, is a reading-task and in the second part certain sentences must be made in response to pictures.

In the first part, that starts on the next page, you see on each page a sentence or a piece of text. You should read out these texts one by one *without* hesitation or stuttering. It is therefore important that, before reading out a new page, you should carefully take this in.

Take care that you do not put stress on unsound positions. In some cases words are printed in italics, e.g.:
It is not just warm here; it is *very* warm here.

The italicized word (here "very") should be stressed.

The first page of the reading-booklet contained the newspaper-like story from which the read-out fragment of Chapter 4 were obtained:

Dutch original

Langs de Duinlaan aan de rand van IJmuiden staan al jarenlang woonwagens. De in meerderheid bejaarde bewoners van de drieëndertig wagens mogen daar niet blijven. Dit is het gevolg van een landelijke beleidsverandering ten aanzien van woonwagenkampen.

Alle woonwagenbewoners in Nederland woonden ooit in kleine kampen, maar vanaf midden jaren zestig werden ze gedwongen te verhuizen naar grote regionale centra. De overheid komt nu terug op deze oude benadering en wil meer integratie van de woonwagenbewoners met de burgerij. Daarom moeten ze in normale buurten, nabij winkels en andere voorzieningen komen te wonen.

In het nieuwe streven naar verdeling

English translation

Along Dune Avenue, on the skirts of IJmuiden, for years on end there have been caravans. The mainly elderly inmates of the thirty-three vehicles are not allowed to stay there. This is the result of a change in the national policy towards caravan-camps.

All caravanners in the Netherlands once lived in small camps, but from the mid-sixties on they were forced to move to large, regional centres. The government now goes back from this old approach and wants more integration of the caravanners with the commonalty. Therefore, they will have to come to live in normal neighbourhoods, near to shops and other facilities.

In the new movement toward spreading

van woonwagens over de wijken is het woonwagencentrum IJmuiden een groot probleem. De animo voor de verhuizing is minimaal en het kostte wethouder Molenaar daarom veel energie om de bewoners te overreden met de verhuizing mee te doen. Zij stemden uiteindelijk in met een voorlopige verhuizing naar drie nieuwe locaties, om aan hun nieuwe woningen en de nieuwe buren te kunnen wennen.

Nu het woord "voorlopig" in de stukken van de gemeenteraad niet meer voorkomt, zijn de woonwagenbewoners boos. Eén van hen noemde de hele onderneming een drama. Aan de mening van de Duinlaanbewoners zou Molenaar maling hebben gehad.

Een woordvoerder van de heer Molenaar deelde mee, dat de wethouder de manier waarop de media de zaak benaderen verre van ideaal vindt. Hij wil niemand dwingen te verhuizen en wil alleen de belangen van de bewoners dienen. De deur naar nieuwe onderhandelingen blijft daarom ruim open.

Inmiddels heeft een deel van de Duinlaan-bewoners verklaard niet meer met de verhuizing mee te doen.

the caravans over the quarters, the caravan-centre IJmuiden constitutes a large problem. Enthusiasm for moving is minimal, and alderman Miller had to go out of his way to persuade the inhabitants to go along with the move. They finally agreed with a provisional moving to three new locations, to be able to get used to their new houses and the new neighbours.

Now that the word "provisional" has disappeared from the documents of the city council, the caravan inmates are angry. One of them called the entire venture a drama. For the opinion of the Dune Avenue inhabitants Miller would not have cared a rap.

A spokesman of Mr. Miller announced, that the alderman found the way in which the media approached the matter, far from ideal. He does not want to force anybody to move, and he only wants to serve the interests of the inhabitants. The door to new negotiations is therefore left wide open.

By now, part of the Dune Avenue inhabitants have stated not to take part in the move any more.

This text was followed by 31 pages with sentences of different length and a section with question-answer stimuli. These sections we will not describe in detail, as most of the utterances were not used in this study. An overview of the utterances that were obtained from the reading-booklet is presented in Appendix D.

Appendix D

Appendix D: Uniformity of pitch contours.

Uniformity of pitch contour for 48 sentences, in the first table for 18 unpremeditated sentences, in the second table for 30 premeditated ones.

Uniformity of pitch contour for 18 unpremeditated sentences. In the first column the sentence types are shown. In the second column the sentences are presented. For each of these sentences ten transcriptions of ten different speakers were considered. The pitch movements that were found in at least eight of these ten transcribed sentences are printed in bold face, immediately after the syllable on which they occurred.

type	sentence
1 picture	Daar zie ik één–**1**– banaan Daar zie ik een lang–**1**– been Daar zie ik ro–**1**–de limonade Daar zie ik een lui–**1**–e leerling Daar zie ik een draai–**1**–ende molen Daar zie ik een ro–**1**–de deur Daar zie ik een da–**1**–lende lijn Daar zie ik een brui–**1**–ne beer Daar zie ik een bre–**1**–de riem Daar zie ik een hui–**1**–lende baby Daar zie ik een blau–**1**–we bloem Daar zie ik een blij–**1**–e boer
2 pictures	Daar zie ik een brui–**1**–ne beer–**B**– en een lui–**1**–e leer–**A**–ling Daar zie ik een draai–**1**–ende molen–**B**– en een da–**1**–lende lijn–**A**– Daar zie ik een hui–**1**–lende baby en een lang–**1**– been–**A**– Daar zie ik één–**1**– banaan–**2**– en ro–**1**–de limona–**A**–de Daar zie ik een bre–**1**–de riem–**2**–**B**– en een ro–**1**–de deur–**A**– Daar zie ik een blij–**1**–e boer–**B**– en een blau–**1**–we bloem–**A**–

Uniformity of pitch contour for 30 premeditated (read) sentences. In the first column the sentence types are shown. In the second column the sentences are presented. For each of these sentences ten transcriptions of ten different speakers were considered. The pitch movements that were found in at least eight of these ten transcribed sentences are printed in bold face, immediately after the syllable on which they occurred.

type	sentence
enumeration	Zo had hij boeken over die**–1–**ren, over bieren, over hielen, over lieden, over mieren, over nieren en over riemen Om er een paar te noe**–1–**men de loon**–1–**arbeiders, de bomenkwekers, de woningbouwers, de bonenboeren en de mo**–A–**lenaars
sports	De No**–1–**ren**–B–** wonnen van de Roeme**–1–**nen**–B–** met drie**–1–** nul**–A–** De Ie**–1–**ren wonnen van de De**–1–**nen met drie**–1–** één**–A–** De De**–1–**nen wonnen van de No**–1–**ren**–B** met één**–1–** nul**–A–**
car	Marjolein**–1–** wil in haar Mi**–1–**ni naar Berlijn**–A–** rijden Annemarie**–1–** wil in haar Renault naar Roeme**–A–**nie rijden Marleen**–1–** wil in haar La**–1–**da naar Alme**–A–**re rijden
single content word	Iemand een oor**–1&A–** aannaaien Hij wil roe**–1&A–**ier worden Hij wil mo**–1&A–**lenaar worden Hij wil wiel**–1&A–**renner worden Hij wil le**–1&A–**raar worden
1 italicized word	Alleen *jouw* mening is van belang Onder *die***–1–** voorwaarden doen we mee**–A–** Aan *haar***–1–** mening hebben we ma**–A–**ling
2 italicized words	Ze zijn *min***–1–***der* mooi en *min***–1–***der* warm**–A–** Ze zijn *helemaal* niet mooier en *helemaal* niet warmer Ze zijn *heel***–1–** mooi en *heel***–1–** warm Ze zijn niet *zo***–1–***maar* mooi**–2–B–** en ze zijn niet *zo***–1&A–***maar* warm**–2–**
one-word sentence	Ja**–A–** Nee**–1&A–**
questions	Woon je daar al weer een jaar**–2–**? Wat zeg je nou? - I Wat zeg je nou? - II Ben jij do**–1–**minee**–2–**? Annemarie**–2–**? En wat doe jij**–A&2–**? Hoe lang woon jij nou al in Breda? Wie wordt er boerin**–2–**?

Appendix E

Appendix E: Actual parameter scores.

To give an impression of the actual parameter scores associated with the extralinguistic factors that were controlled in this study, the mean scores per factor are presented. First the scores found in the fragments material reported on in Chapter 3 are shown: the mean scores and standard deviations of the two speech styles, the two sex groups, the five age groups and the five paragraphs. Next the parameter scores in the sports sentences material reported in Chapter 4 are presented: the mean scores and standard deviations of the two sex groups, the five age groups, the three sentences and the two sessions. Finally we turn to the mean scores of the 50 speakers in the read and the spontaneous fragments and in the sports sentences. Due to space limitations the personal standard deviations are not reported here.
The measures of the parameters are:

F_0MEAN:	Hz	F_0END:	Hz	DURFIL:	ms
CVP:	ms	SEMRI1:	ST	SLODEC:	ST/s
PPQ:	ms	SEMRI2:	ST	SLORI1:	ST/s
PZR:	–	SEMFAL:	ST	SLORI2:	ST/s
CVA:	–	LOWRI1:	ST	SLOFAL:	ST/s
APQ:	–	LOWRI2:	ST	SLOFIL:	ST/s
AZR:	–	LOWFAL:	ST	SYNRI1:	ms
PAU:	%	DURRI1:	ms	SYNRI2:	ms
RATE:	syll/s	DURRI2:	ms	SYNFAL:	ms
VOI:	%	DURFAL:	ms	SYNFIL:	ms

1) Mean scores per speech style for the parameters in the speech fragments:

	read	spontaneous
F0MEAN	156.6 (43.1)	142.3 (41.9)
CVP	.1670 (.0295)	.1399 (.0346)
PPQ	1.074 (.251)	1.157 (.382)
PZR	28.54 (4.72)	32.10 (4.73)
CVA	2.799 (.047)	2.878 (.049)
APQ	.6828 (.2014)	.7696 (.2306)
AZR	39.52 (3.21)	44.87 (4.37)
PAU	15.03 (7.01)	14.40 (9.00)
RATE	5.611 (.625)	4.770 (.785)
VOI	75.07 (6.24)	62.74 (8.26)

2) Mean scores per sex group for the parameters in the speech fragments:

	females	males
F0MEAN	188 6 (21 2)	110 3 (13 6)
CVP	1630 (0338)	1439 (0333)
PPQ	1 154 (274)	1 077 (366)
PZR	33 58 (3 66)	27 06 (4 05)
CVA	2 847 (061)	2 829 (062)
APQ	6445 (1803)	8079 (2272)
AZR	43 61 (4 26)	40 78 (4 64)
PAU	12 11 (7 03)	17 32 (8 20)
RATE	5 059 (746)	5 321 (877)
VOI	68 37 (9 43)	69 44 (9 68)

3) Mean scores per age group for the parameters in the speech fragments:

