Physiological control and behaviour of the voice source in the production of prosody

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Strik, Wilhelmus Albertus Johannes

Physiological control and behaviour of the voice source in the production of prosody / Wilhelmus Albertus Johannes Strik. - [S.l. : s.n.]. - Ill.

Thesis Nijmegen. - With ref. - With summary in Dutch.

ISBN 90-9007689-1

Subject headings: physiology / prosody / voice source

Physiological control and behaviour of the voice source in the production of prosody

Een wetenschappelijke proeve op het gebied van de Natuurwetenschappen

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
maandag 28 november 1994
des namiddags te 1.30 uur precies

door

Wilhelmus Albertus Johannes Strik

geboren op 18 mei 1957
te Rosmalen

Promotores:	Prof. Dr. L.W.J. Boves Prof. Dr. C.C.A.M. Gielen

This research was partially supported by the Linguistic Research Foundation, which is funded by the Netherlands organisation for scientific research, NWO.

voor Ankie

Acknowledgements

The present thesis is based on research carried out at the Department of Language and Speech of the University of Nijmegen. It would not have been possible to complete this study without the help of many other people. Only some of them are mentioned here.

First of all I would like to thank my supervisor Loe Boves, who is also known as Lou Boves, Louis Boves, prof. dr. L.W.J. Boves and professor Bobes. Although he could not be easily reached during normal working hours, he was always prepared to discuss my problems at length in the evenings or weekends.

I am also indebted to Bert Cranen, for his willingness to answer all my questions and for his valuable comments.

The experiments conducted for the present research were, to say the least, very annoying for the subjects. For this reason I am very grateful to Loe Boves, Harco de Blaauw and Paul Tan. I also would like to thank Jan Strik for participating in one of the experiments and for accurately processing part of the data.

In 1985 I spent two months at the Haskins Laboratories in New Haven, where the first experiment was carried out. I would like to express my gratitude to all members of this institute who contributed to making my stay so pleasant and successful. Special thanks are due to Tom Baer for his hospitality and for helping me carry out my first experiment. Furthermore, I would like to mention Hiroshi Muta, who managed to measure all signals I asked him to record.

The Nijmegen experiments were carried out at the Department of Otorhinolaryngology of the University hospital in Nijmegen. I am therefore indebted to all people who made this possible. In particular, I would like to mention Philip Blok, because he was always prepared to come to Nijmegen on Saturdays to position the pressure catheter and the EMG-electrodes. Many thanks are also due to Hans Zondag, who managed to monitor all signals and to record them on tape in a professional way.

A friendly and stimulating atmosphere is essential in order to carry out research. For this reason I would like to thank the (ex-)colleagues and (ex-)students at the Department of Language and Speech. The daily coffee ritual constituted a welcome break and a good occasion for nice chats. Also, I have good memories of the yearly trips of the department, the occasional dinners followed by slide shows and the various journeys abroad (especially in the POLYGLOT days).

Finally, I am very grateful to my parents for their continuous support throughout the years, and to Catia, for everything she has done in the final stage of this study.

Nijmegen, 12 october 1994

Contents

Chapter 1

Prologue

1.1 Introduction

Speech is an important element of human communication. A common situation in human communication is that in which a speaker wants to transmit a message to a listener. The speaker encodes this message into a speech signal, which the listener tries to decode. The way in which the message is transferred from the speaker to the listener can be represented by means of the so-called speech chain. Although many different stages can be distinguished in this chain, here we will confine ourselves to mentioning only the most important ones.

The speaker has a thought or idea that he wants to transmit to the listener. Once he has selected the right words and has organised them according to the rules of the language, appropriate nerve impulses are sent from the speaker's brain to the muscles involved in speech production. When the vocal organs are set into motion, they produce pressure variations in the surrounding air. In this way, a sound wave is generated which eventually reaches the listener's ear. In the case of successful communication, the listener will extract the speaker's message from the speech signal.

From an acoustic point of view, the speech signal produced by the speaker is a continuously varying pressure signal. It is often assumed that this continuous signal is perceived by listeners as a sequence of discrete units like words, syllables or speech sounds. The latter, which are generally divided in vowels and consonants, constitute the segmental level of speech. Each of these units can be described by means of segmental features, as will be explained in the following section.

Over and above the segmental level there is a suprasegmental level in speech. This consists of features that generally extend over units of speech larger than the segment. Suprasegmental features, also called prosodic or melodic features, are e.g. stress, intonation and tempo. These aspects of speech are generally subsumed under the label "prosody". Variables that are important for the perception of prosody are e.g. pitch, loudness and duration.

If an utterance is pronounced while keeping pitch constant, the resulting speech sounds very monotonous, like the speech produced by robots in old films, or that of the first speech synthesizers. In general, people do not speak monotonously, because in normal communication ample use is made of prosodic features. For instance, to suggest that a specific word is important in an utterance, one could stress that word, and by raising the pitch at the end of a sentence, one can indicate that it is a question.

One of the various stages in the speech chain is the transformation of the linguistic message into a speech signal. To produce this signal it is necessary that the muscles involved in speech production receive the right motor commands. These commands encode not only the segmental, but also the prosodic structure of speech. For instance, important prosodic variables such as pitch and loudness can be varied by means of the laryngeal and respiratory muscles.

The physiological mechanisms that are involved in the production of prosody are discussed in more detail in Section 1.2.2. In that section we will show that there is sufficient evidence that a number of well-known physiological mechanisms play a major role in the production of prosody. However, it is still unclear how these physiological mechanisms interact to produce the right prosodic structure in spontaneous speech (see Section 1.2.3).

The research reported on in this book focuses on one aspect of human speech production, viz. the production of some of the prosodic aspects of the speech signal. The main goal of this study is to increase our understanding of the relation between the physiological mechanisms involved in speech production and the prosodic features of speech. In the present chapter we introduce some of the notions and topics that play a central role in this kind of research. The theoretical framework of our research is outlined in Section 1.2. More precisely, Section 1.2.1 deals with prosody and its function in speech. In Section 1.2.2 we describe some of the fundamental physiological mechanisms of speech production. In Section 1.2.3 the relation between physiology and prosody is considered. Section 1.3 deals with the present investigation. The aim of our study is presented in Section 1.3.1. In Section 1.3.2 attention is paid to the methods adopted in our investigation. Finally, Section 1.4 contains an outline of the present book.

1.2 Theoretical framework

1.2.1 Prosody

As mentioned in Section 1.1, a speech utterance can be seen as a sequence of segments such as vowels and consonants. These segments can be described by means of the so-called segmental features. In general, the following three parameters are used to describe consonants: place of articulation, manner of articulation, and presence or absence of vocal fold vibration. Vowels, on the other hand, are often described by referring to the degree of lip rounding, the height of the raised part of the tongue, and its backward or forward position. A representation of the segmental make-up of an utterance is called a segmental transcription.

Together, vowels and consonants make up greater units such as syllables and words. Words, and the relationships existing between them, are important in conveying the meaning of an utterance. Since the identity of a word is determined by the segments it contains, it follows that the segments determine the content of an utterance. For communication purposes the way in which an utterance is pronounced is often at least as important as the content of the utterance. Just think of the many different ways in which the word "no" can be pronounced: as a statement or as a question, with sarcasm, happiness, resolution or nervousness. Although in any of these cases the word will consist of the same two segments, it will sound different. This variation in pronunciation falls under the heading of prosody and is described by means of prosodic or suprasegmental features. The latter are defined on domains that are larger than the segment, such as syllables, words, phrases and sentences. Prosodic features are stress, intonation, loudness, voice quality, rhythm, tempo and duration, of which some have already been mentioned in the previous section.

It is important to note that while segmental features can be defined for each separate segment without considering the adjacent segments, this is not possible for suprasegmental features. As a matter of fact, the latter can only be defined by a comparison of items in sequence. For instance, whether a vowel is stressed or not can only be determined by comparing it with other segments in the sequence.

The properties of the speech signal that are used to realize the prosodic features are fundamental frequency (F_0), intensity level (IL), duration and spectral contents (SC) of the speech signal. Of these acoustic parameters we will only study F_0, IL and SC. In the rest of this book these will be called prosodic parameters. Each of these prosodic parameters has a correlate in the perceptual domain. So, the perceptual counterparts of F_0, IL and SC (on a suprasegmental level) are pitch, loudness and voice quality, respectively.

At this point it may be useful to give a few examples in order to clarify the use of prosody for readers who are not familiar with this subject matter. As mentioned above, prosody can be used to indicate the difference between a statement and a question. In many languages a distinction is made between rising and falling pitch contours. The former are usually associated with statements and the latter with questions.

Furthermore, prosody can be used to signal the syntactic boundaries in an utterance. While in written language this is usually done by means of punctuation marks, in spoken language prosody is used for this purpose. For instance, if you try to pronounce the current sentence, which should not be difficult, you could resort to varying pitch, loudness and/or tempo when pronouncing “which should not be difficult” in order to indicate that it is a subordinate clause.

In pronouncing the above sentence, you might have stressed one of the words, for instance “pronounce”. This is another example of the use of prosody. By stressing a word one can indicate that that word is important. A word can be stressed in many different ways, and one of the possibilities is to increase F_0, IL and duration for (part of) that word.

Finally, we would like to point out that prosody can also be used to express attitude or emotions, like e.g. superiority, submission, anger, sorrow, nervousness, surprise, excitement and boredom. Many persons will speak more loudly when they are angry and with a higher pitch when they are nervous. Voice quality can also be used for this purpose. Breathy voice is often associated with sexual desire, although language-related differences seem to exist. For instance, in Japanese breathy voice appears to be used to convey respect or submission (Crystal, 1987). Most of the times prosody is used unconsciously by the speaker to utter a specific emotional state, but in certain cases, as in acting, it can be used on purpose.

Not all the prosodic properties of the speech signal have received the same degree of attention in the literature. By far the most studied parameter is F_0. This is probably due to the fact that many studies have revealed that F_0 is the most important parameter for the perception of prominence (see e.g. Lehiste, 1970; Clark & Yallop, 1990). This explains why, from a linguistic point of view, F_0 is considered the most important parameter (see e.g. Crystal, 1987).

1.2.1.1 Intonation models

The attention paid to F_0 is reflected by the considerable number of intonation models that are available in the literature. An important feature of intonation models is the way in which they deal with the slow decrease in F_0 that can be observed in many utterances, the so-called downtrend in F_0. Although a complete review of intonation models would be beyond the scope of this chapter, it is useful to give a short presentation of a few of them, just to make it clear that intonation can be modelled in many different ways. Informally, intonation refers to the linguistically relevant aspects of the F_0 contours. Intonation models try to account for those aspects in a precise and formalised manner. Intonation models can be classified on the basis of the following two characteristics:

❒ whether they hypothesise that the F_0 contour can be modelled by one or two components (type 1 and type 2, respectively)

❒ whether they model the whole F_0 contour or only the F_0 targets (type A and type B, respectively)

By combining these possibilities, four different types of intonation models can be identified:

1A. one-component models with whole F_0 contours

1B. one-component models with F_0 targets

2A. two-component models with whole F_0 contours

2B. two-component models with F_0 targets

The type 2 models are described first, because they received greater attention in our research. The models proposed by Öhman (1967, 1968) and Fujisaki (1991) are type 2A models. Figure 1 shows an F_0-contour that has been generated by using the model by Fujisaki (1991). The total F_0-contour (solid line) is obtained by summing up the phrase component (dashed line) and the accent component. A slightly different type 2A model is proposed by ’t Hart et al. (1990). What is modelled here is not the actual F_0 contour but a “close-copy stylization” of it (see ’t Hart et al., 1990). To illustrate the type 2B models, a hypothetical string of high (H) and low (L) tones is displayed in Figure 2. In the type 2B models the tones or F_0 targets are scaled with respect to a declining line. Sometimes the distance between the F_0 targets and the baseline can become smaller (the

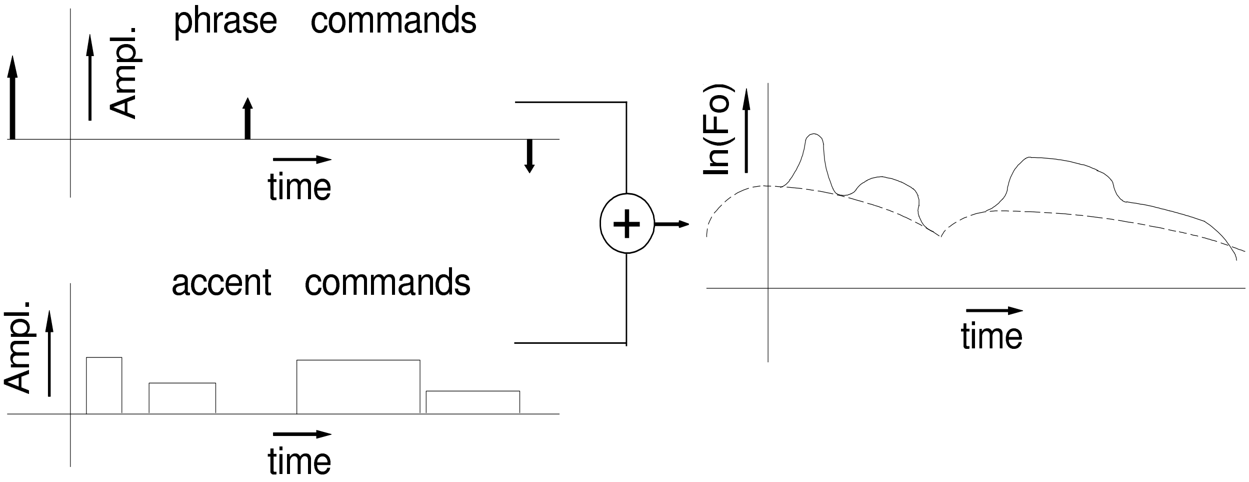


Figure 1. An overview of the type 2A model proposed by Fujisaki (1991). The phrase commands are a set of impulses and they produce the phrase components, i.e. the global component or baseline (dashed line). The accent commands are a set of step-wise functions which define the accent component, i.e. the local variations in F_0. The phrase and accent commands are superimposed to obtain the total F_0 contour (solid line).

so-called downstep); for instance in Figure 2 the distance d2 is smaller than d1. When this happens is determined by phonological downstep rules. Type 2B models are described e.g. by Pierrehumbert (1980), and Pierrehumbert & Beckman (1988).

In two-component models (types 2A and 2B) a distinction is made between a short-term or local component and a long-term or global component. The global component is used to model the gradual decrease of F_0 during an utterance. In one-component models (types 1A and 1B) there is no global component. In type 1A models the downtrend is completely modelled by using falling F_0 contours (e.g. Garding, 1983), whereas in type 1B models this is done by means of downstep (Liberman & Pierrehumbert, 1984; Van den Berg et al., 1992).

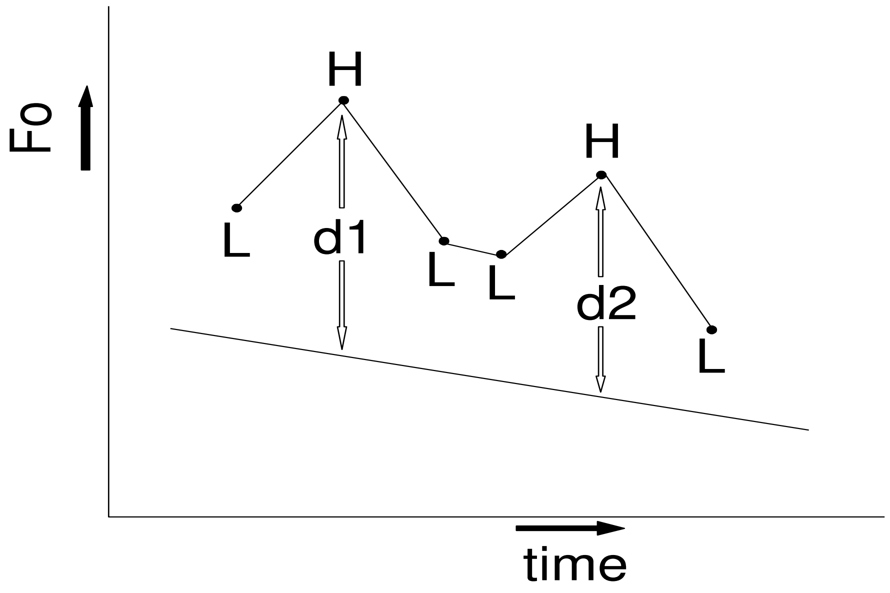


Figure 2. A hypothetical tone string for a type 2B model. The H-tones and L-tones are scaled with respect to a declining reference line. Phonological rules determine when a tone is downstepped, i.e. when its distance to the reference line becomes smaller. For instance, in this example the second H-tone is downstepped, and thus d2 is smaller than d1.

1.2.1.2 Terminology

A note on the terminology used in intonation research is in order here. We have already pointed out that a gradual lowering in F_0 can be observed in many utterances, albeit not in all utterances. In addition to downtrend and downstep, other terms such as downdrift, declination and catathesis have been used in connection with this lowering. Downstep and catathesis are generally used to describe the local behaviour of F_0 (the lowering of the F_0 target mentioned above), whereas downtrend, downdrift and declination are used to describe the global behaviour of F_0.

Initially, declination was a neutral term used to denote the slow decrease in F_0 (see e.g. Cohen et al., 1982; Ladd, 1984), and as such it was employed in our papers. However, the term declination (or declination line) has come to be used to indicate the global component in two-component models. In the latter case declination is only a subcomponent of the total downtrend in F_0. This is especially the case in intonational phonology, in which research field declination is seen as the part of downtrend that cannot be explained by downstep, i.e. the residue (see e.g. Ladd, 1992). At present, using the term declination could therefore imply that a two-component model has tacitly been assumed.

When we began to analyse the various intonation models to decide which one best accounted for our data, we did not want to limit ourselves to two-component models in advance. For this reason, in Chapters 7 and 8 we opted for the more neutral term downtrend, which we also used in describing the various intonation models above. In this book $F_{0,g}$ or $F_{0,global}$ will be used to denote the global component of F_0 in two-component models, and downtrend and declination will be considered as synonyms.

1.2.2 Physiology

Comprehensive descriptions of the anatomy and physiology of the organs involved in speech production can be found in various textbooks (see e.g. Perkell, 1969; Hardcastle, 1976; Catford 1977; Borden & Harris, 1980; Clark & Yallop, 1990). In this section we will therefore limit ourselves to mentioning those physiological processes that are known to play an important role in the production of prosody.

In many publications (see e.g. Ladefoged, 1967; Borden & Harris, 1980; Pickett, 1980; Clark & Yallop, 1990; Ohala, 1990) it is found that the two factors that mainly account for the glottal source variations related to prosody are:

① the characteristics of the vocal folds (i.e. tension, elasticity, length and mass); and

② the subglottal air pressure.

In general, it is assumed that the first factor is mainly responsible for the control of F_0 (Borden & Harris, 1980; Pickett, 1980; Clark & Yallop, 1990; Ohala, 1990), and the second one for the control of IL (Ladefoged, 1967; Lehiste, 1970; Clark & Yallop, 1990; Ohala, 1990). Both factors are almost completely determined by the respiratory and the laryngeal system. These organs are dealt with in the next two sections.

1.2.2.1 The respiratory system

The respiratory system can be considered the source of energy during phonation. It controls the lung volume (V_l) and the pressure in the lungs, and for this purpose many respiratory muscles are used (see e.g. Borden & Harris, 1980; Ohala, 1990). The main goal of the respiratory system is to pump oxygen in and carbon dioxide out of the lungs. During normal breathing, the two phases of the respiratory cycle (i.e. inspiration and expiration) have roughly the same duration. When we speak, however, the duration of the expiratory phase is generally longer than that of the inspiratory phase.

An important element in breathing and speaking is the air pressure. In speech research air pressure is usually expressed in cm H_2O with respect to the atmospheric pressure. Since air moves from places where pressure is high to places where pressure is low, it follows that during the inspiration phase subglottal pressure (P_{sb}, see Figure 3) is lower than the atmospheric pressure (P_{sb} is negative), so that air flows through the vocal tract into the lungs. P_{sb} can be lowered by expanding the chest wall and the abdomen, which has the effect of increasing the lung volume. In contrast, during phonation P_{sb} is generally higher than the external atmospheric pressure (P_{sb} is positive). In this case air will flow out of the lungs and will pass through the opening between the vocal folds, which is called the glottis. The resistance offered by this opening to the airflow is called the impedance of the glottis (Z_g).

When the glottis is open, as in normal breathing, Z_g is small. When Z_g is zero, the pressure above the vocal folds (oral pressure: P_{or}) equals P_{sb}, and the transglottal pressure ($P_{tr} = P_{sb} - P_{or}$) is zero (see Figure 3). In our research we only used the low-frequency component of the pressure signals, i.e. the pressure signals were low-pass filtered and sampled at a 200 Hz rate. Consequently, when discussing pressure signals in this book, we refer to the low-frequency component of these pressure signals.

Z_g is very high in the production of vowels. In this case P_{or} is minimal (it approaches zero) and P_{tr} is maximal (it approaches P_{sb}). Consonants are produced by making a constriction somewhere in the vocal tract. This has the effect of increasing P_{or} and thus of decreasing P_{tr}. At very low values of P_{tr} voiced phonation is no longer possible (Titze, 1992). In the extreme case P_{or} equals P_{sb} and P_{tr} becomes zero, as happens in the production of voiceless plosives.

Of the three pressure signals P_{sb}, P_{or} and P_{tr}, P_{sb} is the most studied one. A positive relation between P_{sb} and F_0 has been found in many measurements (Müller, 1843;

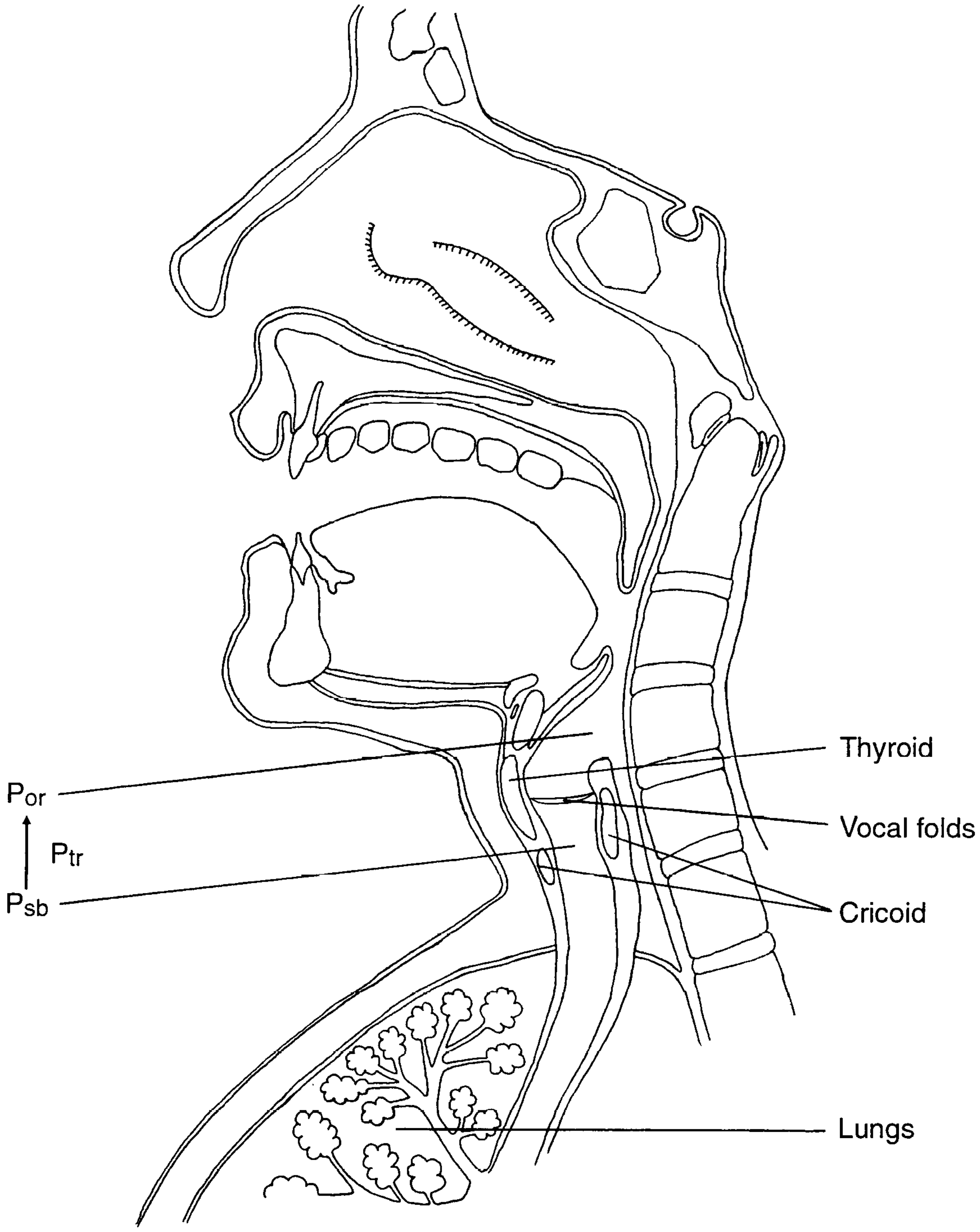


Figure 3. A sagittal section showing the organs that are involved in speech production. Subglottal pressure (P_{sb}) is the pressure below the glottis, oral pressure (P_{or}) is the pressure above the glottis, and transglottal pressure (P_{tr}) is the difference between P_{sb} and P_{or}: $P_{tr} = P_{sb} - P_{or}$.

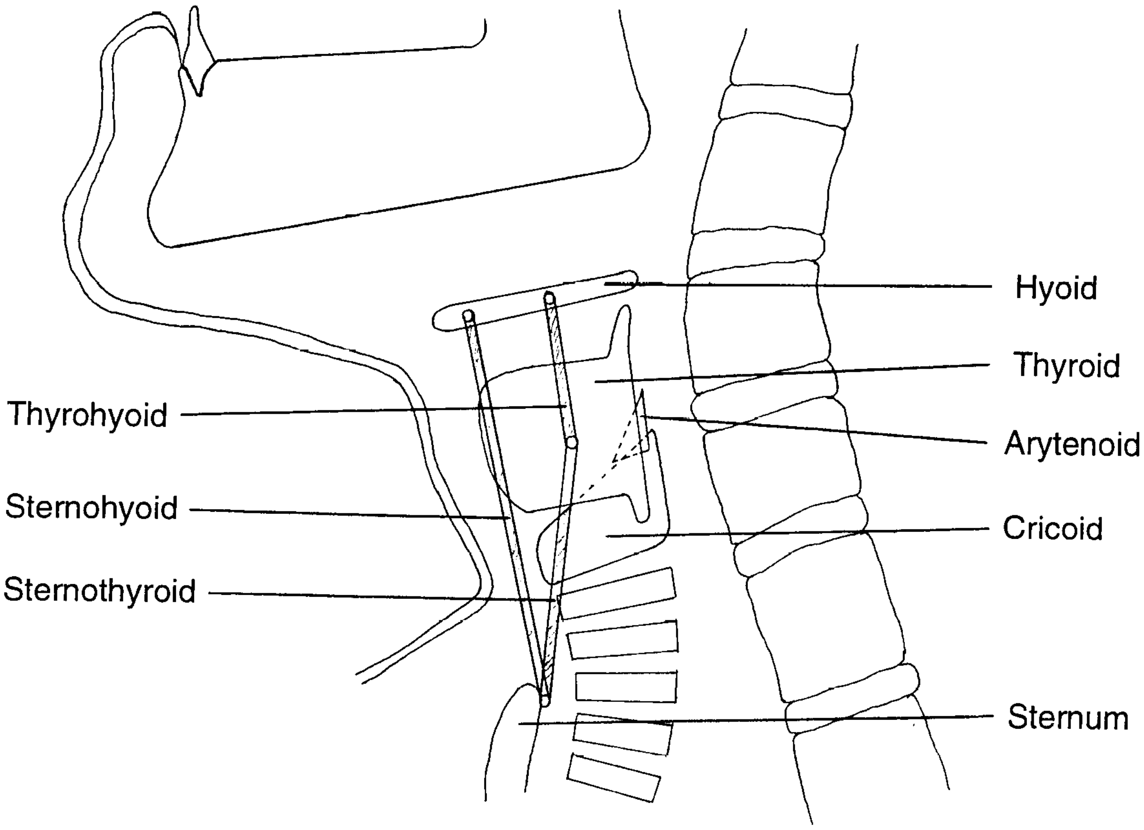


Figure 4. A schematic lateral view of the larynx and three strap muscles, i.e. the sternohyiod (SH), the sternothyroid and the thyrohyoid. The combined activity of these three strap muscles influences the height of the larynx. The sternum is fixed, and thus an increase in the activity of the SH will lower the larynx, ceteris paribus.

Ladefoged, 1967; Lieberman, 1967; Shipp & McGlone, 1971; Hixon et al., 1971; Collier, 1975; Baer et al., 1976; Atkinson, 1978; Baer, 1979; Shipp et al. 1979; Gelfer et al., 1983; Baken & Orlikoff, 1987; Gelfer, 1987; Titze & Durham, 1987). An increase in P_{sb} also leads to an increase in IL (Müller, 1843; Rubin, 1963; Isshiki, 1964; Bouhuys et al., 1968; Baer et al., 1976).

In addition to the pressure signals there are other factors that can influence vocal fold vibration, such as the tension, elasticity, length and mass of the vocal folds. These factors are mainly controlled by the laryngeal muscles.

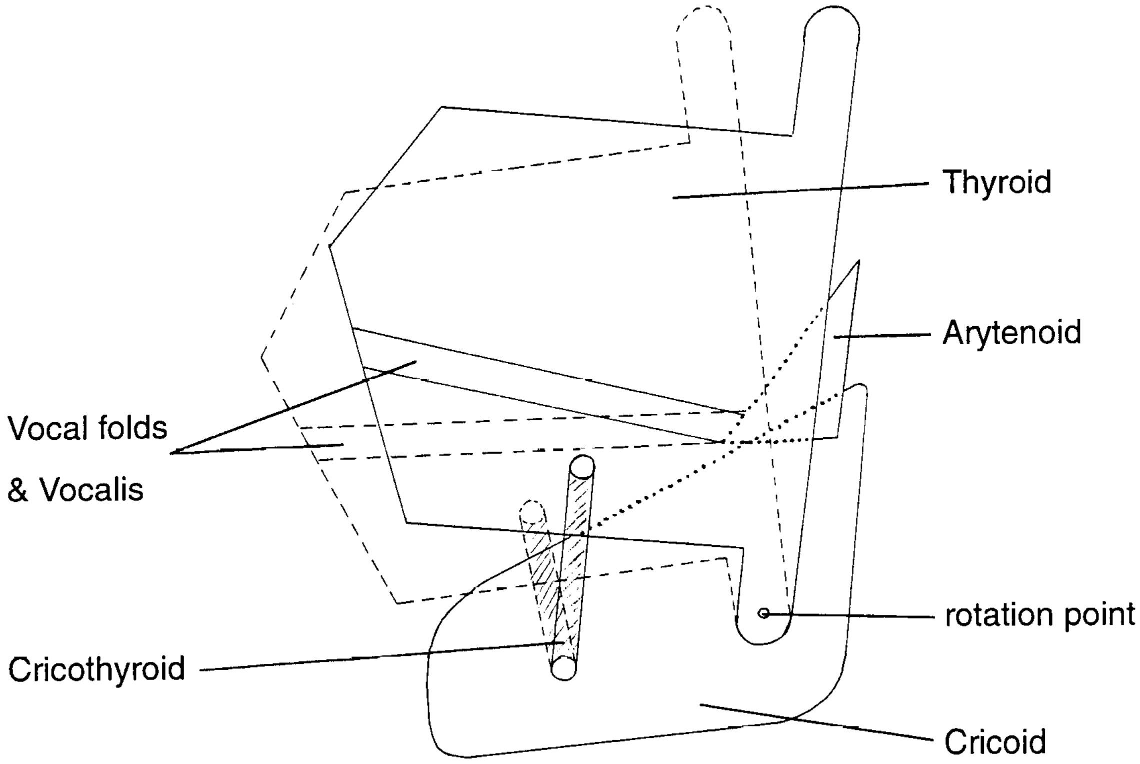


Figure 5. A schematic lateral view of the larynx. The arytenoids are positioned on top of the cricoid, and the vocal folds run from the front part of the arytenoids towards the thyroid. Cricoid and thyroid can rotate with respect to each other. If the CT is shortened by an increase in the activity of the CT, the thyroid will be tilted forward. Consequently, the length and tension of the vocal folds will increase. All other things being equal, this will lead to an increase in F_0.

1.2.2.2 The laryngeal muscles

The larynx is used both as a valve and as a voice source. As a valve it controls the flow of air in or out of the lungs. If necessary, the valve can be closed to block the airflow, as is done during great physical effort. The larynx consists of various cartilages, the vocal folds, and a number of intrinsic laryngeal muscles (see Figures 4 and 5). In addition to the intrinsic laryngeal muscles there are the extrinsic laryngeal muscles (see Figure 4). Both types of muscles are relevant to prosodic research. In the control of F_0 the most important intrinsic laryngeal muscles (see Figure 5) are probably the cricothyroid (CT) and the vocalis (VOC) (Rubin, 1963; Sawashima et al., 1969; Shipp & McGlone, 1971; Gay et al., 1972; Atkinson, 1978; Shipp et al., 1979; Borden & Harris, 1980; Pickett, 1980; Hirose & Sawashima, 1981), whereas the strap muscles (Figure 4) are considered the most important extrinsic laryngeal muscles (Faaborg-Andersen, 1965; Ohala, 1970, 1972; Ohala & Hirose, 1970; Atkinson, 1973, 1978; Collier, 1975; Maeda, 1976, 1980; Atkinson & Erickson, 1977; Erickson et al., 1977; Borden & Harris, 1980; Picket, 1980; Hirose & Sawashima, 1981; Erickson et al., 1983).

The rest of this section deals with the relations between the laryngeal muscles and the prosodic parameters. First we discuss the relation between F_0, CT, VOC and the strap muscles. Subsequently, we go on to examine the relation between these muscles and IL.

The way in which CT can be used to vary F_0 is very clear from a physiological point of view: an increase in the activity of the CT increases the length and tension of the vocal folds (see Figure 5), which has the effect of raising F_0. As a matter of fact, a positive relation between CT and F_0 has often been observed (Rubin 1963; Sawashima et al., 1969; Shipp & McGlone 1971; Gay et al., 1972; Hirose & Gay 1972; Collier 1975; Maeda 1976; Erickson et al., 1977; Atkinson, 1978; Shipp et al., 1979; Hirose & Sawashima 1981; Erickson et al., 1983; Gelfer 1987; Niimi et al., 1987). The conclusion of many studies was that of all physiological factors known to affect F_0, the CT shows the most consistent relation with F_0 (Collier, 1975; Maeda, 1976; Atkinson, 1978; Ohala, 1978; Shipp et al., 1979; Erickson et al., 1983; Gelfer, 1987). Many textbooks also mention CT as the most important muscle in the control of F_0 (see e.g. Borden & Harris, 1980; Clark & Yallop, 1990).

The exact function of the VOC is more controversial than that of the CT. Since the VOC is a subcomponent of the vocal folds (see Figure 5), its effect on F_0 can be twofold. An increase in the activity of the VOC can shorten the vocal folds and therefore lower F_0. However, it can also enhance the internal tension of the vocal folds, thus raising F_0. In spite of this, a positive relation is generally found between F_0 and VOC (Negus, 1928; Katsuki, 1950; Faaborg-Andersen, 1957; Rubin, 1963; Sawashima et al., 1969; Shipp & McGlone, 1971; Gay et al., 1972; Hirose & Gay, 1972 ; Maeda, 1976; Atkinson, 1978; Shipp et al., 1979; Hirose & Sawashima, 1981; Niimi et al., 1987). The VOC probably acts in synergism with the CT in controlling F_0 (Atkinson, 1978; Hirose & Sawashima, 1981; Rubin, 1963; Sawashima et al., 1969; Shipp & McGlone, 1971; Gay et al., 1972; Shipp et al., 1979).

The strap muscles can be used to alter the position of the larynx in the cranial-caudal dimension (see Figure 4). Of all strap muscles the sternohyoid (SH) is probably the most studied one. This muscle seems to be used especially when F_0 is low. As a matter of fact, a negative correlation between SH and F_0 has often been reported (Faaborg-Andersen 1965; Ohala, 1970, 1972; Ohala & Hirose, 1970; Atkinson, 1973, 1978; Collier, 1975; Maeda, 1976, 1980; Atkinson & Erickson, 1977; Erickson et al., 1977; Hirose & Sawashima, 1981; Erickson et al., 1983). Although the relation between SH and F_0 is fairly consistent, it is not completely clear how SH and the other strap muscles can influence F_0. Given that the strap muscles are used to vary the height of the larynx and that many studies have revealed a relation between larynx height and F_0 (Ohala, 1972; Shipp & Haller, 1972; Ewan & Krones, 1973; Shipp, 1975; Hinchcliffe & Harrisson, 1976; Maeda, 1976; Shipp et al., 1979), the following explanation has been advanced: lowering the larynx (for instance owing to the increased activity of the SH) could diminish the tension of the vocal folds in cranial-caudal direction (see Figure 4); since decreased vocal fold tension can lead to lower F_0, it follows that lowering the larynx lowers F_0.

A number of studies in which the relation between IL, CT and VOC was investigated have produced disparate results. A positive correlation between CT and IL was observed by Lindestad et al. (1991), while Rubin (1963) and Hirano et al. (1969, 1970) reported a negative correlation. For CT and IL no correlation at all was found by Faaborg-Andersen (1957) and Sawashima et al. (1969). Moreover, according to Faaborg-Andersen (1957) and Sawashima et al. (1969) VOC and IL were not related either, whereas Rubin (1963) and Hirano et al. (1969, 1970) found a positive correlation between VOC and IL. To sum up, there is great uncertainty about the role of CT and VOC in controlling IL.

1.2.2.3 Physiology and prosodic parameters

Previous research has shown that P_{sb}, CT, VOC and SH affect F_0 and that CT is probably the most important factor. The relation between IL and the above-mentioned laryngeal muscles is not clear. It is probable that the laryngeal muscles play only a minor role in the control of IL and that P_{sb} is the fundamental factor. The relation between the physiological mechanisms and SC has received less attention in prosodic research. This explains why no results regarding this relation were presented in the previous two sections. Everything else being equal, an increase in P_{sb} will lead to an increase in F_0 and IL, but also to a different SC (the spectrum will be less steep). Changes in the activity of the laryngeal muscles will also influence the vibration of the vocal folds thus leading to a different SC. However, it is not known how this happens and which factors play a decisive role in this process.

In Sections 1.2.2.1 and 1.2.2.2 we focused on P_{sb}, CT, VOC and SH, because these physiological mechanisms have received particular attention in previous research. Although other physiological mechanisms have also been studied, for most of these mechanisms no consistent relation with F_0 and IL was found, or their effect on F_0 and IL appeared to be smaller than that of P_{sb}, CT, VOC and SH. It is obvious that these research findings do not exclude the possibility that the prosodic parameters are influenced by other physiological mechanisms than those studied so far.

1.2.3 Relation physiology - prosody

Although it is known that CT, VOC, SH and P_{sb} are involved in the production of prosody, it is not completely clear how these mechanisms cooperate in running speech. There are various reasons for this, the main one being that the production of prosody is a complex process in which many physiological mechanisms are involved. This makes it very difficult to measure all relevant physiological signals simultaneously. For this reason, only a limited number of relevant signals are usually investigated in one experiment. In the majority of studies described in the literature, the measurements concern either the larynx or the respiratory system. However, as both organs are involved in the production of prosody, it is difficult to get a picture of the whole process in this way. Moreover, in many studies attention was paid to singing and sustained phonation. The results of these experiments do not necessarily generalise to spontaneous speech.

So far research has mainly focused on modelling F_0 (intonation models) and its relation with physiological signals. In the following section we will present some of the physiological explanations of intonation that have been proposed in the literature. Before doing this, however, it may be useful to consider the frequency-to-pressure ratio, because this is an important item in the discussion concerning the physiological explanations of intonation.

1.2.3.1 The F_0-P_{sb} ratio

The F_0-P_{sb} ratio (FPR) is the rate of F_0 change resulting from a change in P_{sb}. Apart from P_{sb}, other physiological mechanisms can also influence F_0 (see Section 1.2.2). In order to verify whether an F_0 change could have been caused by a P_{sb} change alone, it is necessary to determine the values of the FPR when P_{sb} is varied and all other factors are held constant. A frequently used method to attain this consists in quickly pressing the chest or the abdomen of a speaker while he is producing a sustained vowel. This has the effect of reducing V_l, thus raising P_{sb} and F_0.

During an experiment of this kind Baer (1979) measured the activity of various laryngeal muscles in order to check whether the other factors did indeed remain constant. He concluded that during the first 50 ms after the external pressure variation no change in the activity of the laryngeal muscles is to be expected. Therefore, during this lapse of time the signals corresponding to F_0 and P_{sb} can be used to determine the FPR. The FPR values observed in these experiments appeared to vary between 2 and 7 Hz/cm H_2O (Ladefoged, 1963, 1967; Baer, 1979; Rothenberg & Mahshie, 1986; Baken & Orlikoff, 1987).

1.2.3.2 Physiology and intonation

The literature offers a number of different physiological explanations of intonation. In this section we intend to review some of the various opinions on this subject. The views presented here can be divided into three groups, depending on the importance attributed to the different factors:

① P_{sb} is the most important factor

② the laryngeal muscles are the most important factor

③ both P_{sb} and the laryngeal muscles are important

Lieberman (1967) made measurements of P_{sb}, but he did not measure the activity of the laryngeal muscles. He observed that F_0 and P_{sb} behaved in a similar way, except at the end of interrogative utterances. At the end of questions there was an increase in F_0, while P_{sb} generally did not increase. His assumption was that the activity of the laryngeal muscles increased at the end of questions, but remained relatively steady otherwise.

Based on this assumption he concluded that, apart from the end of interrogative utterances, F_0 is a function of P_{sb} alone.

The conclusion of Lieberman (1967) can easily be verified by calculating the FPR in his data. According to Lieberman (1967), the FPR is about 20 Hz/cm H_2O in his data, while Ohala (1990) claims that it is even higher. In any case, P_{sb} alone cannot explain all the variation in F_0 in Lieberman's data, and other mechanisms must have been involved. Various investigations have shown that in speech FPR is often higher than 7 Hz/cm H_2O (Ladefoged, 1963; Ohala, 1978, 1990). This is true especially for local stress-related F_0 variations, but also for the global downtrend in F_0.

At the local level the variation in P_{sb} can usually account for no more than 10 or 20% of the F_0 variation. So, it seems that the remaining variation should be ascribed to other physiological mechanisms. After Lieberman (1967) had advanced his model, in which P_{sb} is the main determinant of F_0, several studies have revealed that the laryngeal muscles are involved in local variations in F_0. This is also what is generally claimed in recent publications. The two views expressed below agree that the local F_0 variations are caused by laryngeal activity, but they do not agree on the physiological explanation of declination.

Various authors are of the opinion that F_0 is mainly controlled by the laryngeal muscles, both at the local and at the global level. For instance, this view is clearly expressed by Breckenridge (1977). She states that declination is part of the linguistic code, and since all linguistically significant F_0 distinctions are implemented using the laryngeal muscles, declination must be controlled by these muscles too. Ohala (1978, 1990) too claims that a model in which linguistic aspects of F_0 are completely determined by the laryngeal muscles is much more likely than a model in which respiratory and laryngeal factors interact. The models proposed by Öhman (1967, 1968) and Fujisaki (1991) are also based on this view. In these two-component models both the local and the global components are controlled by the laryngeal muscles.

In addition to the two above-mentioned extreme views there is also an intermediate position which is shared, among others, by Collier (1975) and Atkinson (1978). In the two-component model by Collier the local component is controlled by the laryngeal muscles, while the declination in F_0 is mainly due to the declination in P_{sb}. Just like Collier (1975), Atkinson (1978) also measured P_{sb} and the activity of some of the laryngeal muscles. Atkinson's conclusion was that in statements P_{sb} is the dominant factor in controlling F_0, while activity of the laryngeal muscles predominates at high F_0 and in questions.

From the overview presented above it appears that most researchers now agree that local F_0 variations are due to the activity of the laryngeal muscles. Which physiological mechanisms account for declination is still a moot point, though. This is expressed in the conclusion of a recent paper by Ohala (1990: 42): “It must be concluded that the question of whether F_0 declination is caused by laryngeal or by respiratory activity has still not been answered definitively.”

1.3. The present research

1.3.1 Aim of the present research

In the preceding sections we have reviewed the most important research findings concerning the relation between physiology and prosody. From this overview it appeared that several aspects of this relation are still unclear. The general aim of the research reported on in this thesis was to gain more insight into the behaviour and the physiological control of the voice source in the production of prosody. More specifically, our aim was to provide a physiological explanation of declination.

In trying to achieve this goal we encountered a number of practical and methodological problems, which were mainly due to the lack of adequate research instruments. Consequently, appropriate analysis methods had to be developed within the framework of this research. Initially, these methods were seen as mere instruments to reach our research aim. However, since these instruments turned out to be indispensable in this investigation, they became intermediate goals in themselves.

1.3.2 Methodology

In this section we describe the methodology used to study the relation between physiology and prosody. The most important methodological problems that emerged in this research will also be discussed, together with the various solutions we decided to adopt.

1.3.2.1 Measuring and processing the data

The first step in our research consisted in selecting the physiological signals to be investigated. Previous studies have shown that P_{sb}, CT, VOC and the strap muscles are important physiological signals in the production of prosody (see Section 1.2.2 and 1.2.3). Given that the various strap muscles are probably identical from a functional point of view (Hast, 1968; Baer et al., 1976; Hardcastle, 1976; Maeda, 1976, 1980; Erickson et al., 1977; Erickson et al., 1983), we decided to measure the activity of just one strap muscle. Since SH is most consistently found to correlate with F_0 (see Section 1.2.2), we chose to record the activity of this muscle.

The first problem we encountered concerned the way in which P_{sb}, CT, VOC and SH had to be measured. An experimental setup had to be constructed to record the signals. Moreover, appropriate methods and software had to be developed to analyse the data.

When this project started, researchers at the university of Nijmegen had extensive experience in measuring and processing pressure signals (Boves, 1984; Cranen, 1987), but they had no experience in measuring and processing the electromyographic (EMG) activity of the intrinsic and extrinsic laryngeal muscles. To acquire the necessary skills, the author spent two months at the Haskins Laboratories in New Haven (Connecticut, U.S.A.). During this period the first experiment was carried out. On this occasion the pressure catheter and the EMG electrodes were positioned by dr. Hiroshi Muta.

The experience acquired at the Haskins Laboratories was subsequently used to develop software for analysing the data and to construct a measuring setup at the Department of Otorhinolaryngology of the University Hospital in Nijmegen, where the second and the third experiments were carried out. This time the pressure catheter and the EMG electrodes were positioned by dr. Philip Blok. In addition to the above-mentioned signals, P_{or} was also recorded in these experiments. P_{or} and P_{sb} were used to calculate P_{tr} ($= P_{sb} - P_{or}$). In all three experiments the speech signal, the electroglottogram (EGG), and lung volume (V_l) were also recorded. In the fourth experiment only the speech signal, EGG and V_l were recorded. For details concerning the data recording, the reader is referred to Section 3.2.2.

