Detailed versus gross spectro-temporal cues for the perception of stop consonants

Citation for published version (APA):
Smits, R. L. H. M. (1995). *Detailed versus gross spectro-temporal cues for the perception of stop consonants*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR439117

DOI:
10.6100/IR439117

Document status and date:
Published: 01/01/1995

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Detailed versus gross spectro-temporal cues for the perception of stop consonants

Roel Smits

Detailed versus gross spectro-temporal cues for the perception of stop consonants

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof. dr. J.H. van Lint,
voor een commissie aangewezen door het College van Dekanen
in het openbaar te verdedigen
op vrijdag 23 juni 1995 om 14.00 uur

door

Roland Leendert Henry Marie Smits

geboren te Geldrop

Dit proefschrift is goedgekeurd door de promotoren
prof. dr. R. Collier
prof. dr. Y. Kamp

en de copromotor
dr. L. ten Bosch

"ba, da ga nie"
(Niek Versfeld, personal communication)

Acknowledgments

This thesis would not exist without the contribution of a number of people. I want to thank Lei Willems for creating the project, and for radiating a great feeling of trust from the very start. I thank René Collier for his help and support during the project and the writing of the thesis, communicating was always very easy. Without Louis ten Bosch the project would have taken a totally different turn. I enormously enjoyed our cooperation, it very much shaped my academic interest. I will not easily forget our "deep discussions" and "mutual fertilizations". I want to thank Aditi Lahiri, whose involvement in the project I have very much appreciated, and who has been very supportive and inspiring. I thank Mária Gósy for her help in putting the project on the proper phonetic track, and for our very pleasant cooperation. I thank Werner Verhelst for introducing me to - and getting me interested in - a number of speech processing techniques and issues. Many thanks to Louis Pols and Yves Kamp for thorough and inspiring discussions. I thank B. Yegnanarayana for a very pleasant and inspiring cooperation, and for highly original views on any topic imaginable. I want to thank Berry, Niek, Marc, Paul and Igor for being great roommates and/or fellow coffee addicts. Last but not least, I thank the members of *Zonder Fratsen*, without doubt the most exciting jazz-rock band in the universe.

Contents

Chapter 1

Introduction

The research presented in this thesis addresses a number of aspects relating to the perception of place of articulation of prevocalic stop consonants. This research topic has received much attention over the years, and a vast amount of literature is available. It is the primary purpose of this introductory chapter to summarize the findings of a number of relevant past studies and to reason what the present research hopes to offer in addition to the existing knowledge. To this end, we will first describe the general problem of stop-consonant perception. Next, we will present a very general qualitative model for the perception of stop consonants which meets most of the assumptions and findings reported in the literature. From the angle of the proposed general model, we will discuss 3 major models of speech perception: the Haskins Laboratories' motor theory, Blumstein and Stevens' theory of acoustic invariance, and Massaro and Oden's fuzzy-logical model of speech perception. We will indicate to what extent these theories are relevant to our research. Next, a number of experimental studies will be reviewed, again from the perspective of the general model. Finally, we will formulate the specific objectives of the research presented in this thesis.

1.1 The problem of stop-consonant perception

The problem of stop-consonant perception, or indeed of speech perception in general, has often been described as a problem of *variability and invariance* (e.g. Liberman and Mattingly, 1985; Perkell and Klatt, 1986; Klatt, 1989; Lahiri and Marslen-Wilson, 1991). Utterances which are associated with one and the same linguistic class[1], be it the distinctive feature[2], the phoneme, or larger units, show large acoustic variability due to factors such as phonetic context, syllabic stress, speaker characteristics (vocal-tract length, articulatory habits, etc.), speaking rate, transmission characteristics (background noise, reverberation, filtering, etc.), and coincidence (token-to-token variation). Thus, the mapping of the acoustic signals to the respective linguistic classes is far from simple.