	18-25 yr	26-35 yr	36-45 yr	46-55 yr	56-65 yr
F0MEAN	161 1 (49 5)	148 3 (41 8)	143 5 (39 3)	147 9 (40 9)	146 5 (41 2)
CVP	1425 (0314)	1415 (0303)	1573 (0367)	1569 (0339)	1692 (0341)
PPQ	1 086 (270)	1 029 (234)	1 250 (450)	1 120 (255)	1 093 (333)
PZR	32 00 (4 57)	31 41 (4 89)	30 63 (5 08)	29 30 (4 59)	28 27 (5 19)
CVA	2 835 (059)	2 831 (068)	2 840 (064)	2 846 (061)	2 839 (058)
APQ	7234 (2276)	7153 (2377)	8043 (2499)	6923 (1769)	6957 (1847)
AZR	42 19 (4 40)	42 25 (4 81)	42 44 (4 54)	42 93 (4 65)	41 16 (4 80)
PAU	13 82 (7 05)	14 31 (7 96)	14 03 (8 09)	15 46 (9 20)	15 98 (7 74)
RATE	5 501 (836)	5 122 (871)	5 224 (856)	5 126 (717)	4 979 (744)
VOI	69 87 (9 66)	69 30 (9 48)	66 29 (9 24)	68 57 (9 73)	70 49 (9 25)

4) Mean scores per paragraph for the parameters in the read speech fragments

	par 1	par 2	par 3	par 4	par 5
F0MEAN	157 8 (45 3)	156 0 (43 3)	157 4 (42 5)	156 4 (42 7)	155 2 (42 1)
CVP	1715 (0280)	1685 (0305)	1667 (0299)	1626 (0295)	1655 (0295)
PPQ	1 091 (220)	1 138 (222)	1 087 (240)	1 033 (332)	1 022 (206)
PZR	28 24 (4 97)	29 06 (4 22)	28 18 (4 73)	29 20 (4 70)	28 04 (4 93)
CVA	2 797 (032)	2 844 (030)	2 826 (034)	2 754 (038)	2 772 (035)
APQ	7038 (1690)	8434 (1860)	7535 (1595)	5001 (1261)	6130 (1724)
AZR	40 32 (3 11)	39 93 (3 42)	39 57 (3 33)	38 88 (3 10)	38 91 (2 88)
PAU	13 77 (6 65)	13 26 (7 04)	13 07 (5 97)	18 88 (7 09)	16 20 (6 53)
RATE	5 412 (567)	5 737 (636)	5 635 (605)	5 402 (580)	5 867 (609)
VOI	74 31 (5 34)	71 39 (5 42)	73 93 (5 22)	75 25 (6 02)	80 46 (5 45)

5) Mean scores per sex group for the parameters in the sports sentences:

	females	males
F0MEAN	211.3 (23.4)	127.4 (15.9)
CVP	.2192 (.0489)	.2085 (.0429)
PPQ	1.218 (.337)	1.033 (.298)
PZR	32.81 (3.83)	24.42 (4.02)
CVA	2.681 (.056)	2.682 (.067)
APQ	.2843 (.0964)	.3623 (.1581)
AZR	36.71 (3.81)	34.82 (4.73)
PAU	2.928 (5.281)	3.466 (4.165)
VOI	86.28 (7.02)	85.38 (5.92)
RATE	5.092 (.678)	5.273 (.704)
F0END	154.8 (26.9)	86.9 (13.1)
SEMRI1	8.409 (3.238)	8.767 (2.871)
SEMRI2	6.799 (2.927)	6.180 (2.424)
SEMFAL	7.564 (3.186)	7.811 (2.830)
LOWRI1	4.030 (3.126)	4.930 (3.310)
LOWRI2	.226 (2.552)	1.520 (2.566)
LOWFAL	−.750 (2.929)	−.334 (1.853)
DURRI1	256.3 (98.9)	218.1 (78.4)
DURRI2	209.3 (101.6)	177.7 (76.5)
DURFAL	208.3 (102.6)	226.1 (87.5)
DURFIL	138.9 (84.9)	100.5 (72.7)
SLODEC	−1.91 (.97)	−2.41 (1.15)
SLORI1	36.49 (21.23)	45.76 (27.55)
SLORI2	38.11 (21.68)	39.84 (19.56)
SLOFAL	−43.98 (38.09)	−37.78 (15.43)
SLOFIL	6.61 (34.55)	7.92 (26.60)
SYNRI1	87.21 (60.00)	95.82 (52.30)
SYNRI2	155.8 (85.6)	144.7 (84.8)
SYNFAL	120.3 (91.8)	103.5 (77.0)
SYNFIL	−88.0 (67.5)	−122.7 (72.5)

6) Mean scores per age group for the parameters in the sports sentences·

	18-25 yr	26-35 yr	36-45 yr	46-55 yr	56-65 yr
F0MEAN	181 6 (52 7)	163 6 (45 7)	161 6 (40 2)	170 5 (47 5)	169 3 (44 7)
CVP	2056 (0466)	1962 (0452)	2224 (0488)	2169 (0400)	2282 (0443)
PPQ	1 090 (342)	987 (264)	1 175 (322)	1 216 (273)	1 159 (398)
PZR	29 47 (5 90)	29 08 (6 70)	28 60 (5 12)	28 04 (4 74)	27 89 (6 08)
CVA	2 663 (062)	2 673 (058)	2 692 (064)	2 702 (062)	2 677 (055)
APQ	2826 (1456)	3032 (1180)	3254 (1317)	3944 (1612)	3110 (0919)
AZR	36 41 (4 25)	36 23 (4 27)	35 17 (4 47)	36 77 (4 73)	34 24 (3 86)
PAU	2 427 (3 992)	3 552 (4 666)	3 998 (4 731)	2 698 (4 134)	3 309 (5 980)
VOI	87 34 (5 83)	85 85 (7 14)	84 46 (6 49)	86 20 (6 70)	85 30 (6 13)
RATE	5 473 (630)	5 143 (670)	5 236 (530)	5 277 (760)	4 782 (698)
F0END	139 8 (45 0)	123 4 (40 5)	114 5 (36 8)	116 1 (36 0)	110 0 (35 6)
SEMRI1	7 387 (2 659)	8 092 (2 667)	8 854 (3 687)	9 114 (2 998)	9 496 (2 792)
SEMRI2	5 821 (2 619)	6 013 (2 869)	5 972 (2 368)	6 958 (2 665)	7 685 (2 554)
SEMFAL	6 558 (2 176)	6 289 (2 542)	7 135 (2 513)	9 540 (2 998)	8 915 (3 288)
LOWRI1	3 727 (3 188)	3 598 (2 852)	5 147 (3 123)	4 624 (3 290)	5 301 (3 461)
LOWRI2	075 (1 697)	202 (2 450)	812 (2 435)	1 812 (2 958)	1 464 (3 007)
LOWFAL	− 835 (1 360)	− 130 (1 252)	− 581 (3 045)	− 931 (2 596)	− 232 (3 260)
DURRI1	231 3 (87 7)	254 0 (95 2)	253 2 (98 8)	227 2 (95 0)	220 5 (74 6)
DURRI2	175 5 (79 7)	222 3 (112 9)	190 8 (69 6)	179 0 (97 4)	199 7 (85 2)
DURFAL	211 8 (89 9)	246 3 (115 0)	190 5 (80 7)	212 3 (87 9)	225 2 (95 4)
DURFIL	128 3 (75 6)	129 5 (83 0)	112 0 (71 7)	116 0 (66 3)	112 5 (104 7)
SLODEC	−1 99 (1 05)	1 85 (87)	−2 33 (1 03)	−2 27 (1 24)	−2 37 (1 15)
SLORI1	36 35 (24 03)	36 02 (19 11)	38 34 (19 89)	45 23 (23 22)	49 67 (33 73)
SLORI2	37 96 (21 68)	32 49 (20 90)	35 74 (18 50)	44 73 (19 86)	43 95 (20 01)
SLOFAL	−33 88 (12 4)	28 45 (12 3)	−41 72 (16 0)	−49 98 (19 0)	−50 36 (54 9)
SLOFIL	12 36 (25 76)	1 18 (11 26)	4 80 (20 80)	12 29 (41 11)	5 68 (42 33)
SYNRI1	66 5 (54 5)	78 1 (50 7)	86 9 (52 5)	107 1 (50 7)	118 9 (58 1)
SYNRI2	131 6 (76 5)	163 4 (105 1)	156 7 (80 9)	143 6 (80 2)	155 9 (79 3)
SYNFAL	100 4 (80 7)	138 3 (104 9)	88 2 (66 6)	112 5 (72 9)	120 0 (88 7)
SYNFIL	−111 4 (56 0)	−108 0 (76 2)	−102 3 (69 0)	−99 8 (63 2)	−105 2 (92 4)

7) Mean scores per sentence for the parameters in the sports sentences

	sentence 1	sentence 2	sentence 3
F0MEAN	169 6 (47 0)	170 8 (46 5)	167 5 (46 6)
CVP	2098 (0421)	2356 (0458)	1961 (0421)
PPQ	1 013 (262)	1 292 (366)	1 071 (291)
PZR	28 66 (5 91)	29 62 (5 44)	27 56 (5 74)
CVA	2 666 (059)	2 719 (051)	2 658 (056)
APQ	2935 (1226)	3622 (1366)	3132 (1410)
AZR	35 55 (4 23)	35 98 (4 57)	35 76 (4 40)
PAU	3 066 (4 181)	3 635 (4 656)	2 890 (5 369)
VOI	86 30 (5 62)	81 99 (5 59)	89 20 (6 17)
RATE	5 077 (641)	5 106 (681)	5 364 (734)
F0END	119 5 (40 4)	120 3 (38 9)	122 7 (41 2)
SEMRI1	9 234 (2 435)	7 736 (3 747)	8 795 (2 668)
SEMRI2	5 289 (2 221)	7 218 (2 731)	6 963 (2 724)
SEMFAL	7 162 (2 653)	8 061 (3 073)	7 839 (3 233)
LOWRI1	3 850 (2 510)	5 839 (3 701)	3 751 (3 002)
LOWRI2	1 718 (2 577)	715 (2 673)	386 (2 551)
LOWFAL	651 (2 522)	− 276 (2 093)	− 698 (2 714)
DURRI1	245 8 (96 7)	219 2 (102 4)	246 7 (68 99)
DURRI2	216 8 (96 4)	179 6 (74 15)	184 0 (97 37)
DURFAL	217 7 (87 5)	234 5 (101 5)	199 5 (95 1)
DURFIL	101 6 (70 0)	148 7 (79 7)	108 7 (85 8)
SLODEC	2 26 (99)	−2 29 (1 09)	−1 94 (1 15)
SLORI1	41 91 (16 0)	43 51 (37 2)	38 0 (15 2)
SLORI2	28 23 (17 37)	44 16 (18 22)	44 53 (21 84)
SLOFAL	−40 73 (44 22)	−38 73 (16 89)	−43 18 (17 88)
SLOFIL	7 56 (34 76)	3 48 (16 69)	10 76 (36 73)
SYNRI1	76 9 (35 9)	90 8 (73 0)	106 8 (50 2)
SYNRI2	146 8 (76 9)	158 4 (92 3)	145 6 (86 2)
SYNFAL	120 5 (76 3)	109 5 (84 3)	105 8 (93 7)
SYNFIL	−97 3 (63 1)	−125 0 (80 2)	−93 7 (68 5)

Mean scores per speaker for the 10 TI parameters in the *read* fragments:

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAU	RATE	VOI
F11	234.9	.1834	1.199	36.00	2.781	0.508	40.73	12.24	6.295	80.82
F12	214.1	.1520	1.094	35.72	2.841	0.586	43.02	14.21	5.812	76.72
F13	211.2	.1570	1.072	33.65	2.791	0.445	40.28	11.04	5.801	81.48
F14	219.4	.1725	1.080	33.25	2.772	0.509	44.52	10.25	5.083	76.43
F15	211.1	.1615	1.362	32.48	2.808	0.633	40.44	17.68	6.635	77.08
F21	174.5	.1848	1.175	33.84	2.835	0.740	46.34	8.91	5.669	65.76
F22	200.6	.1837	1.216	32.24	2.822	0.626	38.31	17.09	6.052	78.50
F23	199.0	.1339	1.069	35.56	2.792	0.668	43.41	9.02	5.377	77.31
F24	178.3	.1455	0.985	31.97	2.757	0.484	43.56	12.86	4.787	77.30
F25	214.6	.1426	0.996	35.70	2.796	0.451	41.65	17.69	5.220	77.57
F31	202.2	.1582	1.089	32.32	2.840	0.709	41.59	15.22	5.845	72.27
F32	174.2	.1792	1.068	30.42	2.828	0.603	40.10	8.46	4.975	64.88
F33	186.5	.1322	1.154	36.84	2.781	0.594	43.90	1.97	4.754	71.20
F34	174.8	.2100	1.197	29.44	2.869	0.615	40.10	15.01	5.329	61.81
F35	186.9	.1465	0.953	32.61	2.778	0.652	39.64	10.57	5.027	72.23
F41	177.6	.1510	1.014	28.88	2.813	0.709	39.70	13.44	5.115	71.18
F42	210.2	.1962	1.141	32.11	2.835	0.548	42.80	4.98	5.956	76.17
F43	198.2	.1802	1.000	30.52	2.786	0.447	38.74	13.65	5.265	79.52
F44	193.4	.1868	1.088	30.75	2.836	0.720	38.87	14.05	5.349	70.54
F45	211.3	.1915	1.124	32.97	2.826	0.552	38.87	15.75	4.880	77.77
F51	185.4	.1559	0.984	30.27	2.811	0.655	36.00	10.51	5.233	77.36
F52	160.8	.2283	1.107	24.87	2.804	0.731	37.47	16.52	5.295	76.98
F53	233.1	.2002	1.373	35.60	2.814	0.502	40.25	13.96	5.136	79.15
F54	161.7	.2084	1.006	26.72	2.751	0.482	41.93	7.37	4.898	70.58
F55	187.6	.1804	1.348	31.49	2.822	0.573	40.48	13.36	5.226	71.71
M11	99.6	.1409	1.435	26.58	2.771	0.976	38.56	15.43	5.660	72.05
M12	139.9	.1546	0.754	26.80	2.811	0.716	39.28	16.23	6.550	78.36
M13	128.5	.1571	0.889	26.11	2.799	0.885	38.98	15.89	6.167	80.12
M14	119.6	.1386	0.919	25.21	2.779	0.703	37.41	11.19	5.464	72.22
M15	119.5	.1406	0.798	25.22	2.820	0.771	35.50	20.98	6.002	73.71
M21	100.0	.1461	1.177	24.63	2.766	0.813	36.74	13.68	6.050	74.25
M22	117.6	.1078	0.628	27.77	2.736	0.618	35.57	9.89	4.832	78.29
M23	119.0	.1430	0.950	26.67	2.796	0.876	37.17	23.96	6.565	81.11
M24	134.4	.1600	0.855	24.41	2.766	0.496	38.06	8.45	5.345	70.58
M25	100.5	.1642	1.012	23.61	2.762	0.844	38.07	22.77	6.232	76.86
M31	133.0	.1873	1.173	26.06	2.758	0.792	39.02	15.82	5.931	76.26
M32	110.3	.1694	1.301	24.91	2.810	0.961	36.36	19.59	6.380	74.59
M33	103.9	.1546	1.178	25.49	2.784	0.790	38.96	30.00	6.239	76.95
M34	117.5	.1709	0.947	21.64	2.787	0.789	35.06	16.78	6.375	78.11
M35	102.8	.2393	1.714	26.47	2.815	1.031	42.72	14.00	5.215	67.04
M41	117.5	.1488	1.114	26.37	2.778	0.687	41.94	13.86	5.807	71.00
M42	115.2	.1970	1.316	25.73	2.790	0.725	43.12	17.26	5.491	65.22
M43	123.2	.1826	1.097	22.79	2.809	0.690	36.45	22.18	5.540	79.04
M44	108.6	.1368	0.999	22.41	2.831	0.792	41.25	25.24	5.519	74.88
M45	126.0	.1612	0.977	23.96	2.791	0.743	38.05	19.20	6.303	79.93
M51	124.1	.1667	0.902	25.20	2.824	0.759	38.56	10.86	5.376	75.42
M52	120.2	.1508	0.874	22.93	2.803	0.670	36.89	26.38	6.393	80.24
M53	128.3	.1462	0.993	26.01	2.767	0.748	38.13	17.18	5.313	79.77
M54	117.0	.1840	0.869	20.76	2.801	0.824	34.91	21.07	5.800	78.39
M55	100.5	.1781	0.939	23.21	2.789	0.697	36.67	18.04	4.963	76.64

Mean scores per speaker for the 10 TI parameters in the *spontaneous* fragments:

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAU	RATE	VOI
F11	205.6	.1538	1.269	37.88	2.889	0.695	45.67	7.26	5.679	64.39
F12	198.7	.1390	1.289	38.23	2.927	0.670	47.68	17.07	5.302	62.82
F13	189.5	.1197	0.965	37.16	2.878	0.647	47.58	16.07	4.973	69.37
F14	206.8	.1686	1.112	35.16	2.857	0.686	46.96	12.30	5.024	67.03
F15	191.2	.1434	1.152	33.78	2.893	0.638	43.92	6.28	5.399	63.94
F21	162.6	.1769	1.318	34.93	2.918	0.888	47.75	7.45	4.468	49.09
F22	184.8	.1431	1.151	36.32	2.897	0.556	48.55	12.77	4.198	57.09
F23	179.7	.1106	1.018	37.12	2.896	0.765	46.08	2.55	4.680	64.50
F24	178.4	.1338	0.919	35.18	2.854	0.568	47.82	11.75	4.122	65.05
F25	200.5	.1557	1.061	37.78	2.850	0.609	47.40	13.29	4.226	62.90
F31	191.5	.1513	1.633	36.22	2.883	0.990	48.85	7.69	5.147	55.47
F32	161.4	.1246	1.127	34.53	2.883	0.630	45.72	6.81	3.937	54.71
F33	180.3	.1511	1.104	35.35	2.866	0.701	45.90	7.23	4.401	68.58
F34	176.8	.1711	1.455	35.25	2.956	0.631	47.36	21.34	5.298	57.24
F35	167.5	.1402	0.857	33.76	2.842	0.706	44.88	9.97	4.250	63.91
F41	153.5	.1488	1.371	32.56	2.895	0.736	45.96	18.44	4.888	61.47
F42	185.8	.1155	1.041	36.99	2.852	0.557	51.05	3.60	4.062	58.20
F43	166.9	.1394	0.919	33.77	2.910	0.551	41.63	17.55	4.591	70.39
F44	167.6	.1640	1.335	33.59	2.912	0.865	46.69	5.63	4.655	53.21
F45	189.4	.1781	1.142	35.08	2.928	0.800	44.25	20.87	5.106	63.60
F51	184.9	.1784	1.144	33.94	2.903	0.550	41.81	15.96	4.418	75.41
F52	148.6	.1843	1.064	25.68	2.871	0.966	41.26	15.90	4.866	63.10
F53	219.6	.1923	1.789	37.84	2.872	0.656	48.86	14.12	4.779	64.98
F54	164.2	.1983	1.120	29.41	2.882	0.675	45.98	13.10	4.782	61.96
F55	172.5	.1468	1.462	35.48	2.865	0.743	48.31	14.75	4.707	57.49
M11	89.5	.1185	1.474	32.32	2.832	0.900	47.03	12.07	4.363	52.04
M12	110.2	.0838	0.794	30.15	2.849	0.950	36.54	19.91	5.439	65.49
M13	112.5	.1540	1.162	32.78	2.883	1.096	46.67	12.12	5.707	68.32
M14	112.0	.1159	1.077	31.91	2.879	0.700	43.04	7.59	4.071	57.09
M15	107.3	.0951	0.822	29.61	2.845	0.755	40.05	20.61	4.596	57.92
M21	93.8	.1317	1.323	28.14	2.865	0.870	42.62	14.76	4.611	60.64
M22	110.8	.1007	0.791	33.69	2.909	0.648	44.02	19.65	4.201	74.32
M23	113.2	.1079	0.912	31.20	2.878	1.098	40.74	21.06	5.289	67.33
M24	110.7	.1396	0.821	27.74	2.884	0.627	43.94	12.82	4.901	64.28
M25	92.2	.1176	1.201	29.79	2.837	1.063	37.13	25.70	5.605	63.20
M31	116.4	.1280	1.671	34.44	2.873	1.085	48.96	12.37	5.265	58.78
M32	107.1	.1536	1.104	26.48	2.801	0.821	41.15	14.19	4.898	67.52
M33	85.1	.1123	1.234	28.70	2.854	0.835	41.57	18.97	5.626	68.74
M34	103.3	.1102	0.880	27.51	2.879	0.805	41.29	15.52	4.622	64.01
M35	88.3	.1549	2.152	34.13	2.916	1.345	45.78	19.07	4.963	51.58
M41	113.0	.1117	1.007	29.72	2.807	0.585	49.60	5.09	4.586	67.96
M42	84.8	.1653	1.686	29.67	2.821	0.847	47.86	6.11	4.246	49.45
M43	98.8	.1235	1.084	26.79	2.881	0.799	44.00	21.39	5.312	68.10
M44	106.5	.1517	1.151	25.27	2.873	0.815	45.89	19.60	4.571	61.23
M45	109.5	.1070	0.793	26.11	2.936	0.678	41.97	31.26	5.272	72.44
M51	111.3	.1500	0.935	27.40	2.872	0.829	42.63	15.96	4.370	64.29
M52	113.0	.1277	1.088	28.19	2.877	0.720	47.83	21.15	5.162	63.46
M53	96.3	.1318	1.164	29.72	2.883	0.708	43.39	11.53	4.006	62.35
M54	108.0	.1341	0.719	22.96	2.869	0.735	38.14	30.27	4.624	71.38
M55	93.4	.1416	0.973	27.68	2.895	0.689	43.63	11.52	4.224	59.11

Mean scores per speaker for the 10 *TI* parameters in the sports sentences:

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	PAU	RATE	VOI
F11	250.2	.2178	1.273	36.59	2.659	0.207	36.61	0.32	6.356	89.72
F12	226.6	.2069	1.309	36.65	2.650	0.285	38.71	0.40	5.880	90.15
F13	217.2	.2091	1.037	30.51	2.689	0.173	34.46	0.38	5.634	91.88
F14	238.8	.2495	1.281	34.96	2.634	0.187	41.97	0.94	4.371	90.69
F15	218.8	.1966	1.360	30.97	2.642	0.285	33.60	0.67	6.059	87.51
F21	185.9	.2517	1.044	33.72	2.691	0.262	43.34	4.28	4.401	72.63
F22	222.2	.2545	1.282	34.63	2.691	0.261	35.60	1.51	5.259	87.77
F23	211.2	.1530	1.011	36.67	2.679	0.206	36.27	2.22	4.769	88.46
F24	186.3	.1543	0.979	33.30	2.631	0.235	39.54	1.04	4.890	91.57
F25	223.5	.2156	1.077	34.33	2.704	0.210	37.95	12.16	4.017	87.78
F31	221.6	.2174	1.439	32.34	2.695	0.331	39.61	2.45	5.451	78.85
F32	187.1	.2360	1.070	31.76	2.744	0.259	37.71	3.37	4.827	80.27
F33	192.5	.1663	1.274	35.80	2.658	0.272	37.73	0.00	5.215	90.18
F34	198.6	.2844	1.153	27.97	2.754	0.193	31.67	12.31	4.531	78.08
F35	195.0	.1712	1.035	32.85	2.618	0.331	35.90	0.06	5.374	91.70
F41	192.5	.1919	1.215	31.16	2.678	0.267	34.49	1.62	5.053	86.43
F42	238.4	.2571	1.518	32.20	2.710	0.287	39.20	0.37	5.813	91.43
F43	218.1	.2040	1.184	29.14	2.657	0.230	34.32	0.41	5.001	90.93
F44	209.5	.2189	1.058	30.90	2.724	0.325	35.82	2.78	4.039	82.21
F45	218.6	.2363	1.151	32.50	2.671	0.266	35.83	0.74	4.432	89.16
F51	200.7	.2044	1.181	35.00	2.735	0.245	32.10	4.93	4.309	87.83
F52	173.0	.2823	1.225	27.36	2.647	0.305	35.96	6.78	4.669	87.47
F53	264.7	.2404	1.818	37.72	2.663	0.281	39.38	0.31	4.701	84.75
F54	188.9	.2396	1.009	28.58	2.699	0.270	34.48	12.67	3.512	77.05
F55	201.9	.2218	1.466	32.56	2.694	0.251	35.38	0.49	5.181	82.50
M11	98.2	.1495	1.103	25.16	2.642	0.348	36.13	1.07	5.495	82.36
M12	156.0	.1985	0.764	26.57	2.644	0.215	36.29	9.14	4.718	86.52
M13	149.2	.2287	0.961	26.31	2.731	0.365	37.31	1.40	5.208	89.12
M14	128.7	.1983	0.927	23.22	2.620	0.240	36.10	1.20	5.089	82.14
M15	132.3	.2012	0.883	23.72	2.723	0.266	32.91	8.76	4.686	83.25
M21	109.6	.1652	1.072	21.34	2.682	0.366	35.11	0.79	5.938	87.71
M22	121.4	.1559	0.643	24.13	2.716	0.303	33.67	7.34	4.326	89.32
M23	127.4	.1969	0.937	23.61	2.612	0.305	33.22	3.18	6.045	86.46
M24	144.1	.2191	0.776	29.01	2.667	0.242	34.10	1.25	4.494	85.84
M25	104.1	.1957	1.049	20.03	2.655	0.315	33.52	1.75	5.563	80.91
M31	144.9	.1942	1.248	25.74	2.617	0.381	36.87	1.65	4.852	81.46
M32	119.0	.2175	1.072	21.88	2.706	0.254	32.96	5.86	5.453	88.46
M33	111.2	.2173	1.045	27.01	2.744	0.378	31.36	7.18	4.900	83.19
M34	122.2	.2232	1.012	24.99	2.689	0.216	31.59	1.00	5.648	89.72
M35	123.7	.2968	1.405	25.66	2.694	0.261	36.28	6.11	4.068	82.71
M41	125.5	.1720	1.023	26.89	2.678	0.301	38.69	1.10	6.032	88.81
M42	122.8	.2404	1.506	26.33	2.702	0.417	41.62	9.28	4.900	75.32
M43	130.9	.2272	1.119	24.59	2.700	0.304	35.98	3.81	5.242	84.40
M44	113.5	.1895	1.380	22.85	2.816	0.527	39.45	6.80	4.603	83.10
M45	134.9	.2313	1.007	23.81	2.687	0.486	32.21	0.08	6.265	90.23
M51	151.2	.2475	1.064	23.34	2.681	0.273	34.83	3.81	3.783	82.27
M52	132.1	.1694	0.916	22.66	2.681	0.251	34.79	1.28	4.794	80.56
M53	142.3	.2071	1.009	26.94	2.694	0.295	34.79	1.81	4.517	90.29
M54	117.8	.2086	0.987	19.92	2.634	0.279	30.75	0.81	6.114	90.96
M55	120.7	.2604	0.909	24.82	2.645	0.315	29.89	0.19	4.662	89.30

Mean scores per speaker for the 20 *CB* parameters in the sports sentences:

	F_0END	SEMRI1	SEMRI2	SEMFAL	LOWRI1	LOWRI2	LOWFAL	DURRI1	DURRI2	DURFAL
F11	186.8	8.06	4.64	5.52	4.50	0.17	0.85	176.7	163.3	186.7
F12	201.6	7.81	2.64	3.04	-0.28	-0.01	0.34	230.0	141.7	165.0
F13	179.4	6.71	8.17	7.61	1.63	-1.24	0.74	200.0	166.7	168.3
F14	183.6	9.34	9.08	10.05	2.98	-1.45	2.65	263.3	290.0	340.0
F15	160.9	5.18	5.50	5.59	5.65	0.93	-0.56	201.7	101.7	151.7
F21	116.7	10.17	10.00	9.14	6.79	-0.61	0.16	270.0	281.7	225.0
F22	170.4	8.65	5.19	4.96	3.30	-0.79	0.66	318.3	155.0	161.7
F23	182.1	6.15	4.80	3.68	1.68	-1.32	0.82	265.0	195.0	271.7
F24	144.8	6.52	5.90	6.02	3.69	0.58	-0.80	285.0	166.7	230.0
F25	177.9	9.34	9.07	7.95	2.72	-1.65	0.61	316.7	428.3	433.3
F31	164.2	7.10	5.75	8.90	5.06	-0.55	3.68	218.3	223.3	150.0
F32	140.5	9.79	5.44	5.78	3.66	0.08	0.37	406.7	165.0	191.7
F33	151.5	4.45	5.20	5.67	3.55	2.08	-0.84	253.3	111.7	166.7
F34	137.8	11.00	3.62	5.85	5.36	-0.88	3.28	330.0	245.0	93.3
F35	143.1	6.96	4.94	6.50	4.46	2.02	-0.40	321.7	198.3	183.3
F41	143.7	7.19	6.35	7.83	2.94	-0.03	1.47	190.0	165.0	235.0
F42	167.0	12.44	10.23	9.83	2.30	-0.21	0.06	281.7	203.3	180.0
F43	151.9	7.94	6.73	10.89	4.45	1.94	2.61	193.3	175.0	201.7
F44	147.0	7.08	7.93	9.81	6.15	0.51	1.39	263.3	336.7	308.3
F45	132.8	9.46	6.13	10.59	7.72	4.97	-0.50	336.7	131.7	191.7
F51	128.2	5.59	6.70	5.21	7.71	1.50	-2.59	206.7	315.0	241.7
F52	108.2	12.42	10.26	10.07	5.49	-0.69	0.41	238.3	305.0	211.7
F53	190.5	12.05	8.75	11.62	0.91	-1.09	4.12	188.3	170.0	165.0
F54	118.7	9.48	8.98	9.61	4.57	2.04	-1.37	255.0	205.0	200.0
F55	140.9	9.37	7.99	7.38	3.76	-0.66	1.56	198.3	191.7	155.0
M11	85.4	5.23	5.14	6.02	2.41	-1.30	2.16	218.3	158.3	195.0
M12	107.8	7.92	5.16	6.75	5.70	1.77	0.39	288.3	218.3	270.0
M13	98.7	8.41	7.54	8.44	4.97	1.02	0.36	276.7	171.7	190.0
M14	92.8	6.67	4.52	6.64	6.55	1.02	0.87	230.0	130.0	195.0
M15	100.8	8.53	5.83	5.91	3.16	-0.17	0.55	228.3	213.3	256.7
M21	79.4	8.33	5.73	7.15	3.18	0.87	0.63	155.0	158.3	246.7
M22	95.4	6.34	4.40	5.69	2.89	0.64	0.64	236.7	185.0	228.3
M23	95.9	6.69	5.89	7.48	4.13	1.51	0.11	236.7	223.3	268.3
M24	91.3	8.60	3.45	5.96	5.54	3.83	-1.73	141.7	210.0	241.7
M25	80.3	10.12	5.71	4.86	2.06	-1.06	0.19	315.0	220.0	156.7
M31	99.9	9.39	5.32	7.28	4.59	2.76	-0.42	230.0	176.7	275.0
M32	85.0	8.20	6.56	7.46	5.94	0.02	0.76	150.0	186.7	220.0
M33	79.8	7.60	4.93	5.98	4.26	0.61	-0.03	181.7	188.3	221.7
M34	85.3	9.49	8.21	8.00	4.28	-0.55	0.42	225.0	163.3	205.0
M35	57.9	14.56	9.75	9.93	10.32	2.53	-1.02	215.0	250.0	198.3
M41	87.7	7.40	3.76	5.81	5.65	2.12	-0.06	175.0	85.0	143.3
M42	68.9	13.01	9.54	13.50	5.05	3.51	0.46	156.7	166.7	190.0
M43	90.6	9.17	8.25	9.33	4.73	0.90	0.69	241.7	228.3	253.3
M44	82.7	7.93	6.98	8.14	1.75	0.81	0.60	213.3	186.7	220.0
M45	88.8	9.52	3.67	9.66	5.49	3.61	2.59	220.0	111.7	200.0
M51	89.3	10.16	9.04	13.34	7.83	4.03	0.23	275.0	195.0	386.7
M52	88.7	9.79	5.83	6.69	2.46	2.81	-0.32	186.7	141.7	161.7
M53	89.1	8.62	6.92	8.45	6.38	2.59	-0.41	235.0	166.7	265.0
M54	76.5	8.29	6.82	9.64	5.60	2.38	0.42	206.7	155.0	253.3
M55	74.3	9.19	5.55	7.14	8.32	1.71	0.25	215.0	151.7	211.7