Once the signals mentioned above had been recorded, they had to be processed and made available for further analysis. Processing is a cover term for a number of different activities ranging from A/D conversion to smoothing the EMG signals. One of these activities, i.e. averaging the signals, proved to be problematic, as will be explained below. Further details concerning data processing are described in Section 3.2.3.

In analysing physiological signals related to speech, it is advisable to average the signals pertaining to several repetitions of the same sentence. The aim of the averaging procedure is twofold: (1) to overcome the limitations of low signal-to-noise ratios in physiological signals (especially the EMG signals), and (2) to avoid misinterpretations caused by idiosyncrasies of individual tokens.

In order to obtain multiple realizations of one and the same sentence, the subjects involved in the experiments were asked to repeat each sentence a number of times. However, it turned out that with conventional methods it was not possible to obtain meaningful averages of the signals corresponding to the various repetitions, because the temporal variation between the repetitions was too large. Under these circumstances, one would run the risk of averaging the signal corresponding to a given articulatory event in one repetition with the signal of a different articulatory event in another repetition. Considering that signal averaging was a necessary stage in our research, a new method was developed which makes it possible to average signals in a meaningful way (see Chapter 2).

1.3.2.2 The relation between physiological signals and prosodic parameters

Once the processing and averaging of the signals was made possible, the relation between the physiological signals and the prosodic parameters was investigated. In Section 1.2.1 we already pointed out that F_0 is the most studied prosodic parameter, probably because it is often considered the most important parameter from a linguistic point of view. Although the latter may be true, it should be noted that from the point of view of physics and physiology all prosodic parameters are equally important. In addition to F_0, IL and SC also play a major role. However, these parameters have received little attention so far.

Another important point concerning the prosodic parameters is that they are not independent of each other. For example, all other things being equal, an increase in P_{sb} alone will lead to an increase in F_0, but also to an increase in IL and to a different spectral make-up of the speech signal (the spectral tilt will decrease and the higher harmonics will have a relatively greater contribution). Given this interdependence between the various prosodic parameters, it seems that prosodic research should go beyond a mere study of F_0 and should include other relevant acoustic parameters as well. This is exactly what we have tried to do in our research. Although we paid considerable attention to F_0 in our investigation, we also studied other parameters such as IL and SC.

The three prosodic parameters investigated in this study (F_0, IL and SC) are influenced by P_{sb}, CT, VOC and SH, as explained in Section 1.2.2, but in an indirect way. These physiological signals first influence the voice source signal, and in turn, changes in the voice source signal lead to changes in the speech signal. Considering that this is the natural order of these causal relations, it seems advisable to investigate the effect of the physiological processes on the source signal, and the way in which changes in the source signal cause changes in the speech signal.

In order to study the relation between the physiological processes and the source signal, the source signal has to be computed and parameterized. Once the voice source parameters have been calculated, it is possible to study the relations between the following three groups of parameters:

① physiological signals

② voice source parameters

③ prosodic parameters

Given the considerable amount of data for which the source parameters had to be calculated, an automatic method seemed to be preferable. However, since it turned out that such a method was not available at the beginning of our investigation, an automatic method for determining the voice source parameters in speech was developed within the framework of our research (see Chapters 3, 4, 5 and 6).

1.3.2.3 A physiological model of intonation

The relation between the measured physiological signals and intonation was studied in detail (see Chapter 7). As a first step towards a comprehensive physiological model of intonation, we used a qualitative analysis method to study the relation between the measured physiological signals and intonation. Our goal was to find out whether systematic behaviour could be observed in the data, and in which way this behaviour could be modelled. The proposed physiological model of intonation was based on consistent behaviour found in our own data and in the relevant data available in the literature. Subsequently, we tried to derive a quantitative implementation of this model. To that end "trend lines" were fitted manually to the data (see Chapter 8).

Finally, we investigated the relation between the downtrend in F_0 and the downtrend in P_{sb}. Downtrend, declination and downdrift are used to denote the tendency of F_0 to decrease during the course of an utterance, as was already mentioned in Section 1.2.1. But this is not an operational definition that can be used to determine the downtrend in a given F_0 contour. In the past, trend lines, defined in several different ways, have often been used to model downtrend, and the F_0 values of these trend lines were compared to the P_{sb} values. However, the definition of the "trend" determines the outcome of the analysis to a significant extent. Moreover, besides P_{sb} there are other factors that can influence F_0, and in studying the relation between F_0 and P_{sb} one has to take care that the other factors remain constant, or are properly accounted for. Therefore, we decided to use a statistical analysis method in which a correction is made for some of the other factors.

1.4 Outline of this book

This section gives an outline of the research reported on in this thesis. The Chapters 2, 3, 4, 5, 7 and 8 are presented in chronological order, that is according to their date of first publication. Although the article in Chapter 6 is the most recent one, it is placed immediately after Chapter 5 because thematically it is more related to Chapters 3, 4 and 5.

In Section 1.3.2.1 we alluded to the fact that the large amount of temporal variation between different realizations of the same speech signal made it difficult to average the signals in a meaningful way. To overcome this problem, a method was developed in which a dynamic programming algorithm is used to correct for timing differences. This method is described in Chapter 2.

Subsequently, in order to investigate the relation between the physiological signals, the voice source parameters and the prosodic parameters (as explained in Section 1.3.2.3), an automatic method had to be developed to calculate the voice source parameters. In this method an automatic inverse filtering algorithm (see Chapter 4) is used to obtain an estimate of the voice source signal from the speech signal. The various stages of the development of this method are described in Chapters 3 to 6. In Chapter 3 only those voice source parameters are used that can be calculated from the voice source signal by means of simple mathematical operators, i.e. the amplitude parameters that can be calculated by taking minimum and maximum of a signal. In Chapters 4, 5 and 6 a voice source model is fitted to the voice source signal, in order to obtain estimates of other voice source parameters besides the amplitude parameters.

In Chapters 3 and 4 the relation of the calculated voice source parameters to the physiological signals and the prosodic parameters is studied. In Chapter 5 two automatic parameterization methods are compared, while in Chapter 6 the behaviour of a voice source model in different optimization algorithms is studied. The goal of the research reported in Chapters 5 and 6 is to improve the automatic parameterization method.

The main goal of the research reported in Chapter 7 is to propose a comprehensive physiological model of intonation. A qualitative analysis was used to investigate how P_{sb}, CT, VOC and SH are used in the production of intonation.

As explained in Section 1.3.1, the specific aim of our research was to arrive at a physiological model of declination. To this end, the frequently observed simultaneous downtrend in F_0 and P_{sb} was studied. The goal pursued in Chapter 8 was to determine whether the downtrend in F_0 could be due to the downtrend in P_{sb}.

Chapter 2

A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech

Abstract

In analysing physiological signals related to speech, it is necessary to average several repetitions in order to improve the Signal-to-Noise Ratio. However, in a recent experiment we found considerable differences in the articulation rate of repeated realizations of a medium length utterance, especially for untrained subjects. This makes averaging of related physiological signals a non-trivial problem. A new method of time-alignment and averaging of the physiological signals is described. In this method a dynamic programming algorithm is used, which successfully corrects for the timing differences between the repetitions.

This chapter is a slightly revised version of the following paper: Strik, H. & Boves, L. (1991a) A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech. *Journal of Phonetics*, 19, 367-378.

2.1 Introduction

A quantitative study of the physiological basis of speech production requires the simultaneous measurement of acoustic signals and a number of physiological signals. The usual procedure to overcome the limitations of low Signal-to-Noise Ratios in physiological signals, and to avoid misinterpretations caused by idiosyncrasies of single tokens, is to average multiple repetitions of the "same" utterance (Collier, 1975; Baer et al., 1976; Maeda, 1976; Atkinson, 1978). To allow averaging, the utterances must be lined up in time. To that end, line-up points must be defined in every repetition. Typical choices are distinctive events like the release of a plosive or the onset of voicing, preferably close to the middle of the utterance.

This method of linear time-alignment and averaging cannot be applied to all speech signals in the same way. For instance, no averaging is usually applied to fundamental frequency (F_0) signals, because discontinuities (e.g. at voiceless segments) preclude straightforward averaging. Instead, the F_0 contour of one of the repetitions is chosen to represent the "average" F_0 contour (Collier, 1975; Maeda, 1976; Atkinson, 1978).

The applicability of the method is further limited by the inherent variability of speech production. Two types of variation must be distinguished, viz. variation in articulation and variation in speaking rate. The inappropriateness of averaging over substantially different articulations is obvious: the average will not meaningfully represent any of the qualitatively different articulatory events. Averaging over substantially different rates causes similar problems; if variation in speaking rate is large, the portions of the signals being averaged will correspond to the same articulatory event only very near to the line-up points.

The two kinds of variation are not independent, since extreme changes in speaking rate can effect a change in articulation as well. In an earlier study (Strik & Boves, 1988b) we found considerable differences in the speaking rate for repetitions of a medium length utterance. However, the range of speaking rates was such that rate-induced articulatory variations are unlikely to be a first-order effect. Therefore, we designed a technique to overcome the effects of temporal variation. We propose a novel processing technique in which a Dynamic Programming (DP) algorithm is used to time-align the tokens in a non-linear way, so that meaningful averaging remains possible.

The proposed method corrects for the variation in speaking rate, but then there is still the problem of variation in articulation. However, it is safe to assume that repeated realizations of the same utterance are fairly similar. The data from an experiment with quasi-spontaneous speech of an untrained subject were used to check this assumption a posteriori.

This chapter is organised as follows. First, the method is described. After a brief explanation of the general DP algorithm, an overview is provided of the specific procedure for non-linear time-alignment and averaging of physiological signals related to speech. A more detailed description of the six component stages will then follow. The same data were also used to test the method. Results of analyses carried out both with the method of linear and that of non-linear time-alignment and averaging were compared.

2.2 Method of non-linear time-alignment and averaging

2.2.1 The DP algorithm

We begin with a brief description of the DP algorithm (see, e.g., Sakoe & Chiba, 1978, for a more detailed explanation). DP has been successfully used in speech recognition, where it is often called Dynamic Time Warping (DTW). Our algorithm is based on the flowchart in Sakoe & Chiba (1978). It finds the optimal time registration between two patterns, a reference pattern R of length J and a test pattern T of length I, as illustrated in Figure 1(a). Both patterns are sequences of feature vectors derived from the speech signals by appropriate feature extraction. The frames of the two patterns define a grid of IxJ points (Figure 1a).

A suitable distance metric is used to calculate the distance at point $P_k = (i,j)$ between frame i of test pattern T and frame j of reference pattern R: $d[P_k] = d[T_i,R_j]$. A path P is a sequence of K grid points: $P = P_1, P_2, P_3, ..., P_k, ..., P_K$ (see Figure 1a). The total distance between T and R for a given path P is the weighted sum of the local distances:

$$D_P[T,R] = \sum_{k=1}^{K} C_k * d[P_k].$$

By definition, the optimal path P_o is the path that minimizes $D_P[T,R]$. The path P_o represents a function F, called the warping function, which realizes a mapping from the time axis of T onto that of R. The warping function F, or the optimal path P_o, can be

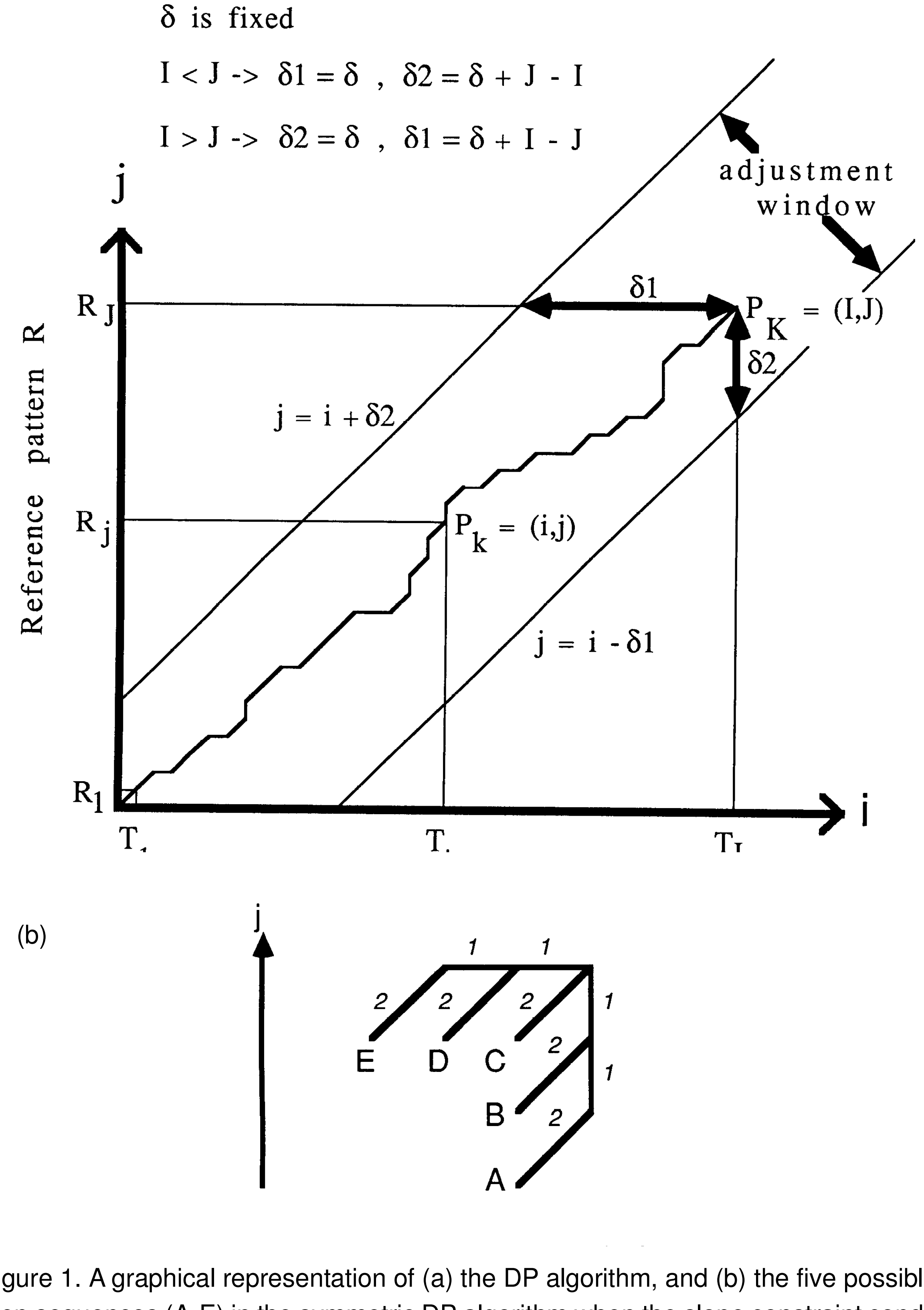


Figure 1. A graphical representation of (a) the DP algorithm, and (b) the five possible step sequences (A-E) in the symmetric DP algorithm when the slope constraint condition is 0.5. The numbers in italics are the weighting coefficients C_k.

used to normalize the time axis of T with respect to the time axis of R. When there are no timing differences between T and R, the path P_o coincides with the line i=j.

The path P is usually constrained to start at $P_1 = (1,1)$, end at $P_K = (I,J)$, and to remain within an adjustment window. In the method proposed in this chapter a slope constraint condition of 0.5 (see Sakoe & Chiba, 1978) is used, which means that a diagonal step can be followed, or preceded, by at most two off-diagonal (i.e. horizontal or vertical) steps. The consequence is that only the five step sequences given in Figure 1b are allowed. That is, each legal path is some concatenation of a set of these five step sequences. This symmetric form DP-matching is used because Sakoe & Chiba found that it gave better results in speech recognition than the asymmetric form.

2.2.2 General overview of the method

The method of non-linear time-alignment and averaging of physiological signals, proposed here, can be split into six successive stages:

① specification of line-up points;

② selection of a reference pattern;

③ calculation of cepstrum coefficients of the acoustic signals;

④ calculation of a warping function for each token (DP);

⑤ mapping of the physiological signals, using the warping function; and

⑥ calculation of median values and variation of time-normalized physiological signals.

A necessary requirement for this method is that all physiological signals be sampled at the same sampling frequency (F_S). For the experiment used for evaluation of the method, F_S is 200 Hz, so the sampling time ($T_S = 1/F_S$) is 5 ms. The individual stages are described below.

2.2.2.1 Specification of line-up points

Even though DP has proved useful in speech recognition, for the purpose at hand some modifications seemed necessary. First of all, in basic speech research one is often interested in the (average) physiological signals before and after an utterance. However, it is difficult to obtain a useful time registration path by comparing silence with silence. Also, it is often desirable to have an exact time-alignment of a particular event in an utterance to study the (average) physiological signals in the neighbourhood of this event. Therefore, our method allows one to define several line-up points in an utterance, which are time-aligned exactly; the DP algorithm is only applied between those line-up points (Figure 2). The first line-up point is interpreted as the beginning of the utterance, and the last one as the end of the utterance. Before the first line-up point, and after the last line-up point, the time registration path runs diagonally (Figure 2).

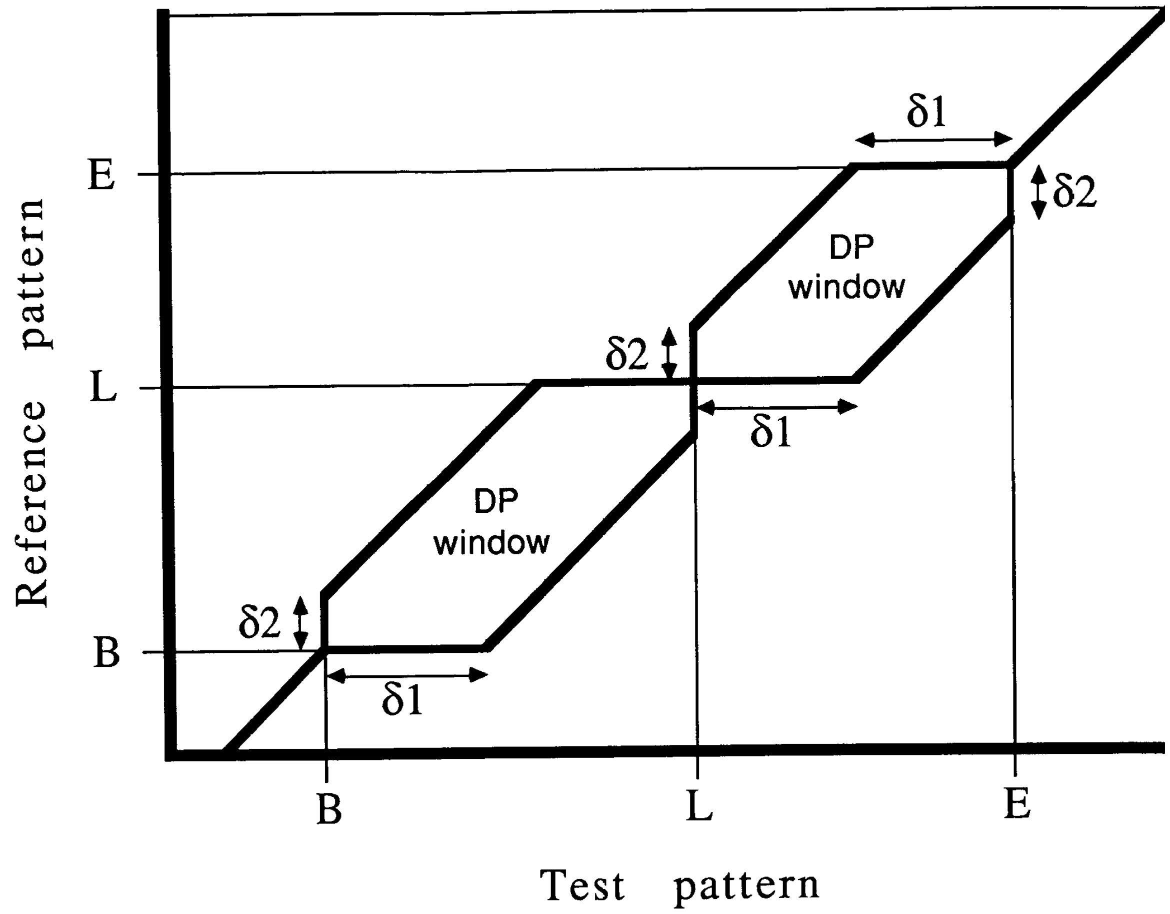


Figure 2. A graphical representation of non-linear time-alignment, when three line-up points are used. B indicates the beginning of the utterance, E the end and L an acoustic event near the middle of the utterance.

2.2.2.2 Selection of a reference pattern

One of the tokens is chosen as a reference for time-normalization of the remaining tokens. The best choice for this reference pattern or template seems to be the token with median length, because it requires the least adaptation in the other tokens.

2.2.2.3 Calculation of feature vectors

The recording conditions during experiments in which several physiological signals are measured are often such, that the Signal-to-Noise Ratio (SNR) of the audio signals is not high. The current method should also be applicable to audio signals with mediocre SNR. Cepstrum coefficients are known to give good results in speech recognition (Davis & Mermelstein, 1980; Paliwal & Rao, 1982). Therefore, the first 12 cepstrum coefficients are used as feature vectors. The speech signals were digitized with a sampling frequency of 10 kHz and submitted to a 12th order LPC analysis using a 250 point Hamming window and a window shift of T_S = 5 ms. The vectors of LPC coefficients were subsequently transformed to vectors of 12 cepstrum coefficients (Markel & Gray, 1976).

2.2.2.4 Determination of optimal time registration path

In the fourth stage the warping function has to be found that minimizes the distance between test pattern and reference pattern. The exact choice of the distance metric does not seem critical for our purpose. A simple Euclidian distance measure proved to be sufficient. However, the definition of the adjustment window is critical. Because there can be a substantial difference in the length of patterns under comparison, we used the adjustment window shown in Figure 1a, which is different from the one given by Sakoe & Chiba (1978).

2.2.2.5 Transformation of the physiological signals

The warping functions computed in the previous stage describe the differences in the temporal structure of all tokens relative to the reference token, i.e. they allow normalization of the time axes of the tokens by mapping them onto that of the reference token. Since the physiological signals are measured on the same time axis as the speech signal, their time axes can be normalized using the warp functions derived from the speech signals.

The time-normalized or warped signal W is computed from the original signal S by using a non-linear function F_n: $W(j) = F_n[S(i)]$. The calculation starts at grid point $P_K = (I,J)$, and backtracks to grid point $P_1 = (1,1)$. Because only the five step sequences given in Figure 1b are allowed, the function F_n has to be defined only for these five partial paths. For time compression, step sequences D and E in Figure 1b, W(j) is obtained by averaging over two and three samples, respectively. For time stretching (step sequences A and B), W(j) and preceding samples are obtained by linear interpolation (Figure 3). And for a single diagonal step, step sequence C, no local transformation of the time-axes is made.

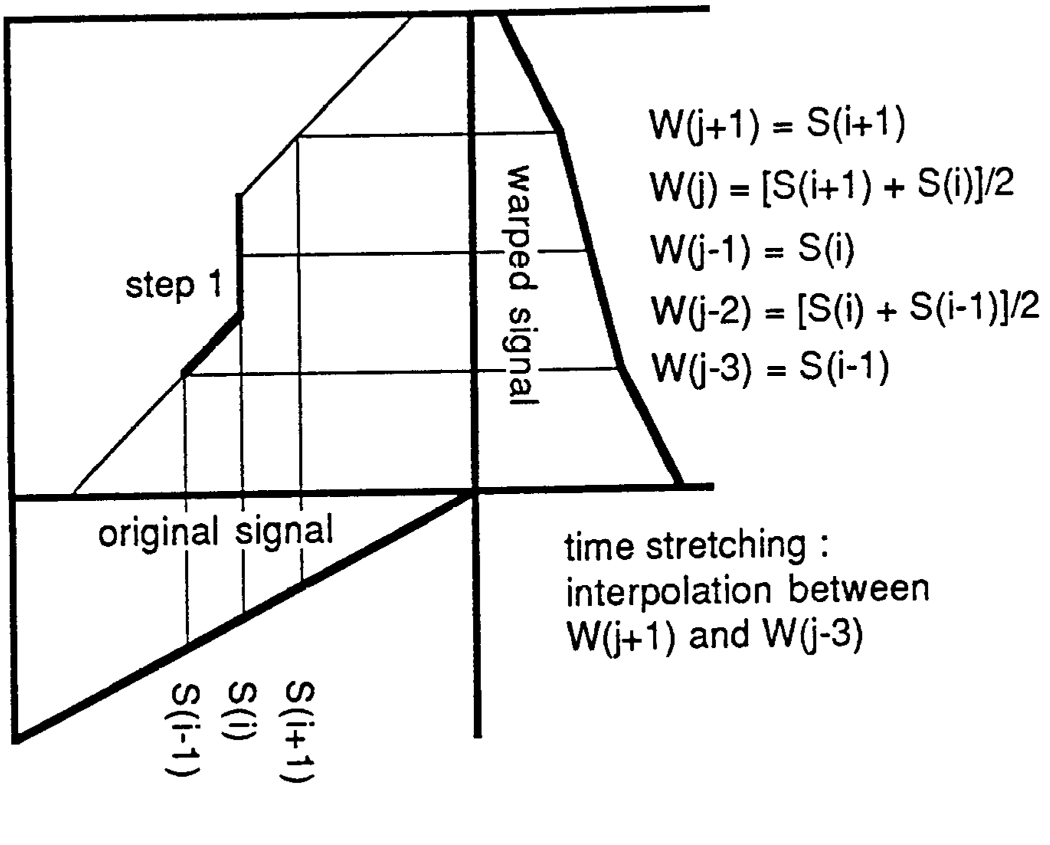


Figure 3. An example of the function F_n for time stretching (step sequence A). In this example a straight line is used as the input signal S, which is drawn at the bottom. Shown is how linear interpolation is used to calculate the five output samples from the three input samples.

The result is a function F_n that is defined in the following way:

Step sequence A	$W(j) = [S(i+1) + S(i)]/2$
	$W(j-1) = S(i)$
	$W(j-2) = [S(i) + S(i-1)]/2$
Step sequence B	$W(j) = [S(i+1) + 2*S(i)]/3$
	$W(j-1) = [2*S(i) + S(i-1)]/3$
Step sequence C	$W(j) = S(i)$
Step sequence D	$W(j) = [S(i) + S(i-1)]/2$
Step sequence E	$W(j) = [S(i) + S(i-1) + S(i-2)]/3$

As it is impossible to determine a meaningful warping function for the silent intervals before and after the utterances, the time structure is left unchanged here by letting the path run diagonally (Figure 2).

2.2.2.6 Averaging

For every physiological process the expected value of the time-normalized signals must be computed. We prefer the median over the arithmetic mean value, since it reduces the effect of outliers. The median signals are then smoothed. In addition to the median value, a measure of the variation around the median (the phonatory or articulatory variation) can also be important. We found that the range spanned by all but the n largest and n smallest values, where n is of course (much) less than half the number of available tokens, is a useful measure of variation.

The method of averaging described above is appropriate for continuous signals. But F_0, one of the signals that has received much attention in speech research, is a discontinuous signal. For unvoiced frames F_0 was set to zero. We found that taking the median value of F_0 gives the appropriate voiced-unvoiced decision and the desired average F_0 value.

2.3 Experimental evaluation

To compare the methods of linear and non-linear time-alignment, we used data from an experiment in which simultaneous recordings were made of the acoustic signal, electroglottogram (EGG), lung volume (V_l), subglottal pressure (P_{sb}), supraglottal or oral pressure (P_{or}), and electromyographic (EMG) activity of the sternohyoid (SH) and vocalis (VOC) muscles. An untrained male subject was asked to produce an utterance spontaneously. His response was: *Ik heb het idee dat mijn keel wordt afgeknepen door die band* ("I have the feeling that my throat is being pinched off by that band"). He was then asked to repeat that sentence 29 times. All physiological signals were then pre-processed to obtain signals with a sampling rate of 200 Hz. This experiment is described in more detail elsewhere (Strik & Boves, 1992a).

The original, spontaneous sentence deviated from the 29 repetitions in that it included a pause of almost half a second when the subject swallowed. In order to minimize the risk that utterances containing different articulatory gestures were averaged, only the last 29 sentences were used for analysis.

2.3.1 Variation in speaking rate

The oscillograms of three audio signals are shown in Figure 4. It is obvious that there are large differences in the durations of the utterances. The mean length of the 29 utterances was 2310 ms (sd = 130 ms), while the maximum and the minimum length were 2615 ms and 2165 ms, respectively.

The release of the /k/ of "*keel*" was used as the line-up point for the method of linear time-alignment. This line-up point was chosen because it is expected to be clearly distinguishable, and it is situated near the middle of the sentence. The mean duration of the first part (from beginning to the line-up point) was 880 ms (sd=80 ms), with a maximum of 1075 ms and a minimum of 780 ms. The mean duration of the last part (from line-up point to the end) was 1430 ms (sd=70 ms); the maximum and minimum values were 1590 ms and 1320 ms. The subject increased his articulation rate as he repeated the utterances more often, but even for the last six sentences the ranges for the first and last parts were 120 ms and 90 ms, respectively. So even after numerous repetitions the variation is still so large that straightforward averaging of the tokens could result in combining physiological signals of different articulatory movements.

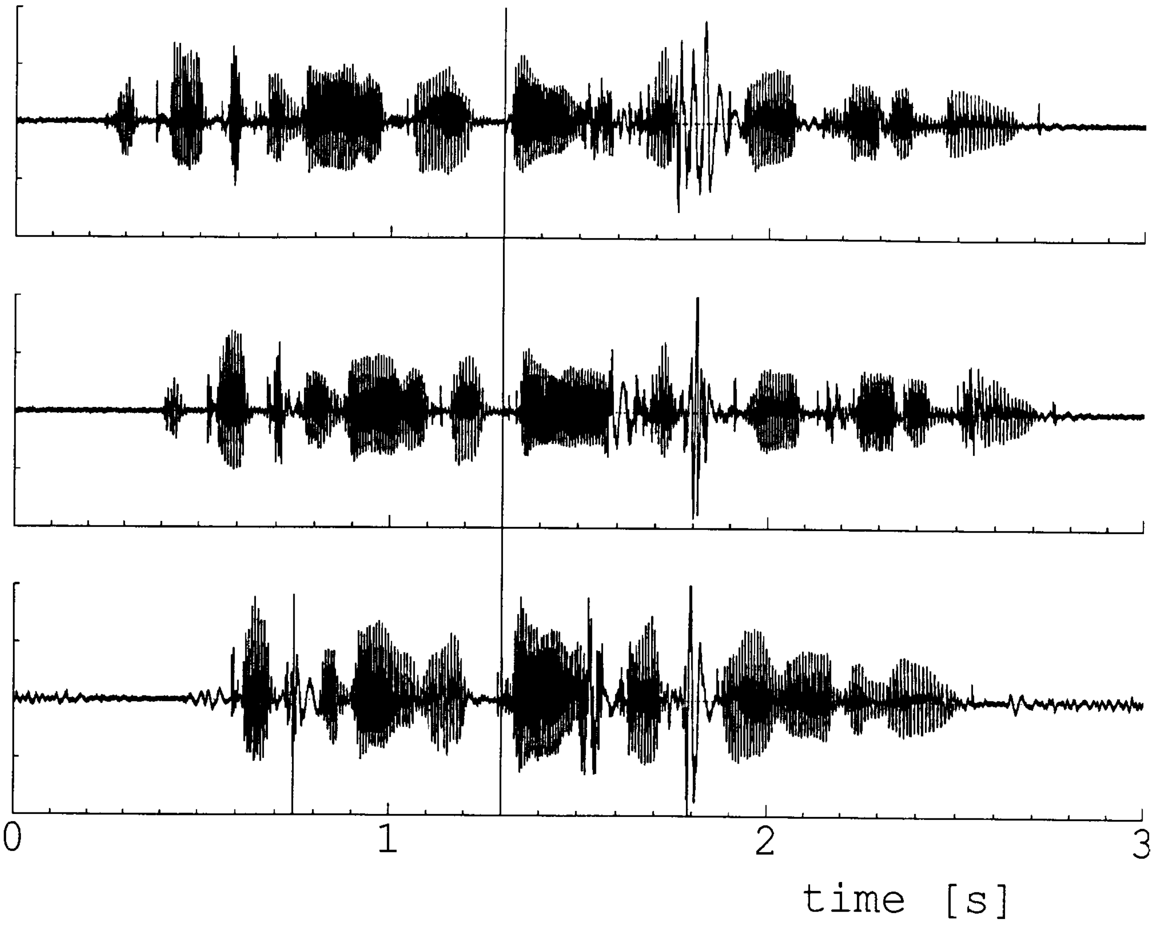


Figure 4. Oscillograms of the audio signals of three repetitions of the same utterance. Here and in Figures 5-9 the straight vertical line at 1.3 s connects the line-up points of the individual signals.

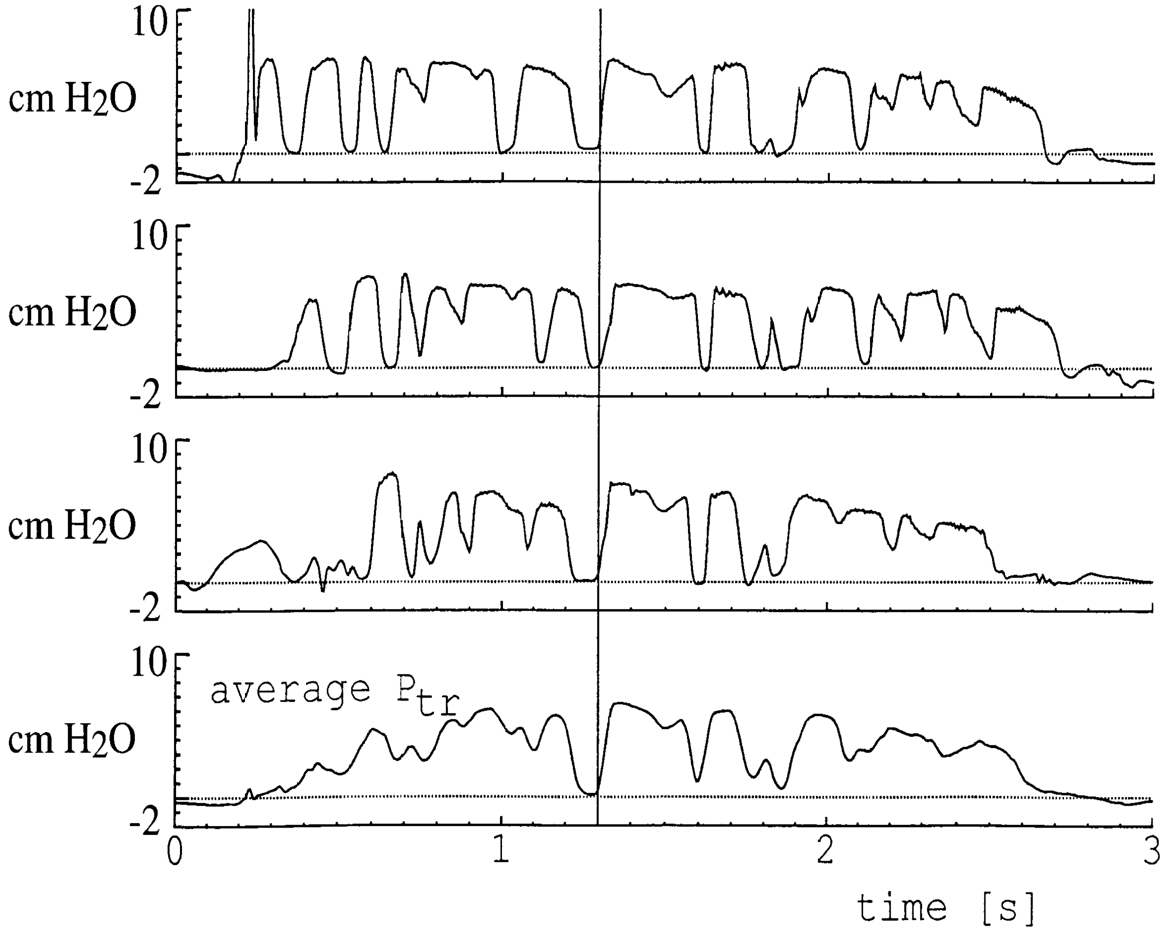


Figure 5. The three upper panels display the transglottal pressure signals of the three utterances given in Figure 4. The lower panel contains the average transglottal pressure signal for 29 repetitions.

2.3.2 Method of linear time-alignment

Although we did not expect linear time-alignment to produce meaningful results, we still wanted to test its viability. In the three upper panels of Figure 5 the time-aligned transglottal pressure (P_{tr}) signals, corresponding to the audio signals of Figure 4, are shown. The timing differences are very large, and the time-alignment is reasonable only just before and after the line-up point, as reflected in the average signal in the bottom trace, which becomes increasingly meaningless towards both beginning and end of the utterance.

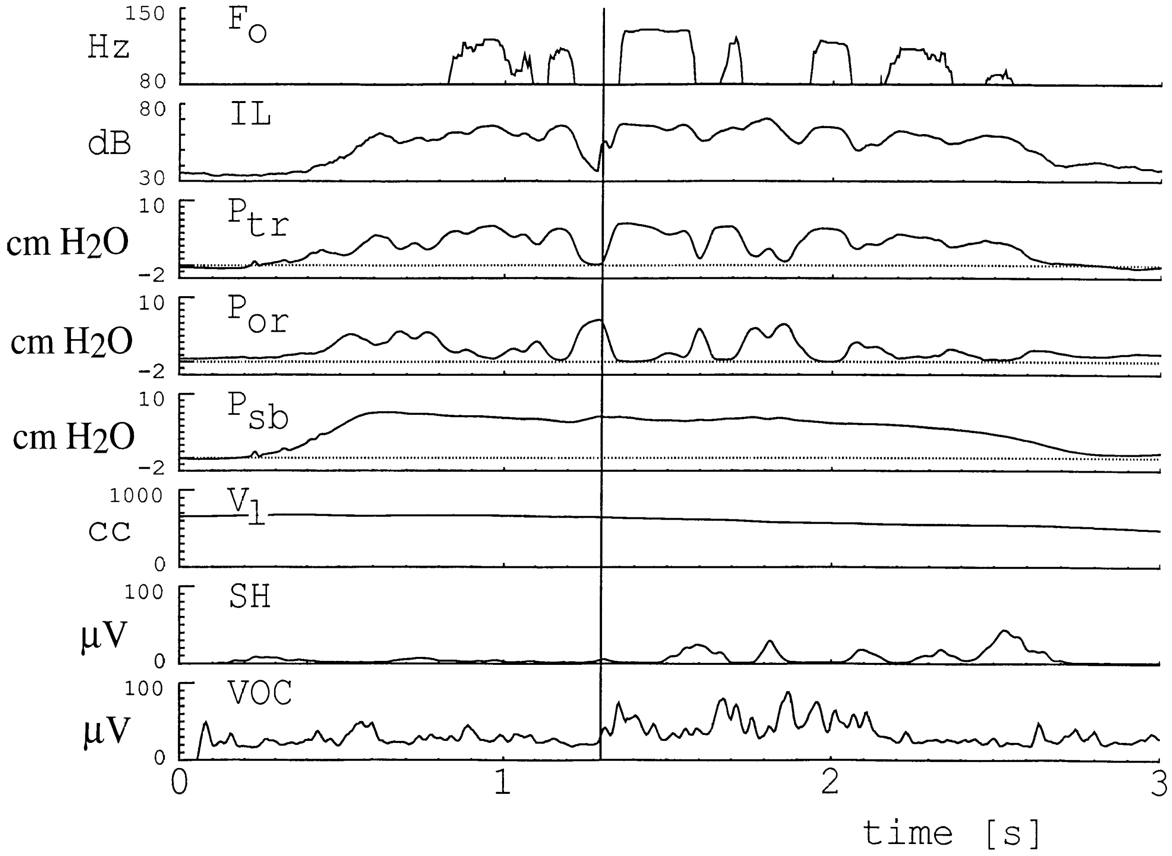


Figure 6. Average physiological signals for fundamental frequency, intensity level, transglottal pressure, oral pressure, subglottal pressure, lung volume, and electromyographic activity of the sternohyoid and vocalis muscles, obtained by the method of linear time-alignment.

Figure 6 shows the average signals for F_0, Intensity Level (IL), P_{tr}, P_{or}, P_{sb}, V_l, SH and VOC. Especially for F_0, IL and the pressure signals, it is apparent that the averages are only meaningful in the immediate neighbourhood of the line-up point.

2.3.3 Method of nonlinear time-alignment and averaging

For the method of non-linear time-alignment and averaging, warping functions were calculated for all tokens using the token with median length (2295 ms) as the template. These warping functions were then used to map the physiological signals. Before averaging the signals, we checked whether the degree of time-alignment obtained by warping the signals was sufficient.

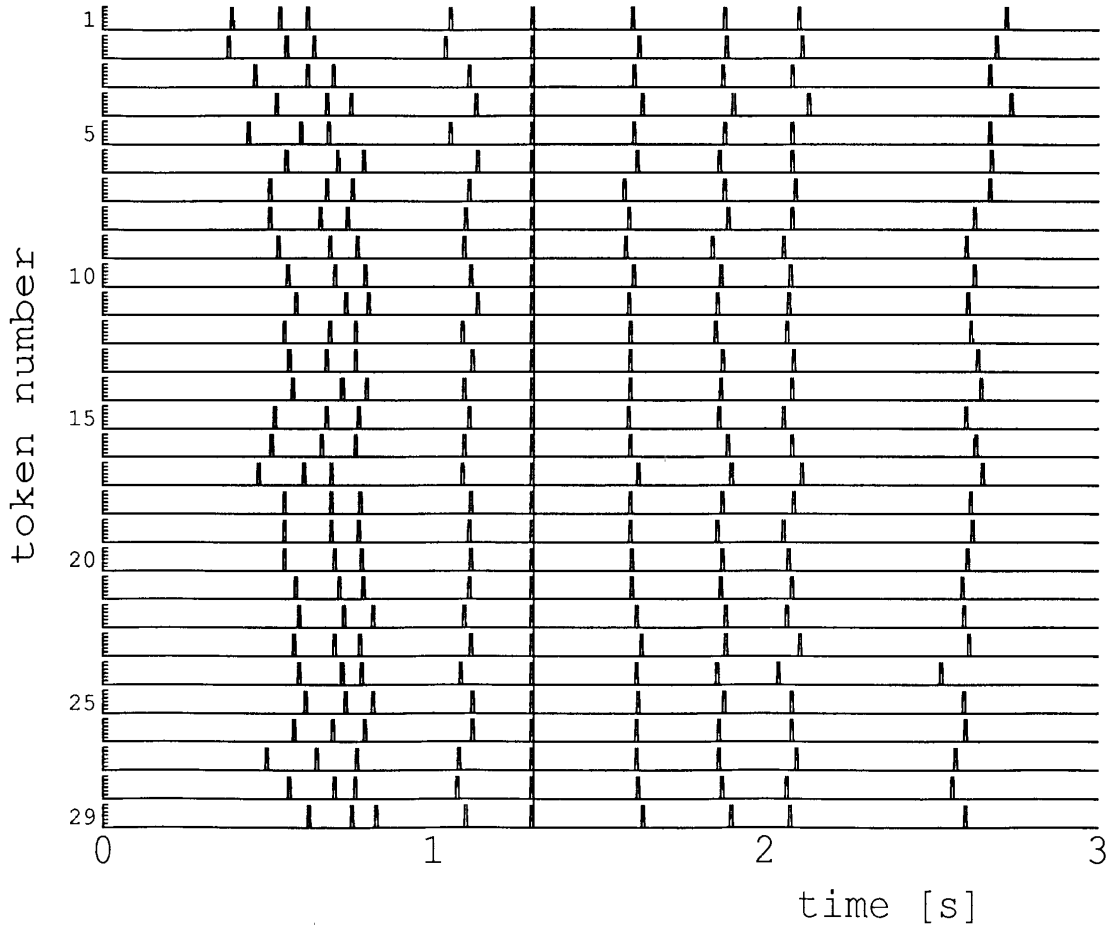


Figure 7. The labels of the 29 utterances after linear time-alignment.

To that end, nine labels were placed manually in all 29 tokens at marked acoustic events. Labelled events were releases of unvoiced plosives, one of them being the /k/ used as line-up point for linear time-alignment. This is shown in Figure 7, where the fifth label is the /k/ used as line-up point. Away from the line-up point the degree of time-alignment among the labels diminishes. As a matter of fact, the timing differences are already fairly large for the two neighbouring labels, 4 and 6. The largest timing differences were found at the beginning of the utterances.

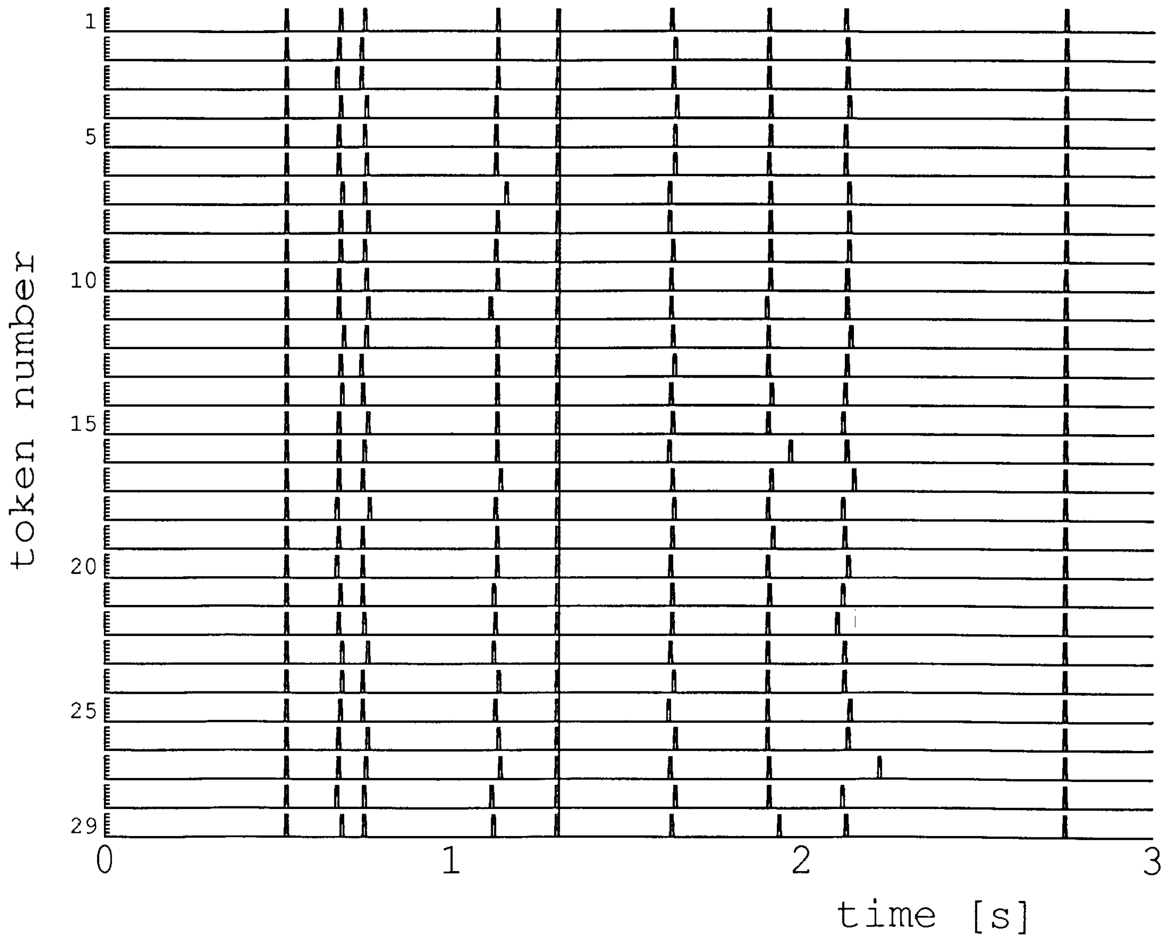


Figure 8. The labels of the 29 utterances after non-linear time-alignment.