The primary problem of variability due to phonetic context has been described in terms of the *overlapping* rather than *sequential* nature of the speech signal (e.g. Liberman *et al.*, 1967; Fant, 1973; Fowler, 1986; Lindblom, 1986). In contrast with, for example, printed text, the speech signal cannot be segmented into a sequence

[1] Only sub-lexical linguistic classes are considered here.

[2] Throughout this chapter we will use the term *distinctive feature* in a general sense, not discriminating between various proposed feature systems (e.g. Jakobson *et al.*, 1952; Chomsky and Halle, 1968; Stevens *et al.*, 1992).

of discrete units, each of which uniquely corresponds to one linguistic class. Instead, due to the continuous and overlapping movement of the articulators, each acoustic segment, however small, generally contains information regarding several neighboring phonemes.

It is in this respect - the acoustic variability with phonetic context - that the stop consonant seems to be the champion of all phonemes. The class of stop consonants is defined in the articulatory domain as those phonemes during the production of which a complete closure of the vocal tract is realized. The stop consonants of the Dutch language are /b, d, p, t, k/. In general, the production of an intervocalic stop will consist of the following phases:

1. Movement of the articulators toward closure position;
2. Closure;
3. Release;
4. Movement of the articulators toward the vowel-target position.

The major acoustic consequences of these 4 phases are:

1. Formant transitions from the previous vowel into closure;
2. Silence or low-frequency low-amplitude vocal murmur;
3. An initial transient, immediately followed or overlapped by a burst of frication noise;
4. Formant transitions, either voiced or aspirated, into the following vowel. (e.g. Fant, 1973; Henton *et al.*, 1992).

As will be discussed extensively later, perception studies and acoustic analyses have shown that (1) the acoustic information in all four phases is relevant for the perception of place of articulation, and (2) a majority of the acoustic structures in the four phases is highly variable with phonetic context. In addition, due to the rapid change of the acoustic properties of the vocal tract immediately prior to closure or after release, the resulting acoustic signal is highly dynamic, that is, its spectral content changes rapidly over time (e.g. Fant, 1973). This makes it relatively difficult to define, let alone measure, a consistent set of acoustic descriptors.

1.2 A general qualitative model for the perception of stop consonants

In this section we present a general qualitative model for the perception of stop consonants. The model is not intended to represent any novel viewpoints, but instead provides the author's synthesis of most of the assumptions and findings reported in the literature on the perception of stop consonants. Although the model is primarily intended to describe the perception of stop consonants, it is generally also applicable to the perception of various other consonants. The general setup of the model is similar to the "low-level" part of the model of lexical access from features (LAFF, e.g. Stevens *et al.*, 1992). It should, however, be stressed that our model does not incorporate any lexical effects.

Figure 1.1 presents a schematic outline of the model. Ovals represent input (bottom) and output (top) information, rectangles represent stages in the model

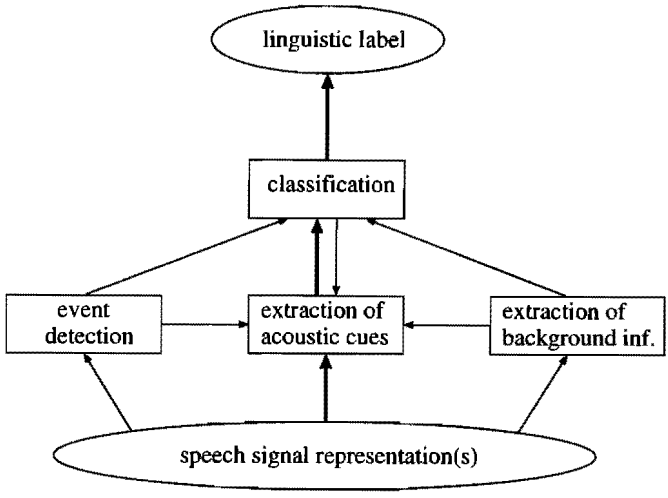

Figure 1.1: Flow diagram of the general qualitative model of stop-consonant perception. For further explanation, see text.

where a distinct type of processing takes place. Thick arrows indicate the main flow of acoustic and linguistic information, which forms the core of the model. Thin arrows indicate additional "control" information, which is used to adjust processes in the central branch of the model. We will describe each of the processing stages below, working from the speech signal representation(s) upwards.