	DURFIL	SLODEC	SLORI1	SLORI2	SLOFAL	SLOFIL	SYNRI1	SYNRI2	SYNFAL	SYNFIL
F11	141.7	-3.28	45.3	29.9	-32.0	7.1	51.0	128.1	72.5	-114.1
F12	181.7	-1.67	52.9	18.9	-21.8	2.5	47.7	104.5	91.8	-73.2
F13	213.3	-1.15	34.3	48.6	-45.9	4.3	95.4	101.8	103.2	-65.1
F14	170.0	-1.34	36.4	32.2	-30.7	17.9	78.6	176.2	183.4	-156.6
F15	126.7	-2.51	25.9	65.2	-40.3	-3.9	36.4	113.0	79.3	-72.3
F21	21.7	-1.95	39.1	42.9	-41.6	1.2	49.0	165.7	153.5	-71.5
F22	191.7	-1.84	32.0	47.5	-31.8	4.0	90.5	120.2	91.8	-69.9
F23	128.3	-1.29	23.2	25.0	-14.3	4.9	60.2	162.0	146.3	-125.4
F24	185.0	-1.90	22.8	34.9	-28.1	-3.8	98.1	133.1	106.2	-123.8
F25	170.0	-1.25	31.9	22.3	-19.8	5.8	110.9	270.8	347.3	-86.0
F31	71.7	-2.88	31.5	26.1	-56.6	9.1	47.2	174.7	110.6	-39.4
F32	115.0	-1.92	23.7	37.5	-33.4	4.8	76.4	142.4	100.1	-91.6
F33	143.3	-1.73	18.7	50.3	-41.7	-3.1	81.8	75.9	84.8	-81.9
F34	103.3	-2.05	34.0	19.5	-59.8	29.8	37.3	258.9	17.3	-76.1
F35	180.0	-2.32	21.9	26.8	-39.2	-1.7	79.2	111.9	88.3	-95.0
F41	125.0	-1.09	39.6	40.9	-34.6	12.0	84.8	156.5	92.6	-142.4
F42	125.0	-2.11	47.0	57.4	-62.9	4.3	109.5	125.9	139.6	-40.4
F43	103.3	-1.63	42.5	42.5	-55.6	48.7	95.7	86.5	43.1	-158.6
F44	140.0	-1.92	27.4	26.6	-34.6	10.7	87.8	271.7	197.6	-110.8
F45	140.0	-2.05	32.0	49.5	-59.6	-3.9	161.0	101.4	88.1	-103.6
F51	190.0	-2.79	29.8	21.8	-24.1	-11.5	116.9	250.9	163.5	-78.2
F52	58.3	-2.02	55.0	36.1	-50.2	-5.2	132.3	236.7	124.3	-87.4
F53	158.3	-0.83	76.4	57.6	-121.5	52.8	138.1	156.3	109.5	-55.5
F54	198.3	-1.64	40.4	48.5	-50.5	-4.5	113.6	153.5	196.3	-103.7
F55	90.0	-2.60	48.5	44.2	-68.8	-17.0	100.7	117.0	77.4	-77.6
M11	93.3	-2.08	29.4	38.8	-31.9	32.1	37.6	127.2	47.9	-147.6
M12	63.3	-2.31	30.6	34.4	-29.8	5.5	82.5	111.1	105.6	-164.4
M13	145.0	-1.68	32.0	44.6	-46.6	2.6	80.9	119.7	59.1	-131.0
M14	85.0	-2.56	35.3	34.8	-34.6	29.8	69.7	133.1	95.1	-99.9
M15	63.3	-1.29	41.4	32.2	-25.2	25.6	85.4	201.7	166.5	-90.2
M21	36.7	-2.48	55.8	42.3	-33.2	1.8	47.0	132.1	90.8	-155.8
M22	118.3	-1.57	27.3	26.9	-25.2	6.1	84.8	47.6	60.9	-167.5
M23	145.0	-1.53	27.8	27.4	-30.1	0.4	45.5	165.8	176.7	-91.7
M24	208.3	-2.75	67.0	25.3	-27.0	-9.2	88.9	257.9	122.0	-119.6
M25	90.0	-2.00	33.2	30.4	-33.4	0.6	105.9	179.2	87.4	-69.2
M31	63.3	-2.20	42.0	36.7	-28.8	4.2	118.3	92.7	122.3	-152.7
M32	143.3	-2.55	54.3	36.3	-35.4	7.6	79.9	152.4	139.9	-80.1
M33	121.7	-1.47	41.0	26.9	-29.2	1.0	120.4	180.0	118.4	-103.3
M34	113.3	-2.34	44.2	54.9	-40.8	7.6	131.8	127.9	60.7	-144.3
M35	65.0	-3.82	72.0	42.4	-52.2	-11.4	97.1	250.6	40.0	-158.3
M41	111.7	-3.77	47.7	47.0	-49.1	-7.1	102.3	107.0	97.5	-45.9
M42	83.3	-2.47	88.1	64.0	-71.2	1.7	113.8	131.6	105.5	-84.5
M43	148.3	-1.91	37.7	46.4	-41.2	4.5	82.9	169.2	146.6	-106.7
M44	81.7	-2.96	41.9	39.5	-39.8	14.6	120.6	186.5	110.7	-109.3
M45	101.7	-2.82	48.4	33.6	-51.2	37.5	112.9	99.5	103.9	-96.1
M51	126.7	-2.90	38.9	52.7	-35.1	17.9	137.1	126.6	220.4	-166.3
M52	33.3	-1.88	76.7	46.7	-48.1	-1.3	100.4	125.7	23.3	-138.4
M53	106.7	-1.77	38.2	41.7	-31.9	-4.2	156.3	135.7	105.1	-159.9
M54	33.3	-3.88	45.4	45.9	-38.6	24.7	76.0	125.1	98.3	-155.1
M55	130.0	-3.39	47.5	44.1	-34.8	5.2	117.7	131.8	82.2	-129.5

Appendix F

Appendix F: Raw percentages of correct identification.

To make the outcomes of the analyses in this book mutually comparable, we reported the percentages of identification exceeding chance (Klecka's tau, see Klecka, 1980), instead of the percentages of correct identification per se. The latter scores are reported below for all LDA's reported in this study The table numbers correspond to the original table numbers, with the addition of the letter "A".

Table 3.7A
Percentages of correct assignment (c a.) overall, within sessions (w s) and the cross-validation IS (c v) in LDA's with speakers as groups, over (1) total material (2) read material (3) spontaneous material (4) females and (5) males

	total	read-out	spontaneous	females	males
c a.	60 9	88 2	70 8	66 4	64 6
w s	72 1	96 6	89 6	76 6	72 2
c v	34 3	54 8	30 8	35 8	33 0

Table 3.8A
Percentages of correct identification in LDA's with all TI parameters *except* mean F_0, and with mean F_0 *only* F_0, with the speakers as groups over (1) the total material (2) the read material (3) spontaneous material (4) the females and (5) the males

	total	read	spontaneous	females	males
c a no F_0	48 9	79 6	57 0	53 8	56 0
c a only F_0	10 1	20 2	17 4	13 6	8 4

Table 3.9A
Percentages of correct identification in LDA's with speakers as groups per sex and per speech style, over (1) read material, female speakers (2) spontaneous material, female speakers (3) read material, male speakers and (4) spontaneous material, male speakers

	female speakers		male speakers	
	read	spontaneous	read	spontaneous
c a	90 4	69 6	88 4	76 4
w s	98 0	89 6	95 2	94 8
c v	57 2	28 4	51 6	29 2

Table 3.10A
Percentages of correct identification, IS within sessions and cross-validation IS in LDA's with the speech styles as groups, over (1) the total material (2) the females and (3) the males

	total	female speakers	male speakers
c a.	93 5	93 8	95 6
w s	94 0	93 8	96 2
c v	93 3	91 8	95 4

Table 3.11A
Percentages of correct identification, IS within sessions and cross-validation IS in LDA's with the sexes as groups, over (1) the total material (2) the read-out material and (3) the spontaneous material

	total	read	spontaneous
c a	98 5	98 8	99 0
w s	98 7	99 0	99 0
c v	98 5	98 8	99 0

Table 3.12A
Percentages of correct identification in LDA's with the sexes as groups over all TI parameters *except* F_0MEAN and for F_0MEAN *only* over (1) the total material (2) the read material and (3) the spontaneous material

	total	read	spontaneous
corr class no F_0MEAN	90 0	95 0	86 0
corr class only F_0MEAN	98 6	98 4	99 2

Table 3.13A
Percentages of correct identification, IS within sessions and the cross-validation IS in LDA's with the age groups as discriminant groups, over (1) the total material (2) the read material, (3) the spontaneous material (4) the females and (5) the males

	total	read	spontaneous	females	males
c a	41 0	44 6	44 4	48 6	50 2
w s	43 1	47 2	46 4	51 4	51 0
c v	37 5	39 2	35 2	44 6	42 8

Table 3.14A
Percentages of correct identification, IS within sessions and the cross-validation IS in LDA's with the paragraphs as groups, over (1) the total material (2) the females and (3) the males

	total	females	males
c a	61 0	65 6	65 2
w s	63 0	63 6	67 6
c v	56 2	54 4	59 6

Table 3.15A
Percentages of correct identification, IS within sessions and the cross-validation IS in LDA's with the sessions as groups, over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males

	total	read	spontaneous	females	males
c a	57 6	61 4	58 6	57 8	60 0

Table 3.16A
Percentages of correct identification in LDA's with speakers as groups and 200 fragments of 75 seconds as the experimental cases (top row), and the percentages for LDA's with 1000 fragments of 15 seconds (bottom row), over (1) the total material (2) the read material (3) the spontaneous material (4) the females and (5) the males

	total	read	spontaneous	females	males
c a 75 s	87 0	99 0	88 0	92 0	82 0
c a 15 s	60 9	88 2	70 8	66 4	64 6

Table 4.5A
Percentages of correct identification, IS within sessions and the cross-validation IS in LDA's with TI parameters as predictors and with speakers as groups, over (1) the total material (2) the females and (3) the males

	total	females	males
c a	81 7	84 0	86 7
w s	92 0	89 3	93 3
c v	42 7	46 0	40 7

Table 4.6A
Percentages of correct identification in LDA's with all TI parameters *except* the F_0MEAN, and with *only* F_0MEAN as the predictors and with speakers as groups, over (1) the total material (2) the females only and (3) the males

	total	females	males
c a no F_0MEAN	68 7	74 7	76 0
c a only F_0MEAN	20 3	24 0	18 0

Table 4.7A
Percentages of correct identification, the IS within sessions and the cross-validation IS in LDA's with the CB parameters as predictors and with speakers as groups, over (1) the total material (2) the females and (3) the males

	total	females	males
c a	86 7	92 0	84 0
w s	86 7	90 7	69 3
c v	43 0	45 3	20 0

Table 4.8A
Percentages of correct identification in LDA's with all CB parameters *except* F_0END, and with *only* F_0END as predictors, with speakers as groups, over (1) the total material (2) the females and (3) the males

	total	females	males
c a no F_0END	67 3	78 0	69 3
c a only F_0END	14 7	20 0	13 3

Table 4.9A
Percentages of correct identification in LDA's with both the TI and the CB parameters as predictors and with speakers as groups, over (1) the total material (2) the females and (3) the males

	total	females	males
c a	97 0	98 0	96 0
w s	99 3	98 0	95 3
c v	52 3	52 0	43 3

Table 4.10A
Percentages of correct identification in LDA's with both the TI and the CB parameters as predictors, but *without* the predictor variables F_0MEAN and F_0END, and with *only* F_0MEAN and F_0END, with speakers as groups, over (1) the total material (2) the females and (3) the males

	total	females	males
c a no F_0MEAN or F_0END	91 3	92 0	91 3
c a only F_0MEAN and F_0FND	35 0	40 0	31 3

Table 4.11A
Percentages of correct identification in LDA's with speakers as groups, over the total material, the females and the males The first data column specifies the percentages of correct identification of LDA's with only TI, the second with only CB, and the third with both parameter types combined In parentheses, the IS within sessions and the cross-validation IS are presented

	time-integrated	contour-bound	both types
all material	81 3 (92 3, 42 7)	86 7 (86 7, 43 0)	97 3 (99 3, 53 3)
females	84 0 (88 7, 40 7)	92 0 (90 7, 45 3)	98 0 (98 7, 53 3)
males	86 7 (94 0, 40 0)	84 0 (69 3, 20 0)	96 0 (95 3, 44 7)