The warping functions were then used to time-align the labels, and the result is shown in Figure 8. Apart from some inaccuracies, all labels (i.e. the corresponding acoustic events) seem to be aligned very well. Because the acoustic events of the whole sentence are time-aligned by non-linear time-alignment, meaningful averaging at this stage seems possible.

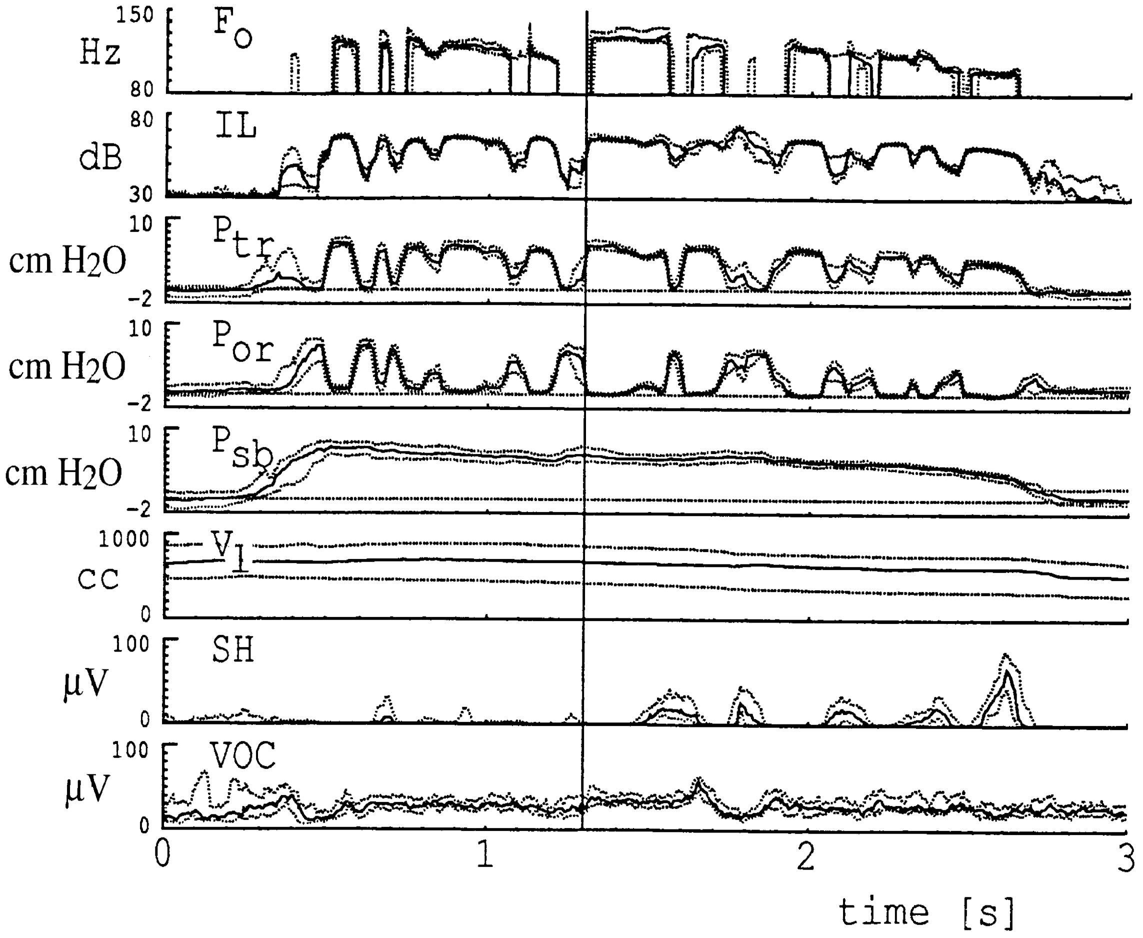


Figure 9. Median physiological signals (solid lines) for fundamental frequency (F_0), intensity level (IL), transglottal pressure (P_{tr}), oral pressure (P_{or}), subglottal pressure (P_{sb}), lung volume (V_l), and electromyographic activity of the sternohyoid (SH) and vocalis (VOC) muscles, obtained by the method of non-linear time-alignment and averaging. The dotted lines are a measure of the amount of variation (see text).

Median signals are plotted in Figure 9. It can be seen that the median signals are not only meaningful near the line-up point, but also towards the beginning and the end of the utterance.

2.3.4 Variation in pronunciation

Non-linear time-alignment seems successful in time-aligning the acoustical events of all utterances to a reasonable degree. However, for meaningful averaging another requirement must be fulfilled; the different realizations of the utterances must be produced with essentially the same articulatory gestures. We cannot test whether the movements of the articulators were very much alike in the different utterances, but we can gauge the amount of variation of some relevant physiological signals.

The dotted lines in Figure 9 give an idea of the range of the middle 20 values at each time instant (see method). From these traces we can infer that, apart from V_l, the amount of variation of the physiological signals among the different realizations of an utterance is within reasonable bounds.

2.4 Discussion and conclusions

Both for trained (Strik & Boves, 1988b) and for untrained subjects (this study), we have found a substantial degree of time variation between repetitions of a medium length utterance. Even after numerous repetitions these timing differences did not disappear. With such differences in temporal structure, linear time-alignment and averaging no longer seems a useful procedure with which to extract meaningful relations.

A possible solution might seem to be the following. Define several line-up points in each repetition, time-align these line-up points, and do linear time-alignment in between. However, the timing differences are not distributed uniformly, and therefore the number of line-up points needed to obtain a reasonable overall time-alignment would be very large.

We have shown that the method of non-linear time-alignment and averaging presented here, works satisfactorily, despite the mediocre SNR of the speech signals and the highly non-stationary character of the noise. Thus, the technique of DP, developed in the framework of automatic speech recognition, can also be a very useful tool in basic research for processing physiological (or comparable) signals related to speech. After time normalization, median values are obtained for all measured physiological quantities. These median values can be used for further analysis.

The method of non-linear time-alignment has some further advantages. In contrast with the method of linear time-alignment, it also yields an average signal for F_0. Furthermore, the technique can be used (semi-) automatically, which makes it very attractive in a research situation that is characterized by the need to handle large amounts of signals. Finally, the method can be used to time-align and average all kinds of signals for which timing differences are apparent.

Chapter 3

Control of fundamental frequency, intensity and voice quality in speech

Abstract

In this chapter the control of fundamental frequency, intensity level of the radiated acoustic signal and voice quality is studied in normal conversational speech. It is shown that the physiological factors that best explain the features of the speech wave measured, depend on the part of the utterance taken into account. Also, it appears that in speech transglottal pressure is more important than subglottal pressure. We conclude that currently available mathematical models that describe the waveform of glottal volume flow lack a number of parameters necessary for a better understanding of the physiological control of the speech parameters investigated in this study.

This chapter is a slightly revised version of the following paper: Strik, H. & Boves, L. (1992a) Control of fundamental frequency, intensity and voice quality in speech. *Journal of Phonetics*, 20, 15-25.

3.1 Introduction

The relation between subglottal pressure (P_{sb}) and laryngeal configurations, on the one hand, and fundamental frequency (F_0), intensity level (IL), and glottal volume flow (U_g) on the other is extremely complex. Moreover, it is difficult to measure P_{sb} and especially the laryngeal configurations and the ways in which they are brought about. Perhaps owing to measurement problems, most investigations of laryngeal control and its effects on the radiated acoustic signal have dealt with sustained vowels produced in widely different ways, rather than with "normal" speech production. In many studies F_0 was varied over several octaves and P_{sb} over a range from approximately 5 cm H_2O to well above 30 cm H_2O. There are several reasons why the results obtained in those studies may not be directly applicable to speech production. In "normal, neutral" speech the ranges are much smaller. Thus some of the control mechanisms needed to span the wide ranges in "phonation experiments" may be much less important in speech. Also, in sustained vowels, oral pressure (P_{or}) may be considered equal to atmospheric pressure. But in speech production, where non-negligible constrictions of the vocal tract occur, P_{or} is much more important. In the present study we have looked into the relation of laryngeal characteristics to IL and F_0 in normal speech. We will touch upon some methodological aspects of the research. Also, we will pay due attention to the role of P_{or} and transglottal pressure (P_{tr}).

3.2 Material and methods

3.2.1 Experimental procedure

The subject in this study was a male native speaker of Dutch, with no experience in phonetics or linguistics and with no history of respiratory or laryngeal dysfunction. During the production of various utterances (sustained vowels, sentences with different intonation patterns) simultaneous recordings of the acoustic signal, electroglottogram (EGG), lung volume (V_l), P_{sb}, P_{or}, and electromyographic (EMG) activity of the sternohyoid (SH) and vocalis (VOC) muscles were obtained. Near the end of the recording session he was asked to produce an utterance spontaneously. He replied by saying (in Dutch): *Ik heb het idee dat mijn keel wordt afgeknepen door die band* ("I have the feeling that my throat is being pinched off by that band"). After having uttered this sentence, he was asked to repeat it 29 times.

3.2.2 Data recording

The speech signal was transduced by a condenser microphone (B&K 4134) placed about 10 cm in front of the mouth, and amplified by a measuring amplifier (B&K 2607). The EGG was recorded with a Fourcin-Abberton laryngograph (Fourcin, 1974).

The pressure signals were recorded using a Millar ® catheter with four miniature pressure transducers, in the way described by Cranen & Boves (1985). The catheter was introduced into the pharynx via the nose, and then into the trachea via the posterior commissure. It did not have a noticeable effect on phonation (Boves, 1984).

The EMG signals were recorded using hooked-wire electrodes (Hirose, 1971). The electrodes were inserted percutaneously, and correct electrode placement was confirmed by audio-visual monitoring of the signals during various functional manoeuvres. The perimeter of chest and abdomen were measured with mercury filled strain-gauge wires (Strik & Boves, 1988d). All signals were recorded on a 14-channel instrumentation recorder (TEAC XR-510), using a bandwidth of 5 kHz.

3.2.3 Data processing

All signals were A/D converted off-line at a 10 kHz sampling rate. The files were stored on a microVAX computer. Because of the sluggishness of the articulators it seems sufficient to use a sampling frequency of 200 Hz. Therefore, the goal of preprocessing is to derive physiological signals which all have a sampling rate of 200 Hz.

F_0 and IL were calculated with the SIF (Standard Inverse Filtering) program of ILS (Interactive Laboratory System). Both values were calculated every 5 ms, resulting in F_0 and IL signals sampled at a 200 Hz rate. Pressure signals, chest and abdomen signals were low-pass filtered and downsampled to 200 Hz. Lung volume was calculated from the low-pass filtered chest and abdomen signals.

The integrated rectified EMG was calculated in the way described by Basmajian (1967): first the signal is full-wave-rectified, and then it is integrated over successive periods of 5 ms. The integrator is reset after each integration. Finally, the signal is smoothed by convolving it with a triangular function (base length 35 ms).

There is a time delay between the change of the electric potential of a muscle and the resulting effect in the acoustic signal (Atkinson, 1978). To overcome this delay, all EMG signals were shifted forward over their mean response times.

After preprocessing, median signals were calculated with the method of non-linear time-alignment and averaging that is described in Strik & Boves (1991), in which the fifth repetition was used as a point of reference.

3.2.4 The parameters of the glottal volume flow

There are a number of different ways to parameterize the glottal volume velocity waveform (Fant, 1986; Cranen & Boves, 1987; Klatt & Klatt, 1990). We will adopt the widely used Liljencrants-Fant (LF) model in this chapter, although we do not believe that it is the best model from a physiological point of view: many of its features seem to be motivated by perception, i.e. by the ease with which they allow one to approximate or explain (spectral) characteristics of the speech wave that are important from a perceptual point of view. On the other hand, most of the parameters can also be related to what is known about the physiology of phonation. Specifically, the LF-model allows one to describe the maximum amplitude of the flow during the open glottis interval (U_0), the duty cycle of the flow pulses, the amount of skewing of the pulses, the amplitude of dU_g at the moment of glottal closure (E_e) and the time delay between the moment of major vocal tract excitation and the instant where the glottal flow becomes quasi-constant (T_a).

3.2.5 Calculation of glottal volume flow

Of course, no direct recordings of the glottal volume flow were made; this signal is derived from the speech waveform by means of inverse filtering. Closed glottis interval covariance LPC was used to estimate the parameters of the inverse filter. In De Veth et al. (1990) it was shown that this procedure outperforms more complicated ones that attempt to estimate the parameters of the inverse filter by means of Robust ARMA analysis.

Inverse filtering yields an estimate of dU_g. Integration of this signal gives the flow signal. For the present chapter we only wanted to measure peak glottal flow U_0, excitation strength E_e and P_{tr} for each voiced period. The value of E_e is obtained by taking the minimum of the differentiated flow in each pitch period. Likewise, U_0 is found by looking for the maximum of the flow signal. P_{tr} is measured at the moment of maximum glottal flow. Its value is obtained from a low-pass filtered pressure signal.

Inverse filtering was done on the fifth utterance, because that is the one used as a point of reference in the method of non-linear time-alignment. Inverse filter results were obtained for all voiced periods, including vowels and voiced consonants.

3.3 Results

3.3.1 Control of fundamental frequency

The relation between P_{sb} and F_0 has been addressed by many experimental (e.g. Collier, 1975; Maeda, 1976; Atkinson, 1978; Strik & Boves, 1989) and modelling studies (e.g. Ishizaka & Flanagan, 1972; Titze & Talkin, 1979). Yet, the details of this relation remain unclear. Estimates of the F_0 to P_{sb} ratio from speech and special phonation tasks resulted in values between 5 and 15 Hz/cm H_2O (Collier, 1975; Maeda, 1976; Strik & Boves, 1989). In another type of experiment pressure variations are induced externally. The F_0 to P_{sb} ratios measured in these experiments tend towards values of 2-7 Hz/cm H_2O (Baer, 1979; Strik & Boves, 1989). Strik & Boves (1989) showed that the ratio of an F_0 change resulting from a P_{sb} change alone is probably the same in both experiments, viz. 2-7 Hz/cm H_2O. In “normal” speech there are other factors that control F_0, especially the laryngeal muscles. Owing to the simultaneous operation of these factors, the ratio of total F_0 change to P_{sb} change in utterances is often larger than 2-7 Hz/cm H_2O. The latter ratio is in agreement with the ratio of 2-3 Hz/cm H_2O that was found by Ishizaka & Flanagan (1972) for their self-oscillating two-mass model.

Furthermore, it seems that in most experiments, and therefore in most presently existing models, the effects of P_{or} on F_0 are not sufficiently taken into account. This is probably due to the fact that most experiments were done with sustained vowel phonation, in which the variation in P_{or} is much smaller than in normal speech. Strik & Boves (1988d) studied the relation of F_0 to P_{sb}, P_{or} and P_{tr} in connected speech.

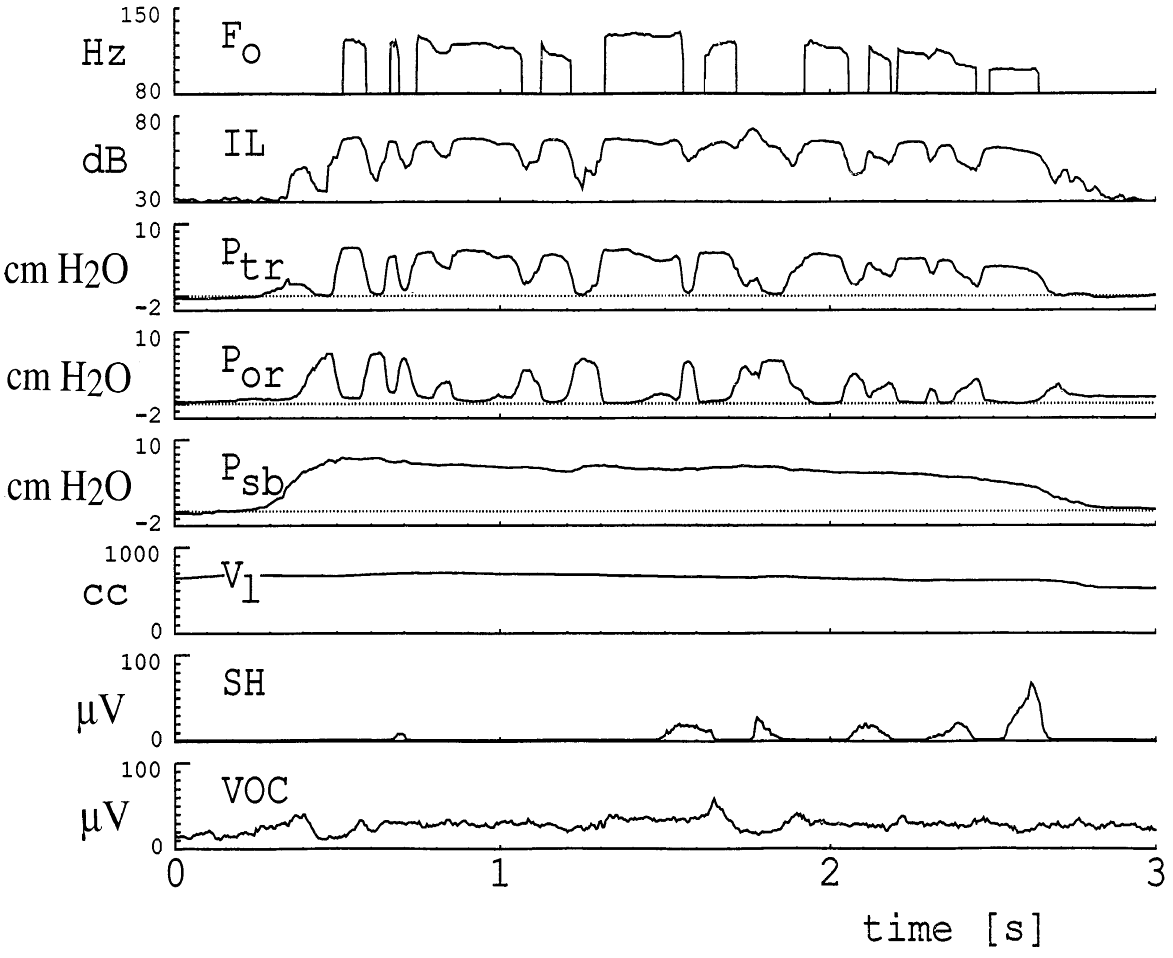


Figure 1. Median physiological signals, obtained by the method of non-linear time-alignment and averaging. Plotted are, from top to bottom, F_0, IL, P_{tr}, P_{or}, P_{sb}, V_l, SH and VOC.

The median signals for the 29 sentence repetitions of this experiment, obtained with the method of non-linear time-alignment, are shown in Figure 1. These signals were used to calculate correlations between the variables of interest. In Table 1 the correlations are given for a long voiced interval (i.e. the third voiced interval in Figure 1), while Table 2 contains the same correlations for all voiced frames.

Table 1. Correlation matrix, means and standard deviations of the median physiological signals for a voiced interval (N=66, if |r|>0.315 then p<0.01).

	F_0	IL	P_{tr}	P_{or}	P_{sb}	mean	SD
F_0	1.000	0.808	0.851	-0.783	0.478	118.58	3.70
IL		1.000	0.960	-0.983	0.111	63.23	3.38
P_{tr}			1.000	-0.968	0.274	5.42	0.88
P_{or}				1.000	-0.054	1.16	0.91
P_{sb}					1.000	6.35	0.16

Table 2. Correlation matrix, means and standard deviations of the median physiological signals for all voiced frames (N=293, if |r|>0.151 then p<0.01).

	F_0	IL	P_{tr}	P_{or}	P_{sb}	mean	SD
F_0	1.000	0.667	0.729	-0.153	0.772	115.87	8.59
IL		1.000	0.923	-0.663	0.492	62.20	4.25
P_{tr}			1.000	-0.638	0.612	4.95	1.17
P_{or}				1.000	0.211	0.89	0.95
P_{sb}					1.000	5.65	0.90

The most important conclusion that can be drawn from the data in Tables 1 and 2 is that the pattern of correlations between P_{sb}, P_{or}, P_{tr} and F_0 depends very much on the part of the utterance over which the measurements are taken. If measurements are limited to a single voiced interval, P_{tr} (and P_{or}) are much better predictors of F_0 than P_{sb} (see Table 1). When measured over a complete utterance, however, P_{sb} and P_{tr} explain essentially the same proportion of the variation in F_0 (see Table 2). This is due to the fact that the range of P_{sb} in individual voiced intervals is rather small (see Table 1). The range spanned by P_{tr}, on the other hand, is much wider, because of the fact that P_{or} varies between P_{sb} in voiceless stops and zero in open vowels. In a complete declarative utterance, on the other hand, the correlation between P_{sb} and F_0 is much enhanced by the fact that both show some amount of declination. The data in the tables were obtained from a single subject and therefore should be verified on a larger population. Yet, from a physiological point of view (as well as on statistical grounds) they seem to be quite plausible.

Our results show that one must be very cautious in interpreting the outcomes of experiments on the physiological control of F_0 (and all other speech parameters, for that matter). Such caution is, of course, the more necessary with respect to single subject studies, like our present study. One must be especially cautious in generalizing the results of experiments to other situations than those under which they were obtained. In fact, only results that can be explained by a fairly comprehensive model may be generalized to situations where a similar model can be assumed, operating in the same regime. We are confident that the conclusion of our investigation are supported by a sufficiently complete model.

3.3.2 Control of intensity and voice quality

Even though the relation between P_{sb} and F_0 has received some attention in the literature, it is important to bear in mind that the effects of P_{sb} and the laryngeal configurations are not limited to F_0; on the contrary, factors like the acoustic power generated at the glottis and the waveshape of the glottal volume flow pulses are also affected. These relations are much less studied. That may, at least in part, be due to the assumption that voice intensity and voice quality are of less importance from a linguistic point of view. However, if it comes to a better understanding of the fundamentals of phonation and of para-linguistic phenomena like voice quality and its variations, radiated intensity and details of the glottal volume velocity waveform become of crucial importance. In the present study we contribute some measurement data related to the control of IL and voice quality obtained from connected speech and show how these data can fit in with modelling research.

3.3.2.1 The relation between IL and pressure

It has long been known that there must be a relation between P_{sb} and IL, if only because P_{sb} is the major source of phonatory energy (cf. Rubin, 1963). However, most measurement data on the relation between P_{sb} and IL seem to stem from in vitro experiments, or at best from experiments where sustained vowels were produced with intensity and pressure variations spanning a range larger than that usually found in conversational speech (e.g. Isshiki, 1964; Bouhuys et al., 1968; Cavagna & Margaria, 1968; Tanaka & Gould, 1983).

In our own investigation of the best predictor of IL in the production of voiced speech sounds, we found that P_{tr} outperforms P_{sb} by far (Strik & Boves, 1988d). The result is true both on a local (i.e. within words or voiced intervals) and on a global level (i.e. looking over complete utterances). In both situations the correlation between P_{tr} and IL exceeds 0.92, while the correlation with P_{sb} is at most 0.49 (when measured over a complete sentence, see Tables 1 and 2). So, at least for this subject, it seems that P_{tr} is more important in the control of IL than P_{sb}.

Isshiki (1964), Bouhuys et al. (1968), Cavagna & Margaria (1968), and Tanaka & Gould (1983) all found high correlations between IL and the logarithm of P_{sb} when subjects produced sustained vowels. For sustained vowel phonation P_{or} is almost constant and close to zero and, as a result, P_{tr} is almost equal to P_{sb}. In our data P_{or}, P_{tr} and IL vary quickly and considerably, while P_{sb} decreases slowly during the course of the utterance (Figure 1). This explains why in our data the relation between IL and P_{sb} is rather weak.

In addition to the correlation coefficient (r), the regression coefficient (slope) is also of importance, because it predicts the amount of change in IL due to a given change in P_{tr}. In order to be able to compare our findings with previous results, we calculated the regression equation between IL and the logarithm of P_{tr}. Based on the 293 voiced frames of the median signals of the current experiment (see Figure 1), we found the following relation:

$$\mathrm{IL} = 41.6 + 30.3 * \log(P_{tr}) \qquad (N = 293,\ r = 0.923)$$

Or, in other words, the intensity (Int) of the radiated speech wave is proportional to P_{tr} to the power 3.03. Interestingly enough, the value of the power in the resulting relation between Int and P_{tr} is quite comparable to results reported in the literature about the relation between Int and P_{sb}. For sustained vowel phonation Cavagna & Margaria (1968) found a value of 3.0 ± 1.0, Isshiki (1964) found a value of 3.3 ± 0.7, and Tanaka & Gould (1983) found a value of 3.18; while Bouhuys et al. (1968) reported a value of 3.0 for singing.

At a first glance it seems strange that comparable regression equations are found for different relations (IL and P_{tr} vs. IL and P_{sb}), obtained for different kinds of speech (normal conversational speech vs. sustained phonation) and different ranges of IL and pressure (2-7 cm H_2O vs. 2-60 cm H_2O). But closer inspection reveals that the two relations are not really different. For sustained vowel phonation and singing of constant tones P_{or} is usually close to zero, and thus P_{sb} and P_{tr} are almost equal. Therefore, for these modes of phonation, the relations between IL and P_{tr} and between IL and P_{sb} are very similar. The conclusion is that the relation between IL and P_{tr} obtained by Isshiki (1964), Bouhuys et al. (1968), Cavagna & Margaria (1968) and Tanaka & Gould (1983) for sustained phonation and large ranges of IL and P_{sb} is comparable to the relation obtained in this experiment for normal conversational speech.

It may still be that P_{sb} is an important factor in the control of IL, certainly if it is varied over ranges that are much wider than those normally found in conversational speech, but that are not unusual in singing or in very loud speech. However, our data suggest that the faster variations of IL related to articulatory manoeuvres are primarily determined by variations in P_{or} that cause similar variations in P_{tr}, whereas the gradual decrease of IL observed during many (declarative) utterances in a large number of languages is caused by a gradual decrease in P_{sb}. The finding that IL is mainly controlled by P_{tr} makes it interesting to further investigate the detailed way in which IL is influenced by P_{tr} via the characteristics of the glottal volume flow.

3.3.2.2 Flow waveform characteristics and P_{tr}

From the literature it is known that the parameters E_e and U_0 in the LF-model have most effect on IL (Gauffin & Sundberg, 1989). Thus we measured E_e and U_0 for all 181 pitch periods of the fifth repetition, for which reliable inverse filter results could be ob-

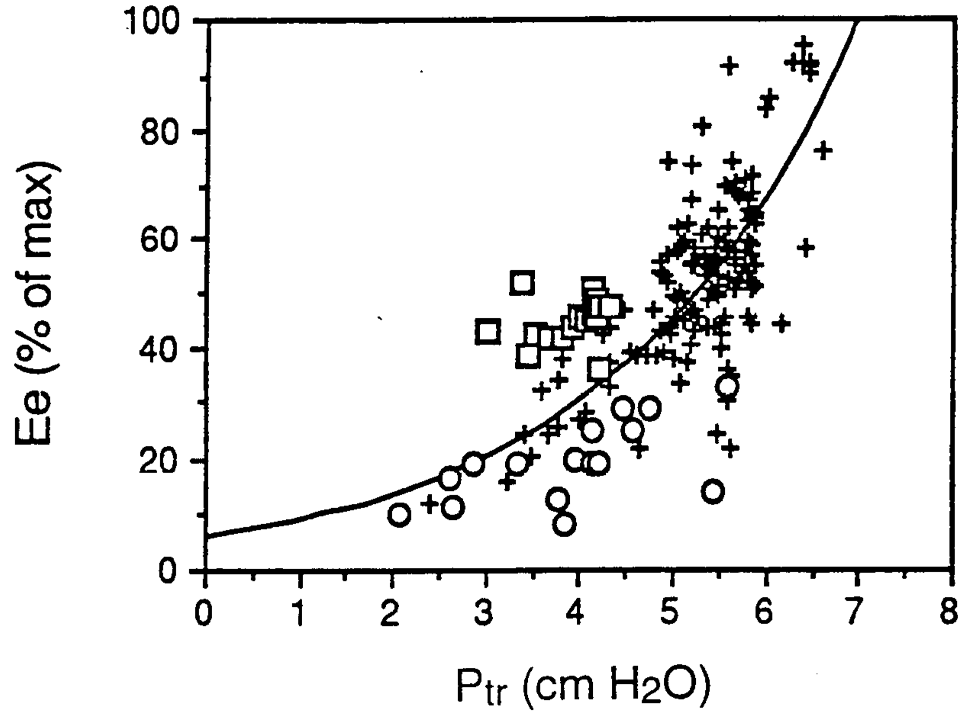


Figure 2. Scatterplot of E_e and P_{tr}. The regression line for the exponential fit for the data of the category "steady phonation" is: $E_e = 6.1 * 10^{0.174 * Ptr}$, r = 0.79, N = 148.

In Figures 2, 3 and 4 the data are divided in three categories: steady phonation (+), V↔UV transitions (O) and /ɑ/ (❐). E_e and U_0 values are always given relative to the maximum observed value for each quantity.

tained. Most of these periods pertained to vowels, but a substantial part comes from voiced consonants. We wanted to examine the relation of E_e and U_0 to P_{tr}.

The relation between E_e and P_{tr} is shown in Figure 2 . It seems as if this relation shows three different regimes. The bulk of the samples (148 out of a total of 181) falls into the category of, what we call, steady phonation. For the data of this category an exponential fit (r = 0.79, see Figure 2) is slightly better than a linear fit (r = 0.73). The second category consists of the pulses in V-UV transitions (i.e. both V→UV and UV→V transitions). For this category E_e is often relatively lower, compared to steady phonation, especially at the beginning of voicing. On the other hand, for the vowel /ɑ/ from the very last syllable of the utterance, E_e is relatively higher (the reasons for treating the utterance final syllable separately are more fully explained in Section 3.3.3).

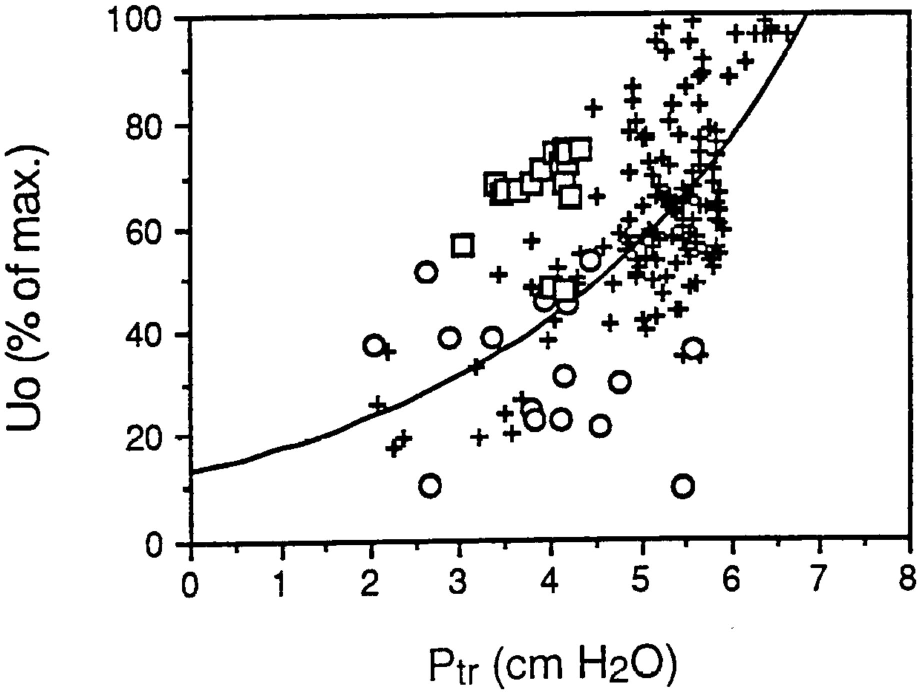


Figure 3. Scatterplot of U_0 and P_{tr}. The regression line for the exponential fit for the data of the category "steady phonation" is: $U_0 = 12.7 * 10^{0.130 * P_{tr}}$, r = 0.72, N = 148.

The correlation between U_0 and P_{tr} is depicted in Figure 3. Again, for the data of the category "steady phonation", an exponential fit (r = 0.72, see Figure 3) is somewhat better than a linear fit (r = 0.65). The data for the vowel /ɑ/ of the last syllable still deviate considerably from the regression line, while the data for V-UV transitions are scattered on both sides of the regression line.

From a look at the spectra of the glottal flow waves it is immediately apparent that the spectral slopes in the three regimes are quite different. In the pitch pulses taken from vowel onsets and from the final stressed syllable the spectral tilt is much steeper than in the central pitch periods of the vowels taken from the beginning and middle of the utterance. The slope difference is more than large enough to have clear perceptual consequences. Thus, the observed differences in voice quality are of sufficient interest to take them into account in the description of speech production and to model them in high quality speech synthesis.

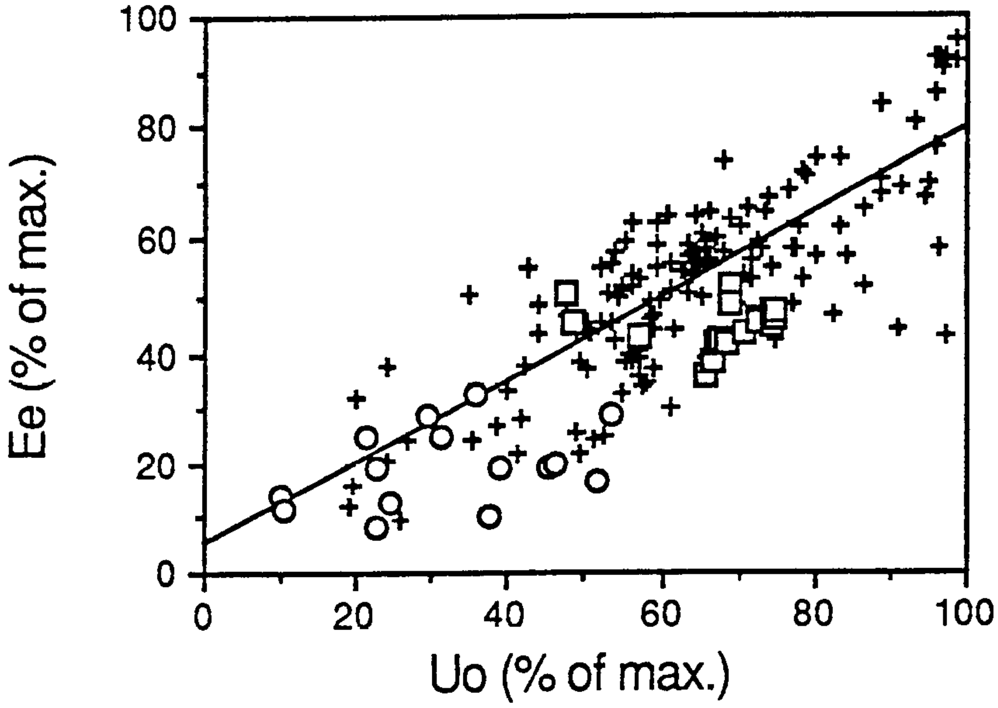


Figure 4. Scatterplot of E_e and U_0. The regression line for the linear fit for the data of the category “steady phonation” is: $E_e = 4.9 + 0.75 * U_0$, r = 0.80, N = 148.

Although E_e and U_0 appear as separate parameters in the LF-model, they are not unrelated themselves, since E_e is dU_g at the moment of major excitation. Thus, if U_0 increases, E_e should also increase, everything else being equal. Therefore, we looked at the relation between E_e and U_0, which is shown in Figure 4. For steady phonation the correlation between E_e and U_0 is very high (viz. 0.80). Apparently the effect of other parameters (like T_0, skewness, and duty cycle) on this relation is not large in “normal” speech.

3.3.3 The utterance final syllable

Towards the end of the utterance F_0, IL, P_{tr} and P_{sb} decrease substantially, while there is a marked increase in the SH activity during the last syllable (see Figure 1). This phenomenon, the so called final fall, has been observed often (Collier, 1975; Maeda, 1976; Strik & Boves, 1989). Presumably, the larynx returns to its rest position, and the lowering of the larynx already starts before phonation has stopped (Maeda, 1976). One would expect that these gross changes in the posture of the larynx should affect the mode of vibration of the vocal folds. This observation motivated a separate study of the glottal flow pulses in the utterance final vowel.

The fact that the characteristics of the vowel in the utterance final syllable deviate from those of the preceding vowels was also observed by Klatt & Klatt (1990). For the last syllable they found increased noise in the F3 region of the spectrum, indicating a greater glottal airflow. But they also found a weaker first harmonic (relative to the amplitude of the second harmonic) in an utterance final syllable, indicative of a pressed voice with a slightly smaller open quotient. Therefore they introduced a novel breathy-laryngealized mode of vibration.

We tried to verify their hypothesis by comparing the data of the (stressed) vowel /ɑ/ of the last syllable (in the word "band"), with the data of the first (unstressed) vowel /ɑ/ in the utterance (in the word "dat"). The spectrum of the utterance final vowel indeed showed increased noise at frequencies above roughly 1.4 kHz. But the amplitude of the first harmonic (relative to the amplitude of the second harmonic) was about 1.5 dB stronger, and the open quotient was approximately 50% in both vowels. Consequently, there is evidence for a breathy mode of phonation at the end of this utterance, but not for laryngealization.

In general, everything else being equal, a decrease in P_{tr} would lead to a decrease in U_0 (Ishizaka & Flanagan, 1972). P_{tr} decreases from 5.5 cm H_2O for the first vowel /ɑ/, to 4.2 cm H_2O for the last vowel /ɑ/, but the amplitude of the AC-component of glottal flow (U_0) increases by roughly 6%. Substantial differences in the degree of adduction are not likely, since the open quotient is about 50% in the two vowels. Presumably, in the utterance final vowel the vocal folds are slackened, either to facilitate the maintenance of voicing with decreased P_{tr}, or due to a general relaxation of muscular activity and a preparation for breathing at the end of an utterance.

In comparing the two vowels /ɑ/ it appears that there is a decrease in E_e of approximately 14% in the last vowel, even though U_0 increases slightly. We found that, generally, the effect of other parameters (like T_0, skewness, and duty cycle) on the relation between E_e and U_0 is not very large for the data of the present experiment (see Figure 4). However, for the last vowel /ɑ/ T_0 is substantially larger than T_0 of the first vowel /ɑ/. After correction for this temporal difference, i.e. when the same number of flow pulses are plotted on the same horizontal scale for both vowels, no major differences in the shape of the glottal volume flow are observed. Consequently, the change in U_0 (+6%) combined with the change in F_0 (-20%) determines the change in E_e (-14%). The fact that, apart from time-stretching, no major differences were found in duty cycle and shape of the glottal pulses between the two vowels /ɑ/, also indicates that the degree of adduction has not changed substantially.

3.4 Conclusion

In this chapter we have shown that the control of F_0, IL and voice quality in normal speech may be somewhat different from what is known from the literature on studies based on sustained vowels or singing. In speech P_{tr} seems to be more important than P_{sb}, mainly because P_{or} cannot be considered as constant and negligible. Also, it was shown that the relative importance of physiological parameters that affect F_0, IL and voice quality depends very much on the nature of the speech from which they are derived. Although the results are based on a single subject study, they fit in very nicely with current models of the physiology of phonation.

Especially from the results concerning the control of IL and voice quality, it became clear that descriptive mathematical models of the glottal flow waveform do not allow one to make the step from description to explanation. High correlations were found for E_e and P_{tr}, and for U_0 and P_{tr}. But in the LF-model there is no relation between E_e and P_{tr}, or between U_0 and P_{tr}, for the simple reason that P_{tr} does not figure in the model. Thus, the LF-model will never allow one to explain these relations, or why several different regimes should exist in the relation between P_{tr} and basic parameters in the model. One will have to resort to models that have a firm physiological basis, like those proposed in Titze (1984) and Cranen (1991).

Chapter 4

On the relation between voice source parameters and prosodic parameters[1] in connected speech

Abstract

The behaviour of the voice source characteristics in connected speech was studied. Voice source parameters were obtained by automatic inverse filtering, followed by automatic fitting of a glottal waveform model to the data. Consistent relations between voice source parameters and prosodic parameters were observed.

This chapter is a slightly revised version of the following paper: Strik, H. & Boves, L. (1992b) On the relation between voice source parameters and prosodic features[1] in connected speech. *Speech Communication*, 11, 167-174.

4.1 Introduction

Modern text-to-speech systems produce speech that is intelligible, but not quite natural. This lack of naturalness is at least in part due to the absence of adequate prosody control. Prosody does not only include fundamental frequency (F_0), intensity (Int) and duration, but it also regards more subtle aspects of the speech signal that can be subsumed under the cover term “voice quality”. Completely satisfactory prosody will therefore require the use of adequate voice source control rules. This opinion is reflected by the fact that many rule-based text-to-speech systems are now being updated, in order to replace a static voice source with a source that can be dynamically controlled. A number of different voice source models have been proposed, each with its own specific advantages and drawbacks. However, it is not our intention to compare different models. Even the most sophisticated voice source model will not improve speech quality if it is not being controlled by the right rules. These rules, on the other hand, cannot be derived without a large amount of data on the behaviour of the voice source in natural speech, or more specifically, of the behaviour of those characteristics of the source that can be mapped onto the model parameters. Fortunately, most modern source models share a large number of parameters, so that most of the results obtained with one model should be easy to generalize to other models.

In a text-to-speech synthesis framework all relevant properties of the voice source can be described in terms of the glottal volume flow signal and its time derivative. These glottal flow signals can be approximated, starting from the acoustic speech signal, via inverse filtering. Model parameters can then be estimated by fitting the model waveform to the inverse filtered waveforms. Inverse filtering and model fitting could in principle be done interactively. However, interactive measurements would take an inordinate

[1] The title of the article was “On the relation between voice source parameters and prosodic features in connected speech”. In this title the term “prosodic features” was used to indicate the acoustic features of the speech signal that are relevant to prosody. In this book the term “prosodic parameters” is used for this purpose, while the label “prosodic features” is reserved for features such as stress, intonation and the like (see Section 1.2.1). In the interests of consistency, the term “features” is replaced by “parameters”.

amount of time, because rule development requires one to process large quantities of speech. Moreover, interactive measurements are difficult to reproduce. For these reasons a procedure was developed to derive the voice source parameters automatically, as will be explained in Section 4.2.

Up to now, most research on voice source characteristics has dealt with sustained vowels produced in different ways. For sustained vowels, which are usually recorded with a high SNR, automatic extraction of the voice parameters is fairly easy. But it is difficult to extrapolate from data acquired from isolated speech sounds to rules for connected speech. Therefore, our aim is to study the behaviour of the voice source in connected, preferably spontaneous speech. In addition to steady state portions of vowels we also want to extract source parameters for voiced consonants, as well as for voiced/unvoiced (V/UV) and UV/V transitions. The results of our work are presented in Section 4.3.

The strategy we adopt to find relations between several voice source parameters on the one hand, and between voice source parameters and prosody on the other, is the following: first we derive general relations by averaging over all data; after that we look for local deviations from these general relations. Special attention is given to the relation between voice source parameters and prosodic parameters like F_0, intensity (Int) and voice quality.

4.2 Method and material

4.2.1 Speech material

To study voice source characteristics data were collected for four male subjects. For all subjects recordings were made of the speech signal, electroglottogram (EGG), subglottal (P_{sb}) and oral (P_{or}) pressure, lung volume and electromyographic activity of some laryngeal muscles (mostly cricothyroid, vocalis and sternohyoid). The signals were stored on wide band FM-tape. All recordings were made at the Department of Otorhinolaryngology of the University Hospital in Nijmegen, in a room in which no special acoustic precautions were taken. For the current chapter only data of one subject were used (Strik & Boves, 1992a). Near the end of a recording session he was asked to produce an utterance spontaneously. His response was: *Ik heb het idee dat mijn keel wordt afgeknepen door die band* ("I have the feeling that my throat is being pinched off by that band"). He then repeated this utterance 29 times. The 30 utterances had an average length of 2.3 seconds. For this chapter inverse filter results of the first four utterances were analysed.

4.2.2 Inverse filtering

The speech signal was transduced by a condenser microphone (B&K 4134) placed about 10 cm in front of the mouth, pre-amplified at the microphone (B&K 1619), and amplified by a measuring amplifier (B&K 2607) using the built-in 22.5 Hz high-pass filter to suppress low-frequency noise. The speech signal was A/D converted off-line at a 10 kHz sampling rate, and processed with a phase correction filter in order to undo the low-frequency phase distortion caused by the high-pass filter.

Closed glottis interval covariance LPC analysis was used to estimate the parameters of the inverse filter. In De Veth et al. (1990) it was shown that this technique for estimating the inverse filter is as powerful as more sophisticated techniques, like Robust ARMA analysis. The moment of glottal closure was determined from the EGG, and it is used to position the analysis window. Inverse filtering yields an estimate of the differentiated glottal volume flow (dU_g); integration of dU_g gives the flow signal (U_g).

Closed glottis interval inverse filtering is a complex process; its implementation requires several choices to be made to fix parameters. The most important parameters are the length and exact position of the analysis window, the pre-emphasis factor, and the order of the analysis. In general, there seems to be no combination of these parameters that is optimal for each individual pitch period in a normal speech utterance. However, a 12th-order LPC analysis with a pre-emphasis factor of 0.95 worked satisfactorily for almost all pitch periods.

Thus window position and window length were left as the parameters to be varied. Instead of trying to formulate criteria that would allow one to determine the unique optimal combination of window length and position for each period, we decided to try a large number of combinations and to leave it to a simple statistical procedure to make the final selection (see Section 4.2.4.).

4.2.3 Voice source parameters

For automatic fitting of a glottal waveform model to inverse filtered flow signals we used a special software package (Jansen et al., 1991). The fit is done pitch synchronously. The periods are defined by the minima in dU_g, because these time points can be located most reliably. This software package makes it possible to use different glottal waveform models, different definitions of the error function and different optimization routines. The choices made for this study are given below.

The so-called LF-model was used, because it seems useful for synthesis, and because it has already been studied in great detail (Fant et al., 1985). The model and its parameters are presented in Figure 1. The relations between the dimensionless wave shape parameters of the LF-model and the spectrum are well-known (see e.g. Fant & Lin, 1988): R_g has a small influence on the amplitude relations of the lower harmonics, R_k influences the spectral balance, and R_a influences the spectral tilt.

The difference between the model and the measured signals can be described by means of an error function. The latter can be defined in the time domain, the frequency domain, or in both domains simultaneously. In this study the error function is based on the time signals of flow and flow derivative. In a pilot experiment it was found that this error definition minimizes the number of discontinuities in the signals fitted to U_g and dU_g. For a given pitch period the error function is calculated by subtracting the modelled signal from the measured signal. The best fitting model waveform is found by adapting the model parameters in such a way that the energy in the error function is minimized.