1.2.1 Speech signal representation(s)

The incoming speech signal undergoes one or several initial transformations before any acoustic properties are extracted. In most studies, some form of spectro-temporal transformation, like the spectrogram, is assumed to provide the basic representation for all subsequent processing. In some studies, on the other hand, several sources of information are assumed to be available, such as the spectrogram plus the speech waveform envelope (Cole and Scott, 1974), or a combination of a spectro-temporal representation which enhances formants and one which enhances sudden changes in energy (Seneff, 1988).

During the last 15-odd years, the representation issue has received much attention from two major angles. In one approach, representations are sought which reflect the properties of the human hearing system more faithfully than the spectrogram. The resulting "auditory" speech representations vary in their degree of sophistication from relatively basic critical-bands-like filterbanks (e.g. Searle *et al.*, 1979; Schouten and Pols, 1981; Suomi, 1985, 1987), to models incorporating phase-locking and adaptation characteristics of the auditory nerve cells (e.g. Delgutte, 1986; Seneff, 1988; Fox and Feth, 1992). In another approach, the suitability of the

traditional quasi-stationary analysis, such as the spectrogram and linear prediction, for the analysis of the rapidly time-varying speech signal is questioned. Moreover, in some cases it is claimed that the apparent lack of acoustic invariance in stop consonants can be attributed to the inherent unfitness of the classical analysis methods for these signals (e.g. Nathan and Silverman, 1991; Velez and Garudadri, 1992). Subsequently, novel non-stationary methods are proposed (e.g. Nathan *et al.*, 1991; Pitton *et al.*, 1994). Surprisingly, despite the extensive research effort, quantitative data on the accuracy of the traditional quasi-stationary techniques are hardly available. Chapter 2 is devoted to this issue and focuses on the accuracy of measuring formant frequencies in highly dynamic speech signals, such as stop consonants.

Finally, we remark that the separation between representation and cue measurement, although conceptually convenient, is formally difficult to make (Fukunaga, 1972). In the process of measuring a cue, several subsequent signal transformations are usually made, like Fourier transformation, spectral envelope calculation and peak picking. It is not evident where the boundary between representation and cue measurement is supposed to be situated. In this thesis we will adopt the pragmatic definition that the boundary between representation and cue extraction lies at the point before which the signal transformations are shared by all cue-measurement procedures, and beyond which signal transformations are specialized for each cue measurement.

1.2.2 Extraction of background information

The background-information module in Figure 1.1 comprises a number of highly complex processes which extract information from the speech signal concerning transmission conditions, speaker characteristics, and speaking rate. As indicated earlier, it is well-known that transmission conditions such as background noise, reverberation and filtering have a large influence on speech perception (e.g. Miller and Nicely, 1955; Duquesnoy and Plomp, 1988). However, the human hearing system seems to be able to partly cancel background noise and reverberation and selectively attend to one voice (e.g. Darwin and Gardner, 1987). As far as speaker characteristics are concerned, experiments have shown that confusion in consonant perception increases with increasing speaker uncertainty (using several speakers within one listening session in random order) (e.g. Mullennix and Pisoni, 1990). This indicates that listeners "tune in" on a certain speaker, that is, cues are measured or interpreted differently for different speakers. Rand (1971) showed that listeners perform a vocal-tract size normalization in the perception of synthetic stops. Finally, a number of acoustic characteristics of stop consonants systematically change with changing speaking rate (e.g. Crystal and House, 1988). Listeners seem to adapt to these changing cues (e.g. Miller, 1981; Sommers *et al.*, 1992).

As indicated in Figure 1.1, the background-information module gets input from the speech representation. It sends output to the cue-extraction module, indicating that the details of the cue extraction may depend on information concerning transmission, speaker characteristics and speaking rate. Similarly, the outgoing connection to the classification module indicates that the classification process may

be adjusted to actual background information, e.g. by shifting category boundaries.