Table 4.12A
Percentages of correct identification in LDA's with speakers as groups, into which only ten parameters were entered LDA's were performed over the total material, the females and the males The first data column specifies the percentages of correct identification in LDA's with only TI, the second with only CB, and the third with both parameter types combined In parentheses, the IS within sessions and the cross-validation IS are presented

	time-integrated	contour-bound	both types
all material	81 7 (92 0, 42 7)	77 3 (86 7, 43 0)	84 7 (93 0, 44 7)
females	84 0 (89 3, 46 0)	86 7 (90 7, 45 3)	84 7 (98 7, 48 0)
males	86 7 (93 3, 40 7)	76 3 (69 3, 20 0)	88 0 (95 3, 45 3)

Table 4.13A
Percentages of correct identification in LDA's with sexes as groups, with (1) all predictors (2) all predictors except F_0MEAN and F_0END, and (3) only F_0MEAN and F_0END The first data column specifies the percentages of correct identification in LDA's with only TI, the second with only CB, and the third with both parameter types combined In parentheses, the IS between sessions and the cross-validation IS are presented

	time-integrated	contour-bound	both TI and CB
all material	98 7 (98 7, 98 7)	99 0 (98 3, 96 0)	99 7 (99 7, 99 3)
all mat , except F_0MEAN and F_0END	88 7 (88 7, 86 7)	69 7 (64 7, 61 0)	89 0 (89 3, 84 7)
only F_0MEAN and/or F_0END	98 7 (98 7, 98 7)	92 7 (93 3, 93 0)	98 3 (98 7, 98 7)

Table 4.14A
Percentages of correct identification in LDA's with age groups as experimental groups, over the total material, the females and the males The first data column specifies the percentages of correct identification of LDA's with only TI, the second with only CB, and the third with both parameter types combined In parentheses, the IS within sessions and the cross-validation IS are presented

	time-integrated	contour-bound	both types
all material	38 0 (33 3, 27 0)	40 3 (38 0, 32 0)	46 7 (38 0, 33 3)
females	52 0 (43 3, 39 3)	65 3 (64 0, 51 3)	73 3 (70 7, 52 7)
males	40 0 (34 0, 22 7)	45 3 (40 7, 28 0)	48 7 (41 3, 27 3)

Table 4.15A
Percentages of correct identification in LDA's with sentences as groups, over the total material, the females and the males The first data column specifies the percentages of correct identification in LDA's with only TI, the second with only CB, and the third with both parameter types combined In parentheses, the IS within sessions and the cross-validation IS are presented

	time-integrated	contour-bound	both types
all material	56 0 (55 0, 52 3)	67 3 (62 0, 65 7)	78 3 (76 0, 70 7)
females	52 7 (52 0, 46 0)	64 7 (57 3, 46 7)	76 7 (61 3, 52 7)
males	62 7 (58 0, 54 7)	67 3 (62 0, 59 3)	81 3 (70 0, 65 3)

Table 4.16A
Percentages of correct identification in LDA's with sessions as groups, over the total material, the females and the males The first data column specifies the percentages of correct identification of LDA's with only TI, the second with only CB, and the third with both parameter types combined

	time-integrated	contour-bound	both types
all material	56 7	62 7	62 7
females	56 0	56 0	56 0
males	58 0	63 3	63 3

Appendix G

Appendix G: Correlations for 21 CB parameters and 10 TI parameters.

Matrix with the correlations of 10 TI parameters and 21 CB parameters with the former ones, in the 300 utterances (critical values of r: |.114| ($p < 5$ %) and |.149| ($p < 1$ %), two-tailed)

	F_0MEAN	CVP	PPQ	PZR	CVA	APQ	AZR	RATE	PAUSE	VOI
CVP	18									
PPQ	32	49								
PZR	75	13	38							
CVA	- 05	27	25	03						
APQ	- 32	05	31	- 18	33					
AZR	24	04	29	33	05	02				
RATE	- 12	- 23	04	- 15	- 21	08	- 01			
PAUSE	- 11	20	- 06	- 10	34	00	- 04	- 19		
VOI	13	- 34	- 28	01	- 37	- 22	- 09	33	- 26	
F_0END	91	- 02	25	72	- 07	- 31	23	01	- 14	- 19
SEMDEC	- 19	25	01	- 13	05	08	- 06	- 15	10	11
SEMRI1	00	38	12	- 12	- 04	- 04	12	- 16	11	08
SEMRI2	12	37	28	02	13	00	11	- 24	11	14
SEMFAL	06	33	18	- 11	05	13	10	- 22	06	15
LOWRI1	- 11	37	03	- 07	10	07	- 12	- 17	08	18
LOWRI2	- 18	00	- 12	- 16	- 07	13	- 03	- 13	- 03	- 01
LOWFAL	- 13	- 01	- 07	00	- 01	01	- 05	- 10	00	- 07
DURRI1	17	02	- 18	12	- 07	- 09	09	- 18	12	- 03
DURRI2	15	19	- 04	07	07	- 10	- 01	- 35	32	15
DURFAL	- 03	03	- 13	- 02	07	02	- 06	- 32	20	02
DURFIL	27	05	07	31	12	- 08	02	- 15	- 13	- 12
SLODEC	- 20	16	04	- 13	- 04	11	- 04	17	- 06	03
SLORI1	- 11	21	24	- 11	07	05	08	- 01	- 02	13
SLORI2	- 01	10	24	- 01	02	04	11	11	- 13	- 02
SLOFAL	17	19	31	06	- 01	02	12	04	- 06	12
SLOFIL	- 04	03	01	07	07	07	04	- 06	03	02
SYNRI1	- 04	08	00	- 07	03	05	02	- 18	06	- 11
SYNRI2	02	22	05	02	20	01	- 06	- 23	29	22
SYNFAL	14	04	07	12	09	- 01	05	- 24	23	06
SYNFIL	20	01	09	17	01	- 04	14	14	00	04

Appendix H

Appendix H: Speaker identification within sex-age groups:

parameter type	groups	c.a. within group	cross-validation within group	overall c.a.	overall cross-validation
TI	F 1	66.7	75.0	81.3	41.5
	F 2	95.9	75.0		
	F 3	70.8	41.7		
	F 4	70.8	8.3		
	F 5	100.0	83.3		
	M 1	95.9	66.6		
	M 2	95.9	50.0		
	M 3	79.2	25.0		
	M 4	87.5	16.7		
	M 5	100.0	50.0		
	mean	86.3	49.2		
CB	F 1	87.5	50.0	86.4	41.8
	F 2	83.4	41.7		
	F 3	75.0	33.3		
	F 4	66.7	16.7		
	F 5	95.9	58.3		
	M 1	33.3	50.0		
	M 2	87.5	0.0		
	M 3	62.5	33.3		
	M 4	79.2	0.0		
	M 5	62.5	0.0		
	mean	73.4	28.3		
TI + CB	F 1	100.0	25.0	93.9	44.9
	F 2	95.9	58.3		
	F 3	87.5	41.7		
	F 4	87.5	33.3		
	F 5	100.0	83.3		
	M 1	95.9	66.6		
	M 2	100.0	50.0		
	M 3	86.1	25.0		
	M 4	75.7	16.7		
	M 5	100.0	50.0		
	mean	92.9	45.0		

Samenvatting (summary in Dutch)

Stemmen verschillen, dat is duidelijk. Veel minder duidelijk is, op welke punten stemmen verschillen. Om een aantal redenen is het interessant vast te stellen in welke spraakeigenschappen de grootste variatie tussen sprekers wordt aangetroffen.

Een eerste reden voor dergelijk onderzoek is, dat kennis omtrent sprekerafhankelijkheid van spraakkenmerken wellicht iets zegt over het belang van deze kenmerken vanuit taalkundig perspectief. Waarschijnlijk staat ons taalsysteem vooral daar grote variatie toe, waar dat het overkomen van de beoogde boodschap niet hindert. Met andere woorden: eigenschappen waarop verschillende sprekers weinig variatie aan de dag leggen zijn waarschijnlijk essentieel voor het overbrengen van de boodschap van spreker naar luisteraar.

Een tweede drijfveer voor het doen van onderzoek naar sprekervariatie is een gevolg van de toegenomen technische mogelijkheden op het gebied van spraakonderzoek. Voor de belangrijkste toepassingen van spraakonderzoek, spraaksynthese en -herkenning, kan kennis met betrekking tot sprekerspecificiteit grote voordelen opleveren. Spraaksynthese is het automatisch hoorbaar maken van een in de computer opgeslagen tekst en spraakherkenning is het omzetten van door een spreker geuite spraak in geschreven tekst. Zowel spraaksynthese als spraakherkenning zijn in toenemende mate gebaat bij informatie met betrekking tot sprekerkenmerken. Zo kan met behulp van dergelijke kennis de automatische weergave van een stuk tekst natuurlijker gaan klinken, en zelfs een eigen "stem" krijgen. Voor automatische spraakherkenning geldt nog steeds dat het succes hiervan deels afhankelijk is van de mate waarin een systeem informatie over specifieke sprekers bevat. Kennis over de eigenschappen waarop spraak tussen sprekers varieert kan bijdragen tot efficiënter maken, en daarmee verkorten, van de trainingsperiode waarin een herkenningssysteem de spraak van een bepaalde spreker leert herkennen.

Tenslotte kan kennis over sprekervariatie in spraakkenmerken gebruikt worden bij het oplossen van allerlei praktische problemen. In de eerste plaats valt hierbij te denken aan forensisch onderzoek, waarbij moet worden vastgesteld aan wie de stem op een bepaalde opname toebehoort (vergelijkbaar met vingerafdrukkenonderzoek). Een ander voorbeeld van een interessante toepassing zijn de zogenaamde elektronische toegangssystemen, waarbij de stem als een soort sleutel toegang geeft tot bijvoorbeeld een beveiligd gedeelte van een bank.

Hoewel tot dusver de toepassingsmogelijkheden van onderzoek naar sprekerkenmerken breed zijn uitgemeten, moet het in dit boek gerapporteerde onderzoek in de eerste plaats gezien worden als een verkennende studie, die niet tot directe toepassingen zal leiden.

Van de vele mogelijke stemeigenschappen die ten aanzien van sprekerspecificiteit bestudeerd zouden kunnen worden, worden in dit onderzoek alleen de *prosodische* kenmerken onderzocht. Van oorsprong werd onder "prosodie" verstaan de leer van de klemtoon, de lengte van lettergrepen en van de metriek, zoals die al sinds de klassieke

oudheid bestaat. In de moderne fonetiek heeft prosodie een nauwere betekenis gekregen. Het betreft daarin vooral de studie van "variaties in toonhoogte, luidheid, tempo en ritme" (Crystal, 1985: 249).

In dit boek werden twee typen prosodische maten onderzocht. In de eerste plaats zijn dat maten die verkregen worden door over een bepaald tijdsinterval te integreren, oftewel TI-maten (Engels: "Time-Integrated"). Een voorbeeld van een dergelijke maat is de spreeksnelheid, waarbij over een bepaald tijdsinterval het aantal uitgesproken lettergrepen geteld wordt, en vervolgens gedeeld wordt door de duur van het betreffende interval.