An adaptive nonlinear least-squares optimization algorithm called NL2SNO (Dennis et al., 1981) was used to find the best fit. The algorithm returns the (minimized) error energy, and the parameters for which that optimum is found. If the minimal error is smaller than a pre-defined threshold, then the fit is said to be good. But if the minimal error remains above the threshold, then all LF-parameters for that pitch period are set to -1 to indicate that the fit is not successful.

4.2.4 Averaging the results

Inverse filtering was done for all 25 combinations of 5 window lengths (33, 34, 35, 36 and 37 samples) and 5 window shifts (-2, -1, 0, 1, and 2 samples relative to the moment of glottal closure). The LF-parameters were obtained for all 25 resulting inverse

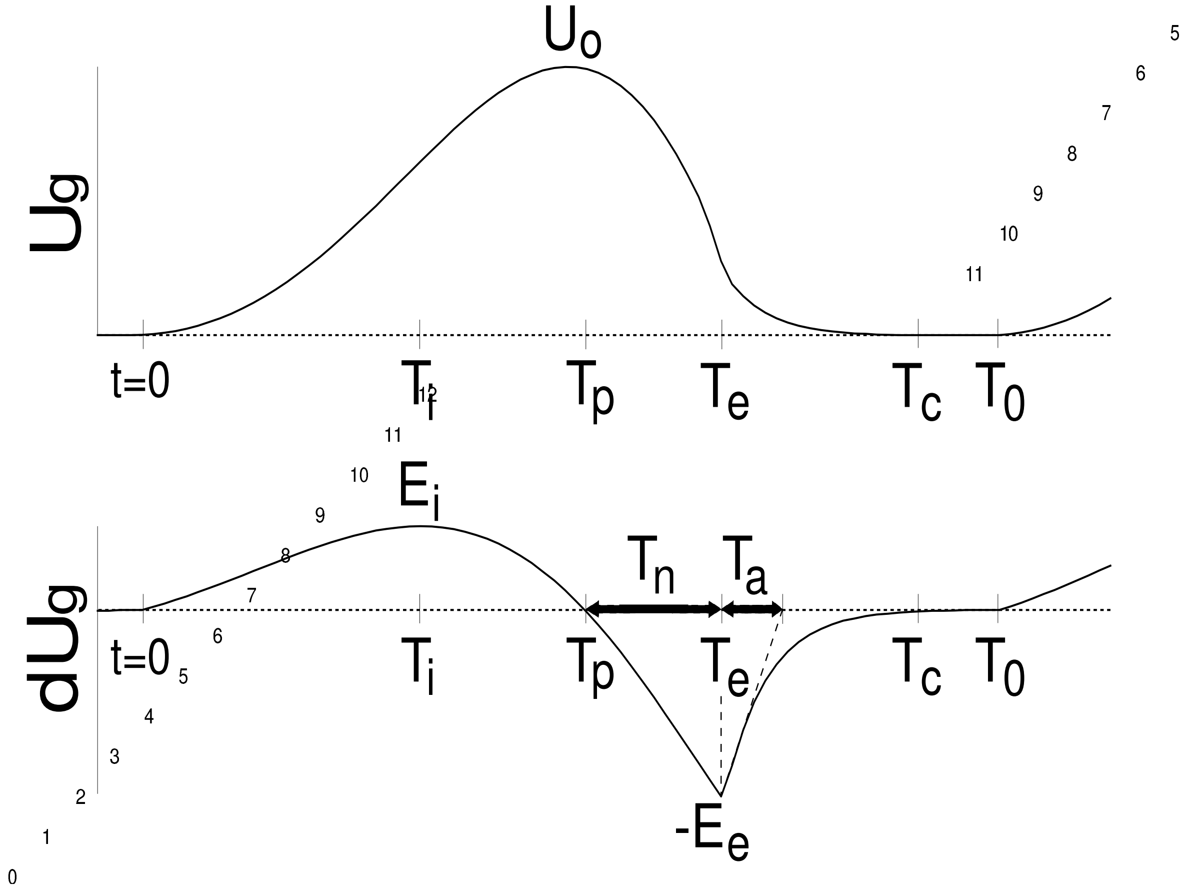


Figure 1. Glottal flow (U_g) and glottal flow derivative (dU_g) with the parameters of the LF-model. U_0: maximum of U_g; E_i: maximum of dU_g; E_e: absolute value of the minimum of dU_g; t = 0: time of glottal opening; T_c: time of glottal closure; T_i, T_p, T_e: time points of E_i, U_0, and E_e, respectively; T_a: the time between T_e and the projection of the tangent of dU_g in $t=T_e$; $T_n = T_e - T_p$. The dimensionless wave shape parameters that can be derived from the LF-parameters are: $R_g = T_0/2T_p$; $R_k = T_e/T_p - 1 = T_n/T_p$; $R_a = T_a/T_0$.

filter signals, by fitting the LF-model to the data. For each pitch period median values for all parameters in the LF-model were calculated.

The median value of a parameter for a pitch period can become negative (-1), if at least 13 of the 25 values of that parameter are equal to -1. This occurs if in more than half of the cases the fit was not successful. The data of all pitch periods in which the median value of one of the LF-parameters is equal to -1 were discarded. In total 128 periods were discarded, and the data of 613 pitch periods were used for further analysis.

The disadvantage of using such a conservative criterion is that a lot of data have to be discarded, but the advantage is that the risk of errors in the final data is reduced. We are convinced that keeping more of the data for the consonants and onsets/offsets would not have changed our results and conclusions.

4.3 Results

Figure 2 shows the audio signal, the automatically calculated inverse filter results, and the automatically obtained fits for five consecutive pitch periods of a vowel /e/. The differentiated flow signals often contain a pronounced ripple. It is clear from this figure that attempts to measure the LF-parameters directly from the raw dU_g or U_g signals would result in noisy estimates. For instance, the maximum of dU_g (E_i) and the place of this maximum (T_i) are to a large extent determined by the ripple. By fitting an LF-model to the data the measurements are made more robust. The fit procedure is almost always able to find a combination of LF-parameters that generates a model signal that closely resembles the measured flow signal.

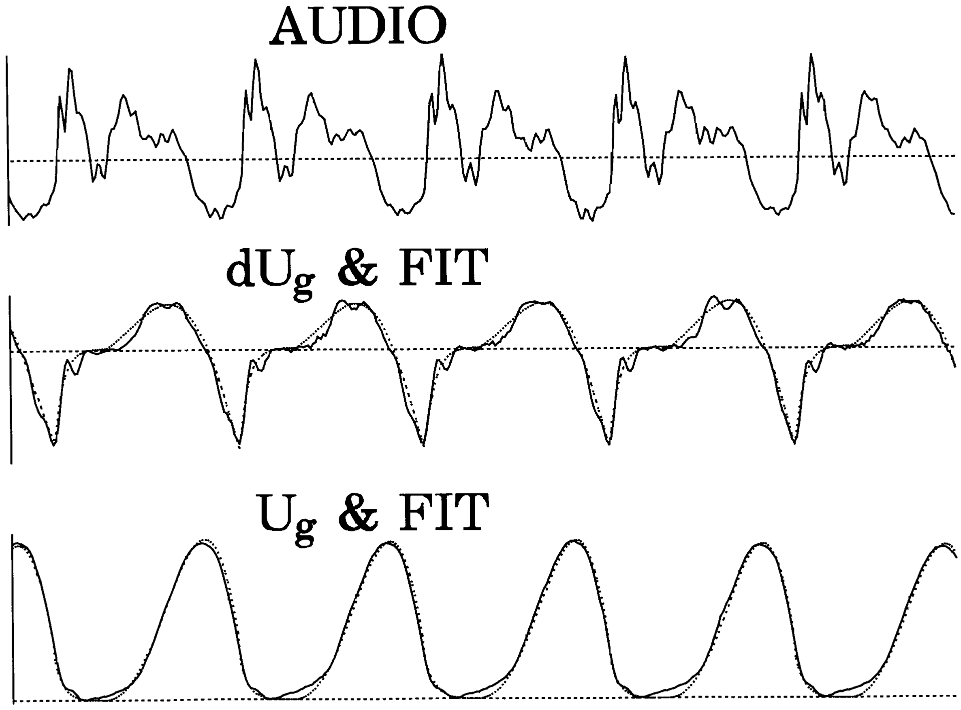


Figure 2. Results of the automatic fitting procedure for five periods of a vowel /e/. Shown are, from top to bottom, audio signal, glottal flow derivative (dU_g, solid line) with fitted signal (dotted line) and glottal flow (U_g, solid line) with fitted signal (dotted line).

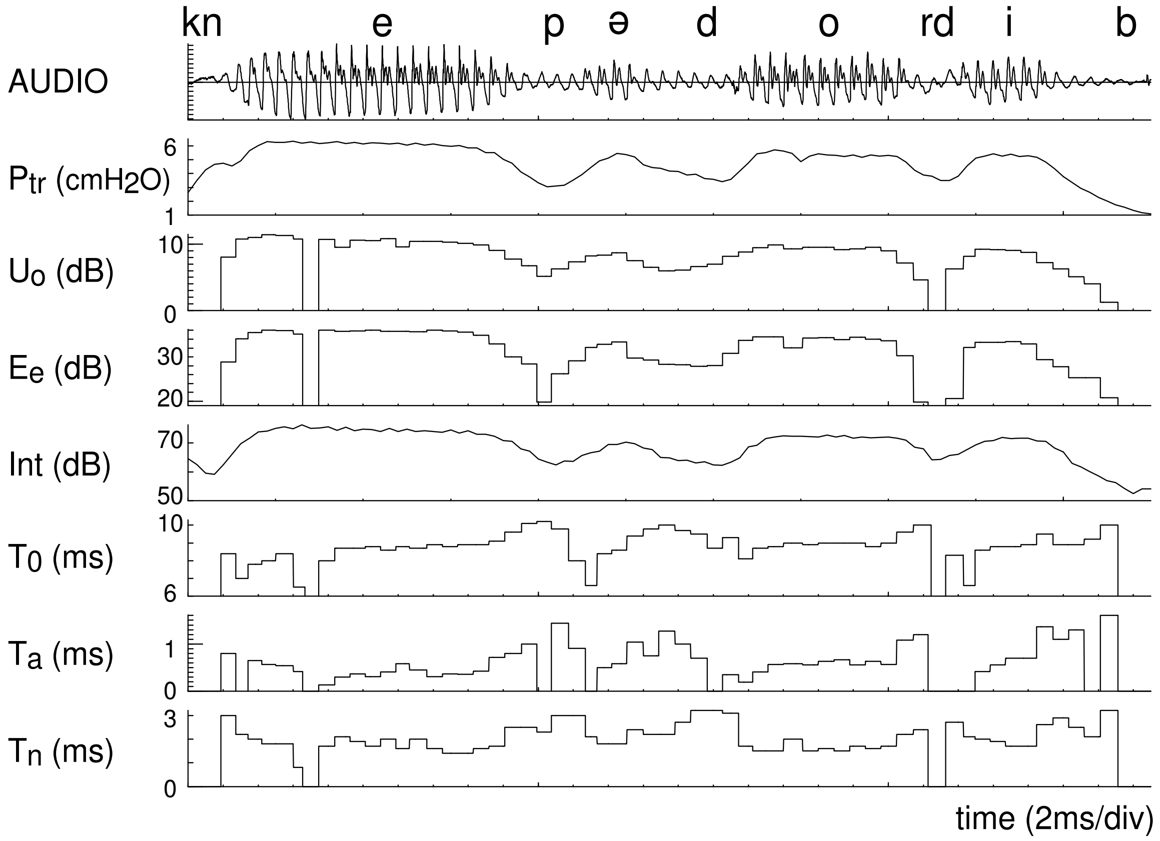


Figure 3. Results for a voiced interval to illustrate the behaviour of the voice source parameters. Given are, from top to bottom, phonetic transcription, audio signal, transglottal pressure (P_{tr}), median values of U_0, E_e, Int, T_0, T_a and T_n. Although /p/ is phonologically an unvoiced plosive, it can be observed that voicing continues in this utterance.

In Figure 3 the median values of the most relevant parameters are given for a voiced interval of one of the utterances. For some pitch periods the median values of all LF-parameters are -1, indicating that for the majority of the 25 combinations the fit was not successful for these periods. There are two factors that could hinder a good fit. Sometimes the estimate of the vocal tract transfer function was not correct, in which case inverse filtering did not yield a flow signal that resembles an LF-pulse even remotely. However, there were also cases in which inverse filtering produced a reasonable estimate of dU_g, but where the optimization routine did not converge. Not surprisingly, estimation problems occurred more often in voiced consonants, and during voice onset and offset (the first and last periods of a voiced segment) than during the steady parts of vowels.

Furthermore, it was observed that estimates of the parameters of the first part of the LF-model (the exponentially growing sine wave, i.e. T_p, T_e, E_e) varied less than those of the return phase (i.e. T_a). In part this is due to the fact that the duration of the first part is longer than the duration of the return phase. But another cause is that the return phase is often not smooth and contains a ripple (see Figure 2). This pronounced ripple often affects the automatic fitting process for the return phase. In many cases a reasonable fit could be reached for the first part of the LF-model, but not for the return phase. The result is that the median value of T_a is often -1, while this is not the case for the other parameters (see Figure 3).

For the moment we do not know whether the failure of the fit procedure to converge to an acceptably small error is due to computational problems or to the failure of the LF-model to approximate all glottal flow pulse forms that occur in real speech.

4.3.1 General behaviour

Typical behaviour of the LF-parameters can be observed in Figure 3. During transitions from vowel to consonant T_0, T_a and T_n generally increase, while transglottal pressure (P_{tr}), U_0, E_e and Int decrease. The consistent reciprocal relation between the parameters in these two sets is reflected in the correlation coefficients (see Table 1),

Table 1. Correlations between four amplitude related parameters (P_{tr}, U_0, E_e, Int) and three time parameters (T_o, T_n, T_a) for 613 voiced periods.

	P_{tr}	U_0	E_e	Int
T_o	-0.44	-0.17	-0.35	-0.45
T_n	-0.41	-0.19	-0.48	-0.42
T_a	-0.31	-0.36	-0.50	-0.36

which are all negative and highly significant ($p<0.0001$). For these and all following correlation coefficients the level of significance for a two-tailed test was calculated (Ferguson, 1987). The correlation coefficient between two sets of 613 samples is said to be significant at the 0.01% level ($p<0.0001$) if its absolute value is larger than 0.16.

The rationale behind this very general behaviour is most probably the following. For vowels the impedance of the glottis is much higher than the impedance of the vocal tract, and thus P_{tr} is almost equal to P_{sb}. For consonants there is a constriction in the vocal tract, causing a pressure build-up above the glottis and a drop in P_{tr}. In order to keep vibration going (with a lowered P_{tr}) during these voiced consonants, some adjustments must be made: the vocal folds are slackened and abducted, and the consequence is that T_a and T_n are raised. Lowering of P_{tr} and slackening of the folds will lower F_0, and thus raise T_0. Although the folds are slackened, the decrease in P_{tr} is such that the amplitude of vibration of the folds decreases, and with it the modulation of the flow (U_0), and eventually E_e and Int.

The observed reciprocal relation provides a natural way of dividing the LF-parameters into two sets. The first set consists of T_i, T_p, T_e, T_n, T_a and T_0, and will be referred to as the “time parameters”, while the second set (P_{tr}, U_0, E_e, Int) will be referred to as the “amplitude related parameters”. Relations within the first set are described in Section 4.3.2, and relations within the second set in Section 4.3.3. The relations between F_0 and other parameters can be derived directly from the relations of these parameters with T_0. Therefore, they are not treated separately, but are part of Section 4.3.2. The behaviour of the wave shape parameters R_g, R_k and R_a is described in Section 4.3.4.

4.3.2 Time parameters

It was already mentioned that during transitions from vowels to consonants T_0, T_a and T_n are generally raised (see Figure 3). The following question then emerges: How does a change in T_0 affect the time parameters, or, in other words, how does the shape of the pulse change with F_0? In this section we try to answer this question by looking at the relations between T_0 and the other time parameters.

The five time parameters T_i, T_p, T_e, T_a and T_n were first plotted as a function of T_0 on a double logarithmic scale, and the best linear fits were calculated. The resulting lines are of the form:

$$\log T_x = \log A + B * \log T_0 \Leftrightarrow T_x = A * T_0^B,\ x \in \{i, p, e, a, n\}$$

The regression lines for T_i, T_p, T_e, T_a and T_n are shown in Figure 4. All correlations between the logarithm of the five time parameters and the logarithm of T_0 are positive and highly significant ($p<0.0001$). So, on average, all time parameters increase with increasing T_0, and the glottal pulse is stretched. However, this stretching is not distributed uniformly over the entire period.

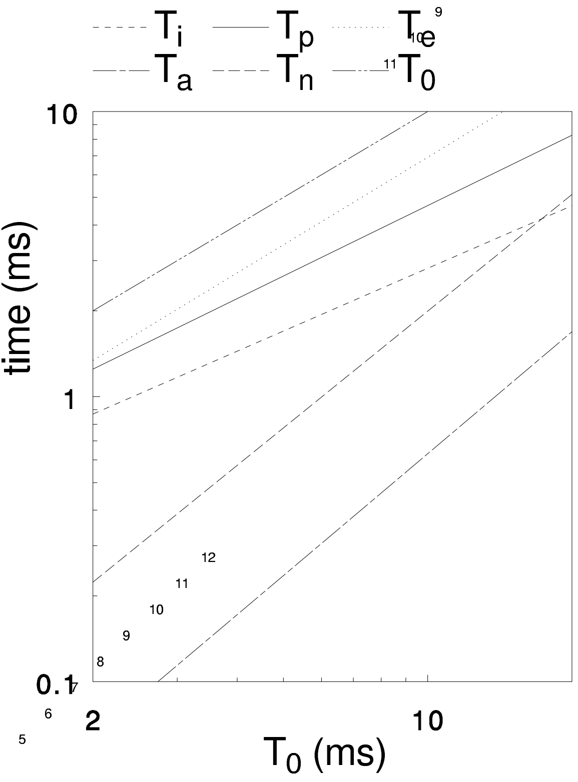


Figure 4. The relation between the time parameters and T_0. Given are the regression lines of the time parameters as a function of T_0. Note that both the horizontal and the vertical axis have a logarithmic scale.

If a time parameter changes linearly with T_0, then its regression line in Figure 4 should have a slope of 1. In that case it would run parallel to the reference line for T_0 that is also given in Figure 4 ($T_0 = 1.T_0^1$), which obviously has a slope of 1. This is the case for T_e, so generally the duration of the first part of the LF-pulse changes linearly with T_0. However, the increase in T_i and T_p is less than linear, and the increase in T_a and T_n ($T_n = T_e - T_p$) is more than linear (see Figure 4). The ordering of the time parameters with ascending power is T_i, T_p, T_e, T_n, T_a. It seems as if the amount of stretching increases when going towards the end of the LF-pulse. With regard to the shape of the LF-pulse, the consequence is that the skewing decreases more than linearly with T_0.

4.3.3 Amplitude related parameters

A constantly high covariance between the amplitude related parameters was found for all data (see Table 2 and Figure 5). At first sight the high covariance of these parameters does not seem surprising, as an increase in P_{tr} alone (everything else being equal) would increase the amplitude of vibration of the vocal folds, and therefore lead to an increase in U_0 and E_e. Increasing U_0 and E_e by roughly the same amount would lift the spectrum (see Fant & Lin, 1988), and thus increase Int. However, our data form a mix of voiced consonants, stressed and unstressed vowels. Thus one might expect large variations, both in the glottis and in the vocal tract. For instance, for voiced consonants T_a and T_n are generally higher than for vowels (see Section 4.3.2). A change in T_a has little effect on Int, but an increase in T_n (i.e. less skewing) combined with a decrease in U_0 would lead to a decrease in E_e that is relatively larger than the decrease in U_0. Given the large variation in articulatory gestures, it is surprising that the covariance between P_{tr}, U_0, E_e and Int is invariably high.

Table 2. Correlations between logP_{tr}, logU_0, logE_e and logInt for 613 voiced periods.

	logP_{tr}	logU_0	logE_e
logU_0	0.60		
logE_e	0.63	0.86	
logInt	0.81	0.78	0.81

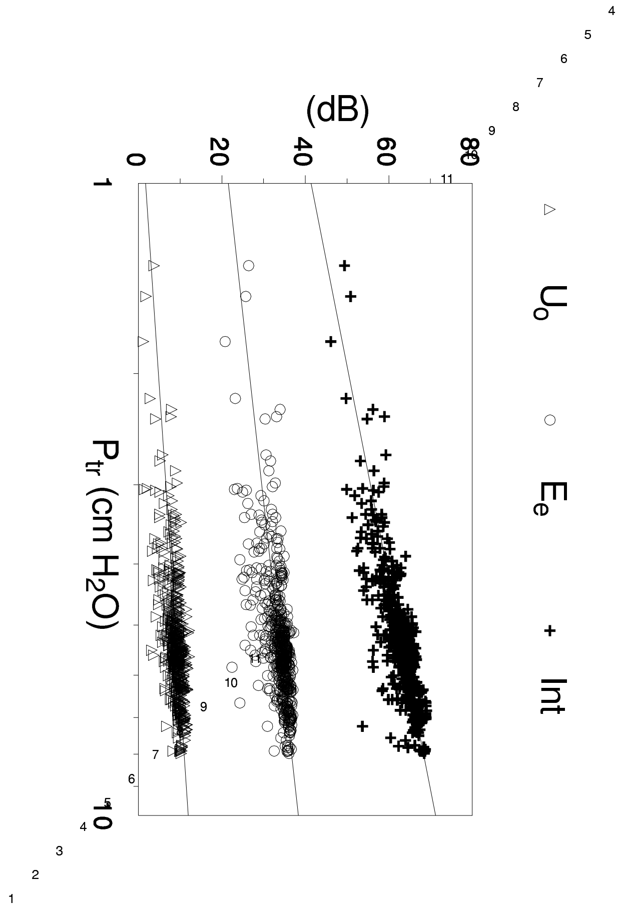


Figure 5. Scatterplots of the amplitude related parameters U_0, E_e and Int as a function of P_{tr}, with regression lines. Note that both the horizontal and the vertical axis have a logarithmic scale.

Regression lines were calculated for the amplitude related parameters. The procedure used was analogous to the procedure used for the time parameters, as described in Section 4.3.2. The regression lines are of the form:

$$\log X = \log A + B * \log P_{tr} \Leftrightarrow X = A * P_{tr}^{B},\ X \in \{U_0, E_e, \mathrm{Int}\}$$

The slope of the regression line for U_0 in Figure 5 is 1.0, indicating that the relation between U_0 and P_{tr} is approximately linear. In the LF-model E_e is a function of U_0 and the skewing of the glottal pulse. The fact that both U_0 and skewing increase with increasing P_{tr} explains why the slope for E_e (of 1.6, see Figure 5) is larger than the slope for U_0. The slope of the regression line for Int (of 3.0) is about twice the value found for E_e, which is not surprising, because the Int of a freely travelling spherical sound wave is proportional to the square of the derivative of the mouth flow (Beranek, 1954). However, without the use of a proper production model it is difficult to unravel the exact underlying relations between the parameters.

4.3.4 Wave shape parameters

For the dimensionless wave shape parameters R_g, R_k and R_a the following general relations can then be derived. R_g is almost constant; the correlation of R_g with T_0 is positive but very small (see Table 3). For the range of R_g values found in this study, the influence of this parameter on the spectrum (and thus on voice quality) is very small. The correlations of R_a and R_k with T_0 (see Table 3) are positive and highly significant ($p<0.0001$), which implies that voice quality changes with T_0 and consequently with F_0. The correlations of R_a and R_k with Int and P_{tr} were even higher (see Table 3), so voice quality also changes with Int. The average values of R_g, R_k and R_a were 108%, 41%, and 6.5%, respectively and are in accordance with the values given by Carlson et al. (1989).

Table 3. Correlations of the wave shape parameters R_g, R_k and R_a with T_0, P_{tr}, and Int for 613 voiced periods.

	T_0	P_{tr}	Int
R_g	0.04	-0.20	-0.17
R_k	0.22	-0.30	-0.28
R_a	0.18	-0.23	-0.29

4.3.5 Deviations from the general behaviour

The fact that we have a large data set in which most parameters display consistent relations allows us to identify the outliers, i.e. the instances that do not fit in with the general pattern. Pitch periods that show different relations between the parameters are mainly found during voice onset and offset, and in the last syllable of an utterance.

The values of U_0 for voice onset and offset generally fall below the regression line of U_0 on P_{tr} that is given in Figure 5, but there are also differences between voice onset and offset. The average P_{tr} during an UV/V transition (5.0 cm H_2O) is higher than the average P_{tr} during a V/UV transition (3.7 cm H_2O). It seems that higher P_{tr} values are needed to initiate vibration of the vocal folds, than to keep vibration going towards the end of a voiced interval. At the beginning of a voiced interval the average values of Int and F_0 (59 dB and 131 Hz) are also higher than those at the end of a voiced interval (57 dB and 120 Hz). Furthermore, a rise in T_a and T_n was found both towards beginning and end of a voiced interval.

Near the end of all 30 utterances there was a substantial decrease in P_{sb}, P_{tr}, Int and F_0. A marked increase in the activity of the sternohyoid was also observed. Moreover, for the final vowel U_0 was relatively high, compared to the general trend. The deviating behaviour of the voice source during the final syllable was also observed by Klatt & Klatt (1990). This is described in more detail in Strik & Boves (1992a).

4.4 Conclusions

In general, the method of automatic inverse filtering and fitting worked satisfactorily. Most problems were encountered with attempts to obtain a good approximation for the T_a parameter in pitch periods taken from consonants. For some glottal periods our method did not succeed in finding a combination of LF-parameters that define an LF-model that closely resembles dU_g. This could be a shortcoming of the inverse filter or the fitting procedure, but also of the LF-model. It remains to be seen if the LF-model can describe all variations in the glottal pulse that occur in different kinds of speech.

Consistent relations were found within the set of the time parameters and the set of amplitude related parameters, but also between the parameters of these two sets. The highest correlations were found between P_{tr}, U_0, E_e and Int. The behaviour of the voice source during voice onset, voice offset and the last syllable was different from the general behaviour. When relating LF-parameters to prosody the general picture is that voice quality is mainly affected by R_k and R_a (or T_n and T_a), and that Int is mainly affected by E_e (or U_0).

All these fluctuations in the voice source parameters are likely to have perceptual consequences. To improve the naturalness of synthetic speech, these effects have to be taken into account.

Chapter 5

Comparing methods for automatic extraction of voice source parameters from continuous speech

Abstract

Two methods are presented for automatic calculation of the voice source parameters from continuous speech. Both methods are used to calculate the voice source parameters for natural speech. However, for natural speech no objective test procedure seems to be available. Therefore, both methods were also tested on synthetic speech.

This chapter is a slightly revised version of the following paper: Strik, H., Jansen, J. & Boves, L. (1992) Comparing methods for automatic extraction of voice source parameters from continuous speech. *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, 1, 121-124.

5.1 Introduction

Modern text-to-speech systems produce speech that is intelligible, but not quite natural. A substantial improvement of the naturalness of synthetic speech can probably be achieved by the use of a properly controlled voice source model. To derive rules for voice source parameters large amounts of data are required. Extracting voice source parameters by hand is time consuming, subjective and thus probably not entirely reproducible. However, automatic extraction of voice source parameters from continuous speech is also far from trivial. Our aim is to develop methods for automatic extraction of voice source parameters from continuous speech. In this chapter we propose and test two generic automatic methods.

In continuous speech the glottal parameters may change from period to period. Thus, the procedure must be pitch-synchronous. In our case, estimates of the frequency response of the vocal tract are based on an analysis of the closed glottis interval only. It is well known that analyses over such short intervals can yield wildly fluctuating results if, for instance, the analysis window is shifted or extended by just one or two samples. Because the causes behind these fluctuations are not understood, it is not possible to determine the optimal window location and length by simple automatic procedures. Therefore, we estimate the vocal tract transfer function for a number of window positions and lengths for each pitch period. The two methods differ in the way in which these multiple estimates are used. Method 1 computes an inverse filtered flow waveform for each individual estimate, and then computes the median value for the parameters describing the glottal waveforms. Method 2 uses the multiple estimates to obtain a single optimal estimate of the vocal tract transfer function, which is subsequently used to obtain a single optimal glottal flow waveform.

Both methods are used to derive voice source parameters from natural speech. But for natural speech it is difficult to evaluate the performance objectively, because the true glottal flow waveforms are not known. The only possible evaluation is a subjective one, viz. checking that the voice source parameters behave the way they are expected to behave. As long as there is no objective test procedure for natural speech, an objective evaluation can only be done with synthetic speech. In this chapter we will first test whether the two methods give plausible results for natural speech. Then we will perform an objective, quantitative test of both methods with synthetic speech.

5.2 Material

5.2.1 Natural speech

To study voice source characteristics data were obtained for four male subjects. For the current chapter only the data of one subject were used. The speech signal was transduced by a condenser microphone (B&K 4134) placed about 10 cm in front of the mouth, pre-amplified at the microphone (B&K 1619), and amplified by a measuring amplifier (B&K 2607). The speech signal was A/D converted off-line at a 10 kHz sampling rate.

5.2.2 Synthetic speech

The synthesis system used to generate the synthetic speech is a serial pole-zero synthesizer (Boves et al., 1987) that uses the DEC-Talk source (Klatt and Klatt, 1990). The main reason for using the DEC-Talk source is its computational simplicity. All synthesis signals have a sampling frequency of 10 kHz.

5.3 Method

Inverse filtering is often used to obtain an estimate of the glottal flow signal. At the base of this method is the assumption that the voice source and the vocal tract filter do not interact. It is known that this assumption is not valid (see e.g. Anantapadmanabha & Fant, 1982), but it is a useful approximation of the human speech system. The approximation is optimal during the closed glottis intervals, because then there is least interaction between the subglottal and the supraglottal cavities. Therefore, the analysis window should preferably be confined to the closed glottis interval. In De Veth et al. (1990) we showed that Closed glottis interval Covariance Linear Predictive (CC-LP) analysis is as powerful as more sophisticated techniques, like robust pole-zero analysis.

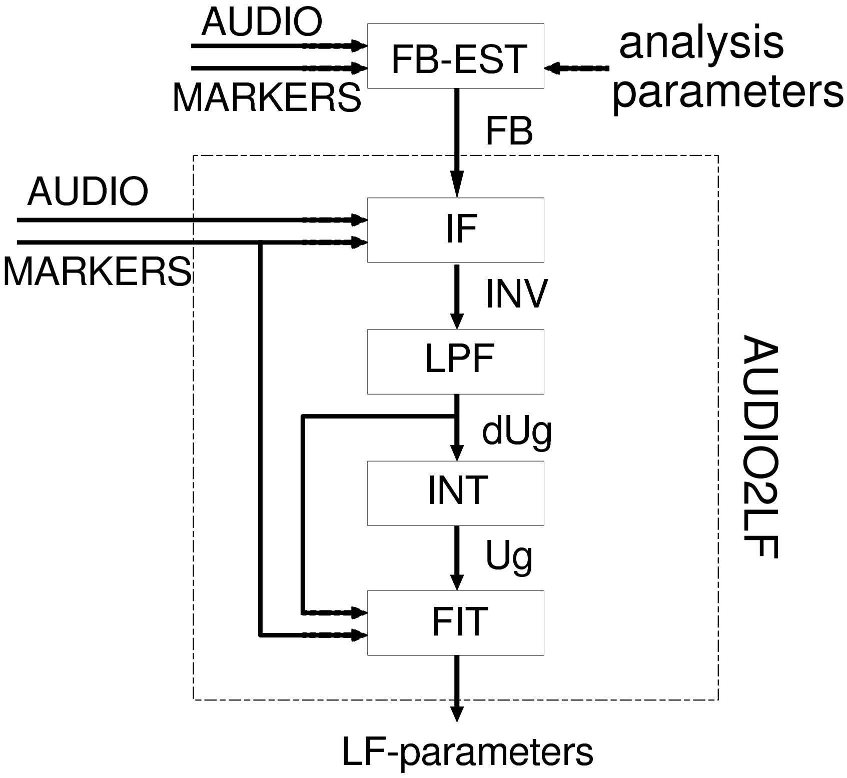


Figure 1. Block diagram of the general method.

5.3.1 General method

A block diagram of the general method is shown in Figure 1. First the vocal tract filter is estimated. This is done by converting the results of the CC-LP analysis to M Formant-Bandwidth pairs. This module is therefore called “FB-EST” (see Figure 1). The problem of finding the correct set of analysis parameters is treated in the next section.

The estimated filter is inverted, and the inverse filter is used to filter the audio signal (module IF in Figure 1). The resulting inverse filtered signal, INV, is a first estimate of the differentiated glottal volume flow. As the inverse filtered signal often contains high-frequency noise, it is low-pass filtered (module LPF in Figure 1). The resulting signal is a new, usually better, estimate of the differentiated glottal volume flow, dU_g. The glottal volume flow, U_g, is obtained by integration of dU_g (module INT).

Voice source parameters, like open quotient and skewing, could be derived directly from dUg and Ug. However, since these signals are often noisy, direct measurements yield unreliable values. Fitting a voice source model to the data is probably a more robust method of obtaining voice source parameters. In our system, the fit is done simultaneously on dUg and Ug (Strik & Boves, 1992b).

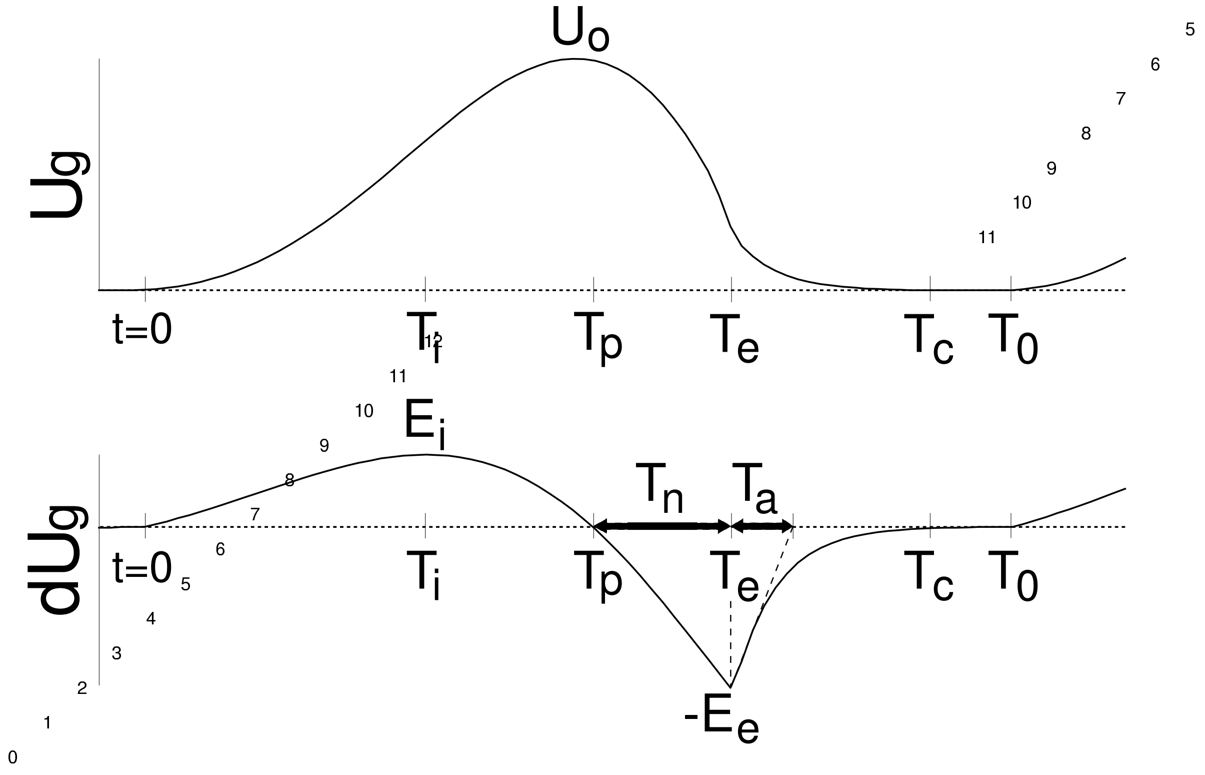


Figure 2. Glottal flow (Ug) and glottal flow derivative (dUg) with the parameters of the LF-model.

The LF-model was used as voice source model because it seems useful for synthesis, and because it has already been studied in great detail (Fant et al., 1985). The model and its parameters are shown in Figure 2. The LF-model is a four-parameter model. There are different combinations of the LF-parameters that uniquely define a flow pulse. The four parameters that are used for the generation of flow pulses during the fit procedure are U_0, T_p, T_e and T_a. Both during analysis and synthesis a fifth parameter is necessary to position the LF-pulses, viz. T_0.

The general method can be split in two parts. In the first part the formants and bandwidths of the vocal tract filter are estimated (FB-EST), and in the second part this filter is used to derive LF-parameters from the audio signal. This second module is therefore called AUDIO2LF (see Figure 1).

5.3.2 Analysis parameters

For CC-LP analysis a number of choices have to be made like position and width of the closed glottis interval, order of the analysis, and pre-emphasis factor. Usually, no combination of choices is optimal for the whole utterance. However, for the natural speech that was used in this study a 12th-order LPC analysis with a pre-emphasis factor of 1.00 (+6 dB/oct) worked satisfactorily for almost all pitch periods. Thus, window position and window length are left as parameters that can be varied.

Generally, the moments of glottal closure are easier to detect than the moments of opening. It is possible to identify the locations of the main excitation from the audio signal (Childers & Lee, 1991). However, we prefer to determine the moments of glottal closure from the electroglottogram (Strik & Boves, 1992b). The signal with the closure markers is called “MARKERS” (see Figure 1). The window position will be determined relative to a closure marker (this parameter is called window shift). Five window shifts (of -2, -1, 0, 1, 2 samples) were used. For synthetic speech the close markers are found by using the same operations (differentiation, low-pass filtering, and peak-picking) for the source signal.

For natural speech it is difficult to obtain an accurate estimate of the instant of glottal opening. Therefore, LP analysis is done for five different, fixed window lengths. If the length of the window is too large, then the last part will be in the open phase. This will perturb the estimates of the formants and especially of the bandwidths. On the other hand, if the length of the window is too short then LP analysis will not be able to make an accurate estimate. As a compromise we used window lengths of 33, 34, 35, 36 and 37 samples.

In both methods the FB-pairs are estimated for the 25 combinations of analysis parameters (see Figure 3). The 25 estimations of the vocal tract filter are used in both methods to obtain a single set of voice source parameters.

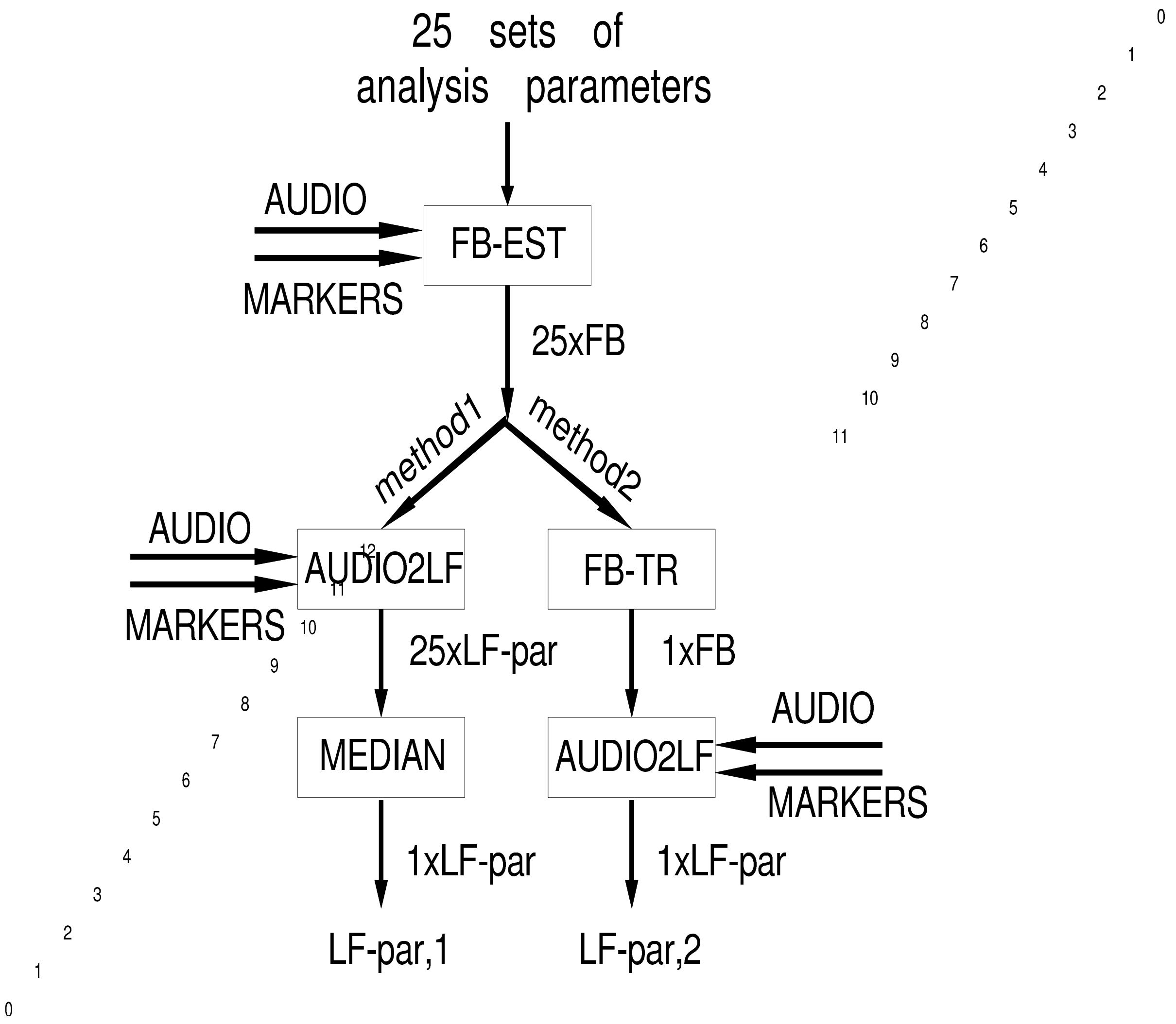


Figure 3. Block diagram illustrating the two methods (for further details, see Section 5.3.3).

5.3.3 The two methods

5.3.3.1 Method 1

In the first method the module AUDIO2LF is used 25 times (see Figure 3). The result is 25 sets of LF-parameters. Finally, median values are calculated for the LF-parameters (module MEDIAN in Figure 3), which results in one set of average LF-parameters. This method is described in more detail in Strik & Boves (1992b).

5.3.3.2 Method 2

In method 2 a Formant-Bandwidth TRacker (module FB-TR in Figure 3) is used to obtain an optimal set of FB-pairs. In this method the optimal filter is used to obtain one set of LF-parameters, which implies that the module AUDIO2LF is applied only once.

5.3.3.3 FB-Tracker

The goal of the module FB-TR is to find the first 4 FB-pairs. Generally, the first 4 FB-pairs are among the first 5 poles that are modelled by the LP analysis. Therefore, all 5 possible combinations of 4 out of 5 poles are generated. This is done for the 25 estimates of the vocal tract filter, and the result is a lattice of 4 FB-pairs with a depth of 125. A Viterbi algorithm is used to find the optimal path in this lattice.

The optimal path is the path with minimal total cost. For calculation of the cost, a transition and a local cost function are used. The transition cost function is the Euclidean distance between the values of the current frame and the values of the previous frame. The local cost function is the Euclidean distance between the values of the current frame and the reference values. The latter are obtained by a correlation LP analysis, using an analysis window of 256 samples.

For higher formants the variation in both frequency and bandwidth values is higher than for lower formants. For FB-tracking the effect would be that the FB-values of higher formants are more important in determining the optimal path. Therefore, all variables are converted to standard normal variables by first subtracting the mean value, and subsequently dividing by the standard deviation.

The total cost function has four contributions, viz. the transition and local costs of the formants and the transition and local costs of the bandwidths. Each of these four contributions can be given a weight. The weights that were used are 4, 2, 1 and 0, respectively. The reference bandwidths obtained by correlation LP analysis are usually larger than the bandwidths obtained by CC-LP. These bandwidths cannot be compared in a straightforward manner, and consequently the weight of the local cost of the bandwidths is set to zero.

5.4 Results

5.4.1 Natural speech

The data used in the current chapter are obtained by applying the two methods to a natural utterance with a length of about 2.5 seconds. In Figure 4 the audio signal of part of the utterance is shown, together with T_0 and the four LF-parameters as calculated by the two methods.

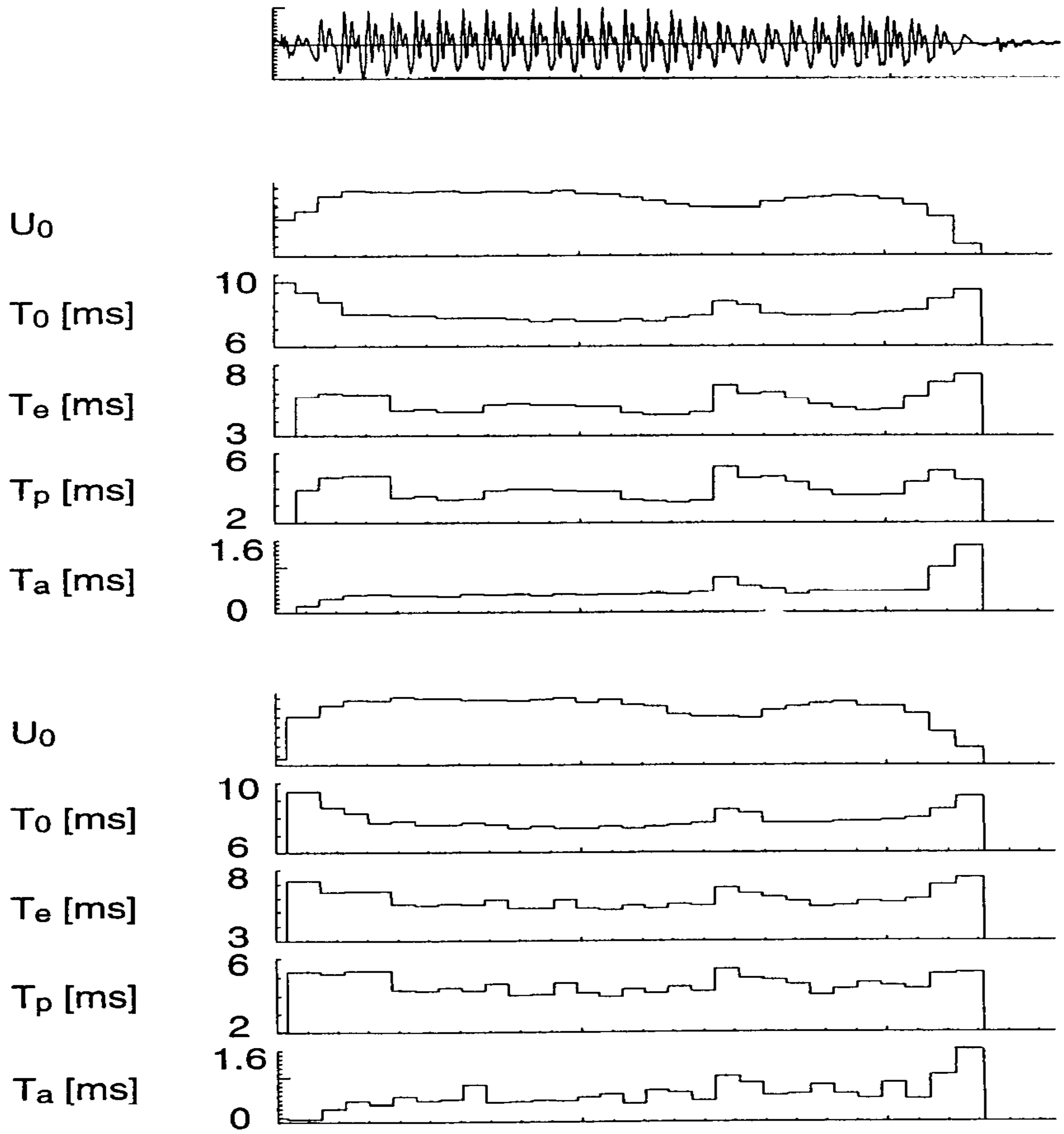


Figure 4. Audio signal and the LF-parameters calculated by the two methods.