1.2.3 Event detection

Before any acoustic cue can be measured on the speech signal in a useful way, it must be indicated *where* (along the time dimension) the measurement is to be made. The event-detection module performs an initial "bootstrap" of the speech signal by locating important acoustic events, such as the onset and offset voicing, plosive burst onset, and moments of maximum spectral change. Although in many acoustic and perception studies it is often implicitly assumed that the perceptual system has the locations of these events at its disposal, some studies make the notion of an event detector explicit (e.g. Searle *et al.*, 1979; Blumstein and Stevens, 1981; Kewley-Port and Luce, 1984; Furui, 1986). The most complete and detailed account of an event detector is, however, given in Stevens *et al.* (1992), Stevens (1994), and Liu (1993). They describe an automatic "landmark" detector, which locates instances in the speech signal which are associated with major articulatory events, such as consonantal closure and release. In addition the landmark detector makes a broad phonetic classification of the detected landmark, such as "abrupt consonantal closure".

In our model, the event detector sends information to the cue extraction module concerning where and which cues should be measured. Which cues should be measured depends on the broad phonetic class of the acoustic event. For instance, different place-cue measurements may be needed for a final stop than for an initial stop. The event detector is also assumed to send information to the classification module, so the classification can be tuned to interpret the incoming cues in accordance with the broad phonetic class at hand. The event detector receives no background information as it is assumed to be able to operate more or less independently of speaker identity, speaking rate, and background conditions.

1.2.4 Extraction of acoustic cues

The cue-extraction module measures a number of acoustic cues on the speech representation(s). A great number of acoustic-phonetic studies have addressed the issue which acoustic structures in stop consonants correlate with place of articulation, and thus form potential perceptual cues. Besides, a large body of perceptual research has been devoted to the issue whether and how these potential place cues are actually used by listeners. These acoustic and perceptual studies will be discussed in detail later in this introductory chapter.

The cue-extraction module is assumed to receive input from the speech representation(s), the background-information module, the event detector, and the classifier. As discussed earlier, background information may be used to adjust details of the cue-measurement procedure. The event detector supplies information concerning where and what cues are to be measured. Finally, the classifier provides "top-down" information concerning phonological context, that is, linguistic labels that have been determined earlier. Just like the broad phonetic information determined

by the event detector, the phonological context may help to determine what cues are to be extracted from the speech signal.

The cue-extraction module outputs the measured cue values to the classifier, which will use this information to assign linguistic labels.

1.2.5 Classification

The classification module determines linguistic labels on the basis of incoming information from the cue extractor, the event detector, and the background-information module. The cue extractor supplies the actual values of a number of acoustic cues, which form an acoustic input vector to some type of classification mechanism. The event detector and background-information module supply additional information which is used to adjust details in the classification process. For example, classification boundaries may be shifted according to whether the event is consonantal closure or opening, or to suit the speaker's vocal tract length or speaking rate.

Many perception experiments have provided descriptive data on the classification process, predominantly by describing the shape of the identification functions on place-of-articulation continua where one or two cues were systematically varied. However, very basic issues in the classification process are still unsettled, such as what formal type of classification strategy is used, and what type of linguistic labels are actually determined. With respect to the first aspect, Kuhl is investigating the possible role of *prototypes* in speech perception. While the prototype seems to be a useful concept in modeling vowel perception (e.g. Kuhl, 1992), preliminary investigations do not give clear support for the use of prototypes in stop-consonant perception (Davis and Kuhl, 1992). Pisoni (1992) recently argued against the assumption of having a single idealized prototype per phonetic category. Instead, he promoted the concept of multiple prototypes, or exemplars, per category, covering variability due to phonetic context and speaker identity. In the research reported in later chapters we will refrain from making assumptions regarding category prototypes. Instead we will make the less restrictive assumption that perception is basically governed by category *boundaries*. This assumption, which is (more or less implicitly) also made by a number of other authors (e.g. Lahiri *et al.*, 1984; Jongman and Miller, 1991), is a classification-theoretic expression of the fundamental axiom that linguistic communication is achieved by transmitting *distinctions* rather than *idealized symbols* (e.g. Jakobson *et al.*, 1952).