Het tweede type prosodische maten wordt gemeten op een bepaald punt in een uiting. Er worden in deze studie alleen metingen verricht aan het begin en eind van uitingen, alsmede op *keerpunten* in de toonhoogtecontour, dat wil zeggen op punten waar toonhoogtebewegingen beginnen of eindigen. Om te bepalen of de door de sprekers geproduceerde toonhoogtebewegingen in fonologisch opzicht vergelijkbaar zijn, werden zij vergeleken met behulp van de intonatiegrammatica van het Nederlands ('t Hart et al., 1990). De realisaties van toonhoogtebewegingen door verschillende sprekers kunnen alleen vergeleken worden op vaste punten in vaste uitingen; daarom spreken wij van contourgebonden maten, oftewel CB-maten (Engels: "Contour-Bound").

Persoonsgebonden spraakkenmerken zijn deels het gevolg van echt sprekerspecifieke eigenschappen, zoals de unieke anatomische en fysiologische structuur van de spraakorganen, maar deels ook van gedragspatronen die gerelateerd zijn aan eigenschappen als geslacht, leeftijd, sociaal-economische achtergrond, enzovoort. De invloed van dergelijke sprekereigenschappen, bijvoorbeeld geslacht en leeftijd, kan zelf ook weer veroorzaakt worden door zowel anatomisch-fysiologische verschillen als door sociaalculturele factoren. Tenslotte kunnen persoonsgebonden verschillen gerelateerd zijn aan de uitgevoerde spreektaak: voorlezen, conversatie, enzovoort. De invloed van een aantal belangrijke spraak- en taakeigenschappen (spreekstijl, geslacht en leeftijd) op de te onderzoeken prosodische maten werd in dit onderzoek systematisch gevarieerd ten einde de invloed van deze factoren op de sprekerherkenning te kunnen controleren.

Voor de twee parametertypen, de TI- en de CB-maten, wilden we vaststellen in hoeverre zij apart of in combinatie gebruikt kunnen worden om sprekers te identificeren. Ook wilden we de sprekerspecificiteit van de parametertypen vergelijken.

In hoofdstuk 2 werden de te gebruiken maten geïntroduceerd. De TI-maten kunnen worden verdeeld in drie groepen: toonhoogte- en amplitudematen en temporele maten. De gebruikte toonhoogtematen waren de gemiddelde toonhoogte, de variatiecoëfficiënt van de toonhoogte, en twee maten voor de duurvariatie van dicht bij elkaar gelegen perioden, oftewel *toonhoogteperturbatiematen*. Ook voor de amplitude werden de variatiecoëfficiënt en twee perturbatiematen bepaald. Er werden drie temporele maten gebruikt: de spreeksnelheid, de hoeveelheid pauze en het percentage stemhebbende spraak.

Op het gebied van de CB-maten werden metingen verricht op verschillende posities in de F_0 contour: bij de begin- en eindtoonhoogte en bij de begin- en eindpunten van bepaalde toonhoogtebewegingen. Deze metingen bestonden in de eerste plaats uit toonhoogtemetingen. De eindtoonhoogte van de uiting werd uitgedrukt in hertz. De laagste punten van de bestudeerde toonhoogtebewegingen werden uitgedrukt in het aantal semitonen verschil met deze eindtoonhoogte, en de hoogste punten in de toonhoogtebewegingen werden weergegeven in het aantal semitonen verschil ten opzichte van de laagste toonhoogte in de betreffende bewegingen. Ook de tijdsduur van de toonhoogte-

bewegingen werd gemeten, evenals, door deling van de toonhoogteverandering door de tijdsduur, de richtingscoëfficiënt. Tenslotte werd voor alle bewegingen het tijdsinterval tussen het begin van de beweging en het begin van de klinker in de betrokken lettergreep vastgesteld (*synchronisatietijd*). Met deze maat werd een brug geslagen tussen de prosodische en de segmentele structuur van de uitingen.

In hoofdstuk 2 werden ook de criteria volgens welke de sprekersgroep is geselecteerd uitgelegd, alsmede de methode waarmee het spraakmateriaal werd verzameld. Voor de meting van de TI-maten verkregen wij spraakfragmenten uit twee taken: een interview en een voorleestaak. Voor het meten van CB-parameters hadden wij zinnen nodig die qua prosodisch gedrag eenvormig zouden zijn. In proefexperimenten werden zinnen gevonden die dit vergelijkbare gedrag enigszins ontlokten; de gebruikte uitingen gaven bij bijna alle sprekers aanleiding tot vergelijkbaar prosodisch gedrag en bevatten toonhoogtebewegingen die door alle sprekers gerealiseerd werden. Hoofdstuk 2 werd besloten met een beschrijving van de gehanteerde opnameprocedure en -apparatuur.

In hoofdstuk 3 werden de sprekeronderscheidende eigenschappen van de TI-maten uitgetest in het daartoe meest geëigende deel van ons materiaal: in tamelijk lange spraakfragmenten. Daar wij het gebruik van redundante maten wilden voorkomen, werd eerst de onderlinge gerelateerdheid van de maten vastgesteld met behulp van een correlatiematrix en een factoranalyse. De tien maten bleken niet al te zeer aan elkaar gerelateerd te zijn en werden daarom alle onderworpen aan een variantieanalyse. In deze analyses werden de mate van geassocieerdheid en de significantie van een aantal factoren, te weten Spreker, Spreekstijl, Geslacht, Leeftijdsgroep, Fragment/Paragraaf en Sessie, bepaald voor ieder van de tien maten. De variantieanalyses hadden twee voordelen. Ten eerste kon de mate van gerelateerdheid van de prosodische maten en de bovengenoemde factoren worden vastgesteld, en in de tweede plaats konden mogelijke redenen worden gevonden voor eventuele verschillen in de bruikbaarheid van de TI-maten voor (vooral) sprekerherkenning. Om bruikbaar te zijn voor sprekeridentificatie is het voor een maat belangrijk dat de interacties van de factor Spreker met andere extralinguïstische factoren, zoals Spreekstijl, Fragment/ Paragraaf en Sessie:

- niet significant zijn of
- een kleiner deel van de variantie verklaren dan de hoofdfactor Spreker.

In alle analyses verklaarde de factor Spreker een groter deel van de variantie dan de interactietermen.

De gecombineerde sprekeridentificatie-mogelijkheden van de TI-maten werden vastgesteld met behulp van lineaire discriminantanalyses (LDA's). Deze analyses werden uitgevoerd voor het totale materiaal en voor deelverzamelingen ervan. In een LDA met alle fragmenten vonden we een identificatiescore van 60 %, hetgeen betekent dat boven de hoeveelheid sprekerherkenning die al door toeval zou worden verkregen (bij 50 sprekers is dat 2 %) nog eens 60 % herkend werd.

In een LDA met sprekers van beide geslachten hadden wij vooraf betere sprekerherkenning verwacht, omdat in zo'n analyse voor ieder toe te wijzen spraakfragment de hoeveelheid potentiële kandidaten kleiner zou zijn dan het totale aantal sprekers. Immers, mannen en vrouwen hebben nogal verschillende spraakkenmerken en daardoor zouden sprekers van verschillend geslacht niet snel verward worden. Desondanks bleek dat het percentage correcte herkenning verhoogd werd door alleen gegevens van mannen of van

vrouwen te analyseren. Een nog grotere toename in de identificatiescore kon worden verkregen door alleen fragmenten van één van de spreekstijlen te analyseren. Vooral de voorgelezen fragmenten waren zeer sprekerspecifiek.

In alle genoemde analyses werd een vooraanstaande rol gespeeld door de gemiddelde toonhoogte, maar sprekerherkenning bleek niet onmogelijk te worden zonder deze maat, terwijl een LDA met alleen de gemiddelde toonhoogte als predictormaat weinig succesvol was.

Sprekerspecifieke maten hebben weinig praktische betekenis als sprekers na verloop van tijd heel andere waarden realiseren. Daarom werd in dit onderzoek veel belang gehecht aan de zogenaamde kruisvalidatie-analyses. Daarin werd spraakmateriaal uit één van de opnamesessies gebruikt om discriminantfuncties te bepalen die vervolgens werden toegepast om het spraakmateriaal uit de andere sessie aan de sprekers toe te wijzen. Helaas bleek de identificatiescore bij kruisvalidatie slechts 33 % te bedragen. In alleen voorgelezen materiaal lag dit percentage hoger: 54 %.

Discriminantanalyses werden ook toegepast om spreekstijlen, geslachts- en leeftijdsgroepen, (voorgelezen) paragrafen en opnamesessies te onderscheiden. Dat de spreekstijlen nogal uitgesproken van elkaar verschilden, bleek al uit de variantieanalyses, waarin voor veel parameters significante verschillen werden gevonden. Het bleek dan ook heel goed mogelijk de stijlen van elkaar te onderscheiden: de identificatiescore bedroeg 87 %, en geen van de parameters was hierbij van uitgesproken belang.

De identificatie van de geslachtsgroepen was geen moeilijke opgave; het bleek zelfs mogelijk een perfecte herkenning te bewerkstelligen. Zoals verwacht kon worden, was vooral de gemiddelde toonhoogte een maat van belang, en het bleek zelfs dat geslachtsherkenning op basis van alleen deze maat succesvoller was dan herkenning met behulp van alle maten uitgezonderd de gemiddelde toonhoogte. Overigens bleken ook deze andere maten een hoog percentage geslachtsherkenning mogelijk te maken.

Een goede herkenning van de leeftijdsgroepen mocht a priori niet verwacht worden, daar de gekozen sprekers zich in een levensperiode bevonden waarin geen grote stemveranderingen plaatsvinden: de sprekers waren de puberteit reeds voorbij en vertoonden waarschijnlijk nog geen ouderdomsverschijnselen. Toch bleek het enigszins mogelijk de vijf groepen te onderscheiden; er werd 26 % boven kans herkend. Dit percentage steeg iets als alleen materiaal van één van de spreekstijlen in beschouwing werd genomen, en steeg nog iets meer als alleen gegevens van mannen of vrouwen werden gebruikt.

Om de invloed van de duur van de fragmenten vast te stellen werd hoofdstuk 3 besloten met analyses waarin het materiaal uit verschillende fragmenten was samengevoegd tot fragmenten van 45 seconden. Het bleek dat bij deze langere fragmenten een duidelijk betere sprekeridentificatie mogelijk was, waarschijnlijk omdat (een deel van) de TI-maten door de langere integratietijden aan stabiliteit wonnen.

Het doel van hoofdstuk 4 was voornamelijk het vaststellen van de sprekeridentificerende eigenschappen van de CB-maten. Bovendien werd het identificerende vermogen van de CB-maten vergeleken met dat van de TI-maten en met een combinatie van de twee typen.

Het in hoofdstuk 4 gebruikte spraakmateriaal bestond uit opnamen van realisaties van drie "sportzinnen". Deze zinnen waren van de vorm *De Ieren wonnen van de Denen met drie-één*. Metingen werden verricht aan de stijging op de eerste lettergreep van de eerste nationaliteit, aan de stijging op het eerste getal van de score en aan de daling op het tweede getal van de score. Deze daling liep niet door tot aan het eind van de uiting. Ook

het resterende deel van de toonhoogtecontour, van het einde van de laatste daling tot het einde van de uiting, beschouwden wij als een soort toonhoogtebeweging. Het bleek, dat het hier meestal een lichte stijging betrof.