The part of the utterance shown in Figure 4 clearly demonstrates the dynamics of the LF-parameters. Generally, it is observed that U_0 covaries with the amplitude of the speech signal. The same holds for the other amplitude related parameters of the LF-model, like E_e and E_i. During transitions from vowels to consonants it is often observed that U_0 decreases, while the time parameters T_0, T_e, T_p and T_a increase. The same effects were found by Bickley & Stevens (1986) for artificial vocal tract constrictions.

The four LF-parameters and T_0 can be used to calculate all other glottal waveform parameters like open quotient, speed quotient, skewing etc. As an example, the 3 dimensionless wave shape parameters have been calculated: $R_g = T_0/2T_p$, $R_k = T_e/T_p - 1$, and $R_a = T_a/T_0$. The average values of R_g, R_k and R_a for all 194 voiced periods of the utterance are 111, 42 and 6.7 for method 1, and 110, 41 and 7.1 for method 2. These values are in accordance with the values given in Carlson et al. (1989).

Although the results of the two methods are slightly different, both methods seem to give plausible results. Given these results, two questions emerged: what is the reliability of the results, and which method is the best one? As there is no objective test procedure for natural speech, we also tested both methods on synthetic speech.

5.4.2 Synthetic speech

The DEC-Talk waveform and the LF waveform are fundamentally different. The LF-parameters derived from synthetic speech by using the two methods cannot be compared directly to the source signal used during synthesis. For evaluation of the test results a reference was required. This reference was obtained by performing the same operations on the source signal as on the inverse filter results, i.e. low-pass filtering, integration and fitting of an LF-model.

The rationale behind this is that the differentiated source signal should really be compared to the inverse filtered signals. All operations that are necessary to obtain LF-parameters from the inverse filtered signals should therefore also be applied to the differentiated source signal. In Figure 5 the low-pass filtered and the fitted signal are shown. Apart from the fitted signal, the fit also yields the four LF-parameters for each pitch period. These reference parameters are used below for evaluation. The utterance used for evaluation was a random concatenation of all vowels, liquids and glides that are used in the synthesis system.

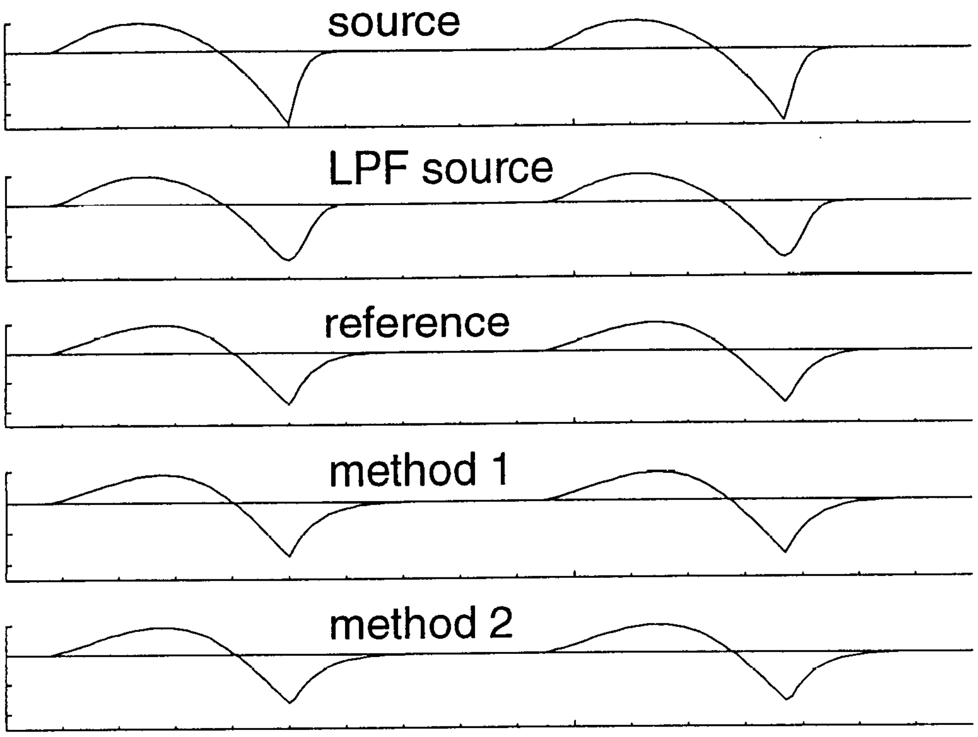


Figure 5. Shown are from top to bottom: two periods of the DEC-talk voice source for a consonant /l/, the low-pass filtered source signal, the reference signal, and the source signals calculated from the synthetic speech signal by methods 1 and 2.

In Figure 5 it can be seen that low-pass filtering and fitting mainly influences the excitation strength and the return phase. The effects are clear for a synthetic glottal pulse (as in Figure 5), but the same effects occur for the inverse filtered signals derived from the speech signals. In order to calculate the true values of the voice source parameters, a correction is mandatory after the two methods have been applied. The amount of the two corrections can be calculated from synthetic speech and should be verified for natural speech. In the following part we will only test the similarity of the reference parameters (set 0) and the parameters obtained by means of the two methods (sets 1 and 2).

In Figure 5 one can see that the reference signal and the source signals obtained with the two methods from the synthetic speech signal are very much alike. A regression analysis on the voice source parameters was done to test the degree of resemblance. The results are given in Table 1.

Table 1. Results of regression analysis for different combinations of the four LF-parameters. For each combination of two variables Y and X, a straight line is fitted through the data: $Y = intercept + slope*X$. The correlation coefficient r indicates the goodness of the fit (N = 232). In the upper box the results of the two methods (subscript 1 and 2) are compared with the point of reference (subscript 0). In the lower box the results of the two methods are compared with each other.

Y	X	intercept	slope	r
$U_{0,1}$	$U_{0,0}$	-9.7	1.03	0.96
$U_{0,2}$	$U_{0,0}$	-11.4	1.05	0.97
$T_{e,1}$	$T_{e,0}$	-0.33	0.95	0.75
$T_{e,2}$	$T_{e,0}$	0.07	1.01	0.76
$T_{p,1}$	$T_{p,0}$	0.25	0.96	0.60
$T_{p,2}$	$T_{p,0}$	-0.09	1.05	0.62
$T_{a,1}$	$T_{a,0}$	0.05	0.95	0.69
$T_{a,2}$	$T_{a,0}$	0.00	1.00	0.64
$U_{0,2}$	$U_{0,1}$	3.8	0.99	0.98
$T_{e,2}$	$T_{e,1}$	-0.05	1.01	0.97
$T_{p,2}$	$T_{p,1}$	-0.09	1.02	0.96
$T_{a,2}$	$T_{a,1}$	0.02	0.90	0.80

The results of the comparisons between the two methods and the reference point are given in the upper part of Table 1. All slopes are about 1, and all intercepts are small, which means that the parameters calculated by the two methods have, on average, the same value as the reference parameters (the intercept of U_0 has a larger absolute value, but the value of the intercept relative to the average value of U_0 is even smaller than those of T_e, T_p and T_a). The correlation coefficients in Table 1 for U_0 are almost 1, and those of T_e, T_p and T_a are somewhat smaller but still highly significant. This means that the values calculated by the two methods for T_e, T_p and T_a closely resemble the reference value, while the calculated values for U_0 and the reference values are almost identical. Generally, the results of method 2 are slightly better than those of method 1.

In the lower part of Table 1 the two methods are compared. For U_0, T_e and T_p the intercept is small, and the correlation coefficient and the slope are almost 1. This means that for U_0, T_e and T_p the results of the two methods are very much alike, whereas the values of T_a for the two methods are less similar.

5.5 Conclusions

In this chapter two methods are proposed for the automatic extraction of voice source parameters from continuous speech. For natural speech the two methods produce comparable and reasonable results. There is a need for an objective procedure to test the reliability of the extracted parameters. As long as this test is not available, the best alternative is to test these methods on synthetic speech.

The various operations carried out during the extraction of the voice source parameters from speech, influence the magnitude of these parameters. In order to re-estimate the true magnitude of the parameters these effects have to be corrected.

From the tests on synthetic speech it appeared that both methods succeed in estimating the voice source parameters quite accurately. The results obtained for the amplitude parameter U_0 are better than those of the time parameters T_e, T_p and T_a. For synthetic speech the second method is slightly better than the first one.

Chapter 6

Automatic estimation of voice source parameters

Abstract

Voice source parameters can be estimated by fitting a voice source model to the glottal flow signal which is obtained by means of inverse filtering. In this chapter we investigate the behaviour of the LF-model in a number of non-linear parameter estimation procedures. It is concluded that (1) the parameter estimates are robust against additive (white and narrow band) noise in the flow waveforms, (2) simplex search algorithms perform better than steepest descent algorithms, provided that (3) the LF-pulse is generated with an algorithm that treats all parameters as real numbers.

This chapter is a slightly revised version of the following paper: Strik, H. & Boves, L. (1994) Automatic estimation of voice source parameters. *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, 1, 155-158.

6.1 Introduction

In recent years an increasing need for automatic techniques to extract voice source parameters has been observed. A technique of this kind is described in Strik et al. (1992) and Strik et al. (1993). For natural speech, where the input is unknown, this method gave plausible results (Strik et al., 1992), while for synthetic speech, where the input is known, the parameters could be re-estimated with a reasonable accuracy (Strik et al., 1992; Strik et al., 1993). At the moment we are testing and trying to improve this automatic method.

In the proposed method the speech signal is inverse filtered, and the resulting estimate of glottal flow is parameterized by fitting a voice source model to the waveforms. Inverse filtering has been studied in detail and it seems unlikely that the procedure can recover the input flow waveform exactly. Thus, it is necessary to know how a parametric model behaves when it is fitted to a corrupted flow signal. As a step in that direction we investigated the behaviour of one specific voice source model, i.e. the LF-model (Fant et al., 1985), in non-linear fit procedures. LF-pulses were fitted to flow signals to which several kinds of white and narrow band noise had been added. In the current chapter we investigate whether the performance of the fit depends on the way in which the LF-pulses are computed. We also study the difference between two classes of fit procedures, viz. simplex search and steepest descent algorithms.

6.2 Method

6.2.1 Fit procedure

In our work non-linear fit procedures are used to estimate LF-parameters from inverse filtered and therefore essentially noisy glottal flow signals (Strik et al., 1992; Strik et al., 1993). The signal resulting from inverse filtering is an estimate of differentiated glottal flow (dU_g). dU_g is then used to calculate the voice source parameters. For each pitch period an LF-model is fitted to dU_g (see Strik et al., 1993). The fit procedure consists of three stages:

① initial estimate

② simplex search algorithm

③ Levenberg-Marquardt algorithm

Non-linear optimization algorithms require that an initial estimate is computed to start the procedure. The impact of the initial estimate on the final result is discussed in Strik et al. (1993). In the second stage of the fit procedure the simplex search algorithm of Nelder & Mead (1964) is used. Of the several optimization algorithms that were tested, the simplex search algorithm usually came closer to the global minimum than the gradient algorithms. Probably, discontinuities in the error function cause the gradient algorithms to get stuck in local minima more often than the simplex search algorithm. However, in the neighbourhood of a minimum, the simplex algorithm may do worse (see Nelder & Mead, 1964). Therefore, the Levenberg-Marquardt algorithm (a gradient algorithm) is used after the simplex algorithm.

In our work we have chosen the LF-model (Fant et al., 1985) for parameterization of dU_g. In the fit procedure five LF-parameters, viz. E_e, t_o, t_p, t_e and T_a are estimated for each pitch period (Figure 1). The goal of the fit procedure is to find the LF-parameters that minimize the difference between the LF-pulse and dU_g. This is done by minimizing the RMS-error of the difference between the LF-pulse and dU_g, i.e. the cost function is defined in the time domain. For all signals a sampling frequency of 10 kHz and an amplitude resolution of 12 bits were used.

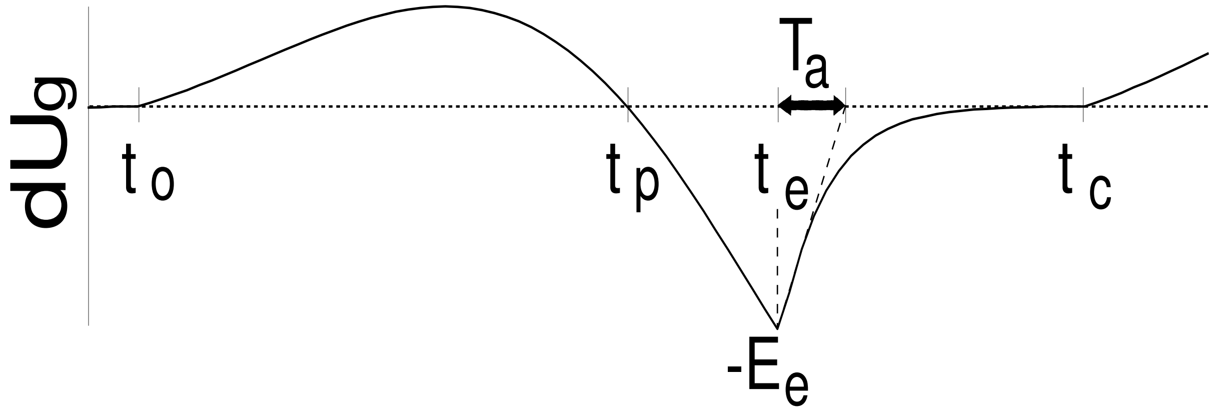


Figure 1. The LF-model and the LF-parameters.

[Note: The time parameters t_p, t_e and t_c in Figure 1 are used to indicate time points. Here a lower case "t" is used to distinguish these time parameters from those in Chapters 4 and 5 (i.e. T_p, T_e and T_c), which refer to time intervals. The relation between the two sets of parameters is: $T_x = t_x - t_o$.]

6.2.2 Test method

Tests were performed to evaluate in which way disturbances on dU_g would affect the estimated parameters. LF-pulses were perturbated in a controlled way, and the LF-parameters of the disturbed signals were estimated by means of the fit procedure. The errors in the estimated parameters were calculated in the following way:

- ❐ $ERR(X) = 100\%*abs(X_{est} - X_{inp})/X_{inp}$, for $X = E_e$
- ❐ $ERR(Y) = abs(Y_{est} - Y_{inp})$, for $Y = t_o$, t_p, t_e and T_a

Subsequently, the absolute values of the errors were averaged. In Figures 2 to 5 mean errors are shown. In the upper row are the mean errors in the estimations of E_e (unit: %), and in the middle and lower rows are the errors in the estimates of t_o, t_p, t_e and T_a (unit: µsec or ms).

The effects of the perturbations cannot always be studied by a single, isolated LF-pulse. For instance, a formant ripple will be present in dU_g when formant and bandwidth values are not estimated correctly. To calculate the first samples of dU_g for the current pitch period, the speech signal resulting from the previous excitation is used. This speech signal is dependent on the amplitude and the shape of the previous flow pulse. Thus, the formant ripple at the beginning of the current pitch period will depend on the amplitude and the shape of the flow pulse in the previous pitch period.

Therefore, we used sequences of three LF-pulses, and each time the voice source model was fitted to the (perturbated) pulse in the middle. Furthermore, 11 LF-pulses with different shapes were used. These 11 standard LF-pulses will be called the basic pulses. The perturbated LF-pulses will be called the test pulses.

6.3 Tests

6.3.1 Quantization

For the fit procedure an algorithm is needed which calculates an LF-pulse for each combination of the five LF-parameters. Initially we used the algorithm described in Lin (1990). In Lin's algorithm E_e and t_p can change continuously, but t_o and t_e are always rounded off to the nearest integer (i.e. the nearest sample point). The consequence is that the error defined as a function of t_o or t_e is a step function. This is certainly

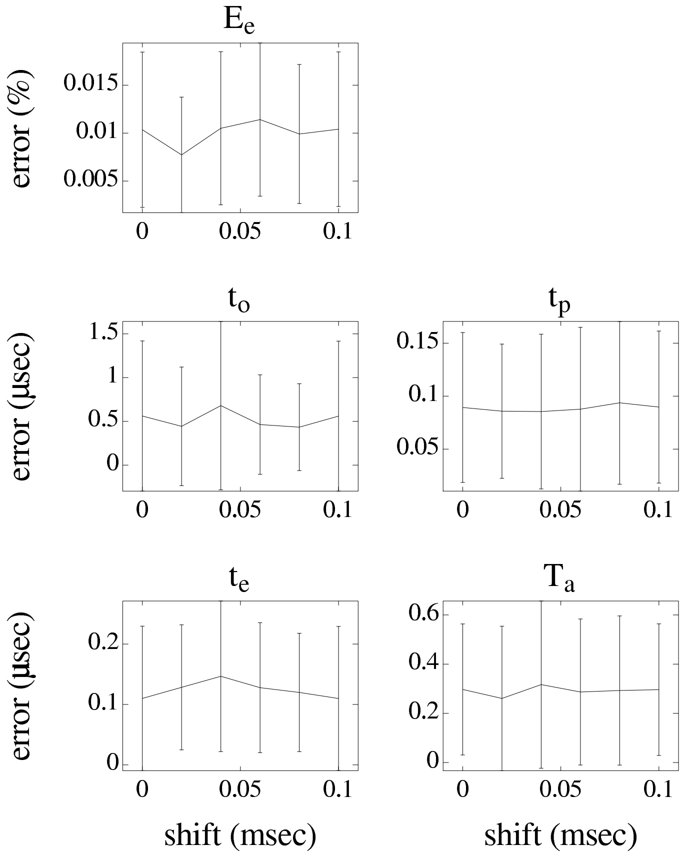


Figure 2. Mean and standard deviation of the error in the estimated LF-parameters for different values of shift.

problematic for gradient algorithms, but also for the simplex algorithm it often resulted in a false convergence to a local minimum. In order to make it possible to have non-integer estimates of t_o and t_e, we adapted Lin's algorithm. In the adapted algorithm the analytic expression of the LF-model is used to calculate the LF-pulse in continuous time, and the LF-pulse is then sampled. With the new algorithm, the fit procedure converged to the global minimum more often, and the resulting errors in the estimated parameters were considerably smaller.

The new algorithm was investigated with test pulses generated by shifting the 11 basic pulses in steps of 0.02 ms (= 0.2 sample), from 0.0 to 0.1 ms (6 values for shift). The amplitude E_e was varied from -1025 to -1023 in steps of 0.2 (11 values for E_e). For the resulting 726 test pulses (11 x 6 x 11) a fit was done, and the estimated parameters were compared with the input values.

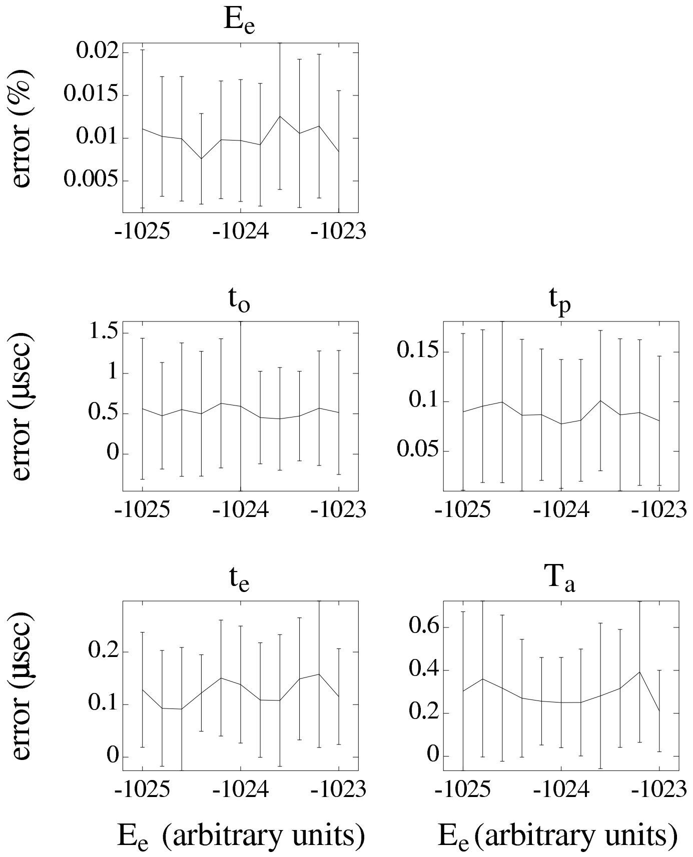


Figure 3. Mean and standard deviation of the error in the estimated LF-parameters for different values of E_e.

In Figure 2 the results for the 121 E_e values (11 x 11) were pooled for each value of shift, and mean and standard deviation were calculated. Similarly, in Figure 3 the 66 values of shift (11 x 6) were pooled for each value of E_e, and again mean and standard deviation were calculated. In Figures 2 and 3 it can be seen that for none of the values of shift and E_e the results deviate significantly. Also, no trend in the errors can be observed. Thus, it appears that with the proposed fit procedure and algorithm for calculation of the LF-pulse it is possible to get accurate estimates for all parameters.

6.3.2 Noise

In practice the voice source signal will always contain noise. In this section we want to study the way in which noise affects the estimates. The noise present in the inverse

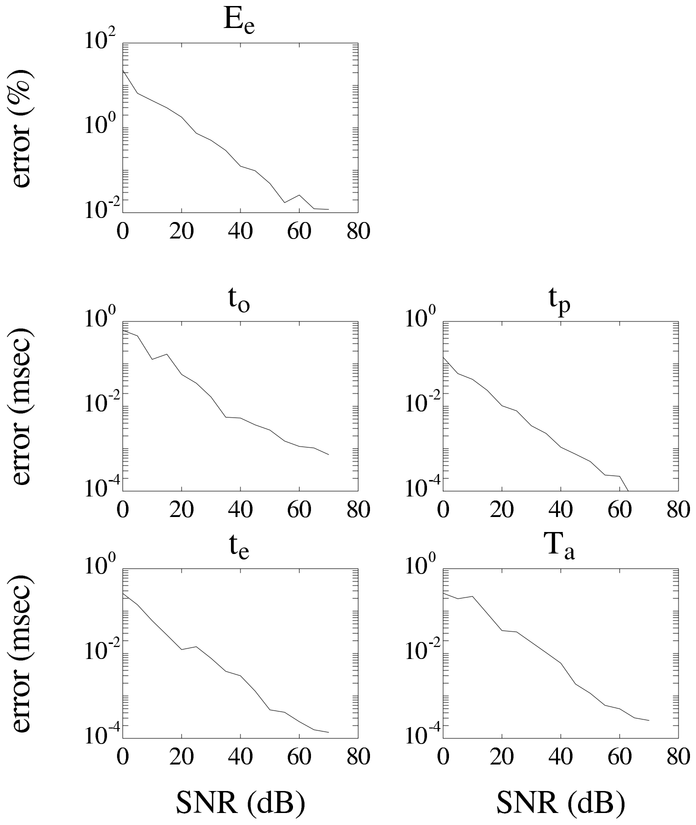


Figure 4. Mean error in the estimated LF-parameters for different values of SNR.

filtered signal can come from many different sources, among which are noise sources at the glottis or in the vocal tract, background noise during the recordings, and quantization noise during A/D conversion. The amplitude and the spectrum of those noise sources may be difficult to establish. To simplify matters we chose to use additive white noise. Although real noise may be different, it is still meaningful to test the influence of additive white noise on the estimated parameters.

Noise with different amplitudes was added to the 11 basic pulses. The amplitudes were chosen so that the Signal-to-Noise Ratio (SNR) varied from 0 to 70 dB, in steps of 5 dB (15 SNR values). For calculation of the SNR the energy of the LF-pulse and the noise signal were integrated over the whole pitch period. The LF-model was fitted to the 165 test pulses (11 x 15) and mean errors for the estimated parameters were calculated (Figure 4).

It can be seen in Figure 4 that the mean error decreases with increasing SNR. If we compare the mean error for the various time parameters we observe that it is largest for t_o. The explanation is probably that near t_o the signal generally changes more slowly than at other time points. Therefore, the noise has more effect on the fit in the neighbourhood of t_o than on other places.

6.3.3 Formant ripple

Errors in the estimation of formant (F) and bandwidth (B) values used in inverse filtering will result in a formant ripple in dU_g. In this section we focus on the effect that a formant ripple will have on the estimated parameters. Test pulses were obtained by fil-

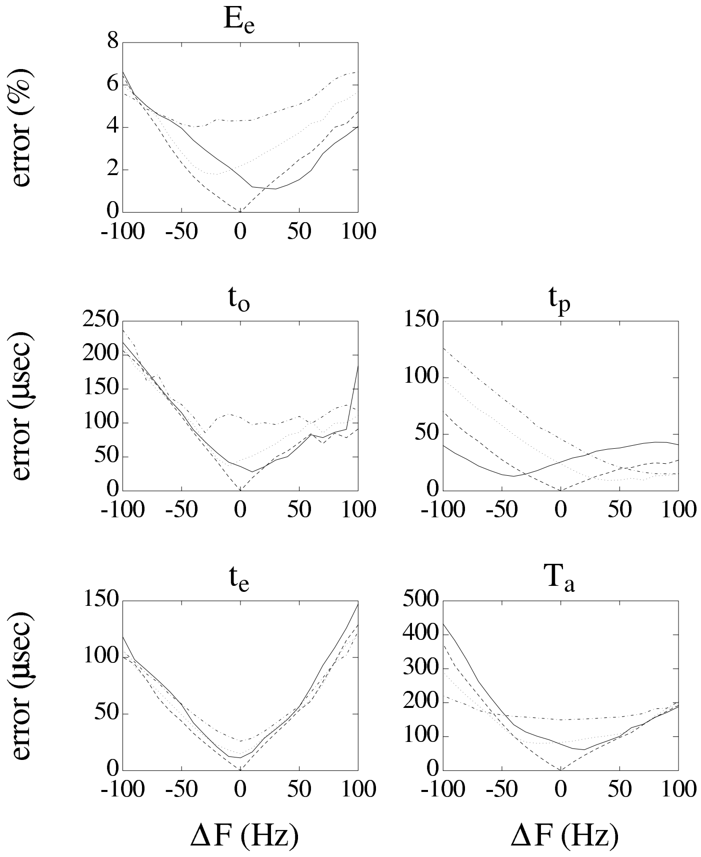


Figure 5. Mean error in the estimated LF-parameters for different values of ΔF and ΔB (F = 500 Hz, B = 80 Hz). The different line types refer to different values of ΔB: -50% (solid), 0% (dashed), +50% (dotted), and +100% (dashed-dotted).

tering the 11 basic pulses with a filter consisting of 1 pole and 1 zero. The pole was kept constant at the "correct" value of F and B, while the F and B values of the zero were varied. The error in F (ΔF) was varied between -20% and 20% of F, and the error in B (ΔB) between -50% and 100% of B.

The results for F = 500 Hz and B = 80 Hz are shown in Figure 5. The error in the estimates of the LF-parameters is smallest when there is no formant ripple ($\Delta F = \Delta B = 0$), as was to be expected. For large values of ΔF and ΔB, and consequently a large ripple in the test pulses, the errors in the estimates remain remarkably small. Probably, the explanation is that the best fit is determined for the whole period, and thus a local ripple does not necessarily have a drastic effect on the global fit.

The mean error in T_a is larger than the mean error in t_o, t_p and t_e. This is no surprise, as the formant ripple is most pronounced just after the main excitation. If B is not too large, the ripple will still be present at the beginning of the next pulse. This ripple can be problematic for the estimation of t_o, especially because the signal changes relatively slowly around t_o. This could be the reason why the mean errors in t_o are larger than the mean errors in t_p and t_e.

The experiment was repeated for other values of F and B. For formant values of 400 and 600 Hz and a bandwidth of 80 Hz the resulting errors were comparable with those in Figure 5. However, for higher formant values (in the range of the second and third formant) the errors were substantially smaller. The ripple caused by errors in the estimation of the second or third formant have higher frequencies and thus will be less problematic for the fit procedure than a ripple with a low frequency caused by an error in the estimation of the first formant. Furthermore, the bandwidth of higher formants will generally be larger than the bandwidth of the first formant, and the corresponding ripple will damp out more quickly. This ripple will have less effect on the estimates of the following period.

6.4 Conclusions

In Section 6.3.1. we showed that with the proposed method it is possible to get accurate estimations of time points that do not coincide with sample points. The algorithm used to calculate the LF-pulse proved to be very important, i.e. with an adapted version of Lin's algorithm (Lin, 1990) the errors in the estimations were considerably smaller than with the original version. The adapted algorithm is more time-consuming, but this is no problem for a fit procedure.

The magnitude of the errors resulting from additive white noise increases with decreasing SNR. With a 12-bit quantization it is usually possible to have an SNR of at least 30 dB. If the SNR is higher than 30 dB the error in E_e is less than 0.5%, and the error in the time parameters is less than 0.02 ms (i.e. 0.2 sample). For most applications this is probably acceptable.

Errors in the estimation of the first formant have more effect on the estimated parameters than errors in the higher formants. For this reason, special attention should be given to the estimation of the first formant during inverse filtering.

For the disturbances tested in the present chapter the fit procedure gave satisfactory results. However, other disturbances could be present in the inverse filtered signal, and these disturbances should also be tested. Furthermore, in all tests the perturbations were applied to LF-pulses. However, it is still unclear whether the LF-model can give a sufficiently accurate description of all glottal flow pulses. This is also a topic for future research.

Chapter 7

A physiological model of intonation

Abstract

In the literature different views are expressed regarding the relation between intonation in running speech and the underlying physiological mechanisms. The goal of the present research was to clarify this relation. Simultaneous measurements were made of laryngeal and respiratory activity for two subjects. On the basis of these data and a considerable amount of comparable data from the literature, a physiological model of intonation is proposed. In the resulting model the global downtrend in F_0 is accounted for by the slow decline in subglottal pressure, while the local variations in F_0 are caused mainly by local variations in subglottal pressure and activity of cricothyroid and vocalis muscles. The role of the strap muscles remains unclear.

This chapter is based on the following paper: Strik, H. & Boves, L. (1993) A physiological model of intonation. *Proceedings of the Department of Language and Speech, Nijmegen University*, 16/17, 96-105.

7.1 Introduction

Several studies dealing with the relation between fundamental frequency (F_0) and the underlying physiological processes show that subglottal pressure (P_{sb}) and the activity of the cricothyroid (CT), vocalis (VOC) and sternohyoid (SH) muscles are important factors in the control of F_0 (Rubin, 1963; Shipp & McGlone, 1971; Collier, 1975; Baer et al., 1976; Maeda, 1976; Atkinson, 1978; Shipp et al., 1979; Hirose & Sawashima, 1981; Gelfer, 1987). However, a substantial part of the data described in these papers pertain to singing and sustained phonation; and many of these investigations are based on measurements concerning either the respiratory system or the activity of the laryngeal muscles. Relatively few studies make use of simultaneous registrations of respiratory and laryngeal activity in running speech. This may be one of the reasons why it is not completely clear yet how these factors cooperate in the regulation of F_0 for running speech.

However, intonation is not the same as F_0. Intonation refers to the linguistically relevant aspects of the F_0 contours. Two of these aspects have received particular attention in the literature, viz. phrase level contours and prominence. The lack of agreement on the physiological processes underlying intonation is reflected by the fact that at least three different accounts of the physiology of intonation have been given in recent publications.

In Fujisaki's model (Fujisaki, 1991) the phrase and accent components are controlled by two functionally different parts of the CT, the CT pars obliqua and the CT pars recta, respectively. In the model by 't Hart et al. (1990) the slow decrease in F_0 is brought about by a slow decrease in P_{sb}, whereas all other pitch movements are controlled by the CT. Finally, there are phonologically oriented intonation models, in which intonation contours are viewed as concatenations of targets. Beckman & Pierrehumbert (1992) state that the High targets are accounted for by the CT only. According to these authors the physiological explanation of the Low targets is more complex. They suggest a number of possibilities such as CT relaxation, P_{sb} lowering, strap muscle activity and activity of the cricopharyngeus muscle.

The purpose of the research reported on in this chapter is to clarify the relation between intonation and the physiological mechanisms for running speech. To this end simultaneous measurements of laryngeal and respiratory activity were made for two subjects. Our ultimate goal is to propose a quantitative model for the physiological control of intonation. As a first step towards a quantitative model we used a qualitative analysis method for the following reasons:

(1) At present there seems to be no realistic voice source - vocal tract model that can be used for a quantitative analysis of the relation between the measured physiological signals and F_0. For instance, it is not possible to predict F_0 on the basis of EMG activity of a laryngeal muscle. At the most, one can predict the direction of F_0 changes as a function of EMG activity.

(2) For a quantitative analysis of the relation between the physiological signals and intonation, an intonation model should be used. However, it is not clear which of the various models proposed in the literature is the most appropriate. To establish this, one would have to compare the different models. It will be very difficult, though, to define one measure with which all models could be evaluated, because the various models use different entities: e.g. tones or F_0 targets in phonologically oriented models, F_0 movements in the model proposed by 't Hart et al., (1990), or phrase and accent commands in Fujisaki's model (1991). However, such a comparison would require a detailed study of its own.

(3) From our own data, as well as from the literature (Gay et al., 1972; Ladefoged, 1967), it appears that there are differences between subjects in the physiology underlying intonation. A qualitative analysis made it possible to use data available in the literature, in addition to our own data. Especially the data found in Ladefoged (1967), Lieberman (1967), Collier (1975), and Gelfer (1987) proved to be useful.

Our goal was to find out whether systematic behaviour can be observed in the data, and in which way this behaviour can be modelled.

The outline of this chapter is as follows. Section 7.2 describes the material and the method used in our research. The results for running speech are presented in Section 7.3. The physiological model resulting from our investigations is described in Section 7.4. In Section 7.5 we discuss our model and its relation to previous research. Finally, in Section 7.6 we will present the conclusions.

7.2 Material and method

For two Dutch male subjects recordings were made of the audio signal, electroglottogram, lung volume (V_l), P_{sb}, SH and VOC. In addition to these signals, the CT was also measured for subject LB, and oral pressure (P_{or}) for subject HB. In the latter case transglottal pressure (P_{tr}) was calculated by taking the difference of P_{sb} and P_{or}: $P_{tr} = P_{sb} - P_{or}$.

The measurements were made while the subjects produced sustained vowels and meaningful Dutch sentences with different intonation patterns. The sentences spoken by subject LB were "Piet slikte zijn pillen met bier" (SU: Short Utterance); and "Piet slikte gisteren zijn vierentwintig gele pillen liever in stilte met bier" (LU: Long Utterance). The sentences produced by subject HB were "Heleen wil die kleren meenemen" (SU: Short Utterance); "Heleen en Emiel willen die kleren liever wel weer meenemen" (LU: Long Utterance); and "Indien Emiel die kleren wil meenemen, willen wij ze eerst wel even zien" (SWC: Sentence With Comma). These sentences contain mainly high vowels, in order to minimize the involvement of the SH in articulatory gestures.

Our goal was to obtain utterances with different intonation patterns. Therefore, the subjects were instructed to produce the utterances in various ways. The following instructions were given: early stress (HB-SU1); early and late stress, lower F_0 between the 2 stressed syllables (HB-SU2, LB-SU2 and LB-LU2); early and late stress, keep F_0 high between the 2 stressed syllables (HB-SU3, LB-SU1 and LB-LU1); and question intonation (HB-SU4, HB-LU4, LB-SU3 and LB-LU3). Subject HB did not receive any instruction on how to produce the sentence HB-SWC. Within the recorded sentences there were no inspirations (resets of V_l), nor any resets of F_0 or P_{sb}.

The subjects repeated each sentence 5 to 8 times. The signals of these repetitions were used to calculate average signals. All EMG signals were first shifted forward over their mean response times (Atkinson, 1978). Next, the method of non-linear time-alignment and averaging (Strik & Boves, 1991a) was used to average the signals. An advantage of this method is that it also yields an average F_0 contour, while in previous studies one of the F_0 contours was usually chosen to represent the “average” F_0 contour. All signals shown in the present chapter are average signals which are time-aligned with the audio signal. The procedures used for recording and processing the data are described in more detail in Strik & Boves (1992a).

7.3 Running speech

A speaker can use many different physiological mechanisms to control F_0. Therefore, it is remarkable that the within-subject variation between the signals of repetitions of the same utterance is relatively small. This suggests that speakers have a good notion of the manner in which they want to produce an utterance, and that they have good control over these mechanisms. Consequently, meaningful averaging of the data is possible (Strik & Boves, 1991a).

Although there are some individual differences, consistent behaviour between subjects can be observed in the data. Ladefoged (1967) also noted that for his subjects "the results obtained so far are sufficiently consistent to suggest the general pattern of the relationships involved". This consistency led to the discovery of general patterns in the behaviour of P_{sb}, CT, VOC and SH, which we want to describe in this section.

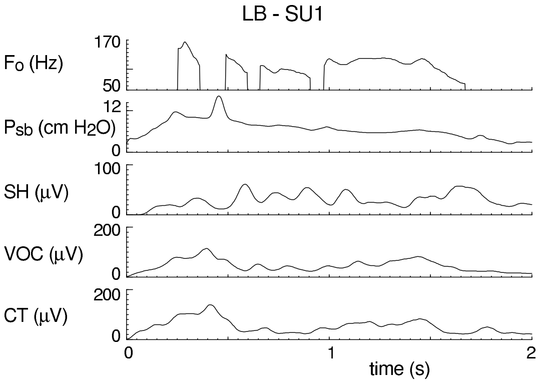
From our data it appears that both F_0 and the physiological signals have two components, viz. a global and a local one. This was also noticed by Maeda (1980) and Gelfer (1987) for their physiological data. The intonation models described in 't Hart et al., (1990), Fujisaki (1991), and Beckman & Pierrehumbert (1992) also have two components. In Strik & Boves (1992a) we showed that this qualitative observation has a quantitative statistical basis: on a global level P_{sb} explains most of the observed variance of F_0, while on a local level the laryngeal muscles become more important. Our treatment focuses on the linguistically significant aspects of F_0, especially those connected with stress, phrasing and the question-statement distinction. Initial rise and final lowering of F_0 are treated separately for reasons that are explained in Section 7.3.2.5.

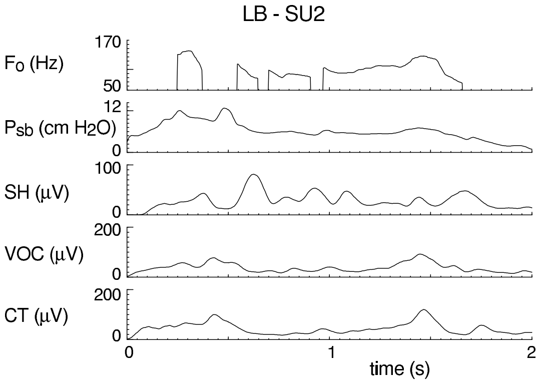
7.3.1 Global level

From the recordings shown in Figures 1 and 2 it is apparent that P_{sb} has a global and a local component. The global pattern of P_{sb} will be called $P_{sb,g}$. In most sentences there is a gradual lowering of $P_{sb,g}$. This can also be seen in the data of Lieberman (1967), Collier (1975) and Gelfer (1987).

Figure 1. Average physiological signals for 6 utterances of subject LB. SU: "Piet slikte zijn pillen met bier"; LU: "Piet slikte gisteren zijn vierentwintig gele pillen liever in stilte met bier"

Figure 2. Average physiological signals for 6 utterances of subject HB. SU: "Heleen wil die kleren meenemen"; LU: "Heleen en Emiel willen die kleren liever wel weer meenemen"; SWC: "Indien Emiel die kleren wil meenemen, willen wij ze eerst wel even zien"


LB - SU1
Fo (Hz)
170
50
Psb (cm H2O)
12
0
SH (µV)
100
0
VOC (µV)
200
0
CT (µV)
200
0
0
1
2
time (s)


LB - SU2
Fo (Hz)
170
50
Psb (cm H2O)
12
0
SH (µV)
100
0
VOC (µV)
200
0
CT (µV)
200
0
0
1
2
time (s)

The shape of $P_{sb,g}$ differs among speakers. For subject LB the shape is concave (see e.g. Figure 1b): the slope is steep initially, and it gradually becomes more flat towards the end. The same pattern is observed for the two subjects in the studies of Collier (1975) and Gelfer (1987), and for speaker 2 in the study of Lieberman (1967). However, for subject HB the pattern is more convex (see e.g. Figure 2a): $P_{sb,g}$ decreases slowly in the beginning, and more rapidly near the end. The same pattern is also found for speakers 1 and 3 in the study of Lieberman (1967). In spite of the differences between subjects, the behaviour is relatively consistent within subjects.

The global reference level of CT, VOC and SH seems to be constant (see e.g. Figure 1f). In other words, these laryngeal muscles usually do not have a global component. Consequently, it seems that the global behaviour of F_0 is generally determined by $P_{sb,g}$. This relation is studied in more detail in Strik & Boves (submitted). The global component of F_0 will be called $F_{0,g}$. A downtrend in $P_{sb,g}$ will result in a downtrend in $F_{0,g}$. This downtrend in $F_{0,g}$ has been observed in declarative utterances of many languages (Breckenridge, 1977).

7.3.2 Local level

F_0, P_{sb} and the laryngeal muscles have a local component. If there are local variations in F_0, these normally seem to coincide with local variations in P_{sb}, CT, VOC and SH (see e.g. Figures 1f, 2d, 2f). This can also be seen in the data of Ladefoged (1967), Lieberman (1967), Collier (1975) and Gelfer (1987).

7.3.2.1 Cricothyroid

The conclusion of many studies was that of all physiological factors known to affect F_0, the CT shows the most consistent relation to F_0 (Collier, 1975; Maeda, 1976; Atkinson, 1978; Shipp et al., 1979; Erickson et al., 1983; Gelfer, 1987). Also in our data we see that for local variations in F_0 there is usually a local variation in the activity of the CT. It is generally agreed that the CT is an important factor in the control of F_0. The local variation in CT explains (at least) part of the local variation in F_0.

7.3.2.2 Vocalis

For local F_0 movements we usually observe a covariance of F_0 and VOC in our data. This relation was also studied by Maeda (1976) and Atkinson (1978) for sentences with various intonation patterns. No direct relation between VOC and the F_0 movements was found by Maeda for the subject used in his study. However, Atkinson did find a positive correlation between VOC and F_0 for his subject. VOC is used to control F_0 for sustained phonation and singing (Rubin, 1963; Sawashima et al., 1969; Shipp & McGlone, 1971; Gay et al., 1972; Shipp et al., 1979). Hirose & Gay (1972) observed an increase in the activity of CT and VOC for stressed vowels in isolated words. Probably VOC acts in synergy with CT in the control of F_0.

7.3.3.3 Subglottal pressure

Both measurements (Ladefoged, 1967; Lieberman, 1967; Collier, 1975; Baer et al., 1976; Atkinson, 1978; Baken & Orlikoff, 1987; Gelfer, 1987) and modelling (Titze, 1989) have shown that a change in P_{sb} will affect F_0, ceteris paribus. During local F_0 movements a covariation of F_0 and P_{sb} is often observed in our data, and in the data of Ladefoged (1967), Lieberman (1967), Collier (1975) and Gelfer (1987). Part of this local P_{sb} variation might be due to a change in the impedance of the glottis which, in turn, results from changes in the activity of the laryngeal muscles (e.g. the changes in CT and VOC, as noted above). However, part of the P_{sb} variation could also be due to changes in pulmonic activity. For instance, increased activity of the respiratory muscles for stressed syllables was found by Ladefoged (1967) and Van Katwijk (1974). Whatever the cause of a P_{sb} variation, the result is a change in F_0.

7.3.2.4 Sternohyoid

The function of the SH in the control of F_0 is not completely understood. Erickson & Atkinson (1976), Maeda (1976) and Erickson et al., (1983) postulated that F_0 falls are initiated by a relaxation of the CT, which is followed by increased activity of the SH. Collier (1975) argued that SH cannot be the primary effector of an F_0 fall. Atkinson (1978) found a high negative correlation between SH and F_0, while Erickson et al., (1977) concluded that the SH has “a slightly negative relation to F_0”. For some sentences in our data there is also a small negative correlation between SH and F_0. However, this negative correlation is mainly brought about by the increase in SH and the lowering of F_0 at the end of many utterances (the so-called final lowering, see Section 7.3.2.5). Of course, final lowering will affect the correlation coefficient to a greater extent if the utterances are short, like those used by Atkinson (1978). The SH is probably used in some F_0 lowerings, but it is also used for articulatory gestures such as jaw lowering, tongue lowering and retraction. Therefore, the relation between the SH and F_0 is probably complex. This is illustrated in Figure 2c. During the F_0 lowering there is a peak in the activity of SH, and in this case the SH could have assisted in lowering F_0. But similar peaks can be observed also when F_0 increases or remains steadily high. Anyhow, no consistent, transparent relation can be found in our data nor in the data of Collier (1975) and Gelfer (1987).

7.3.2.5 Initial rise and final lowering

High values of F_0, CT, VOC and P_{sb} are often observed at the beginning of utterances, both in our data and in the data of Collier (1975), Maeda (1976) and Gelfer (1987). This effect shows up more prominently in the utterances of subject LB, especially in the longer ones, while it is less evident in the utterances of subject HB. In questions this initial rise is slightly reduced compared to the statements.

Towards the end of many utterances F_0 and P_{sb} often decrease substantially, while there is a marked increase in the SH activity. Final lowering has also been observed by Collier (1975) and Maeda (1976). Increased SH activity and the large drop in P_{sb} usually take place before phonation has stopped. However, in interrogative sentences both changes are often delayed till after the utterance. Furthermore, the small drop in P_{sb}, which sometimes remains, is counterbalanced by a large increase in the activity of CT and VOC. Therefore, the final lowering of F_0 is rarely observed in questions.

The initial rise is probably the result of laryngeal adjustments that are needed to start phonation (prephonatory tuning), while the final lowering could be a preparation for the next inhalation (Wyke, 1983). Both kinds of local F_0 variations could therefore be seen as the by-product of physiological manoeuvres that are necessary for speech production.

Initial rise and final lowering are not generally used to signal stress, but still they could be linguistically significant. Prosody plays an important role in communication. It is used, among other things, to mark the boundaries between phrases (Breckenridge, 1977; Cooper & Sorensen, 1981). Pierrehumbert (1979) suggested that the downtrend in F_0 and IL may be important in the perception of phrasing. The F_0 fall that results from the downtrend in F_0, is often enlarged by initial rise and final lowering. Consequently, both effects could assist in the signalling of boundaries. In interrogative utterances the indications of a linguistic control of both phenomena are especially clear. In these utterances initial rise and final lowering were often reduced. Of course, a high F_0 at the beginning, and especially a lowering of F_0 at the end of an utterance would interfere with the desired rising intonation.