Another issue which appears to be far from settled is the type of labels which form the output of the classifier. Among the proposed linguistic units are distinctive features, allophones, segments, and syllables (e.g. Wickelgren, 1976). It is evident that the choice of the output units will have important implications for the actual structure of the classifier. In our study, it is assumed that the output of the classifier is in terms of the place-of-articulation features [labial], [dental], [velar], or in terms of the segments /b, d, p, t, k/. Note that for our purpose these two options coincide, as we exclusively deal with the perception of place of articulation.

1.3 Review of three theories of speech perception

In this section, three models of speech perception are discussed against the background of the general qualitative model described earlier (see Figure 1.1). First, we will discuss two models which have been highly influential in research on speech perception, the *motor theory* (Liberman *et al.*, 1967; Liberman and Mattingly, 1985), and the *theory of acoustic invariance* (Blumstein and Stevens, 1981; Stevens and Blumstein, 1981). The third model that will be addressed is the *fuzzy-logical model of speech perception* (Oden and Massaro, 1978; Massaro and Oden, 1980). Although this model has been less influential than the motor theory and the invariance theory, it is discussed here because it will help to put our own research in proper perspective.

1.3.1 The motor theory

The Haskins motor theory has been expounded in two major theoretical papers: Liberman *et al.* (1967), and Liberman and Mattingly (1985). The primary inspiration to the original formulation of the theory was the finding, in the early Haskins research, that the acoustic structures that cue one and the same phoneme can be vastly different in different phonetic contexts (e.g. Cooper *et al.*, 1952; Liberman *et al.*, 1954). Furthermore, early experiments showed that a sequence of context-free acoustic segments, each corresponding to a single phoneme, could not convey an intelligible message to listeners, unless the rate of information transmission was greatly reduced, like in Morse code (e.g. Cooper, 1950; Harris, 1953). This seems to indicate that speech is not just coarticulated due to certain articulatory restriction, it *needs* to be coarticulated in order to be interpretable for listeners. These observations led to the concept of the "speech code": in the process of transforming a message into sound, the message is *encoded* in acoustic structures in a non-trivial way; perception of speech is the process in which the complex acoustic code is *decoded* and the original message is recovered (Liberman *et al.*, 1967).

The motor theory explicitly makes the following major claims.

1. Perception and production of speech are intimately related through common processing strategies and representations.
2. The human nervous system contains a special "speech module", which plays a central role in the perception as well as production of speech. Perception of speech takes place essentially by correlating the incoming neural patterns from the auditory system with outgoing neural patterns that control the articulators.
3. The central linguistic unit of speech perception is the phoneme.
4. Invariant structures corresponding to the phoneme do not exist at the acoustic level.
5. Invariant structures corresponding to the phoneme exist at the level of the neuro-motor commands (original formulation of the theory, Liberman *et al.*, 1967).

6. Invariant structures corresponding to the phoneme exist at the level of the intended articulatory gestures (revised formulation of the theory, Liberman and Mattingly, 1985).

Before we discuss these claims from the angle of our general model we make two critical comments. First of all, it is important to realize that part of the claims made in the motor theory refer to the *neuro-physiological* level of speech perception, rather than the *signal-analytical* level, which causes confusion. On the one hand, it is claimed that the problem of the acoustic-to-linguistic mapping cannot be solved at a signal-analytical level. On the other hand, however, it is claimed that this problem is apparently solved by the special speech module. Thus, as Klatt (1989) aptly reasons, the difficult problem of the acoustic-to-linguistic mapping is avoided by stating *where* it takes place, rather than *how* it takes place. The signal-analytical formulation of the problem remains open.