We gebruikten 21 CB-maten: de eindwaarde van de toonhoogte, het verschil in toonhoogte tussen de laagste punten van de toonhoogtebewegingen en de eind-toonhoogte, het toonhoogteverschil tussen het begin en het einde van de uiting, het toonhoogteverschil tussen hoogste en laagste punt in de toonhoogtebewegingen, de duur van de toonhoogtebewegingen, de richtingscoëfficiënt van de declinatie en van de toonhoogtebewegingen, en de tijdsduur tussen het begin van de toonhoogtebeweging en het begin van de klinker in de betreffende lettergreep.

De structuur van hoofdstuk 4 was gelijk aan die van hoofdstuk 3 en wij bestudeerden weer eerst de onderlinge gerelateerdheid van de 21 CB-maten. De correlatie tussen de richtingscoëfficiënt en het toonhoogteverschil van de declinatie was zo hoog dat wij besloten alleen de richtingscoëfficiënt te gebruiken in de verdere analyses. Een factoranalyse toonde aan dat de overblijvende CB-maten niet al te sterk aan elkaar gerelateerd waren.

Vervolgens werden de 10 TI-maten en de 20 CB-maten onderworpen aan variantieanalyses. De Spreker × Sessie-interactie, belangrijk voor het verkrijgen van hoge kruisvalidatiescores, was slechts voor twee CB-maten significant. Voor deze maten was de hoeveelheid verklaarde variantie voor het interactie-effect kleiner dan voor het hoofdeffect. Daarom werd niet aangenomen dat de interacties sprekerherkenning onmogelijk zouden maken.

Het sprekeronderscheidend vermogen van de TI- en de CB-maten werd vastgesteld in LDA's die werden uitgevoerd op zowel het gehele materiaal als op deelverzamelingen ervan. In het totale zinnenmateriaal leidden de CB-maten tot ongeveer dezelfde mate van sprekerherkenning als de TI-maten: 81 % voor de TI-maten en 86 % voor de CB-maten. Dit betekent dat metingen die werden verricht op maar een klein aantal keerpunten in de toonhoogtecontour tot dezelfde identificatieprestatie leidden als metingen die het gevolg waren van integratie over enkele seconden spraak. Het combineren van de twee typen maten verhoogde de sprekeridentificatie tot 97 % correct.

Omdat met een groter aantal parameters al snel betere resultaten worden geboekt, werden ook analyses verricht waarbij het maximum aantal parameters dat tot de analyse werd toegelaten voor alle analyses gelijk werd gesteld aan 10. De herkenningsscore van de CB-maten lag in deze analyses iets lager dan die van de TI-maten: 77 % voor de CB-maten en 81 % voor de TI-maten. Gecombineerd was de score niet veel hoger: 84 %.

In zowel de LDA met de TI-maten als in de analyse met beide typen parameters was de gemiddelde toonhoogte de belangrijkste sprekerherkenningsvariabele. In de analyses met de CB-maten werd een vergelijkbaar belangrijke rol gespeeld door de eindtoonhoogte van de zin. Het weglaten van de gemiddelde toonhoogte uit de TI-analyses en van de eindtoonhoogte uit de CB-analyses leidde tot een daling in de identificatiescore van ongeveer tien procent. Voor zowel eindtoonhoogte als voor gemiddelde toonhoogte geldt dus dat het belangrijke maten zijn, maar dat sprekeridentificatie er niet volledig van afhankelijk is.

Voor de spraakfragmenten in hoofdstuk 3 vonden we dat het percentage correcte herkenning enigszins verhoogd kon worden door het onderverdelen van de gegevensverzameling in deelverzamelingen van mannelijke en vrouwelijke sprekers. Voor de TI-maten in de zinnen werd dit resultaat bevestigd (hoewel het verschil tussen de scores van

de gecombineerde analyse en de analyse per geslachtsgroep klein was). Voor de CB-maten bleek uitsplitsen naar geslacht nauwelijks positief uit te pakken: de score voor mannen was daar 83 %, voor vrouwen 92 % en in de gecombineerde analyse 86 %.

Ook in hoofdstuk 4 waren de resultaten in de kruisvalidatie-analyses slecht: 43 % voor zowel de CB- als de TI-maten en 53 % voor de combinatie van deze twee typen maten. Ook met de in de sportzinnen verkregen gegevens lijkt toepassing van prosodische maten in praktische situaties niet mogelijk.

Er werden ook discriminantanalyses uitgevoerd voor de karakterisering van geslacht, leeftijdsgroep, zin en opnamesessie. Met betrekking tot de distinctiviteit van de geslachten vonden wij in hoofdstuk 3 bijna perfecte geslachtsherkenning op basis van alleen gemiddelde toonhoogte. Hoge geslachtsherkenning bleek echter niet alleen van verschillen in gemiddelde toonhoogte afhankelijk te zijn. In hoofdstuk 4 vonden wij opnieuw een vrijwel perfecte geslachtsherkenning, zowel in analyses waarin alleen TI- of CB-maten werden gebruikt als in de combinatie van de beide typen. In de analyse met de CB-maten speelde eindtoonhoogte een rol die vergelijkbaar is met gemiddelde toonhoogte in de analyses met de TI-maten. LDA's met alle maten behalve de twee toonhoogtematen konden de geslachten ook tamelijk goed onderscheiden, maar in een LDA met alle CB-maten behalve eindtoonhoogte werd een veel lager percentage geslachtsherkenning bereikt: slechts 67 % was correct.

De herkenningsprestatie van LDA's voor de vijf leeftijdsgroepen over de sportzinnen was ongeveer gelijk aan die van de spraakfragmenten in hoofdstuk 3. De TI-maten en de CB-maten onderscheidden de leeftijdsgroepen ongeveer even goed (respectievelijk 23 % en 25 %), en de combinatie van de twee typen maten resulteerde in iets betere resultaten (33 %).

Op basis van de gerapporteerde gegevens is onze eindconclusie dat de gebruikte prosodische maten wellicht een bruikbare bijdrage kunnen leveren aan de herkenning van sprekers en van een aantal andere spreker- en spreektaak-gerelateerde factoren. Een groot voordeel van het gebruik van prosodische maten is, dat deze maten relatief ongevoelig zijn voor verzending van gegevens, bijvoorbeeld over telefoonlijnen. De percentages correcte herkenning van sprekers, waar het in dit onderzoek vooral om ging, bleken echter niet bijzonder hoog, en de identificatiescores in de zo belangrijke kruisvalidaties waren zelfs laag. Daarom lijken de gebruikte prosodische maten alleen een ondersteunende rol te kunnen spelen bij sprekerherkenning. Wellicht kan hun bruikbaarheid vergroot worden door betere analysetechnieken te gebruiken. Ook kunnen hogere identificatiescores bereikt worden door de variabiliteit in de gegevens omlaag te brengen, bijvoorbeeld door alleen gegevens te gebruiken die betrekking hebben op één specifieke spreekstijl of sexe. Belangrijke vooruitgang in de toepasbaarheid van prosodische maten zal waarschijnlijk worden bereikt als meer kennis beschikbaar komt met betrekking tot de onderliggende factoren die de realisatie van toonhoogtebewegingen bepalen.

Curriculum vitae

Hans Kraayeveld werd geboren op 8 juli 1962 in Dordrecht. In 1980 behaalde hij het diploma VWO-B bij de Gemeentelijke Scholengemeenschap "Noordendijk" in Dordrecht. Aansluitend hieraan begon hij de studie Psychologie aan de Rijksuniversiteit Leiden. In 1984 behaalde hij het kandidaatsexamen, waarna hij de doctoraalstudie Functieleer volgde. Tijdens de doctoraalfase volgde hij de bijvakken Sociaal-Wetenschappelijke Informatica en Algemene Taalwetenschap. Stage en scriptie werden in 1985 en 1986 verricht bij het Max-Planckinstituut voor psycholinguïstiek in Nijmegen. In februari 1988 behaalde hij het doctoraalexamen.

Van december 1987 tot juni 1989 vervulde hij zijn vervangende dienstplicht bij de Vakgroep Sociaal-Wetenschappelijke Informatica van de Rijksuniversiteit Leiden. In deze periode begeleidde hij computerpractica, vertaalde en bewerkte een handleiding voor een kennissysteem-shell en verzorgde een cursus "Kennissystemen". De laatste paar maanden van deze periode verrichtte hij onderzoek naar de effecten van de structurering van computermenu's op de zoektijd in deze menu's.

Van 1 september 1989 tot 31 augustus 1993 was hij als Onderzoeker In Opleiding (O.I.O.) in dienst van de Nederlandse organisatie voor Wetenschappelijk Onderzoek (N.W.O.) en was hij werkzaam bij de vakgroep Taal en Spraak van de Katholieke Universiteit Nijmegen. In deze periode werd het onderzoek uitgevoerd waarover in dit proefschrift verslag wordt gedaan.

Na een omscholing tot telecommunicatiebeheerder door de bedrijven Intercai Nederland B.V. en TeleTalent B.V., beide gevestigd in Utrecht, is Hans Kraayeveld tegenwoordig werkzaam bij het laatstgenoemde bedrijf. Momenteel is hij gedetacheerd bij Delta Lloyd Verzekeringsgroep N.V. in Amsterdam.

Stellingen bij het proefschrift

idiosyncrasy in prosody

Hans Kraayeveld

1 Voor de praktische toepassing van contourgebonden maten is het van groot belang dat er betrouwbare hulpmiddelen ontwikkeld worden voor automatische prosodische transcriptie en stilering

2 De mogelijkheden van toepassing van prosodische maten in een forensische context lijken begrensd omdat de meest sprekerspecifieke maat, gemiddelde toonhoogte gemakkelijk door sprekers gemanipuleerd kan worden

3 Het is een bekend gegeven dat sprekers in een gesprek zich aan elkaar aanpassen Gezien het feit dat alle sprekers in het beschreven onderzoek door de schrijver werden geïnterviewd is het beter te spreken van 'spraak-jegens-Hans Kraayeveld' dan van 'spontane spraak'

4 Uitgaande van de hypothese dat de meeste sprekervariatie gevonden wordt in spraakeigenschappen waaraan het linguistisch systeem weinig beperkingen oplegt (zie pagina 1 van dit proefschrift) lijkt het niet zo belangrijk welke toonhoogtecontouren gebruikt worden bij het voorlezen van een zin

5 Natuurbeheer is een contradictio in terminis

6 Omscholing van uitgerangeerde wetenschappers is een nuttige vorm van recycling Uitgaande van het principe 'de vervuiler betaalt' zou de overheid niet moeten bezuinigen op omscholingen

7 Onvrede met het AIO-systeem is voor veel promoverende AIO's aanleiding tot het formuleren van een stelling dienaangaande in hun proefschrift

8 Mensen die bijzonder willen zijn zijn dat zelden Mensen die bijzonder zijn doen daar meestal geen moeite voor

9 Groeicurves maken baby's steeds dikker

10 De meest ingrijpende veranderingen in het persoonlijk leven in de komende decennia zullen voortkomen uit ontwikkelingen op het gebied van de telecommunicatie