From our data it is not manifest whether initial rise and final lowering are linguistically controlled variables, or if they are primarily the by-product of physiological gestures that are needed in speech production. That is the reason why these local F_0 movements are treated separately from the other local F_0 movements which obviously do have a linguistic purpose.

7.4 A physiological model of intonation

The ultimate goal of our research is to derive a quantitative physiological model of intonation. For the reasons mentioned in the introduction, we prefer to use a qualitative analysis method first. The results of our qualitative analysis are presented in Section 7.3. The qualitative model proposed in this section is a purely descriptive model, and should therefore be seen as a first step towards a comprehensive quantitative model. This qualitative model is based on consistent behaviour of P_{sb}, CT, VOC and SH that was observed in the data of various subjects (see Section 7.3).

Intonation and its physiological control take place at two levels, viz. a global and a local level. In general, CT, VOC and SH do not have a global component, while F_0 and P_{sb} do. Therefore, the global component of F_0 ($F_{0,g}$) seems to be determined by $P_{sb,g}$. $P_{sb,g}$ has a tendency to decline. The downtrend in $P_{sb,g}$ will lead to a downtrend in $F_{0,g}$ (Strik & Boves, submitted).

Although the SH is consistently used in final lowerings, no transparent relation was found between the SH and other local F_0 variations. Therefore, in our model the SH does not play a role in the control of the latter type of local variations.

At the beginning of utterances CT, VOC and P_{sb} may have extra high values (initial rise). At the end of utterances SH often shows an increase while P_{sb} drops sharply. If these effects occur during voiced sounds at the end of the utterance, final lowering is observed. Alternatively, SH activity and P_{sb} release may be delayed until after the last voiced sound, in which cases final lowering is absent. The initial rise and final lowering of F_0 will add to the F_0 fall that results from the downtrend in F_0.

Besides initial rise and final lowering, other local variations in F_0 often occur. These local variations in F_0 are generally caused by variations in CT, VOC and P_{sb}. F_0 can be raised by increasing CT, VOC and P_{sb}, and F_0 can be lowered by decreasing P_{sb} and relaxing CT and VOC.

7.5 Discussion

7.5.1 The proposed model

The SH is usually involved in final lowering of F_0. In our model the other F_0 lowerings are brought about by a relaxation of CT, VOC and P_{sb}, i.e. the same mechanisms used to raise F_0 are also used to lower it. According to our data and the data of Collier (1975) and Gelfer (1987), no separate mechanism (like SH) seems to be needed to produce these low tones. The strap muscles are probably used to produce very low tones, as during final lowering. It is possible that these extra low tones do not occur often in those parts of utterances that precede final lowering. This would imply that the role of the SH in the control of F_0 in running speech is limited.

The reference line of a laryngeal muscle is defined as the activity observed when the muscle shows minimal activity. Consequently, the activity of the CT and VOC cannot be lower than the reference level, and it can only be lowered if it has been raised previously. For local variations of P_{sb} it is also observed that P_{sb} is first raised, relative to $P_{sb,g}$, and then lowered again. Thus it seems that a local lowering of CT, VOC and P_{sb} is always preceded by a local rise. The question is what happens if a sentence starts with a high F_0 that is part of the intonation contour proper (i.e., it is not an initial rise). As there is no such intonation pattern in our data nor in the data of Collier (1975) and Gelfer (1987), we can only speculate on the answer. In this case we would expect CT, VOC and P_{sb} to rise before phonation has started, and to remain high until the first F_0 lowering.

In previous intonation studies the term baseline was used regularly. In general it is defined as a line “drawn near or through the low values of F_0 occurring in an utterance”

(Cooper & Sorensen, 1981). This baseline will resemble $F_{0,g}$, although they are not identical. In our model $F_{0,g}$ is the global component of F_0, i.e. the component that remains after all local effects have been removed. Initial rise, final lowering and the rise at the end of questions are considered to be local effects, and thus are not part of $F_{0,g}$. According to the definition given above, they probably are part of the baseline. The baseline also differs from $F_{0,g}$ when F_0 is lowered by F_0-lowering mechanisms (e.g., the strap muscles). In that case the baseline will drop below $F_{0,g}$.

7.5.2 Data from the literature

Our physiological model of intonation is based on our own data and on the data of Lieberman (1967), Ladefoged (1967), Collier (1975) and Gelfer (1987). However, some of the conclusions that were expressed in these articles are different from our conclusions.

Lieberman (1967) made measurements of P_{sb}, but he did not measure the activity of the laryngeal muscles. He observed a resemblance in the behaviour of F_0 and P_{sb}, except at the end of interrogative utterances. At the end of questions there was an increase in F_0, while P_{sb} generally did not increase. His assumption was that the activity of the laryngeal muscles increased at the end of questions, but remained relatively steady otherwise. Based on this assumption he concluded that, apart from questions, F_0 is a function of P_{sb} alone. This conclusion can easily be verified by calculating the frequency-to-pressure ratio in his data. The rate of F_0 changes that result from a change in P_{sb} alone should be in the range 2-7 Hz/cm H_2O (e.g. Ladefoged, 1967; Baer, 1979). According to Lieberman (1967: 97) this ratio is about 20 Hz/cm H_2O in his data, while Ohala (1990) claims that it is even larger. In any case, P_{sb} alone cannot explain all the variation in F_0, and other mechanism must have been involved. It is likely that the laryngeal muscles were involved, not only at the end of questions, but also in other parts of the utterances. Although we do not agree with the above-mentioned conclusion of Lieberman, our model fits the general pattern in his data: $P_{sb,g}$ gradually declines, and local variations in P_{sb} explain part of the local variations in F_0.

The conclusion of Ladefoged (1967) that both vocal cord tension and P_{sb} contribute to stress is in agreement with our model. He presents data for utterances with stress on the last word, and part of the utterances is also produced with a rising intonation (questions). In his data it can be seen that P_{sb} has a local component, for P_{sb} generally increases for stressed words and at the end of questions. This is also in line with our model. Owing to the presence of these P_{sb} increases at the end of most of his utterances, P_{sb} is about level or slightly increases in these utterances. This seems to be in contradiction with our claim that $P_{sb,g}$ generally decreases. However, to study the behaviour of $P_{sb,g}$, the local variations in P_{sb} have to be removed. After this is done, it is likely that $P_{sb,g}$ will decline, also in Ladefoged's data.

The conclusions of Collier (1975) are based on the data of one subject. The majority of the physiological data presented in Gelfer (1987) concern the same subject, while she also shows data for one other subject. Although Collier and Gelfer do not offer an explicit model, their main conclusions are similar: P_{sb} controls the gradual falling baseline, while local F_0 movements are controlled by the CT. They both observed local variations in CT and P_{sb} for local F_0 movements, and found that the frequency-to-pressure ratio for these movements is higher than the expected 2-7 Hz/cm H_2O. They argued that as P_{sb} cannot explain all the variation in F_0, it must be the CT that is the most important factor in the control of F_0. However, one can calibrate the F_0-P_{sb} ratio, but it is virtually impossible to calibrate the F_0-EMG ratio for a laryngeal muscle. An important reason is that the magnitude of an EMG signal depends on many factors that are difficult to control (for instance, the magnitude is dependent on the exact place of the electrode in the muscle). The conclusion is that one can check whether P_{sb} explains all of the variance in F_0, but the same check cannot be made for a laryngeal muscle. Besides CT other factors could be involved. In fact, P_{sb} and VOC (and probably other factors) are usually involved in the local F_0 movements. Because it is difficult to calibrate the F_0-EMG ratio, it is hardly possible to decide on quantitative grounds which factor is the most important.

7.6 Conclusions

Research into the relation between physiology and intonation is complicated by a number of factors. First of all, there seems to be no realistic quantitative voice source - vocal tract model that can be used to study this relationship. Second, although various models have been proposed in the intonation literature, it is not clear a priori which model is the most suitable. Finally, differences in the physiological control of intonation have been found among subjects. This makes it difficult to draw conclusions on the basis of observations concerning one or two subjects. On the other hand, the measurements necessary for this kind of research are so invasive that investigators refrain from measuring great numbers of subjects. For this reason we completed our own data with data from the literature.

Compatible behaviour has been found in the data of Dutch, British and American subjects. The physiological model of intonation proposed in Section 7.4 describes this behaviour. In this model the downtrend in F_0 is accounted for by the slow decrease in P_{sb}. The total F_0 fall is often augmented by an initial rise and a final lowering. Other local variations in F_0 are generally caused by simultaneous variations in CT, VOC and P_{sb}.

Chapter 8

Downtrend in F_0 and P_{sb}

Abstract

In the present chapter we examine the simultaneous downtrend in fundamental frequency and subglottal pressure that is often observed for running speech. In particular, we will test the hypothesis that the downtrend in fundamental frequency is caused by a gradual decrease in subglottal pressure during the course of an utterance. In the literature various ways to model the downtrend in fundamental frequency have been proposed. Our conclusion is that whether the hypothesis stated above is true depends on the model of downtrend adopted.

This chapter is a slightly revised version of the following paper: Strik, H. & Boves, L. (submitted) Downtrend in F_0 and P_{sb}. *Journal of Phonetics*

8.1 Introduction

A simultaneous downtrend in fundamental frequency (F_0) and subglottal pressure (P_{sb}) has often been observed for running speech (Lieberman, 1967; Ohala, 1970; Collier, 1974, 1975; Atkinson, 1978; Gelfer, 1987; Strik & Boves, 1993). As it is known that changes in P_{sb} will affect F_0, everything else being equal (Titze, 1989), it seems plausible to assume that both downtrends are related. However, a considerable deal of controversy surrounds the relation between the two downtrends (see e.g. Ohala, 1978, 1990; Cohen et al., 1982; Ladd, 1984).

Research on the relation between the downtrend in F_0 and P_{sb} is impeded by the fact that there is still no consensus on the correct way to model the downtrend in F_0. In the literature various models have been proposed. Many of these models consist of two components: a short-term or local component and a long-term or global component. In these models the global component is used to model the downtrend in F_0. Only some of these models provide a physiological explanation of both components. Öhman (1968), Collier (1975) and Fujisaki (1991) agree that the local component is controlled by the laryngeal muscles, but they do not agree about the control of the global component. According to Öhman (1968) and Fujisaki (1991) downtrend is also controlled by the laryngeal muscles, while according to Collier (1975) it is controlled by P_{sb}.

In Strik & Boves (1993) the relation between F_0 and some of the physiological mechanisms that are known to be important in the control of F_0 is studied by means of a qualitative analysis. Based on our own data and data from the literature it was concluded that from a physiological viewpoint the following hypothesis is plausible: the downtrend in F_0 is due to the downtrend in P_{sb}. However, this hypothesis is not unchallenged. In this chapter we will discuss the two main counterarguments:

① the lowering in P_{sb} cannot explain all of the decrease in F_0 (Section 8.4.2); and

② downtrend is part of the linguistic code, and thus it must be controlled by laryngeal muscles and not by P_{sb} (Section 8.4.3).

The fact that this issue is still controversial is expressed in the conclusion of a recent article by Ohala (1990): "It must be concluded that the question of whether F_0 declination is caused by laryngeal or by respiratory activity has still not been answered definitively." The purpose of this chapter is to clarify the relation between the downtrend in F_0 and P_{sb}.

In the literature different models of intonation are available, which are motivated both by phonetic and phonological considerations. The primary goal of the present chapter is to study the relation between the downtrend in F_0 and P_{sb}. For this reason we look primarily at intonation from a physiological point of view. As a consequence, we try to avoid theory-laden terms like e.g. 'downdrift', 'declination' and 'baseline' as much as possible. Instead we predominantly use the more neutral term 'downtrend'. In some sections we refer to previous studies in which the term 'declination' is generally used. In these cases we will also use the term 'declination'. In this chapter 'downtrend' and 'declination' are seen as synonyms, and are used to denote the gradual lowering of a signal during a whole utterance.

The outline of the present chapter is as follows. In Section 8.2 material and method are described. Each experiment consisted of two parts. In part one the subjects were instructed to sustain vowels, and in part two they produced meaningful sentences. The results for 'sustained phonation' are described in Section 8.3. These results are then used in the argumentation of Section 8.4, in which the results for 'running speech' are presented. In Section 8.4.1 our physiological model of intonation is described. Subsequently, the two counterarguments mentioned above are discussed in Sections 8.4.2 and 8.4.3, respectively. Section 8.5 contains a general discussion. Finally, some conclusions are drawn in Section 8.6.

8.2 Material and method

Recordings were made of the audio signal, electroglottogram, lung volume (V_l), P_{sb}, and the activity of the sternohyoid (SH) and vocalis (VOC) muscles for two Dutch male subjects. Both subjects had normal phonation and hearing, but had not received special voice training. In addition to these signals, the activity of the cricothyroid (CT) muscle was also measured for subject LB, and oral pressure for subject HB. The electromyographic (EMG) signals of the laryngeal muscles were high-pass filtered, full-wave-rectified, and integrated over successive periods of 5 ms. All EMG signals were shifted forward over their mean response times, using the procedure described in Atkinson (1978).

The measurements were made while the subjects produced sustained vowels and meaningful Dutch sentences with different intonation patterns. The sentences spoken by subject LB were "Piet slikte zijn pillen met bier" (SU: Short Utterance); and "Piet slikte gisteren zijn vierentwintig gele pillen liever in stilte met bier" (LU: Long Utterance). The sentences produced by subject HB were "Heleen wil die kleren meenemen" (SU: Short Utterance); "Heleen en Emiel willen die kleren liever wel weer meenemen" (LU: Long Utterance); and "Indien Emiel die kleren wil meenemen, willen wij ze eerst wel even zien" (SWC: Sentence With Comma). These sentences contain mainly high vowels, in order to minimize the involvement of the SH in articulatory gestures.

Our goal was to obtain utterances with different intonation patterns. Therefore, the subjects were instructed to produce the utterances in various ways. The following instructions were given: early stress (HB-SU1); early and late stress, lower F_0 between the 2 stressed syllables (HB-SU2, LB-SU2 and LB-LU2); early and late stress, keep F_0 high between the 2 stressed syllables (HB-SU3, LB-SU1 and LB-LU1); and question intonation (HB-SU4, HB-LU4, LB-SU3 and LB-LU3). Subject HB did not receive any instruction on how to produce the sentence HB-SWC.

Some sentences were also produced in reiterant form, using either the syllable /fi/ or /vi/. The subjects repeated each sentence 5 to 8 times. The raw signals of these repetitions were used to calculate median signals for each intonation contour. The method of non-linear time-alignment and averaging was used to average all signals, including F_0 (Strik & Boves, 1991a). The procedures used for recording and processing the data are described in more detail in Strik & Boves (1992).

8.3 Sustained vowels

Before the actual measurements of the physiological signals were made, our subjects were trained to produce prolonged vowels for different combinations of F_0 and intensity level (IL). When the subjects were asked to sustain a given vowel, a gradual lowering of F_0 and IL was generally observed. Subsequently, when they were explicitly instructed to keep F_0 and IL constant, the downtrend in F_0 and IL diminished, but it was usually still present. Finally, the subjects were given on-line visual feedback of F_0 and IL. In this condition they often managed to keep both F_0 and IL fairly constant during the production of a vowel.

After the training sessions actual measurements of the physiological signals were obtained. The subjects were given on-line visual feedback and were again instructed to keep F_0 and IL constant for a sustained vowel. This task was repeated for different combinations of F_0 and IL. The measurements show that the subjects usually managed to keep F_0 and IL at the target values. At the beginning of the utterances some variation in P_{sb} and the activity of the laryngeal muscles was observed, probably to reach the target levels for F_0 and IL. Apart from the initial variation the physiological signals usually remained constant for the rest of the utterance. Different combinations of F_0 and IL were achieved by different levels of P_{sb}, SH, CT and VOC. The results of this part of the experiment are described in more detail in Strik & Boves (1987).

This experiment shows that subjects who had no special voice training can keep F_0, IL and P_{sb} constant during a simple utterance (a sustained vowel), but only if they are supported by visual feedback. Subjects report that keeping F_0 and IL constant requires more effort than allowing a gradual decline, and feels less natural. Without visual feedback F_0 and IL (and probably also P_{sb}) tend to fall gradually during the course of an utterance, even if subjects are instructed to keep F_0 and IL constant. The results obtained for sustained phonation will be used to support the argumentation in the next section on running speech.

8.4 Running speech

8.4.1 A physiological model of intonation

In Strik & Boves (1993) we proposed a qualitative model of F_0 control in running speech. Our model describes consistent behaviour of P_{sb}, CT, VOC and SH that was observed in the data of subjects LB and HB, and in other data presented in the literature. Figures with the average signals for the recorded utterances of subjects LB and HB can be found in Strik & Boves (1993). Here we will only display the average signals of a typical utterance (see Figure 1), in order to illustrate our model.

The four physiological signals mentioned above were chosen because it is known that they are important in the control of F_0. In our model, intonation and its physiological control take place at two levels, viz. a global and a local level. This is in accordance with other physiological models of intonation proposed in the literature (like Öhman, 1968; Collier, 1975; and Fujisaki, 1991).

Short-term variations in F_0, P_{sb}, SH, VOC and CT have often been observed (see e.g. Figure 1), i.e. all five signals clearly have a local component. But it is not immediately clear whether all of these five physiological signals also have a global component.

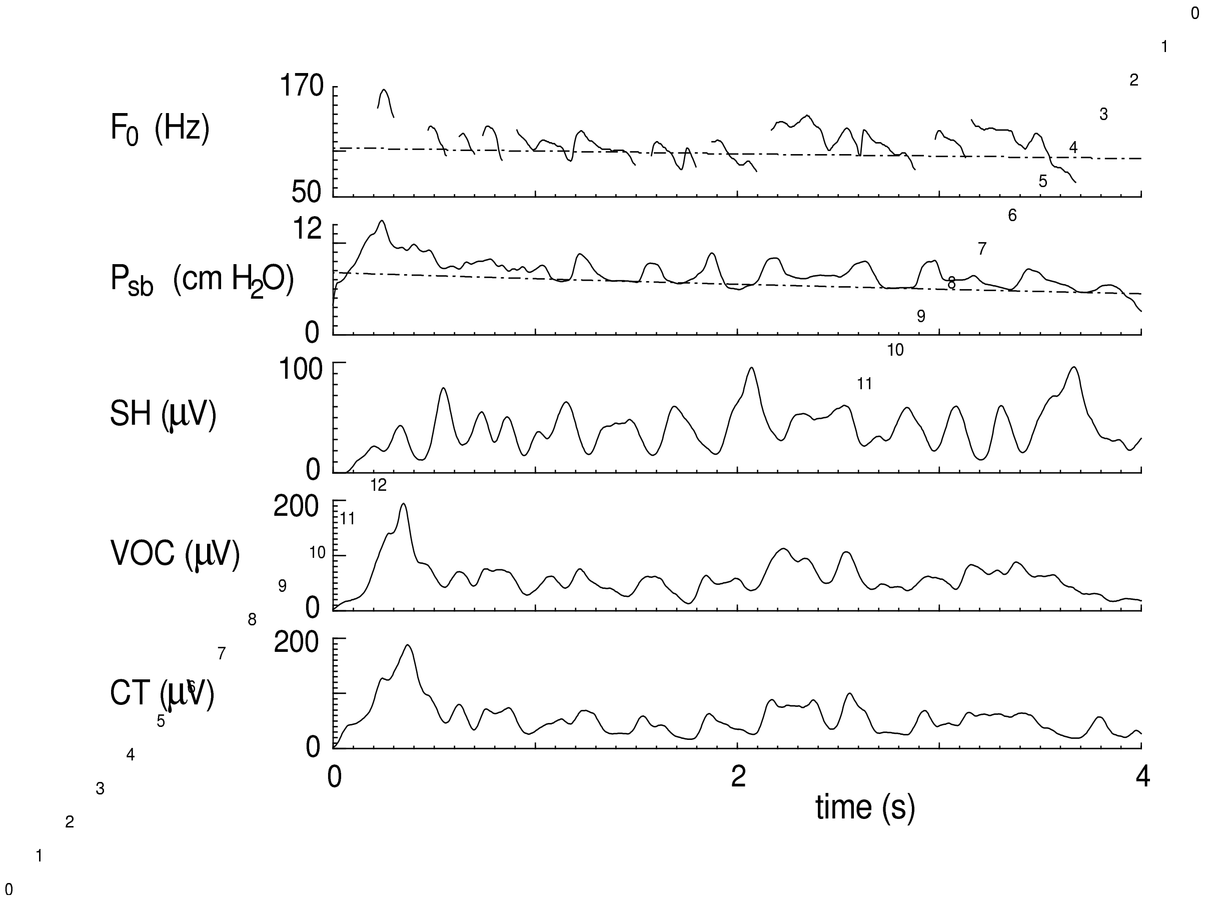


Figure 1. Average physiological signals for the Dutch utterance “Piet slikte gisteren zijn vierentwintig gele pillen liever in stilte met bier” (LU1) spoken by subject LB. Also shown in the first and second panel are the global trend lines $F_{0,g}$ and $P_{sb,g}$, respectively (dashed-dotted lines).

Global level

A gradual lowering of P_{sb} and F_0 during the course of a major syntactic constituent is often observed (see e.g. Lieberman, 1967; Ohala, 1970; Collier, 1974, 1975; Atkinson, 1978; Gelfer, 1987; Strik & Boves, 1993). The domain in which the downtrends in F_0 and P_{sb} occur has previously been given many different names, among other things "breath group" (Lieberman, 1967), "intonation group" (Breckenridge, 1977), "utterance" (Pierrehumbert & Beckman, 1988), "clause or clause complexes" (Clark & Yallop, 1990), or "major phrase" (Honda & Fujimura, 1991). In this chapter we will use the term utterance. Within the recorded sentences there were no inspirations (resets of V_l), nor any resets of F_0 or P_{sb}.

Our definition of a global component is a gradual change spanning the total duration of an utterance. Therefore, in our model P_{sb} and F_0 have a global component. The global component of F_0 and P_{sb} in our model will be called $F_{0,g}$ and $P_{sb,g}$, respectively. In this chapter the terms $F_{0,g}$ and $P_{sb,g}$ will be used for the global components of our model alone. Global components of other models will be denoted otherwise.

The model presented in Strik & Boves (1993) is a qualitative model. To illustrate our model, a possible quantitative decomposition of F_0 and P_{sb} in a global and a local component is shown in Figure 1. $P_{sb,g}$ was obtained by manually fitting an exponential function through most of the valleys of P_{sb} (Figure 1). Because it is assumed that F_0 varies linearly with P_{sb} (Titze, 1989), $F_{0,g}$ was defined in the following way: $F_{0,g} = B_0 + B_1 * P_{sb,g}$. The values of B_0 and B_1 that gave a satisfactory result for this utterance were 70 Hz and 5 Hz/cm H_2O (Figure 1), respectively. We would like to note that the manually fitted trend lines are only presented here to illustrate our qualitative model, and to give an example of a procedure that can be used to obtain the global and local components of P_{sb} and F_0. These manually fitted trend lines are not used for further analysis in the present chapter. Instead we will use a more objective statistical method in the following section.

A gradual change in the activity of SH, VOC or CT during a whole utterance was not observed in any of our recordings nor in published data of other researchers (as far as we know). Sometimes the activity of these three laryngeal muscles varied slowly during part of the utterances, but no instance of a slow increase or decrease during the whole utterance (just like P_{sb} and F_0) was found. It must therefore be concluded, both from our own data and the data presented in various other papers, that in general SH, VOC and CT do not seem to have a global component.

Local level

At the beginning of utterances CT, VOC and P_{sb} may have extra high values, and the result will be a so-called ’initial rise’ of F_0 (Figure 1). At the end of utterances SH activity often increases while P_{sb} drops sharply. If these effects occur during voiced sounds at the end of the utterance, final lowering of F_0 is observed (Figure 1). Alternatively, increased SH activity and P_{sb} release may be delayed until after the last voiced sound, in which cases final lowering is absent (e.g. in most interrogative utterances). The initial rise and final lowering of F_0 will add to the F_0 fall that results from the downtrend in $F_{0,g}$ alone (Figure 1).

The local component of P_{sb} ($P_{sb,l} = P_{sb} - P_{sb,g}$) is generally positive. SH, VOC and CT only have a local component, which is always positive because these signals can never become negative (see Section 8.2). Finally, the local component of F_0 ($F_{0,l} = F_0 - F_{0,g}$) is positive when the effect of the F_0-raising mechanisms (VOC, CT and $P_{sb,l}$) is larger than the effect of the F_0-lowering mechanisms (SH), and $F_{0,l}$ becomes negative when the net effect of F_0-raising and F_0-lowering mechanisms is negative.

Hypothesis

To conclude this section, in our physiological model of intonation SH, VOC and CT do not have a global component, while F_0 and P_{sb} do have a global component. A two-component model was chosen, because from a physiological point of view this seems to be the model that best describes the data. Because a downtrend in $F_{0,g}$ and $P_{sb,g}$ is often observed, the following hypothesis seems likely: The downtrend in $F_{0,g}$ is due to the downtrend in $P_{sb,g}$. This hypothesis has been challenged for different reasons. Two frequently adduced counterarguments are discussed in the next two sections.

8.4.2 The F_0-P_{sb} ratio

8.4.2.1 Counterargument 1

An argument used against the above-mentioned hypothesis is that the variation in $P_{sb,g}$ cannot explain the total variation in $F_{0,g}$, because the F_0-P_{sb} ratio (FPR) observed in running speech is often larger than 7 Hz/cm H_2O (e.g. Maeda, 1976; Ohala, 1978). Studies of the rate of F_0 change resulting from a change in P_{sb} alone (generally by externally induced pressure variations) have revealed that the FPR should be in the range 2-7 Hz/cm H_2O (e.g. Ladefoged, 1967; Baer, 1979). In the present chapter this range will be called the FPR-range. Because the FPR obtained for utterances often seems to exceed the FPR-range, the hypothesis is either rejected totally (Ohala, 1978), or an additional mechanism is invoked to explain (part of) the decrease in F_0 (the tracheal pull mechanism of Maeda, 1976).

Indeed, there seem to be no reasons to assume that the FPR obtained in experiments with externally induced pressure variations differs from the FPR in running speech. But the problem is that the FPR obtained for running speech depends on the way in which the downtrend in F_0 and P_{sb} is defined and modelled.

8.4.2.2 Modelling the relation between F_0 and P_{sb}

In the literature several methods have been proposed to model the downtrend in F_0, such as the difference between F_0 at the beginning and at the end of an utterance (see method 1 below), the baseline of Maeda (1976) and the bottomline and topline of Cooper & Sorensen (1981). Baseline, bottomline and topline are trend lines which are generally fitted manually, just like $P_{sb,g}$ and $F_{0,g}$ in Figure 1. Most probably, the fitting is done manually because it is difficult to define a mathematical error function that could be used to derive the trend lines with an optimization algorithm.

We have done a number of experiments to determine the parameters of the downtrend components. The results of two experiments, in which different definitions of downtrend were used, are presented below. For this aim six utterances of subject LB and six utterances of subject HB were used. For each subject there are four declarative and two interrogative utterances (see Table 1). All signals, including the F_0 signals, are average signals (Section 8.2). Figures with the average signals for these twelve utterances can be found in Strik & Boves (1993). The average signals for one utterance of subject LB are shown in Figure 1.

Method 1

In this method the F_0 and P_{sb} values are taken at two instances, one near the beginning (T_1) and one near the end (T_2). The following values are then calculated: $dF_0 = F_0(T_1) - F_0(T_2)$, $dP_{sb} = P_{sb}(T_1) - P_{sb}(T_2)$, $FPR_1 = dF_0/dP_{sb}$. The total fall in F_0 and P_{sb} from T_1 up to T_2 (dF_0 and dP_{sb}, respectively) is used to model the downtrend in F_0 and P_{sb}, respectively. Basing dF_0 on two F_0 values is error prone. In some studies the F_0 values are obtained from a trend line (e.g. the baseline in Maeda, 1976), while in other studies the F_0 values are taken from a single, representative F_0 contour (e.g. Collier, 1975; Gelfer et al., 1983; Collier, 1987). Our data processing procedure allowed us to average the F_0 curves of all repetitions of a given sentence, thus making the estimation procedure more reliable. In previous studies various choices of T_1 and T_2 have been made, based on different motives (see e.g. Gelfer et al., 1983). In this study T_1 is the first voiced frame, and T_2 the last voiced frame of each utterance. These instants of T_1 and T_2 were mainly chosen because the values of F_0 and P_{sb} at these time-points can be determined very easily for each utterance. Given this choice of T_1 and T_2, all relevant values were calculated for the twelve utterances of subjects LB and HB (see Table 1).

Table 1. Listed from top to bottom are: utterance type, number of voiced samples (N), length of the utterance ($T = T_2 - T_1$) in s, F_0 values of first ($F_0(T_1)$) and last ($F_0(T_2)$) voiced sample in Hz, total fall of F_0 ($dF_0 = F_0(T_1) - F_0(T_2)$) in Hz, average rate of change of F_0 (dF_0/T) in Hz/s, P_{sb} values for first ($P_{sb}(T_1)$) and last ($P_{sb}(T_2)$) voiced sample in cm H_2O, total fall of P_{sb} ($dP_{sb} = P_{sb}(T_1) - P_{sb}(T_2)$) in cm H_2O, average rate of change of P_{sb} (dP_{sb}/T) in cm H_2O/s, $FPR_1 = dF_0/dP_{sb}$ in Hz/cm H_2O and the regression coefficient between F_0 and P_{sb} (FPR_2) in a multiple regression equation, also in Hz/cm H_2O (for explanations, see also the text).

	subject LB						subject HB					
	declarative utterances				questions		declarative utterances				questions	
utt	SU1	SU2	LU1	LU2	SU3	LU3	SU1	SU2	SU3	SWC	SU4	LU4
N	234	226	558	524	222	490	314	342	288	680	260	435
T	1.42	1.41	3.46	3.40	1.31	3.18	1.66	1.78	1.54	3.62	1.39	2.40
$F_0(T_1)$	150	136	147	136	121	138	119	113	121	132	118	114
$F_0(T_2)$	65	67	66	79	167	169	102	106	102	104	200	188
dF_0	85	69	81	57	-46	-31	17	7	19	28	-82	-74
dF_0/T	60.1	49.1	23.4	16.7	-35.2	-9.7	10.2	3.9	12.3	7.7	-59.0	-30.8
$P_{sb}(T_1)$	9.58	9.92	11.64	11.82	8.44	10.95	6.13	6.47	6.29	5.86	5.83	6.04
$P_{sb}(T_2)$	3.44	3.50	4.82	4.57	4.36	5.10	2.33	1.64	1.42	1.77	4.10	3.96
dP_{sb}	6.14	6.42	6.82	7.25	4.08	5.85	3.80	4.83	4.87	4.09	1.73	2.08
dP_{sb}/T	4.34	4.57	1.98	2.13	3.12	1.84	2.29	2.71	3.16	1.13	1.24	0.87
FPR_1	13.9	10.8	11.9	7.87	-11.3	-5.30	4.47	1.45	3.90	6.84	-47.5	-35.5
FPR_2	3.97	7.63	2.30	4.58	6.48	4.42	3.20	3.02	4.79	3.78	6.25	4.12

In all utterances dP_{sb} is positive (Table 1). For subject LB dP_{sb} is always larger than for subject HB. For both subjects dP_{sb} for the interrogative utterances is smaller than dP_{sb} for the declarative utterances. At the end of each question there is a marked increase in F_0, and consequently dF_0 is negative for the questions. But for all declarative utterances dF_0 is positive. For the declarative utterances, dF_0 of subject LB is always larger than dF_0 of subject HB. Partly this can be explained by the larger dP_{sb} for subject LB, as noted above. In addition, for subject LB the CT and VOC often show increased activity at the beginning of an utterance, which causes an initial rise in F_0, and the SH is increased at the end of the utterance during the final lowering of F_0. Both effects will cause dF_0 to be larger than the fall in F_0 resulting from dP_{sb} alone, i.e. both P_{sb} and the laryngeal muscles participate in dF_0.

The values of FPR_1 can be seen in Table 1. Only three of the twelve FPR_1 values are within the accepted FPR-range. FPR_1 for the four questions is negative because dF_0 is negative, four of the eight values of FPR_1 for the statements are larger than 7 cm H_2O and one is smaller than 2 cm H_2O. Based on these FPR_1 values one could conclude that the downtrend in P_{sb} cannot explain all the downtrend in F_0, and thus other factors should contribute to the downtrend in F_0. If downtrend is defined in this way, then this conclusion is correct. After all, dF_0 does depend on both dP_{sb} and the activity of the laryngeal muscles (especially for subject LB, as explained above).

The FPR-range is obtained from experiments with externally induced pressure variations (e.g. Ladefoged, 1967; Baer, 1979). The goal of these experiments was to determine the FPR for F_0 changes that result from P_{sb} changes alone, i.e. one tried to keep other processes that influence F_0 (like the laryngeal muscles) constant (see e.g. Baer, 1979). In these studies the points in a scatterplot for F_0 as a function of P_{sb} could usually be fitted reasonably by a straight line. In Figure 2 an F_0-P_{sb} scatterplot is given for a short utterance of subject LB. Clearly, in this scatterplot the points are not grouped around a straight line. The reason is that during this utterance the other factors which influence F_0 are not constant. Drawn in Figure 2 is the straight line that connects the first and the last voiced frame. FPR_1 is the slope of this line. In Figure 2 one can see that the FPR obtained in this way depends heavily on the exact choice of T_1 and T_2. To sum up, method 1 has two important drawbacks:

1. other factors that can affect F_0 are not constant over the course of an utterance; and
2. because the other factors are not constant it is hazardous to make estimates of the FPR which are based on the values of F_0 and P_{sb} at two instants only.

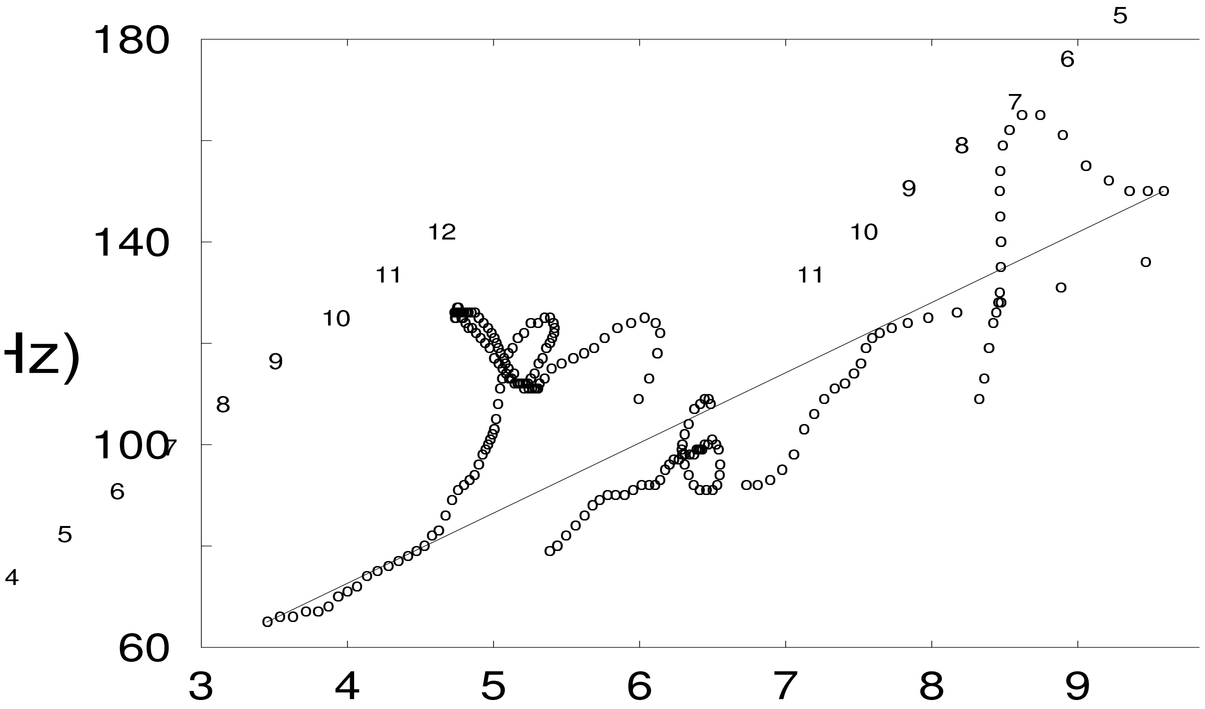


Figure 2. F_0 as a function of P_{sb} for the Dutch utterance “Piet slikte zijn pillen met bier” (SU1) spoken by subject LB. The straight line is the line connecting the first and the last voiced frame. FPR_1 is the slope of this line.

Method 2

In method 2 a multiple regression analysis is used, in which F_0 is the criterion and P_{sb}, VOC and SH are the predictors [1]. The outcome of the regression analysis are the coefficients A_i of the regression equation: $F_0 = A_0 + A_1{*}P_{sb} + A_2{*}VOC + A_3{*}SH$. The FPR is the regression coefficient between F_0 and P_{sb}: $FPR_2 = A_1$. This method does not have the drawbacks of method 1 because a correction is made for some important other factors which influence F_0, and the regression coefficient is based on the data of all voiced frames.

[1] For subject LB the correlation between CT and F_0 is generally larger than the correlation between VOC and F_0, and thus CT is a better predictor of F_0. But because the behaviour of CT and VOC is almost identical for subject LB, and because the activity of the CT was not measured for subject HB, we have chosen the VOC as a predictor in the regression analysis for both subjects.

The multiple regression analysis decomposes F_0 into four components: A_0, A_1*P_{sb}, A_2*VOC and A_3*SH. The first component is the constant A_0. VOC and SH do not have a global component either (Section 8.4.1), and thus in this statistical model the downtrend in F_0 is due to the downtrend in P_{sb} alone. This is in line with the physiological model presented in Section 8.4.1, except for one essential difference. In method 2 P_{sb} is not decomposed into a global and a local component. However, because there are no reasons to assume that the FPR is different on a global and a local level, this does not seem to be a problem. Consequently, the P_{sb} component in the regression analysis (A_1*P_{sb}) contains both the slow downtrend in F_0, and the part of the local variations in F_0 which is due to the local variations in P_{sb}. The other part of the local variations in F_0 is in the VOC and SH components (A_2*VOC and A_3*SH), respectively.

Instead of using the multiple regression analysis we could have based our estimates of the FPR on the global trend lines $P_{sb,g}$ and $F_{0,g}$. To that end, $P_{sb,g}$ and $F_{0,g}$ should have been determined in the way described in Section 8.4.1, i.e. by making manual fits for all utterances. This is certainly possible, but we prefer to use objective, statistical methods (like the multiple regression analysis described in the current section) instead of more subjective methods in which trend lines are fitted manually.

For all voiced frames of the twelve utterances a multiple regression analysis was performed in which F_0 was the criterion and P_{sb}, VOC and SH were the predictors. The resulting FPR_2 values (i.e. the A_1 values) can be seen in Table 1. The resulting values of A_0, A_2 and A_3 were not used for further analysis. Of the 12 FPR_2 values, 11 are in the FPR-range, and one is slightly larger than the maximum of the FPR-range. If the CT had been used as a predictor instead of the VOC for subject LB, then FPR_2 would have been 6.44 Hz/cm H_2O for this utterance, and thus it would have been within the FPR-range (see footnote on previous page). Also for the interrogative utterances FPR_2 is always within the FPR-range, while this was never the case for FPR_1. The rise of F_0 at the end of questions is usually due to an increase of CT, VOC and P_{sb}. In method 2 a correction is made for the increase in VOC, and the result is that the FPR_2 is within the FPR-range. The rapid increase in P_{sb} at the end of the questions is part of P_{sb}, and will also explain part of the end rise in F_0.

To conclude this section, comparison of FPR_1 and FPR_2 values for sentences has shown that the actual values obtained are crucially dependent on the way in which the F_0-P_{sb} ratio is defined. In our opinion FPR_1, which has been used to refute the above-mentioned hypothesis, is not a fair measure because it isolates P_{sb}, but at the same time ignores all other factors affecting F_0. If some important additional influences are factored out of F_0 by means of a multiple regression analysis, as is done with FPR_2, a completely different picture emerges, which is compatible with the hypothesis that the downtrend in P_{sb} explains the downtrend in F_0. Even though the way in which the influence of the laryngeal muscles on F_0 is modelled is extremely crude (the true relation between the activity of the laryngeal muscles and F_0 is very likely to be non-linear) FPR_2 is a much fairer measure than FPR_1. According to this measure the variation in P_{sb} can explain all the variation in F_0, and no additional mechanisms are necessary. Therefore, too large a total F_0 drop does not seem a reason to reject the hypothesis. Also, and perhaps even more important, arguments about the relation between F_0 and P_{sb} depend fully on the way in which the two downtrends are modelled. As long as the model of F_0 downtrend does not partition out effects not related to P_{sb}, it may remain a valid definition of its own, but it should no longer be used in arguments involving P_{sb}.

8.4.3 Control of F_0 and P_{sb}

8.4.3.1 Counterargument 2

At the basis of the second counterargument is the idea that the laryngeal muscles can be controlled linguistically, while this is not possible for the respiratory muscles and thus the downtrend in P_{sb} is a passive process. Subsequently, this idea is used as an argument against the above-stated hypothesis: because the downtrend in F_0 is (at least partially) linguistically controlled it cannot result from an automatic process like the downtrend in P_{sb}. The fact that some authors use this argument in the discussion about the physiological causes of declination was also noted by Cohen et al. (1982).

The second argument against the hypothesis is expressed most clearly by Breckenridge (1977). She states that declination is part of the linguistic system, and therefore it must be controlled by the laryngeal muscles just as other linguistically significant aspects of F_0 are. A similar line of reasoning is followed by Ohala (1978, 1990). In Ohala (1978, 1990) three possible causes for declination are mentioned: (1) tracheal pull (Maeda, 1976); (2) downtrend in P_{sb} (Collier, 1974, 1975); and (3) graded activity in the laryngeal muscles. According to Ohala the first two causes are automatic, non-purposive physiological causes. Because declination is not automatic but controlled, he argues that a model in which linguistic aspects of F_0 are completely determined by actions of the laryngeal muscles is much more likely than a two-component model in which respiratory and laryngeal factors interact.

Clear opinions about the control of the downtrend in P_{sb} can also be found in Gelfer et al. (1983), Ladd (1984) and 't Hart et al. (1990). Gelfer et al. (1983) studied whether declination is actively controlled. They noted a similar downtrend in F_0 and P_{sb}. They argue that if the declination in F_0 is due to the declination in P_{sb}, then this would suggest that declination is a passive phenomenon. In Ladd (1984) three physiological causes of declination are discussed: (1) the downtrend in P_{sb} (Collier, 1975); (2) the tracheal pull (Maeda, 1976); and (3) F_0 rises are harder to produce than F_0 falls (Ohala & Ewan, 1973). According to Ladd, the downtrend in P_{sb} and the tracheal pull are automatic mechanisms. Finally, according to 't Hart et al. (1990) the muscular activity involved in the regulation of V_l and P_{sb} is subject to an automatic control system. In their view declination should be seen mainly as an automatic by-product of respiration.

The examples given above clearly illustrate that there seems to be a widespread notion that the downtrend in P_{sb} is an automatic process. If the downtrend in P_{sb} is a completely passive process, then this could indeed be used as a counterargument against the above-mentioned hypothesis, because there are many indications that declination is under linguistic control, at least to some extent. However, it is not sure that the downtrend in P_{sb} is a passive mechanism. On the contrary, there are many reasons to believe that P_{sb} is controlled. This will be discussed in the next section.

8.4.3.2 Respiratory system

There are three factors which may affect P_{sb} (see e.g. Ladefoged, 1967):
① passive forces, like elastic recoil and gravitational forces;
② active forces, resulting from contractions of respiratory muscles; and
③ the resistance to the air-stream, both at the glottis and in the vocal tract (Z_g).

The pressure that results from passive forces alone is generally called the relaxation pressure (P_{rel}), while the pressure change brought about by active muscle contractions is called the muscular pressure. For a speaker who remains in the same position (usually upright) the gravitational forces are roughly constant and thus P_{rel} would depend on V_l alone. If expiration during speech production were a truly passive process, then the muscular pressure should be zero and P_{sb} should be a function of V_l and Z_g alone. Several observations reveal that this is not the case:

❒ Our data show that for repetitions of the same sentence the amount of inspiration before the utterance was not always the same. Consequently, the V_l traces run essentially parallel (see e.g. Figure 3), while Z_g can be assumed to be reasonably constant. Although the differences in V_l are large, the P_{sb} contours are very much alike (Figure 3).

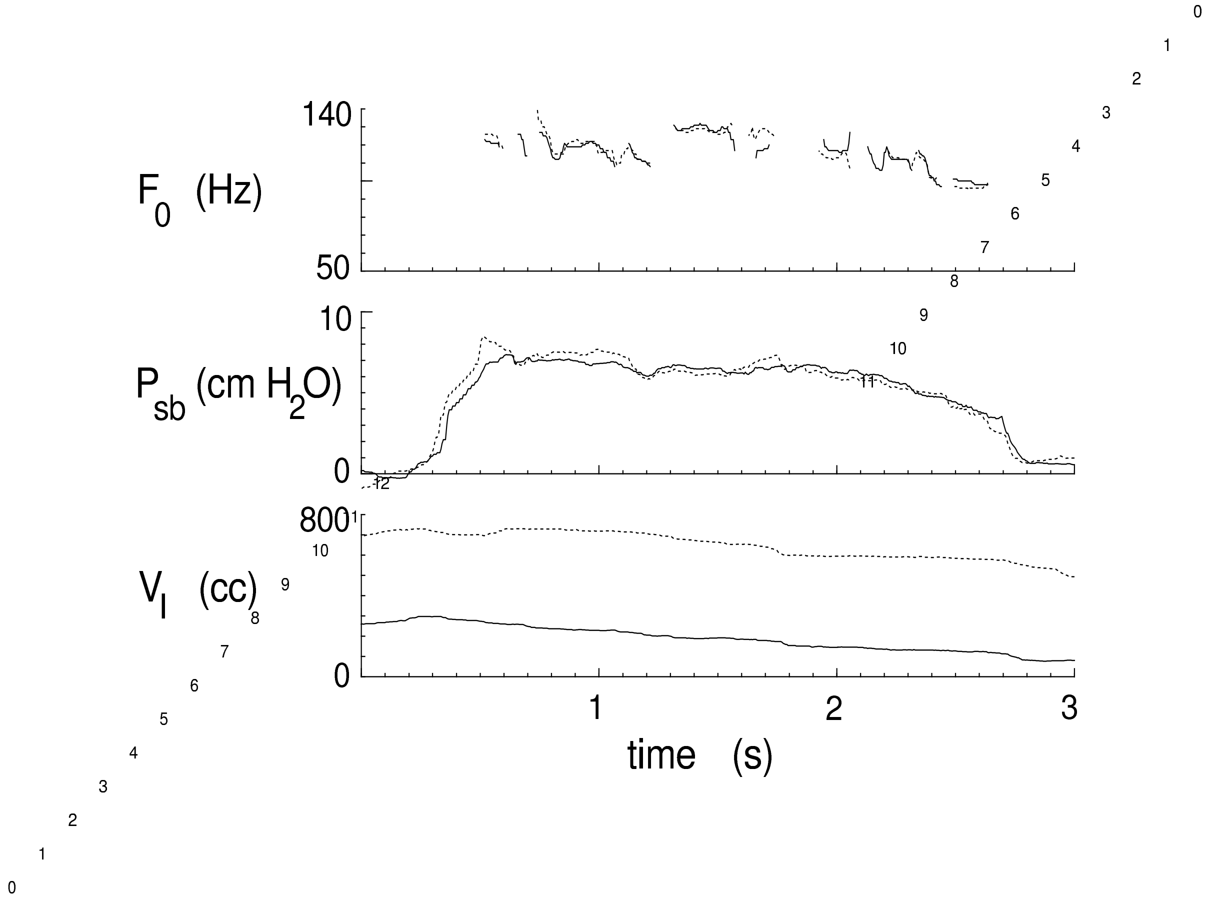


Figure 3. F_0, P_{sb} and V_l signals for two repetitions of a spontaneous sentence spoken by subject HB. The average difference for V_l is 470 cc, and for P_{sb} it is 0.05 cm H_2O.

❒ Some of the sentences were also produced in reiterant form, using either the syllable /fi/ or /vi/. The slopes of the V_l traces of these two types of utterances are different, but also in this case the P_{sb} contours showed much resemblance (see e.g. Figure 4). This was also found by Gelfer (1987).

Figure 4. Average F_0, P_{sb} and V_l signals for two utterances produced with reiterant speech: /vi/ (dashed) and /fi/ (solid).

❒ Speakers can keep their P_{sb} constant during the production of a long sequence of /ma/ syllables (Collier, 1987), and during sustained phonation (Section 8.3). In both cases the activity of the measured laryngeal muscles also remained constant, so Z_g was probably constant. The fact that speakers can keep P_{sb} constant while V_l is decreasing also proves that P_{sb} is not simply a function of V_l and Z_g alone.

❒ During phonation P_{sb} should not become smaller than a threshold value below which phonation is not possible (the so-called phonation threshold pressure, see Titze, 1992). Furthermore, the loudness of the speech is determined to a large extent by P_{sb}, and thus P_{sb} should be kept within a certain range to produce speech with the desired loudness. After inspiration at the beginning of an utterance P_{rel} is often larger than the desired P_{sb}, while at the end of an utterance P_{rel} is often lower than the desired P_{sb} (see e.g. Ladefoged, 1967). If the respiratory muscles were not used, then

P_{sb} and the loudness would decrease rapidly; soon P_{sb} would be smaller than the phonation threshold pressure and phonation would stop. To prevent this, the inspiratory muscles are used at the beginning of an utterance to keep P_{sb} lower than P_{rel}, while expiratory muscles are used when P_{rel} is lower than the desired P_{sb} (Ladefoged, 1967).

The arguments given above force one to assume that the respiratory muscles are used to control P_{sb} during speech production. The following question then arises: How are the respiratory muscles used to control P_{sb}? According to Ladefoged (1967) and Ohala (1990) the amount of control is limited, i.e. they claim that these muscles are only used to keep P_{sb} reasonably constant above some minimal level. However, many measurements show that in general P_{sb} is not constant, but has a tendency to decline, both in sustained phonation (Section 8.3) and in running speech (Lieberman, 1967; Ohala, 1970; Collier, 1974, 1975; Atkinson, 1978; Gelfer, 1987; Strik & Boves, 1993). Furthermore, P_{sb} contours for repetitions of a sentence appear to be very similar in shape as well as in amplitude (see e.g. Figure 3), too similar to assume that P_{sb} has just a convenient (more or less random) value above its minimum.

If the respiratory muscles are under voluntary control, then they can be used to control P_{sb} during speech production. Active control of the respiratory muscles and P_{sb} in speech production seems likely, given the following arguments:

❒ The way the respiratory muscles are used during speech production differs from the way they are used in normal breathing. In normal breathing the duration of inhalations and exhalations is about equal, while in speech production the inspiratory phase is much shorter. Furthermore, it has been observed that the posturing of the respiratory system for speech production (the prephonatory posturing of the chest wall) is different from the posturing for normal breathing (Hixon, Goldman & Mead, 1973; Baken et al., 1979; Baken & Cavallo, 1981).

❒ Breathing pauses occur mainly at major constituent breaks (Winkworth et al., 1994). Breathing pauses can also occur at minor constituent boundaries, but as speaking rate increases they are eliminated from these minor breaks (Grosjean & Collins, 1978). Grosjean & Collins (1978) conclude that "it would appear that breathing in speech depends to a large extent on the speaker’s preplanned pause patterns", and thus breathing would be linguistically controlled.

❒ The amount of air inspired and the V_l at the beginning of sentences was found to be significantly larger for longer utterances compared to shorter ones, and for major syntactic breaks compared to more minor ones (Winkworth et al., 1994). According to Winkworth et al. (1994) these findings indicate that speakers pre-plan their V_l and the volume inspired. It should be noted that this study concerned reading, and therefore their results suggest that the respiratory muscles are under linguistic control during reading.

❒ Indications of extra respiratory activity (i.e. increased lung volume decrement) for stressed syllables were found by Ohala (1977), while Ladefoged (1967) and Van Katwijk (1974) actually measured increased activity of respiratory muscles for stressed syllables. Although not all stressed syllables are probably accompanied by extra activity of the respiratory muscles, these results indicate that linguistic control of the respiratory muscles is possible, at least at a local level. If active control of the respiratory muscles is possible at a local level, then it is likely that it is also possible at a global level.

❒ Loudness is a prosodic, i.e. a linguistic variable. If speakers are asked to increase loudness, they tend to initiate speech at higher lung volumes (Hixon et al., 1973). Winkworth et al. (1994) also found that louder utterances within the "comfortable loudness" range are generally associated with higher lung volumes. According to Weismer (1985) it is more efficient to start at higher lung volumes for loud speech, because larger values of P_{sb} are needed to generate loud speech. So, not only is this an example of linguistic control of the respiratory muscles, it is also an indirect indication of linguistic control of P_{sb}. But there are also more direct indications of voluntary control of P_{sb}.

❒ In addition to P_{sb}, a speaker can use many different physiological mechanisms to control F_0, and thus a given F_0 contour could be produced in various ways. Still, the amount of variation between physiological signals (including P_{sb}) of repetitions of the same utterance is relatively small (Strik & Boves, 1991a; Strik & Boves, 1993). The finding that the inter-repetition variation in P_{sb} and the other physiological signals is small suggests that speakers have a notion of the manner in which they want to produce an utterance, and that they have a good control over P_{sb} and the other mechanisms.

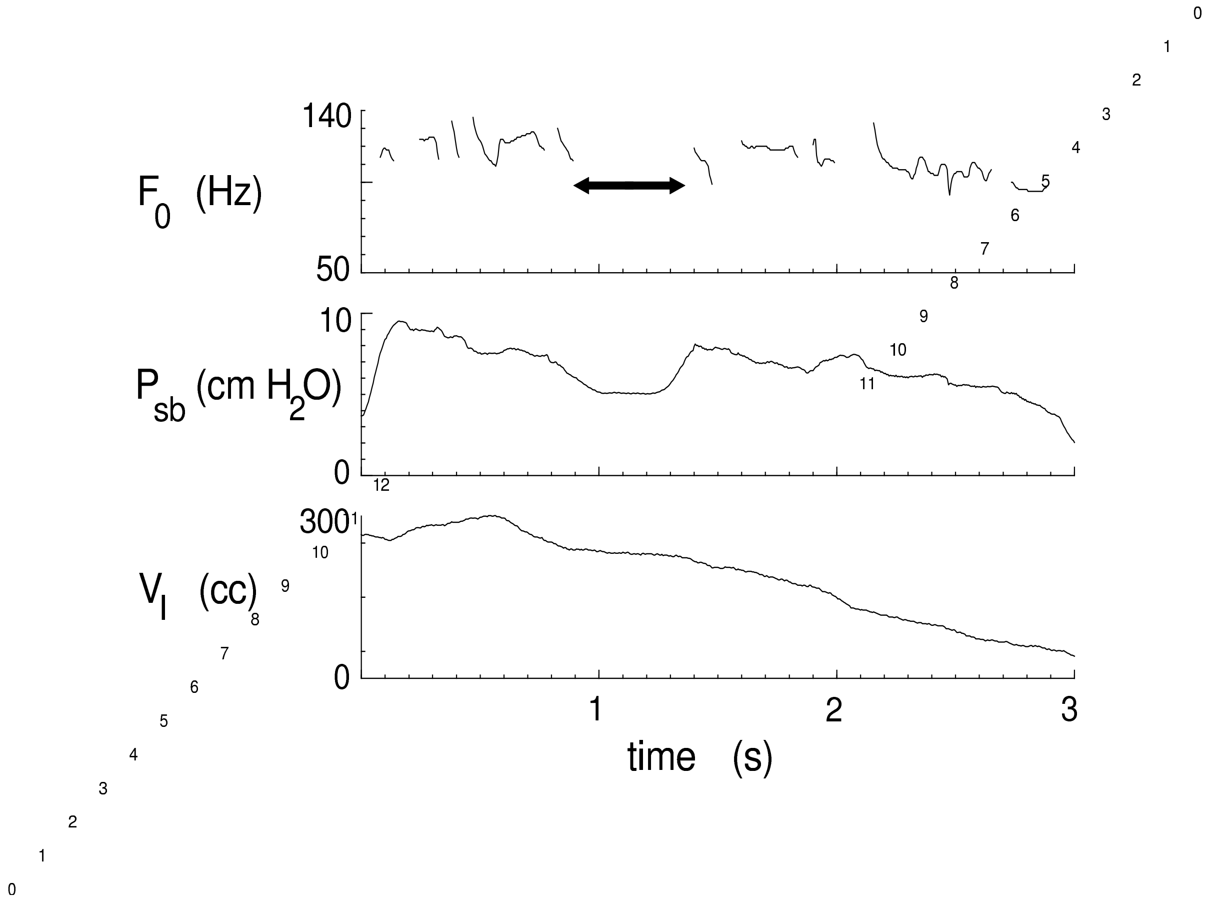


Figure 5. F_0, P_{sb} and V_l signals for a spontaneous utterance spoken by subject HB. The arrow marks the interruption of about 0.5 sec.

❐ Another indication that P_{sb} is actively controlled can be seen in Figure 5. In the middle of a spontaneous utterance subject HB made a swallowing gesture, probably because the pressure catheter was bothering him. During this interruption P_{sb} suddenly drops to about 5 cm H_2O. For subject HB phonation with such a level of P_{sb} is possible, because comparable and even lower values of P_{sb} were found at the beginning of many voiced intervals of the repetitions of the same utterance. If the subject's only intention was to provide a P_{sb} above some minimal level at which phonation is possible, he could have kept P_{sb} at approximately 5 cm H_2O. However, before he resumed phonation, P_{sb} was raised to approximately the value it had before the interruption, and from that point it started declining again.

❐ Finally, after the two subjects in our study had received instructions they were able to keep P_{sb} fairly constant at different levels (Section 8.3), i.e. their P_{sb} was under voluntary control.

The conclusion of this section is that there are several reasons to believe that the respiratory muscles and P_{sb} are actively controlled. If this is the case, then also the second counterargument (specified above) cannot be used to refute the hypothesis that the lowering of $F_{0,g}$ is generally due to a decrease in $P_{sb,g}$.

8.5 Discussion

In this chapter we have argued in favour of a major role for P_{sb} in the control of the ubiquitous downtrend in F_0 contours. The role of P_{sb} has been called into question by a number of authors, and for a number of different reasons. The two most important counterarguments centre around the claim that the total F_0 fall in most published data seems to exceed the range that should be expected from the fall in P_{sb}, and the claim that the respiratory system is not suited for so precise a control as needed for the linguistic, communicative purpose served by F_0 downtrend. These counterarguments have been discussed in Sections 8.4.2 and 8.4.3, respectively.

Before proceeding to a summary of these discussions we would like to address one additional argument. Ohala (1990) claims that there are examples in the literature that show a gradual downtrend of the activity of CT. It appears that these examples are limited to the contours 11 and 15 in Collier (1974). In these registrations a gradual decline of CT activity can indeed be seen, but only in the second half of the utterances. To the best of our knowledge there are no data showing a gradual variation of CT, VOC or SH over complete utterances. But there are numerous examples of P_{sb} decline that span a complete utterance. Thus, we fully acknowledge the possibility that laryngeal muscles contribute to the total fall of F_0 over the course of an utterance, but the available data more or less force us to accept the conclusion that the contribution of P_{sb} to the control of F_0 downtrend (as the concept is defined in our model) is much more important. For this reason, we think that the physiological validity of the models proposed by Öhman (1968) and Fujisaki (1991), which do not acknowledge a role for P_{sb}, is debatable. Speakers can exploit a large array of physiological means to reach a certain goal, and it would be surprising if some of these means would never be exploited. After all, there is no valid reason to suppose that all subjects should always behave in exactly the same way. But individual examples attesting a possible way of control should not be generalized. For the time being, the data speak in favour of P_{sb}.

Coming back to the arguments related to the F_0-P_{sb} ratio, it must be concluded that fair estimates of that ratio are extremely difficult to obtain from sentence material. In all naturally produced utterances laryngeal muscles affect F_0 in addition to P_{sb}. In order to obtain a fair estimate of FPR these additional contributions must be factored out. That is certainly not done by defining dF_0 and dP_{sb} as the difference between the values observed at the beginning and at the end of an utterance, not even when these values are averaged over a large number of tokens, simply because the F_0 values are affected by laryngeal muscle activity.

A fundamental problem in studying the physiological causes of downtrend is that the literature abounds with definitions of F_0 downtrend. Downtrend, declination or downdrift have been used to denote the tendency of F_0 to decrease during the course of an utterance. This qualitative definition can be interpreted in many different ways, and is hardly suitable for studying the relation between physiology and F_0 downtrend. Therefore, a more precise definition of downtrend is needed. Some of the definitions used in the literature are illustrated in Figure 6. Figure 6 shows hand-fitted estimates of a top line, a bottom line, a line connecting the first and last voiced sample in addition to $F_{0,g}$, which was derived in the way described in Section 8.4.1 (this is the same trend line as the one shown in Figure 1). It can easily be seen that the slopes of these lines differ considerably. There is less literature on the definition of downtrend in P_{sb}. Yet, it is clear that the existence of several essentially different definitions or models of F_0 downtrend makes it impossible to discuss 'the' relation between downtrend in P_{sb} and F_0: the outcome of such a discussion is certain to depend on the exact definition of downtrend that is assumed.

According to our definition of a global component, F_0 and P_{sb} do have a global component while CT, VOC and SH generally do not have a global component. The quantitative statistical analysis has shown that, after correcting for the influence of VOC and SH, the variation in P_{sb} can explain all the variation in F_0 (i.e. the FPR is usually within the correct range). Consequently, in our physiological two-component model the downtrend in F_0 can be explained completely by the downtrend in P_{sb}. However, it is always possible that other (unknown) factors also contribute to the downtrend in F_0. That is a possibility which cannot be ruled out.

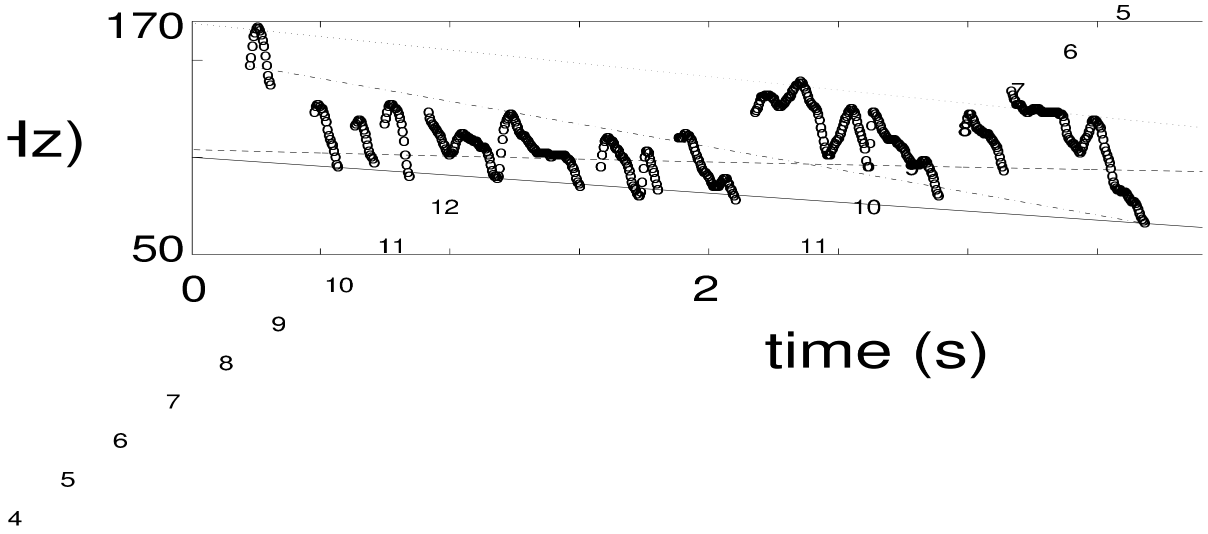


Figure 6. Average F_0 signal and trend lines for utterance LU1 spoken by subject LB (The average F_0 signal is the same signal as in the upper panel of Figure 1). The following trend lines are shown: $F_{0,g}$ (dashed), the line connecting the first and the last voiced frame (dashed-dotted), topline (dotted) and bottomline or baseline (solid).

This physiological two-component model was chosen because it seems to be the model which best describes the physiological data. If, for some reasons, someone prefers another definition of the global component, like for instance the top- or bottomline in Figure 6, the conclusion should indeed be that the downtrend in F_0 cannot be determined entirely by the downtrend in P_{sb}, because top- and bottomline are determined to a large extent by the activity of the laryngeal muscles.

To sum up, in our model the downtrend in F_0 could be entirely due to the downtrend in P_{sb}. For other definitions of downtrend this does not have to be the case, i.e. these downtrend could be determined partially by the activity of the laryngeal muscles. However, the downtrend in P_{sb} will always explain part of the downtrend in F_0.

Ideally, trend lines should not be determined by means of hand fitting, but rather by means of formal, mathematical procedures. However, each and every mathematical fit procedure requires the definition of an error (or cost) function, to quantify the discrepancy between the observed data and the model curve. For the time being, such an error function is almost impossible to define, because it is not possible to reach agreement on the weight of details in the deviations. To a considerable extent, these weights depend on one's theoretical opinions about which details in F_0 curves are linguistically relevant and which are not. Another factor complicating the construction of a completely quantitative model of the control of F_0 in running speech has to do with the lack of knowledge about the relation between EMG activity of the laryngeal muscles and elastic properties of laryngeal tissue. In our own models we have assumed a simple linear relationship, but that is not more than a very crude first approximation. Thus, we have to be content with models that contain non-quantitative or non-realistic quantitative components for some time to come.

8.6 Conclusions

In this chapter we have investigated the relation between downtrend in F_0 and P_{sb}, an issue that has been undecided despite considerable discussion in the recent literature. The most important conclusion of our own experiments and a detailed analysis of data published in the literature is that the issue is genuinely not decidable, unless there is agreement about the way in which downtrend in F_0 and P_{sb} are defined. In our model of F_0 control presented in this chapter we take the view that F_0 and P_{sb} both have a global component, and that these components must be related. Other models or definitions of F_0 downtrend, like a line fitted through the F_0 peaks (the topline), include effects of other factors affecting F_0 besides P_{sb}; therefore, these definitions (or models) of F_0 downtrend do not allow a direct link with downtrend in P_{sb}. Also, we have presented data and arguments from our own experiments and from the literature in favour of a tight and precise control of P_{sb} and the underlying respiratory system. Therefore, the phonetic implementation component of any intonation model should include a role for P_{sb}.

Chapter 9

Epilogue

9.1 Introduction

Our research focused on the relations between (1) physiological measurements in the respiratory and phonatory systems, (2) prosodic parameters (F_0, IL and SC) and (3) intonation. To this end recordings were made of the acoustic speech signal, electroglottogram, lung volume and some physiological mechanisms that are known to be important in the control of prosody such as P_{sb}, P_{or}, P_{tr}, CT, VOC and SH. These recordings were made for three subjects, while for a fourth subject only the speech signal, electroglottogram and lung volume were recorded. The papers collected in this book are based on data of two of these subjects, namely subjects LB and HB. The data of the other two subjects were used in presentations and publications which are not included in the present book (e.g. Strik & Boves, 1988a; Strik & Boves, 1989).

The fact that few subjects were used for the current research may be surprising, but is easy to explain. Invasive measurements of the type made for this research are impossible with large groups of subjects. And even if the raw signals can be recorded, processing of data remains extremely time consuming, despite all the work done to automate all procedures to the largest extent possible.

The results of our research have been presented in the previous chapters. In this chapter we will review the main findings in Section 9.3. In Section 9.4 some remarks will be made on possible further research on the relation between physiological signals and prosody. However, first we will discuss some methodological aspects that were important for the current research.

9.2 Methodological aspects

In our research we considered using a voice source - vocal tract model to study the relation between physiological signals and prosodic parameters, and an intonation model to study the relation between physiological signals and intonation. In this section both types of models are discussed.

9.2.1 Voice source - vocal tract model

Part of our research concerned the relation between the physiological signals measured and the prosodic parameters of the speech signal. For realistic modelling of this relation a voice source - vocal tract model is a prerequisite. This should be a kind of articulatory synthesizer, in which not only the vocal tract but also the voice source is modelled in a physiologically meaningful way. For a given set of physiological signals this model should predict the values of the prosodic parameters, or even better, the speech signal itself. Clearly this is a futuristic model, which does not exist at the moment.

It is questionable whether a useful model can be designed without prior breakthroughs in our knowledge and understanding of phonation. First of all, the number of parameters in such a model would be enormous. Most likely, a substantial part of these parameters cannot be measured for a living human subject. For instance, it will be very difficult to measure the anatomical and physiological properties of the subject, like the length, thickness and stiffness of the vocal folds, and the way in which these properties are influenced by the length and tension of various laryngeal muscles.

Moreover, it is extremely difficult to quantify the relation between the length and tension of a muscle, and the measured EMG signal for that muscle. In our research the EMG signals were recorded with hooked-wire electrodes that were positioned in the laryngeal muscles. In this case the recorded EMG signal depends on the exact position and the impedance of the electrode, or in other words, it depends on which and how many motor units are recorded. Getting the electrode into the correct laryngeal muscle is already very difficult, but finding out where exactly the electrode is in the muscle would be even more difficult.

Consequently, an F_0-EMG ratio, say e.g. an F_0-CT ratio, can only be calibrated within one experiment, if one assumes that everything remains constant during the experiment. To calibrate this ratio, the activation of the specific muscle should be varied (probably externally) while a person is phonating and all other things remain constant. Because a positive correlation between F_0 and CT is often observed, and because CT can be used to lengthen the vocal folds (see Section 1.2.2), it is most likely that there is a causal relation. But, as explained above, it is difficult to calibrate this relation, and this is not done in general. Consequently, if the F_0-CT ratio is not calibrated, one can never find out which part of the F_0 change can be explained by the change in the activity of the CT. One can only conclude that an increase in CT is often observed when F_0 is raised.

To conclude this section, a voice source - vocal tract model which has the physiological signals as its input and F_0, IL and SC as its output is extremely complex, most probably non-linear and comprises a large number of parameters. At present such a model does not exist, and therefore realistic quantitative modelling of the relation between the physiological signals measured in the present research and the prosodic parameters is not possible.

9.2.2 Intonation model

To study the relation between the physiological signals and intonation, an intonation model could have been used. Our reason to use an intonation model is that it would allow a quantitative modelling of the relation between F_0 and intonation, or more precisely, that it could transform an F_0 contour into a number of discrete units. Depending on the model used, these units could be phrase and accent commands in Fujisaki's type 2A model (1991), a declination line and F_0 movements in the type 2A model of 't Hart et al., (1990), or tones in type B models (see Section 1.2.1). Subsequently, the relation between these discrete units and physiological signals could be investigated. It is obvious that if the discrete units are not correct (i.e. if the way in which the F_0 contour is modelled by the intonation model is not correct), then the results of a study of the relation between these discrete units and the physiological signals are questionable. Therefore, great care should be taken to select a suitable intonation model.

Various intonation models have been proposed in the literature (see Section 1.2.1), inter alia the models mentioned above, but no consensus has been reached yet on what the most appropriate one is. In order to find out which model fits the data best, the different intonation models should be compared. It turned out, though, that such comparisons are not straightforward. First of all, all models discussed in Section 1.2.1 require that some kind of intonation transcription be carried out first. In more mundane terms: human intervention is needed to determine how many and which units are present in an utterance. For each intonation model this transcription will consist of the discrete units acknowledged by that model. Even within a given "preferred" model this is not without problems, as it is well-known that making an intonation transcription is skilled work, and that differences are often observed between intonation transcriptions made by different researchers.

Moreover, comparisons are difficult to make because the various model types do not model the same thing. Specifically, type B models model F_0 targets, most type A models model the actual F_0 contour, and the type 2A model of 't Hart et al., (1990) models a stylized version of the F_0 contour that is perceptually equivalent to the original. Because they model different things it will be difficult, if not impossible, to define one measure with which all model types can be evaluated.

Therefore, we had to abandon the idea of comparing the models available in the literature in order to choose the most appropriate one. Given that comparing was not possible, we could have simply taken one of the already extant intonation models. For Dutch, two intonation models are in general use, viz. the type 2A model of 't Hart et al., (1990) and the type 1B model proposed by Van den Berg et al., (1992). Besides these two models, we also considered using Fujisaki's model (1991), because it has been successfully applied to many different languages. However, the problem is that the choice of the intonation model can determine the outcome of our analyses to a large extent. First of all, the discrete units used in these three intonation models are different. Therefore, the results of a study on the relation between the physiological signals and these discrete units would probably yield different results for each model. Furthermore, if a one-component intonation model is chosen, it is impossible to arrive at a physiological model with two components. On the other hand, if a two-component model with an obligatory global component is used, the outcome will always be a physiological model with two components.

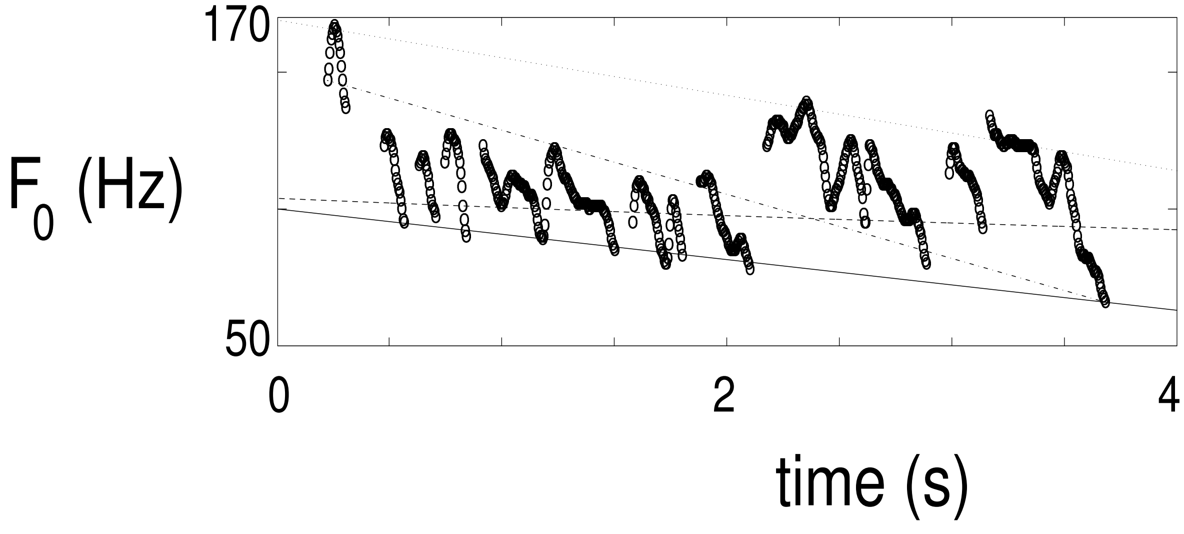


Figure 1. Average F_0 signal and trend lines for an utterance spoken by subject LB. The following trend lines are shown: $F_{0,g}$ (dashed), the line connecting the first and the last voiced frame (dashed-dotted), topline (dotted), and bottomline or baseline (solid). This is the same figure as Figure 6 of Chapter 8.

Finally, using an intonation model could obscure the relation between the physiological signals and intonation, rather than clarify it, as will be explained by means of the following example. In Figure 1 an F_0 contour is shown, together with some global trend lines. In many two-component models the global component of F_0 is a line running below the F_0 values, in which case it is sometimes called a baseline, a bottomline or simply a reference line. Suppose that, according to a (hypothetical) intonation model, the solid bottomline in Figure 1 is the global component of F_0 ($F_{0,global}$). The next step in our analysis would be to study the relation between $F_{0,global}$ and the physiological signals. Because in the measurements for the example utterance P_{sb} has a global component ($P_{sb,global}$), while this is not the case for CT, VOC and SH (see Figure 1 of Chapter 8), the result of the analysis could be that $F_{0,global}$ is due to $P_{sb,global}$. From a physiological point of view this is not completely correct: $F_{0,global}$ is not the F_0 due to $P_{sb,global}$ alone. However, if the F_0-P_{sb} ratio (FPR) is verified, the FPR would probably be larger than 7 Hz/cm H_2O, and the conclusion could then be that $F_{0,global}$ cannot be due to $P_{sb,global}$ alone. In that case, the conclusion would be that $F_{0,global}$ is due to P_{sb} and laryngeal muscle activity. Although this explanation is probably correct, the point is that in this way a much simpler physiological explanation of intonation is obscured, i.e. a physiological model in which the global component in F_0 is due to the global component in P_{sb}, and the local component in F_0 is due to the local variations in P_{sb}, CT, VOC and SH.

9.3 Discussion and conclusions

9.3.1 Method of non-linear time-alignment and averaging

Before the relation between the physiological signals and prosody could be studied, the raw data had to be processed. An essential part of processing consisted of averaging the signals that were measured when the subjects repeated the same sentence. Because a substantial variation in speaking rate was observed in these repetitions, straightforward averaging would have the effect of averaging parts of the signals that are related to different articulatory events. To overcome this problem, the method of non-linear time-alignment and averaging was developed.

This method, which is described in Chapter 2, successfully corrects for the timing differences among the different realizations of the utterances. After time normalization of the physiological signals of the repetitions, it was found that the amount of variation between the signals was within reasonable limits. This means that after time normalization it is possible to obtain meaningful average signals. Further advantages of our method of non-linear time-alignment and averaging are that it also yields an average F_0 contour, that it can be used semi-automatically and that it can be used to time-align and average all kinds of signals for which timing differences are apparent.

9.3.2 P_{sb}, P_{or} and P_{tr}

In many publications in which the relation between physiological signals and prosodic parameters (especially F_0) is discussed, P_{sb} is mentioned as an important factor in the control of prosody (see Sections 1.2.2 and 1.2.3), while P_{or} and P_{tr} are generally not mentioned. The main reasons probably are that in this field of research P_{sb} has been studied more often than P_{or} and P_{tr}, and that in many studies the subjects produced sustained vowels. For sustained vowels P_{or} is small and almost constant. But in natural speech P_{or} varies rapidly, and therefore it affects the prosodic parameters, as was shown in Chapter 3.

The general behaviour of P_{sb}, P_{or} and P_{tr} is as follows. A gradual decrease in P_{sb} is often observed during the utterance. Sometimes P_{sb} also shows some local variations, but the range of these local variations is generally much smaller than the range of the local variations in P_{or} and P_{tr}. P_{or} and P_{tr} vary between zero and P_{sb}. For vowels P_{or} is almost zero (P_{tr} is roughly equal to P_{sb}) and for some consonants P_{or} is almost equal to P_{sb} (P_{tr} is about zero).

The result of this behaviour is that the relation of the pressure signals with the prosodic parameters depends on the interval over which the analysis is carried out. For short intervals containing vowels and voiced consonants, the amount of variation in P_{or} and P_{tr} is generally much larger than the variation in P_{sb}. Although P_{sb} remains almost constant, the increase in P_{or} leads to a decrease in P_{tr}, F_0 and IL. Therefore, for a short interval P_{or} and P_{tr} are more important than P_{sb} in the control of F_0 and IL.

For complete utterances the variation in F_0 and IL is a combination of local variations in F_0 and IL, and the downtrend in F_0 and IL during the utterance. The downtrend in P_{sb} can explain the latter part of the variation in F_0 and IL, and thus for a long interval the correlation coefficient of P_{sb} with both F_0 and IL becomes larger than for a short interval (see Chapter 3).

In general the correlations of P_{tr} with F_0 and IL are higher than the correlations of these prosodic parameters with P_{sb} and P_{or}. Both for short and long intervals the correlation coefficient between P_{tr} and IL is very high, and therefore P_{tr} can be considered as the driving force of the vocal folds. The fact that the correlation of P_{tr} with IL is so high, indicates that other factors (especially the laryngeal muscles) have little effect on IL. The correlations of F_0 with P_{tr} are lower, because the laryngeal muscles have a significant effect on F_0.

9.3.3 Voice source parameters

An ancillary aim of our research was to develop a method for automatic extraction of the voice source parameters from continuous speech. In our method an estimate of the derivative of the glottal flow (dU_g) is obtained by inverse filtering (IF) the speech signal. Integration of dU_g yields an estimate of the glottal flow signal (U_g). In Chapter 3 some voice source parameters are calculated directly from dU_g and U_g by means of simple mathematical operators: excitation strength $E_e = \min(dU_g)$ and peak glottal flow $U_0 = \max(U_g)$.

With simple mathematical operators it was difficult to get reliable estimates of the time parameters, though. Especially estimating the moment of glottal opening (t_o) proved to be a problem, mainly because the glottal flow signals are small and change slowly near t_o. In order to enable reliable estimates of all parameters, a voice source model (the LF-model) was fitted to the data. Because in continuous speech the voice source parameters probably change from period to period, the fit procedure must be pitch-synchronous. Therefore, for each pitch period the LF-parameters are calculated by fitting the LF-model to the glottal flow signal. In Chapter 4 this fit procedure was used to calculate the LF-parameters for four repetitions of a spontaneous utterance.

Subsequently, we tried to improve our automatic parameterization method. Two methods were compared in Chapter 5. In the method that proved to be better, a formant-bandwidth tracker is used to find the optimal traces for the given formant frequencies and bandwidth values. An inverse filter is then computed from the formant and bandwidth values of these traces. In Chapter 6 we studied the sensitivity of the fit procedure to different kinds of disturbances that could be present in dU_g. The most important conclusions of the experiments in Chapters 5 and 6 are the following. An important aspect of the fit procedure is the routine to calculate the LF-pulse. A substantial improvement of the estimated voice source parameters was obtained by using an implementation in which all LF-parameters can change continuously. The optimization algorithm used is also important. In general, the Simplex search algorithm performed better than gradient algorithms. Finally, the method used to low-pass filter the signals is also important. Low-pass filters are often applied because the glottal flow signals are usually noisy. First of all, it should be noted that for filtering the pulse-like glottal flow signals, it is important to use a low-pass filter with a ripple-free impulse response. However, low-pass filters alter the shape of the flow pulse and thus influence the estimates of the voice source parameters. Our solution to this problem was to filter both the flow pulse and the LF-pulse used in the fit procedure. The conclusion is that automatic calculation of the voice source parameters for continuous speech is possible. Our experiments showed that the automatic parameterization method is robust against the disturbances tested. In general, the errors in the estimates of t_o and T_a were larger than the errors in the estimates of the other time parameters.

9.3.4 Physiological signals and prosodic parameters

The general aim of our study was to clarify the relation between some physiological signals and prosody. Since the measured physiological signals influence the voice source signal first and then the speech signal (via the voice source signal), we decided to include the voice source parameters in our investigation (see Sections 1.3.2.2 and 9.3.3).

In Chapters 3, 4 and 5 the voice source parameters were computed as described above. This made it possible to study the relations between the following three groups of parameters:

① physiological signals

② voice source parameters

③ prosodic parameters

The voice source parameters were divided into two groups, viz. the time parameters (T_i, T_p, T_e, T_n, T_a and T_0) and the amplitude parameters (P_{tr}, U_0, E_e and IL). The reason was that high positive correlations were observed within the two groups, while a reciprocal relation was found between parameters of the two groups.

During a transition from a vowel to a voiced consonant, a decrease in P_{tr}, U_0, E_e and IL, and an increase in T_0, T_n and T_a were usually observed. The behaviour of the voice source parameters during the last syllable of an utterance and during voice onset and off- set deviated from the general behaviour that was observed for the other data (see Chapters 3 and 4). The general behaviour is described in this section.

9.3.4.1 Time parameters

During a transition from a vowel to a voiced consonant T_0 usually increases, as mentioned above, and thus F_0 decreases. The average behaviour of all time parameters in relation to T_0 was studied. It turned out that all time parameters increase when T_0 increases, but the amount of increase differs. The duration of the first part of the LF-pulse (T_e) varies linearly with T_0. However, the increase in T_i and T_p is less than linear, while the increase in T_n and T_a is more than linear. T_n is inversely related to the skewing of the pulse, and therefore it can be concluded that the skewing of a pulse generally decreases during the transition of a vowel to a voiced consonant.

9.3.4.2 Amplitude parameters

In section 9.3.2 we concluded that IL is mainly controlled by P_{tr}. The regression equation calculated for our data in Chapter 3 is:

$$IL = 41.6 + 30.3 \log P_{tr} \qquad (N = 293,\ r = 0.923)$$

This finding makes it interesting to investigate the way in which IL is influenced by P_{tr} via the voice source parameters. The relations between P_{tr}, U_0, E_e and IL were studied for four repetitions of a spontaneous utterance. The number of pitch-periods for which these parameters were calculated (N) is 613 (see Chapter 4), and the results were:

$$U_0 \sim P_{tr}^{1.0} \qquad (N = 613,\ r = 0.60)$$

$$E_e \sim P_{tr}^{1.6} \qquad (N = 613,\ r = 0.63)$$

$$Int \sim P_{tr}^{3.0} \qquad (N = 613,\ r = 0.81)$$

Although the data form a mix of voiced consonants, stressed and unstressed vowels, the correlations are very high, indicating the consistent relations between these parameters. The power of the relation between peak glottal flow (U_0) and P_{tr} is 1.0, i.e. U_0 varies linearly with P_{tr}. E_e is a function of U_0 and the skewing of the pulse. The fact that both U_0 and the skewing increase with increasing P_{tr}, explains why the power of the relation between E_e and P_{tr} is larger than the power of the relation between U_0 and P_{tr}.

Finally, the relation between Int and E_e was studied. The Int of a freely travelling spherical sound wave is proportional to the square of the derivative of the mouth flow (D_m): Int ~ D_m^2 (Beranek, 1954). Our results show that Int is also about proportional to the square of E_e, which is the amplitude of the derivative of glottal flow: Int ~ $E_e^{1.9}$ (1.9 = 3.0/1.6). This suggests that D_m varies roughly linearly with E_e. However, a proper voice source - vocal tract model should be used to unravel the exact nature of the underlying relations.

9.3.4.3 Spectral contents

Above we have described the behaviour of the time parameters and the amplitude parameters of the voice source signal. Here we will focus on the aspects related to the spectral contents (SC) of the speech signal. To describe the relation of the LF-parameters with SC it is better to use E_e and the dimensionless wave shape parameters R_g, R_k and R_a (which are defined in Chapter 4). R_g and R_k influence the amplitudes of the two or three lowest harmonics, R_a the overall spectral slope and E_e the overall intensity of the spectrum.

In Chapter 4 R_g, R_k and R_a were calculated for 613 pitch-periods. The amount of variation in R_g was small, and thus its influence on changes in SC is small. Apart from changes in overall intensity, which are caused by changes in E_e, the changes in the spectrum are mainly determined by R_k and R_a.

The correlations of R_a and R_k with T_0 are positive, and highly significant ($p < 0.0001$). Given the relation obtaining between T_0 and F_0, it follows that the correlations of the wave shape parameters R_a and R_k with F_0 are negative. The correlations of R_a and R_k with IL are negative, and even more significant. This means that R_a and R_k covary with F_0 and IL. Probably, the physiological mechanisms that alter F_0 and IL also influence the shape of the glottal wave, and thus SC and voice quality. Because F_0, IL and SC are not independent, it seems advisable to study all three prosodic parameters simultaneously, as was already advocated in Section 1.3.2.2.

In general, during a transition from a vowel to a voiced consonant F_0 and IL decrease, while R_a and R_k increase. The result of increasing R_a is a steeper slope of the spectrum, which is often associated with a more breathy mode of phonation. The increase in R_k would lead to a higher first harmonic, which indicates a less pressed voice. Both findings are in agreement with previous studies. Therefore it can be concluded that the wave shape parameters provide a powerful tool for quantifying SC and voice quality.

9.3.5 A physiological model of intonation

In Chapters 7 and 8 the relation of P_{sb}, CT, VOC and SH with intonation was studied. To that end we first used a qualitative analysis method, for the following reasons: (1) There is no voice source - vocal tract model that allows a realistic quantitative analysis of the relation between these physiological signals and F_0 (see Section 9.2.1). (2) For a quantitative modelling of the relation between F_0 and intonation various models have been described in the literature (see Sections 1.2.1 and 9.2.2). However, it is not evident which model is the most suitable. Furthermore, the choice of the model could influence the results of the analysis, as was explained in Section 9.2.2. This is especially problematic because comparing the different intonation models is hardly possible (see Section 9.2.2). (3) Finally, differences in the physiological control of intonation have been found between subjects. This makes it dangerous to draw conclusions on the basis of observations concerning one or two subjects. An additional advantage of using a qualitative method was that we could study both our own data and the data from the literature. The qualitative analysis method was used to ascertain whether consistent relations between the physiological signals and intonation were present, and also to find out in which way these relations can be modelled.

In our own data and the data from the literature, a gradual lowering of F_0 and P_{sb} during the whole utterance was often observed, while this is generally not the case for CT, VOC and SH. Furthermore, local variations in F_0, P_{sb}, CT, VOC and SH were found. Therefore, from a physiological point of view, the most likely model is a two-component model that separates global and local effects. In the physiological model of intonation proposed in Chapter 7, the global component of F_0 ($F_{0,g}$) is due to the global component of P_{sb} ($P_{sb,g}$), while the local variations in F_0 ($F_{0,l}$) are caused by the local variations in P_{sb} ($P_{sb,l}$), CT, VOC and SH.

In Chapter 8 we provided a quantitative implementation of our model in which $P_{sb,g}$ and $F_{0,g}$ were obtained by manual fitting. In this model $F_{0,g}$ is the F_0 due to $P_{sb,g}$ alone, i.e. when CT, VOC and SH are at their minimal level. Because it is likely that local increases in P_{sb}, CT and VOC can raise F_0 above $F_{0,g}$, and that local increases in SH can result in an F_0 that is lower than $F_{0,g}$, it can be deduced that $F_{0,g}$ should run between the peaks and the valleys of the F_0 contour. Furthermore, if we assume that the relation between F_0 and P_{sb} is linear, and if we also assume that the F_0-P_{sb} ratio (FPR) is between 2 and 7 Hz/cm H_2O, then we can deduce from the slope of $P_{sb,g}$ in which range the slope of $F_{0,g}$ should fall. The dashed line drawn in Figure 1 meets both requirements: it runs between the minima and maxima of F_0, and the FPR is 5 Hz/cm H_2O.

In our model the gradual decline in P_{sb} ($P_{sb,g}$) causes a gradual lowering in F_0 ($F_{0,g}$). In order to check whether this assumption of the model is correct, we tested the following hypothesis: the downtrend in F_0 is caused by the downtrend in P_{sb}. The two main counterarguments are discussed in Chapter 8. The first counterargument is that it is unlikely that P_{sb} is under voluntary control, and because declination is used for linguistic purposes it cannot be controlled by P_{sb}. In Chapter 8 we provided several arguments to show that it is likely that the respiratory muscles and P_{sb} are actively controlled. If this is the case, then this counterargument cannot be used to refute the hypothesis mentioned above.

The second counterargument is that the gradual lowering of P_{sb} cannot explain all of the declination in F_0, because the F_0-P_{sb} ratio (FPR) is too high. However, calculation of the FPR in running speech is not without problems, because apart from P_{sb} there are other physiological mechanisms that influence F_0. We showed that some of the methods used in the past did not give fair estimates of the FPR, because they did not correct for these other factors.

To calculate the FPR we used a statistical method in which a correction was made for some of these other factors. The resulting values of FPR were generally within the allowed range of 2-7 Hz/cm H_2O. The conclusion therefore is that the downtrend in P_{sb} can explain all of the downtrend in F_0, depending on how the “downtrend in F_0” is defined. The explanation holds if “downtrend in F_0” is defined in such a way that effects of laryngeal muscles are excluded.

Downtrend, declination or downdrift are used to denote the tendency of F_0 to decrease during the course of an utterance (see Section 1.2.1). This is a rather vague, qualitative definition, which can be interpreted in many different ways. Therefore, when discussing or studying this issue it is necessary to give a precise, preferably quantitative, definition of downtrend, declination or downdrift.

9.4 Further research

In Chapter 7 we have proposed a qualitative physiological model of intonation that describes the data of a number of subjects in a satisfactory way. A quantitative implementation of this model was provided in Chapter 8. In the latter model a linear relation between the physiological signals and F_0 was assumed. However, this is not more than a very crude first approximation, because this relation is unlikely to be linear. Furthermore, our model is incomplete, especially with respect to the role of the strap muscles (like e.g. SH). It is therefore legitimate to wonder what should be done to arrive at a more complete and more realistic model.

New physiological measurements could lead to a more complete model. For instance, new measurements could throw some light on the role of the strap muscles, in particular on their contribution to lowering F_0. It is also possible that new measurements could reveal that other physiological mechanisms, besides P_{sb}, CT, VOC and the strap muscles, are important in the control of intonation. After all, in some studies it has been mentioned that other physiological mechanisms are involved in the control of intonation; e.g. the tracheal pull mechanism proposed by Maeda (1976), or activity of the cricopharyngeus muscle (Honda & Fujimura, 1991). However, even if we can make the model more comprehensive by making new recordings of important physiological mechanisms, the model will still remain essentially qualitative. Obtaining further physiological measurements concerning F_0 is not enough to bridge the gap between our present position and a true quantitative and predictive model of the physiology of intonation. In order to reach such a comprehensive model more research is needed on both intonation models and voice source - vocal tract models, as will be described below.

To study the relation between the physiological signals and intonation, an intonation model can be used as a kind of analysis instrument which should provide a specification of intonation (see Section 9.2.2). In order to be useful as an analysis instrument this specification should be detailed and preferably quantitative. Another requirement for an analysis instrument is that it be reliable. In order to use an intonation model, an intonation transcription has to be made, as explained in Section 9.2.2. However, differences have been observed in intonation transcriptions made by different persons. Therefore, agreement indices should be used to get an idea of the degree of accuracy of an intonation transcription, as is generally done in research fields in which segmental transcriptions are used as an analysis instrument. Furthermore, it should be possible to evaluate the performance of an intonation model, and to compare its performance to that of other intonation models. More research is needed to fulfil the requirements specified above.

Apart from the relation between intonation and F_0, the relation between F_0 and the physiological signals should also be studied. However, the physiological signals will not only affect F_0, but also IL and SC. Although F_0 may be considered the most important linguistic factor (see Section 1.2.1), the three prosodic parameters are of equal importance from a physical and physiological point of view. Therefore, in order to gain more insight in the physiological control of prosody, it seems advisable to study all three prosodic parameters. For instance, a high correlation between IL and P_{tr} was generally found, and thus by measuring IL one has a rough idea of the behaviour of P_{tr}. On the other hand, SC could give an indication of the mode of phonation, e.g. the amount of adduction of the vocal folds. For this purpose, the voice source parameters are probably useful, because they can be used to quantify vocal quality (we will come back on this point below).

For a realistic analysis of the relation between the physiological signals and the prosodic parameters new voice source - vocal tract models should be developed. Building such models will be an enormous task. First of all, we need a comprehensive biomechanical model of the larynx with all its relevant muscles and tissues. Second, we need a trustworthy model that predicts not only muscle behaviour from measured EMG activity, but also the consequences of the muscle activity for the elastic properties of the vocal fold tissues. Existing models are inadequate for various reasons. For instance, the model proposed by Ishizaka & Flanagan (1972) is too stylized to allow realistic physiological interpretation. Alternatively, the model by Titze & Talkin (1979) is so complex that the number of parameters to estimate or simulate is too large.

Taking the complete speech production system as a black box, with physiological measurement data as input and prosodic parameters as output, might be unnecessarily complex. The physiological signals will first affect the voice source signal; in turn changes in the voice source signal will influence the prosodic parameters. Therefore, it could be better to study the two relations separately. This is possible with the method developed in our research (see Chapters 4, 5 and 6), which computes voice source parameters from the speech signal.

In the proposed method a voice source model that parameterizes the glottal flow is used to obtain voice source parameters for each pitch period. In our research we used the LF-model for this purpose. However, this might not be the optimal model to study the relation between the physiological signals and the voice source parameters. In order to find out which model is most suitable, the relation between the physiological signals and the voice source parameters should be studied for different models.

The voice source parameters could also be employed to study their relation with the prosodic parameters (i.e. F_0, IL and SC), or, optionally, with the prosodic features. Because this research does not require any invasive measurements, and because the voice source parameters can be obtained automatically, large amounts of data can be acquired. In this way databases can be built, and these databases can subsequently be used to study the behaviour of the voice source parameters, and, if possible, to formulate rules for the dynamic behaviour of the voice source parameters. In turn, these rules can be used to improve the quality of synthetic speech, in the same way as the rules for the dynamic behaviour of formants are used in a formant synthesizer. By using rules for all voice source parameters it will be possible to control not only F_0, but also IL and SC.

Voice source parameters could also provide a powerful tool for quantifying voice quality, as was described in Section 9.3.4.3. At present it is difficult to specify voice quality in quantitative terms, and therefore in many studies it is specified in qualitative terms like e.g. breathy and creaky. The relation between these qualitative definitions and the voice source parameters could be studied, and in this way it might be possible to define regions for the voice source parameters (i.e. clusters in a multi-dimensional space) that are related to the qualitative definitions of voice quality. Once this relation is established for some voice source model, this knowledge and the voice source model can subsequently be used to measure voice quality. Further research should reveal whether this is possible.

References

Anantapadmanabha, T.V. & Fant, G. (1982) Calculation of true glottal flow and its components. *Speech Communication*, 1, 167-184.

Atkinson, J.E. (1973) *Aspects of intonation in speech: Implications from an experimental study of fundamental frequency.* Unpublished Ph.D. thesis, Univ. of Connecticut, Storss, CT.

Atkinson, J.E. (1978) Correlation analysis of the physiological features controlling fundamental voice frequency. *Journal of the Acoustical Society of America*, 63, 211-222.

Atkinson, J.E. & Erickson, D. (1977) The function of the strap muscles in speech: Pitch lowering or jaw opening? *Haskins Laboratories Status Report on Speech Research* SR-49: 97-102.

Baer, T. (1979) Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes. *Journal of the Acoustical Society of America*, 65, 1271-1275.

Baer, T., Gay, T. & Niimi, S. (1976) Control of fundamental frequency, intensity and register of phonation. *Haskins Laboratories Status Report on Speech Research*, SR-45/46, 175-185.

Baken, R.J. & Cavallo, S.A. (1981) Prephonatory chest wall posturing. *Folia Phoniatrica*, 33, 193-203.

Baken, R.J., Cavallo, S.A. & Weissman, K.L. (1979) Chest wall movement prior to phonation. *Journal of Speech and Hearing Research*, 22, 862-872.

Baken, R.J. & Orlikoff, R.F. (1987) Phonatory responses to step-function changes in supraglottal pressure. In: T. Baer, C. Sasaki & K.S. Harris (eds.), *Vocal Fold Physiology: Laryngeal function in phonation and respiration*, Boston: College-Hill Press, 273-290.

Basmajian, J.V. (1967) *Muscles Alive, their functions revealed by electromyography*, Baltimore: The Williams & Wilkins Company.

Beckman, M. & Pierrehumbert, J. (1992) Comments on chapters 14 and 15. In: G.J. Docherty & D.R. Ladd (eds.), *Gesture, segment, prosody*, Cambridge: Cambridge University Press, 387-397.

Beranek, L. (1954) *Acoustics*, New York: McGraw-Hill Book Company.

Bickley, C.A. & Stevens, K.N. (1986) Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies. *Journal of Phonetics*, 14, 373-382.

Borden, G.J. & Harris, K.S. (1980) *Speech science primer*, Baltimore: The Williams & Wilkins Company.

Bouhuys, A., Mead, J., Proctor, D.F. & Stevens, K.N. (1968) Pressure-flow events during singing. In: M. Krauss, M. Hammer & A. Bouhuys (eds.), *Annals of the New York Academy of Sciences*, 155, 165-176.

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

Boves, L., Kerkhoff, J. & Loman, H. (1987) A new synthesis model for an allophone based text-to-speech system. *Proceedings of the European Conference on Speech Technology*, Edinburgh, 2, 385-388.

Breckenridge, J. (1977) Declination as a phonological process. *Bell Labatories Technical Memorandum*, Murray Hill, NJ.

Carlson, R., Fant, G., Gobl, C., Granstrom, B., Karlsson, I. & Lin, Q. (1989) Voice source rules for text-to-speech synthesis. *Proceedings of the International Conference on Acoustic Speech Signal Processing*, 1, 223-226.

Catford, J.C. (1977) *Fundamental problems in phonetics*, Edinburgh: Edinburgh University Press.

Cavagna, G.A. & Margaria, R. (1968) Airflow rates and efficiency changes during phonation. In: M. Krauss, M. Hammer & A. Bouhuys (eds.), *Annals of the New York Academy of Sciences*, 155, 152-164.

Childers, D.G. & Lee, C.K. (1991) Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90(5), 2394-2410.

Clark, J. & Yallop, C. (1990) *An introduction to phonetics and phonology*, Oxford: Basil Blackwell.

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

Collier, R. (1974) Laryngeal muscle activity, subglottal air pressure, and the control of pitch in speech. *Haskins Laboratories Status Report on Speech Research*, SR-39/40, 137-170.

Collier, R. (1975) Physiological correlates of intonation patterns. *Journal of the Acoustical Society of America*, 58, 249-255.

Collier, R. (1987) F_0 declination: the control of its setting, resetting, and slope. In: T. Baer, C. Sasaki & K.S. Harris (eds.), *Laryngeal function in phonation and respiration*, Boston: College-Hill Press, 403-421.

Cooper, W.E. & Sorensen, J.M. (1981) *Fundamental frequency in sentence production*, New York: Springer-Verlag.

Cranen, L.I.J. (1987) *The acoustic impedance of the glottis: Measurements and Modelling*, Unpublished Ph.D. thesis, University of Nijmegen.

Cranen, B. (1991) Simultaneous modeling of EGG, PGG and glottal flow. In: J. Gauffin & B. Hammarberg (eds.), *Phonatory mechanisms: physiology, acoustics, and assessment*, San Diego: Singular Publishing Group, 57-64.

Cranen, B. & Boves, L. (1985) Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production. *Journal of the Acoustical Society of America*, 77, 1543-1551.

Cranen, B. & Boves, L. (1987) The acoustic impedance of the glottis: modeling and measurements. In: Th. Baer, C. Sasaki & K. Harris (eds.), *Laryngeal function in phonation and respiration*, Boston: College-Hill Press, 203-218.

Crystal, D. (1987) *The Cambridge encyclopedia of language*, Cambridge: Cambridge University Press.

Davis, S.B. & Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28, 357-366.

Dennis, J.E., Gay, D.M. & Welsch, R.E. (1981) An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7, 348-368.

De Veth, J., Cranen, B., Strik, H. & Boves, L. (1990) Extraction of control parameters for the voice source in a text-to-speech system. *Proceedings International Conference on Acoustic Speech Signal Processing*, 1, 301-304

Erickson, D. & Atkinson, J.E. (1976) The functions of the strap muscles in speech. *Haskins Laboratories Status Report on Speech Research*, SR-45/46, 205-210.

Erickson, D., Baer, T. & Harris, K.S. (1983) The role of the strap muscles in pitch lowering. In: D.M. Bless & J.H. Abbs (eds.), *Vocal Fold Physiology*, San Diego, CA: College-Hill Press.

Erickson, D., Liberman, M. & Niimi, S. (1977) The geniohyoid and the role of the strap muscles. *Haskins Laboratories Status Report on Speech Research*, SR-49, 103-110.

Ewan, W.G. & Krones, R. (1973) A study of larynx height in speech using the thyroumbrometer. *Journal of the Acoustical Society of America*, 53, 345 (A).

Faaborg-Andersen, K. (1957) Electromyographic investigation of intrinsic laryngeal muscles in humans. *Acta Physiologica Scandinavica, 41*, suppl. 140.

Faaborg-Andersen, K. (1965) Electromyography of laryngeal muscles in humans: Techniques and results. In: F. Trojan (ed.), *Aktuelle Problemme der Phoniatrie und Logopaedie*, 3, Basel: Karger.

Fant, G. (1986) Glottal flow: Models and interaction. *Journal of Phonetics*, 14, 393-400.

Fant, G., Liljencrants, J. & Lin, Q. (1985) A four-parameter model of glottal flow. *Speech Transmission Laboratory, Quarterly Program & Status Report*, 4, 1-13.

Fant, G. & Lin, Q. (1988) Frequency domain interpretation and derivation of glottal flow parameters. *Speech Transmission Laboratory, Quarterly Program & Status Report*, 2-3, 1-21.

Ferguson, G.A. (1987) *Statistical analysis in psychology and education*, Singapore: McGraw-Hill Book Company.

Fourcin, A.J. (1974) Laryngographic examination of vocal fold vibration. In: B. Wike (ed.), *Ventilatory and phonatory control systems*, London: Oxford University Press, 315-326.

Fujisaki, H. (1991) Modeling the generation process of F_0 contours as manifestation of linguistic and paralinguistic information. *Proceedings of the XIIth International Congress of Phonetic Sciences*, supplement, 1-10. Aix-en-Provence.

Garding, E. (1983) A generative model of intonation. In: A. Cutler and D.R. Ladd (eds.), *Prosody: Models and Measurements*. Berlin (etc.): Springer-Verlag.

Gauffin, J. & Sundberg, J. (1989) Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, 32, 556-565.

Gay, T., Hirose, H., Strome, M. & Sawashima, M. (1972) Electromyography of the intrinsic laryngeal muscles during phonation. *Annals of otology, rhinology and laryngology* 81 (8), 401-409.

Gelfer, C.E. (1987) *A simultaneous physiological and acoustic study of fundamental frequency declination*. Unpublished Ph.D. thesis, City University of New York.

Gelfer, C., Harris, K., Collier, R. & Baer, T. (1983) Is declination actively controlled?. In: I.R. Titze & C. Scherer (eds.), *Vocal Fold Physiology*, Denver: The Denver Center for the Performing Arts, Inc., 113-125.

Grosjean, F. & Collins, M. (1979) Breathing, pausing and reading. *Phonetica*, 36, 98-114.

Hardcastle, W.J. (1976) *Physiology of Speech Production*, London: Academic Press.

Hart, J. ’t, Collier, R. & Cohen, A. (1990) *A perceptual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge: Cambridge University Press.

Hast, M.H. (1968) Studies on the extrinsic laryngeal muscles. *Archives of Otolaryngology*, 88: 273-278.

Hinchcliffe, R. & Harisson, D. (1976) *Scientific foundations of otolaryngology*, London: William Heinemann Medical Books ltd.

Hirano, M. & Ohala, J. (1969) The function of laryngeal muscles in regulating fundamental frequency and intensity in phonation. *Journal of Speech and Hearing Research*, 12 (3), 616-627.

Hirano, M., Vennard, W. & Ohala, J. (1970) Regulation of register, pitch and intensity of voice. *Folia phoniatrica*, 22, 1-20.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and techniques. *Haskins Laboratories Status Report on Speech Reasearch*, SR-25/26, 73-86.

Hirose, H. & Gay, T. (1972) The activity of the intrinsic laryngeal muscles in voicing distinction. *Phonetica*, 25, 140-164.

Hirose, H. & Sawashima, M. (1981) Functions of the laryngeal muscles in speech. In: K.N. Stevens & M. Hirano (eds.), *Vocal Fold Physiology*, Tokyo: University of Tokyo Press.

Hixon, T.J., Goldman, M.D. & Mead, J. (1973) Kinematics of the chest wall during speech production: volume displacements of the rib cage, abdomen, and lung. *Journal of Speech and Hearing Research*, 16, 78-115.

Hixon, T.J., Klatt, D.H. & Mead, J. (1971) Influence of forced transglottal pressure changes on vocal fundamental frequency. *Journal of the Acoustical Society of America*, 49, 105 (A).

Honda, K. & Fujimura, O. (1991) Intrinsic vowel F_0 and phrase-final F_0 lowering: Phonological vs. biological explanations. In: J. Gauffin & B. Hammarberg (eds.), *Phonatory mechanisms: physiology, acoustics, and assessment*, San Diego: Singular Publishing Group, 57-64.

Ishizaka, K. & Flanagan, J.L. (1972) Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Systems Technical Journal*, 51, 1233-1268.

Isshiki, N. (1964) Regulatory mechanisms of voice intensity variation. *Journal of Speech and Hearing Research*, 7, 17-29.

Jansen, J., Cranen, B. & Boves, L. (1991) Modelling of source characteristics of speech sounds by means of the LF-model. *Proceedings of Eurospeech '91*, 1, 259-262.

Katsuki, Y. (1950) The function of the phonatory muscles. *Japanese Journal of Physiology*, 1, 29-36.

Klatt, D.H. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820-857.

Ladd, D.R. (1984) Declination: a review and some hypotheses. *Phonology Yearbook*, 1, 53-74.

Ladd, D.R. (1992) An introduction to intonational phonology. In: G.J. Docherty & D.R. Ladd (eds.), *Gesture, segment, prosody*, Cambridge: Cambridge University Press, 321-334.

Ladefoged, P. (1963) Some physiological parameters in speech. *Language and Speech*, 6, 109-119.

Ladefoged, P. (1967) *Three areas of experimental phonetics*. Oxford: Oxford University Press.

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

Liberman, M.Y. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In: M. Aronoff & R.T. Oehrle (eds.), *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, Cambridge, MA: MIT Press, 157-233.

Lieberman, P. (1967) *Intonation, Perception and Language*. Cambridge, MA: MIT Press.

Lin, Q. (1990) Speech production theory and articulatory speech synthesis. Unpublished Ph.D. thesis, KTH, Stockholm.

Lindestad, P., Fritzel, B. & Persson, A. (1991) Influence of pitch and intensity on cricothyroid and thyroarytenoid activity in singers and nonsingers, In: J. Gauffin & B. Hammarberg (eds.), *Phonatory mechanisms: physiology, acoustics, and assessment*, San Diego: Singular Publishing Group, 175-182.

Maeda, S. (1976) *A characterization of American English intonation*, Unpublished Ph.D. thesis, MIT, Cambridge.

Maeda, S. (1980) On the F_0 control mechanism of the larynx. In: L. Boë, R. Descout & B. Guérin (eds.), *Larynx et parole*, Grenoble: Institut de Phonétique, 243-257.

Markel, J.D. & Gray Jr., A.H. (1976) *Linear prediction of speech*, Berlin: Springer-Verlag.

Müller, J. (1843) *Elements of physiology*, Philadelphia: Lea and Blanchard.

Negus, V.E. (1928) *The mechanisms of the larynx*, London: William Heinemann Medical Books, Ltd.

Nelder, J.A. & Mead, R. (1964) A simplex method for function minimization. *The Computer Journal*, 7, 308-313.

Niimi, S., Horiguchi, S., Kobayashi, N. & Yamada, M. (1987) Electromyographic study of vibrato and tremolo in singing. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo*, 21, 153-164.

Ohala, J.J. (1970) Aspects of the control and production of speech. *UCLA Working Papers Phonetics*, 15, 1-192.

Ohala, J.J. (1972) How is pitch lowered? *Journal of the Acoustical Society of America*, 52, 124 (A).

Ohala, J.J. (1977) The physiology of stress. In: L.M. Hyman (ed.), *Studies in stress and accent. Southern California Occasional Papers in Linguistics*, 4, 145-168,

Ohala, J.J. (1978) Production of tone. In: V. Fromkin (ed.), *Tone: a linguistic survey*, New York: Academic Press, 5-39.

Ohala, J.J. (1990) Respiratory activity in speech. In: W.J. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling* Netherlands: Kluwer Academic Publishers, 25-53.

Ohala, J. & Hirose, H. (1970) The function of the sternohyoid muscle in speech. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 4, 41-44.

Ohala, J. & Ewan, W.G. (1973) Speed of pitch change, *Journal of the Acoustical Society of America*, 53, 345 (A).

Öhman, S.E.G. (1967) Word and sentence intonation: A quantitative model. *Speech Transmission Laboratory, Quarterly Program & Status Report*, 2-3, 20-54.

Öhman, S.E.G. (1968) A model of word and sentence intonation. *Speech Transmission Laboratory, Quarterly Program & Status Report*, 2-3, 6-11.

Paliwal, K.K. & Rao, P.V.S. (1982) Evaluation of various linear prediction parametric representations in vowel recognition. *Signal processing*, 4, 323-327.

Perkell, J.S. (1969) *Physiology of speech production: results and implications of a quantitative cineradiographic study*, Cambridge, MA: MIT Press.

Pickett, J.M. (1980) *The sounds of speech communication*, Baltimore: University Park Press.

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

Pierrehumbert, J. (1980) *The phonetics and phonology of English intonation*. Ph.D. thesis, MIT (distributed 1988, Bloomington: IULC).

Pierrehumbert, J. & Beckman, M.E. (1988) *Japanese Tone Structure*, Linguistic Inquiry Monograph Series, 125, Cambridge, MA: MIT Press.

Rothenberg, M. & Mahshie, J. (1986) Induced transglottal pressure variations during voicing. *Journal of Phonetics*, 14, 365-371.

Rubin, H.J. (1963) Experimental studies on vocal pitch and intensity in phonation. *The Laryngoscope*, 8, 973-1015.

Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-26, 43-49.

Sawashima, M., Gay, T.J. & Harris, K.S. (1969) Laryngeal muscle activity during vocal pitch and intensity changes. *Haskins Laboratories Status Report on Speech Research*, SR-19/20, 211-220.

Shipp, T. (1975) Vertical laryngeal position during continuous and discrete vocal frequency change. *Journal of Speech and Hearing Research*, 18, 707-718.

Shipp, T., Doherty, E.T. & Morrissey, P. (1979) Predicting vocal frequency from selected physiologic measures. *Journal of the Acoustical Society of America*, 66, 678-684.

Shipp, T. & Haller, R.M. (1972) Vertical larynx height during vocal frequency change. *Journal of the Acoustical Society of America*, 52, 124 (A).

Shipp, T. & McGlone, R.E. (1971) Laryngeal dynamics associated with voice frequency change. *Journal of Speech and Hearing Research*, 14, 761-768.

Strik, H. & Boves, L. (1987) Regulation of intensity and pitch in chest voice. *Proceedings 11th International Congress of Phonetic Sciences*, VI, Tallinn, 32-35.

Strik, H. & Boves, L. (1988a) Control of fundamental frequency and intensity in running speech. *Journal of the Acoustical Society of America*, 84, S82 (A).

Strik, H. & Boves, L. (1988b) Averaging physiological signals with the use of a DTW algorithm. *Proceedings SPEECH'88, 7th FASE Symposium*, Edinburgh, Book 3, 883-890.

Strik, H. & Boves, L. (1988c) Simultaneous control of fundamental frequency and intensity in speech. *Proceedings SPEECH'88, 7th FASE Symposium*, Edinburgh, Book 3, 1115-1121.

Strik, H. & Boves, L. (1988d) Data processing of physiological signals related to speech. *Proceedings of the Department of Language and Speech, Nijmegen University*, 12, 41-56.

Strik, H. & Boves, L. (1989) The fundamental frequency - subglottal pressure ratio. *Proceedings of Eurospeech-89*, 2, 425-428.

Strik, H. & Boves, L. (1991a) A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech. *Journal of Phonetics*, 19, 367-378.

Strik, H. & Boves, L. (1991b) On the relation between voice source characteristics and prosody. *Proceedings EUROSPEECH 91*, Genova, 3, 1145-1148.

Strik, H. & Boves, L. (1992a) Control of fundamental frequency, intensity and voice quality in speech. *Journal of Phonetics*, 20, 15-25.

Strik, H. & Boves, L. (1992b) On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11, 167-174.

Strik, H. & Boves, L. (1993) A physiological model of intonation. *Proceedings of the Department of Language and Speech, Nijmegen University*, 16/17, 96-105.

Strik, H. & Boves, L. (1994) Automatic estimation of voice source parameters. *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, 1, 155-158.

Strik, H. & Boves, L. (submitted) Downtrend in F_0 and P_{sb}. Submitted to *Journal of Phonetics*

Strik, H., Cranen, B. & Boves, L. (1993) Fitting an LF-model to inverse filter signals. *Proceedings of the 3rd European Conference on Speech Technology*, Berlin, 1, 103-106.

Strik, H., Jansen, J. & Boves, L. (1992) Comparing methods for automatic extraction of voice source parameters from continuous speech. *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, 1, 121-124.

Tanaka, S. & Gould, W.J. (1983) Relationships between vocal intensity and noninvasively obtained aerodynamic parameters in normal subjects. *Journal of the Acoustical Society of America*, 73, 1316-1321.

Titze, I.R. (1984) Parameterization of glottal area, glottal flow, and vocal fold contact area. *Journal of the Acoustical Society of America*, 75, 570-580.

Titze, I.R. (1989) On the relation between subglottal pressure and fundamental frequency in phonation. *Journal of the Acoustical Society of America*, 85, 901-906.

Titze, I.R (1992) Phonation threshold pressure: a missing link in glottal aerodynamics. *Journal of the Acoustical Society of America*, 91, 2926-2935.

Titze, I.R. & Durham, P.L. (1987) Passive mechanisms influencing fundamental frequency control. In: T. Baer, C. Sasaki & K.S. Harris (eds.), *Laryngeal function in phonation and respiration*, Boston: College-Hill Press, 304-319.

Titze, I.R. & Talkin, D.T. (1979) A theoretical study of various laryngeal configurations on the acoustics of phonation. *Journal of the Acoustical Society of America*, 66, 60-74.

Van den Berg, R., Gussenhoven, C. & Rietveld, A.C.M. (1992) Downstep in Dutch: implications for a model. In: G.J. Docherty & D.R. Ladd (eds.), *Gesture, segment, prosody*, Cambridge: University Press, 321-334.

Van Katwijk, A. (1974) *Accentuation in Dutch: An experimental linguistic study*. Assen: Van Gorcum.

Weismer, G. (1985) Speech breathing: contemporary views and findings. In: R.G. Daniloff (ed.), *Speech Science*, San Diego: College Hill Press, 47-72.

Wyke, B. (1983) Neuromuscular control systems in voice production. In: D.M. Bless & J.H. Abbs (eds.), *Vocal Fold Physiology*, San Diego: College-Hill Press, 71-76.

Winkworth, A.L., Davis, P.J., Ellis, E. & Adams, R.D. (1994) Variability and consistency in speech breathing during reading: lung volumes, speech intensity, and linguistic factors, *Journal of Speech and Hearing Research*, 37, 535-556.

Samenvatting (Summary in Dutch)

In de onderstaande samenvatting heb ik geprobeerd om mijn onderzoek in zo eenvoudig mogelijke termen te beschrijven, met als voornaamste doel dat het begin ook voor niet-ingewijden begrijpelijk zou zijn. Naarmate de samenvatting vordert, wordt het verhaal noodzakelijkerwijs gedetailleerder en specialistischer, en daardoor ook complexer. Toch hoop ik dat deze samenvatting ook de leek een idee geeft van het onderzoek dat beschreven is in dit proefschrift.

Op het moment dat u dit leest communiceren we met elkaar via het schrift. Geschreven taal is een belangrijk hulpmiddel bij communicatie, net zoals gesproken taal. Als twee mensen met elkaar communiceren, zal het vaak voorkomen dat de ene persoon (de schrijver of de spreker) een boodschap wil overbrengen naar de andere persoon (de lezer resp. de luisteraar). Indien voor de communicatie spraak gebruikt wordt, dan codeert de spreker deze boodschap in een spraaksignaal en de luisteraar probeert de boodschap te extraheren uit het spraaksignaal. De manier waarop een boodschap overgebracht wordt van een spreker naar een luisteraar kan weergegeven worden met een zogenaamde spraakketen. Ofschoon verschillende stadia onderscheiden kunnen worden in deze spraakketen, zullen we ons hier voornamelijk beperken tot spraakproduktie, d.w.z. het omzetten van een linguïstische boodschap in een spraaksignaal. Hoe dat in zijn werk gaat wordt hier kort uitgelegd.

Het perifere spraakproduktiesysteem kan onderverdeeld worden in drie onderdelen, namelijk (1) het ademhalingssysteem, (2) de stembanden en de spieren die invloed hebben op de eigenschappen van de stembanden, en (3) keel-, mond- en neusholte (zie figuur 3 op blz. 11). Het ademhalingssysteem kan beschouwd worden als de energiebron tijdens spraakproduktie. Verschillende ademhalingsspieren worden gebruikt om het longvolume en de druk in de longen te controleren. De belangrijkste functie van het ademhalingssysteem is om zuurstof in en kooldioxyde uit de longen te pompen. Bij normale ademhaling duurt in- en uitademen ongeveer even lang, maar als we spreken is de fase waarin ingeademd wordt meestal veel korter dan de fase waarin uitgeademd wordt. Dit is nuttig omdat in de meeste talen spraak uitsluitend geproduceerd wordt tijdens het uitademen.

Een belangrijk begrip bij ademhalen en spraakproduktie is de luchtdruk. Lucht verplaatst zich van plaatsen met een hoge druk naar plaatsen met een lage druk. Tijdens inademen willen we dat er lucht de longen instroomt. Dit kunnen we bewerkstelligen door de ademhalingsspieren te gebruiken om het volume van de longen te vergroten. Hierdoor wordt de druk in de longen kleiner dan de druk in de atmosfeer, en indien de weg tussen buitenlucht en longen niet afgesloten is, zal lucht de longen instromen. Bij uitademen zullen we de ademhalingsspieren juist gebruiken om het longvolume te verkleinen. Hierdoor ontstaat een overdruk in de longen, en bijgevolg kan lucht de longen uitstromen.

Bij normaal ademen zal meestal niet veel geluid geproduceerd worden. Alleen bij heel diep of heel snel ademen zal een ruisachtig geluid te horen zijn. Bij spraakproduktie wordt de uitstromende lucht gebruikt om spraakgeluid te produceren. Hierbij kan onderscheid gemaakt worden tussen stemhebbende en niet-stemhebbende spraak. Bij niet-stemhebbende spraak trillen de stembanden niet. Hierbij kan bijvoorbeeld gedacht worden aan de niet-stemhebbende wrijfklanken (fricatieven), zoals f en s; of de niet-stemhebbende plofklanken (plosieven), zoals p, t en k. In dit onderzoek zijn met name de stemhebbende klanken onderzocht, en daarom zullen we ons verder beperken tot stemhebbende spraak.

De stembanden zitten in het strottehoofd (larynx), ongeveer op de hoogte waar bij mannen de adamsappel zit (zie figuur 3 op blz. 11, en figuur 4 op blz. 12). De trillingsfrequentie van de stembanden wordt de stembronfrequentie of grondtoonfrequentie genoemd. Hiervoor wordt verder het symbool F_0 gebruikt. De trillingsfrequentie van de stembanden bepaalt de toonhoogte van het spraakgeluid.

Als de stembanden gesloten zijn, is de luchtstroom minimaal, en als de stembanden open gaan, neemt de luchtstroom langzaam toe (zie figuur 2 op blz. 65, het onderste signaal). Door het openen en sluiten van de stembanden ontstaan de zogenaamde stembandpulsen, die ook wel luchtstroompulsen of luchtproppen genoemd worden. De ruimte tussen de stembanden, waar de lucht doorheen stroomt, wordt stemspleet of glottis genoemd, en daarom worden deze stembandpulsen soms ook de glottale pulsen of de glottale luchtstroom (U_g) genoemd. Een voorbeeld van deze glottale luchtstroom (U_g) is te zien in figuur 2 op blz. 65 (het onderste signaal). Te zien is dat deze luchtstroom periodiek toe- en afneemt. De frequentie waarmee dit gebeurt is F_0. Deze periodiciteit is ook waarneembaar in het spraaksignaal (bovenste signaal).

Bij het bespreken van de resultaten hieronder zullen een aantal begrippen en symbolen gebruikt worden, die daarom hier geïntroduceerd worden. De luchtdruk net onder de glottis wordt de subglottale druk (P_{sb}, subglottal pressure) genoemd, en de druk net boven de stembanden is de orale druk (P_{or}, oral pressure). Het drukverval over de glottis is de transglottale druk (P_{tr}, transglottal pressure): $P_{tr} = P_{sb} - P_{or}$. Het is bekend dat P_{sb} van invloed is op de trillingsfrequentie van de stembanden; een hogere P_{sb} leidt tot een hogere F_0. Dit kunt u zelf makkelijk uittesten door iemand die een aangehouden klank produceert snel op de borst of buik te duwen. Door deze duw wordt het longvolume kleiner, en bijgevolg nemen P_{sb} en F_0 toe.

Naast P_{sb} zijn er andere factoren die van invloed zijn op de trillingsfrequentie van de stembanden, namelijk de eigenschappen van de stembanden: spanning, stijfheid, lengte en massa. Een toename van de spanning en de stijfheid, en een afname van de lengte en de massa zorgen voor een hogere F_0. Bijvoorbeeld bij vrouwen en kinderen zijn lengte en massa meestal kleiner dan bij mannen. Daarom hebben vrouwen en kinderen vaak een hogere F_0 dan mannen. Maar behalve verschillen tussen personen, kunnen de eigenschappen van de stembanden ook variëren voor een en dezelfde persoon. De eigenschappen van de stembanden kunnen beïnvloed worden door de verschillende spieren rondom de larynx: de larynxspieren (zie figuur 4 op blz. 12).

De werking van de larynxspieren die in dit onderzoek bestudeerd zijn, wordt hier kort toegelicht. De larynxspier die waarschijnlijk de grootste invloed heeft op F_0 is de m. cricothyroideus (CT). Deze spier loopt van het ringvormig kraakbeen (cricoid) naar het schildvormig kraakbeen (thyroid, zie figuur 5 op blz. 13). Door contractie van de CT worden de stembanden verlengd, en hierdoor wordt F_0 verhoogd (zie figuur 5 op blz. 13). Twee andere larynxspieren die ook van invloed zijn op F_0 zijn de m. vocalis (VOC) en de m. sternohyoideus (SH). De VOC ligt in de stembanden.

Door contractie van de VOC neemt de interne spanning van de stembanden toe, wat leidt tot een verhoging van F_0. De SH bepaalt samen met enkele andere spieren de hoogte van de larynx (zie figuur 4 op blz. 12). In het algemeen is gevonden dat een toename van de activiteit van de SH een verlaging van F_0 tot gevolg heeft. Een mogelijke verklaring is de volgende: door het samentrekken van de SH zakt de larynx; hierdoor neemt de spanning van de stembanden in de verticale richting af, en dit leidt weer tot een verlaging van F_0.

F_0 is de trillingsfrequentie van de stembanden, zoals hierboven al verteld is. Het trillen van de stembanden zorgt er voor dat de geproduceerde spraak stemhebbend is. Wat voor klank geproduceerd wordt, is afhankelijk van de vorm van de ruimtes tussen de stembanden en de buitenlucht: het spraakkanaal, dat bestaat uit keel-, mond- en neusholte (zie figuur 3 op blz. 11). Deze ruimtes fungeren als een akoestisch filter. Zij filteren het geluid dat bij de stembron ontstaat, en zorgen daarmee voor de klankkleur (timbre) van het spraaksignaal. Als u zelf een aangehouden (monotone) klank produceert, en langzaam verandert van de ene naar de andere klinker, zult u merken dat hierbij met name de positie van kaak, tong en lippen verandert. Kaak, tong en lippen zijn belangrijke articulatoren die van invloed zijn op de vorm van het spraakkanaal, en daardoor op het filter dat het stembron geluid filtert.

Hierboven is in het kort uitgelegd hoe spraak geproduceerd wordt. De aldus geproduceerde zinnen bestaan uit woorden, woorden bestaan uit lettergrepen, en lettergrepen bestaan weer uit fonemen. Fonemen kunnen onderverdeeld worden in klinkers en consonanten (mede-klinkers). Fonemen kunnen beschouwd worden als de segmenten van spraak. Deze segmenten hebben bepaalde eigenschappen, die door middel van segmentele kenmerken beschreven kunnen worden. Grofweg kan gezegd worden dat de segmenten de inhoud van een zin bepalen.

Maar vaak is de manier waarop een zin uitgesproken wordt minstens even belangrijk als de inhoud van de zin. Denk hierbij bijvoorbeeld aan de vele verschillende manieren waarop het woordje "nee" uitgesproken kan worden: als een bewering of als een vraag, sarcastisch, blij, resoluut of nerveus. In al deze gevallen bevat het woord dezelfde twee segmenten, maar toch zal het in de genoemde gevallen meestal heel anders klinken en is de boodschap die de spreker over wil brengen naar de luisteraar verschillend. Deze variaties in het spraaksignaal vallen in de categorie prosodie, en kunnen beschreven worden met supra-segmentele oftewel prosodische kenmerken. Prosodische kenmerken zijn bijvoorbeeld klemtoon, intonatie, luidheid, stemkwaliteit, ritme, tempo en duur.

Wat zijn nu de eigenschappen van het spraaksignaal die van belang zijn voor prosodie? Hierboven is al vaak F_0 genoemd, de herhalingsfrequentie van het spraaksignaal. Het is niet toevallig dat F_0 al vaak genoemd is. F_0 is tot nu toe veruit het meest onderzocht en hierover is dan ook het meest bekend. Maar er zijn ook andere eigenschappen van het spraaksignaal die belangrijk zijn zoals intensiteit, spectrale inhoud en duur. In het onderzoek dat in dit boek beschreven wordt, zijn naast F_0 ook intensiteit (IL: Intensity Level) en spectrale inhoud van het spraaksignaal onderzocht. Deze drie parameters worden hier de prosodische parameters genoemd. Duur is niet onderzocht.

Het doel van het onderhavig onderzoek was om meer duidelijkheid te krijgen over de manier waarop prosodie in spraak geproduceerd wordt. De twee belangrijkste fysiologische factoren bij de produktie van prosodie zijn waarschijnlijk: (1) de eigenschappen van de stembanden, met name de spanning, stijfheid, lengte en massa van de stembanden, en (2) de luchtdruk onder de stembanden (P_{sb}).

In veel handboeken staat dat CT, VOC, SH en P_{sb} belangrijke factoren zijn voor de controle van F_0, en dat P_{sb} de belangrijkste factor is in de controle van intensiteit. Over de invloed van de larynxspieren op intensiteit, en over de manier waarop de spectrale inhoud van het spraaksignaal gevarieerd wordt, is veel minder bekend. Verder zijn de bovenstaande relaties meestal onderzocht voor aangehouden klinkers. Hierdoor weten we bijvoorbeeld wel dat CT, VOC, SH en P_{sb} allemaal F_0 kunnen beïnvloeden, maar het is niet (geheel) duidelijk hoe deze fysiologische mechanismen samenwerken bij de produktie van F_0 voor lopende spraak. Daarom is ook dit laatste aspect onderzocht.

Voor het onderzoek werden enkele van de bovengenoemde fysiologische signalen (zoals CT, VOC, SH, P_{sb}, P_{or} en P_{tr}) gemeten bij drie proefpersonen. Naast de genoemde fysiologische signalen werden ook nog een aantal andere signalen opgenomen tijdens de experimenten, namelijk het spraaksignaal, het longvolume, en het zogenaamde electroglottogram. Van een vierde proefpersoon zijn alleen de drie laatstgenoemde signalen geregistreerd.

Tijdens de experimenten werd iedere uiting een aantal malen herhaald. Alvorens de relatie tussen de bovengenoemde signalen onderzocht kon worden, moesten de signalen gemiddeld worden. Rechtstreeks middelen van de signalen was niet mogelijk, om redenen die genoemd zijn in hoofdstuk 2. Daarom is een methode ontworpen die het mogelijk maakt om een zinvol gemiddelde te bepalen voor de gemeten signalen. Deze methode is gebaseerd op het "dynamic time warping" algoritme (zie hoofdstuk 2).

De gemeten fysiologische signalen (namelijk activiteit van de larynxspieren en de druksignalen) beïnvloeden in eerste instantie het trillingsgedrag van de stembanden. Deze veranderingen leiden weer tot veranderingen in de glottale pulsen, die op hun beurt het spraaksignaal beïnvloeden. Gezien deze causale verbanden ligt het voor de hand om eerst te onderzoeken wat het effect is van de fysiologische signalen op het stembronsignaal. Om dit te kunnen doen moet het stembronsignaal bepaald en geparametriseerd worden. Het stembronsignaal kan geschat worden uit het spraaksignaal met behulp van een techniek die “invers filteren” heet. Het is mogelijk om de stembronparameters rechtstreeks te berekenen uit dit stembronsignaal, maar omdat het stembronsignaal vaak stoorsignalen bevat, zal de fout in de aldus berekende parameters vaak erg groot zijn.

Daarom is in het kader van dit onderzoek een andere methode ontwikkeld voor het parametriseren van het stembronsignaal. In deze methode wordt gebruik gemaakt van een model van de stembandpulsen. Dit model is te zien in figuur 1 op blz. 64. In deze figuur is het bovenste signaal U_g (de glottale puls), en het onderste signaal de afgeleide van U_g (dU_g). Dit stembronmodel wordt het LF-model genoemd omdat het voorgesteld is door Liljencrants en Fant. De bijbehorende stembronparameters worden de LF-parameters genoemd, waarvan er enkele hier kort besproken worden.

Op tijdstip t=0 gaan de stembanden open en neemt de luchtstroom langzaam toe. Op tijdstip T_p is de luchtstroom (de grootte van de puls U_g) maximaal en heeft de waarde U_0. Daarna gaan de stembanden weer sluiten en neemt de luchtstroom weer af. Op tijdstip T_e is dU_g minimaal en heeft de waarde $-E_e$. Op tijdstip T_c zijn de stembanden weer geheel gesloten, en is de luchtstroom weer minimaal. De parameter T_n is een maat voor de scheefheid van puls U_g (zie figuur 1 op blz. 64). T_0 ten slotte is de lengte van een volledige periode ($F_0 = 1/T_0$).

Door gebruik te maken van dit LF-model kunnen de LF-parameters als volgt berekend worden. Voor iedere periode wordt een combinatie van LF-parameters gezocht die er voor zorgen dat het LF-model zo goed mogelijk lijkt op het dU_g (zie figuur 2 op blz. 65, het middelste signaal). Van periode tot periode kan de vorm van de glottale puls (dU_g) verschillen, en dan zullen ook de optimale LF-parameters verschillen. Om deze optimale LF-parameters automatisch te kunnen bepalen wordt gebruik gemaakt van mathematische optimalisatie algoritmen. De ontwikkeling van deze methode is beschreven in de hoofdstukken 3 tot en met 6.

Deze methode is gebruikt om de LF-parameters te berekenen voor alle periodes van een aantal uitingen. Vervolgens werden deze data gebruikt om de relatie tussen fysiologische signalen, stembronparameters en de prosodische parameters te onderzoeken. De resultaten van dit onderzoek worden beschreven in hoofdstuk 3 en 4. De belangrijkste resultaten zullen hier kort samengevat worden. Voor de duidelijkheid dient nogmaals gezegd te worden dat de stembanden alleen trillen bij stemhebbende spraak, en dat daarom alleen voor stemhebbende spraak de LF-parameters bepaald kunnen worden (dat wil zeggen zowel voor klinkers als voor stemhebbende consonanten).

Het algemene gedrag van de onderzochte variabelen is als volgt (zie figuur 3 op blz. 66). Voor klinkers is P_{or} ongeveer nul, en bijgevolg is P_{tr} ongeveer gelijk aan P_{sb}. Bij consonanten is er meestal een vernauwing in het spraakkanaal, en daardoor wordt P_{or} groter en P_{tr} kleiner. Als P_{tr} kleiner wordt, wordt de kans dat de stembanden ophouden met trillen groter. Om er voor te zorgen dat de stembanden blijven trillen tijdens de stemhebbende consonanten (bij een lagere P_{tr}), worden daarom vaak een aantal aanpassingen gedaan met behulp van de larynxspieren: de spanning in de stembanden wordt verlaagd en de afstand tussen stembanden wordt vergroot (abductie). Als gevolg hiervan worden de LF-parameters T_a en T_n groter. Doordat P_{tr} en de spanning in de stembanden afneemt zal F_0 kleiner worden, en dus zal T_0 (= $1/F_0$) groter worden. Ofschoon de spanning in de stembanden kleiner geworden is, is de daling in P_{tr} meestal zo groot dat de amplitude van de trilling van de stembanden afneemt. Hierdoor zullen ook U_0, E_e en IL afnemen.

De LF-parameters zijn onder te verdelen in twee groepen, te weten de tijdsparameters en de amplitudeparameters. Binnen beide groepen zijn de correlaties tussen de variabelen positief, en tussen beide groepen zijn de correlaties negatief. Hieronder worden eerst de relaties tussen de tijdsparameters beschreven, en vervolgens de relaties tussen de amplitudeparameters.

Gedurende een transitie van een klinker naar een stemhebbende consonant neemt F_0 meestal af, zoals hierboven beschreven is, en dus neemt T_0 toe. Gemiddeld nemen samen met T_0 alle andere tijdsparameters ook toe, maar de mate waarin dat gebeurt verschilt. De duur van het eerste gedeelte van de LF-puls (T_e) neemt lineair toe met T_0. Maar de toename in T_i en T_p is minder dan lineair, terwijl de toename in T_n en T_a meer dan lineair is. De scheefheid van de glottale puls is evenredig met $1/T_n$, en daarom zal deze scheefheid gemiddeld afnemen tijdens een transitie van een klinker naar een stemhebbende consonant.

De relatie tussen IL en P_{tr} die gevonden is in hoofdstuk 3 is:

$$IL = 41.6 + 30.3 \log P_{tr} \qquad \text{(aantal punten = 293, r = 0.923).}$$

Omdat de correlatiecoëfficiënt tussen IL en P_{tr} erg hoog is, en omdat de invloed van de gemeten larynxspieren op IL erg klein is, kan worden geconcludeerd dat P_{tr} veruit de belangrijkste factor is voor IL. In hoofdstuk 3 is verder aangetoond dat voor de controle van F_0 P_{tr} ook belangrijker is dan P_{sb}. Het drukverval over de glottis (P_{tr}) kan daarom beschouwd worden als de drijvende kracht tijdens stemgeving (fonatie). Voor de volledigheid dient vermeld te worden dat de invloed van de larynxspieren op F_0 groter is dan op IL.

Het resultaat dat IL vrijwel geheel door P_{tr} bepaald wordt, maakt het interessant om te bestuderen hoe IL beïnvloed wordt door P_{tr} via de stembronparameters. De relaties tussen P_{tr}, U_0, E_e en IL werden daarom bestudeerd voor vier herhalingen van een spontane uiting. In totaal werden de LF-parameters berekend voor 613 perioden. Voor deze data werden hoge correlaties gevonden tussen de vier amplitudeparameters P_{tr}, U_0, E_e en IL.

Nadat de relatie tussen de fysiologische signalen, de stembronparameters en de prosodische parameters onderzocht was, werd vervolgens bestudeerd op welke manier de gemeten fysiologische signalen samenwerken bij de produktie van intonatie. Intonatie kan beschouwd worden als het verloop van F_0 gedurende een uiting, waarbij de nadruk ligt op de linguïstisch relevante aspecten van F_0. De resultaten van dit onderzoek zijn beschreven in hoofdstuk 8 en 9. Bij dit onderzoek hebben we niet alleen eigen data gebruikt, maar ook gegevens uit de literatuur.

In de data zijn vaak snelle veranderingen in F_0, P_{sb}, CT, VOC en SH te zien. Met andere woorden, het is duidelijk dat al deze signalen een lokale component hebben. Hierover zijn de meeste onderzoekers het nu wel eens. Maar een omstreden punt is nog steeds of deze signalen ook een globale component bezitten. Een langzame daling in F_0 gedurende een uiting wordt namelijk vaak gevonden, en men is het niet eens over de fysiologische verklaring van deze langzame daling in F_0 (die ook wel declinatie genoemd wordt). Een langzame daling in P_{sb} werd vaak waargenomen, terwijl een gelijksoortige langzame variatie in de activiteit van de larynxspieren gedurende een uiting meestal niet gevonden werd. Daarom hebben de larynxspieren CT, VOC en SH geen lokale component in ons fysiologisch model van intonatie, en P_{sb} wel. Bijgevolg is in het door ons voorgestelde model de langzame daling in F_0 het gevolg van de langzame daling in P_{sb}, terwijl de lokale variaties in F_0 veroorzaakt worden door lokale variaties in P_{sb}, CT, VOC en SH.

Curriculum vitae

Helmer Strik was born on 18 May 1957 in Rosmalen, the Netherlands. In 1976 he finished secondary school (Atheneum-β). Subsequently, he studied physics at the University of Nijmegen. In April 1982 he passed the Bachelor of Science examination and in December 1985 he obtained his Master of Science degree (major subjects were general mechanics and electrodynamics). For his M.Sc. thesis he did research on the dimensional and dynamical aspects of the auditory neurons in the midbrain (torus semicircularis) of the grassfrog (Rana temporaria L.), at the Department of Medical Physics and Biophysics of the University of Nijmegen.

From September 1984 till August 1985 he received practical training in order to obtain a (first degree) qualification to teach physics. Subsequently, he taught physics in secondary school until December 1985.

On 16 December 1985 he started a three-year research project on the physiological control of intonation, which was carried out at the Department of Language and Speech of the University of Nijmegen. Part of this study was conducted at the Haskins Laboratories in New Haven. The present thesis reports on the research which was carried out within the framework of this project.

From May 1989 till August 1992 he worked on the ESPRIT (European Strategic Programme for Research and development in Information Technology) project POLYGLOT. This project comprised research on speech recognition and speech synthesis for seven languages of the European Community. Within this project he did research on knowledge-based speech recognition.

In August 1992 he was appointed at the Department of Language and Speech of the University of Nijmegen. Since then he has carried out research on voice source modelling and has been teaching phonetics, speech recognition and signal processing.