Secondly, in contrast to the two theories which will be discussed next, the motor theory is an entirely *qualitative* theory. The theory only explains or predicts general trends rather than quantitative data.

Let us now briefly review the motor theory from the angle of the proposed general qualitative model. First of all, we note that in their discussion of the acoustic structure of the speech signal, the motor-theory proponents have always concentrated on relatively "detailed" spectro-temporal properties, like formants, spectral peaks in the release burst, etc. Obviously, this (implicit) assumption is relevant to the cue-measuring module in our model.

Secondly, it is claimed that the acoustic cues are not directly mapped onto the linguistic class, that is, the phoneme, but are instead mapped onto the intended articulatory gesture (in the revised motor theory). The phoneme identity is only subsequently derived from the intended gestures. In terms of our model this means that the classification module actually performs two classifications: one from the acoustic cues to the intended gestures, and one from the intended gestures to the phoneme.

Finally, the high variability of acoustic cues with phonetic context and the resulting inevitably high complexity of the decoding process has always been a central issue in the motor theory. We can only tentatively translate this view into three assumptions in terms of our general model. First of all, it is claimed that no single acoustic structure has an invariant relation with the phoneme. Obviously, then, more than one acoustic structure, perhaps many, are measured by the cue-extraction module in order to identify the phoneme. Next, the emphasis on phonetic context may be translated into a large time window within which the cues to a single phoneme are extracted. Finally, the complexity of the decoder can be translated into the complexity of the classification module, e.g. in terms of the shapes of the classification boundaries or multiple prototypes per category.

1.3.2 The theory of acoustic invariance

The theory of acoustic invariance was introduced in a series of experimental studies (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980) and two theoretical papers (Blumstein and Stevens, 1981; Stevens and Blumstein, 1981). One of the novelties in the approach was the introduction of a new set of acoustic cues which capture certain integrative spectro-temporal properties of the speech signal. It was argued that the preoccupation in previous studies with detailed acoustic structures like release burst and formant transitions was guided by the implicit assumption that "... the perceptual mechanism operates on the speech signal in much the same way as the eye examines attributes of the sound spectrogram", Blumstein and Stevens (1980, p. 648). Based on the acoustic theory of speech production (Fant, 1960), and previous observations by, amongst others, Halle *et al.* (1957) and Zue (1976), gross spectral characteristics of the initial 20-odd ms after consonantal release were hypothesized to contain context-independent information for place of articulation of stop consonants. The onset spectra of labial, coronal, and velar stops were assumed to have *diffuse falling*, *diffuse rising*, and *compact* characteristics, respectively. The terms *compact* and *diffuse* refer to the spectrum having respectively one, or no, pronounced energy concentrations. The terms *falling* and *rising* indicate that the spectral energy decreases or increases, respectively, with increasing frequency.

The invariance theory makes the following claims.

1. The distinctive feature, rather than the phoneme, is viewed as the central linguistic unit for perception.
2. An invariant acoustic property is associated with each value of each of the distinctive features. The term *invariant* was originally defined in the most far-reaching sense, that is, independent of phonetic context, speaker identity, speaking rate, transmission conditions and language.
3. The invariant acoustic structures are of an "integrative" nature, which means that for the measurement of an acoustic property, a relatively large window is used across the time as well as frequency dimension. Acoustic structures such as formant onset frequencies typically do not qualify as integrative properties, and an explicit distinction between release burst and voiced formants is not made. Although the acoustic properties associated with place of articulation in stop consonants are the most elaborated, acoustic properties for other features have been proposed as well (e.g. Stevens, 1980; Mack and Blumstein, 1983).
4. The acoustically invariant properties are the *primary* perceptual cues. The term "primary" has two components, namely innate and most important. The infant is born with the ability to extract the invariant cues and is thus able to cope with the interpretation of linguistic distinctions in all languages. During infancy, the listener learns to use, beside the invariant cues, context-dependent cues, such as formant transitions, which in natural speech always accompany the invariant cues. The context-dependent cues are, however, merely "sec-

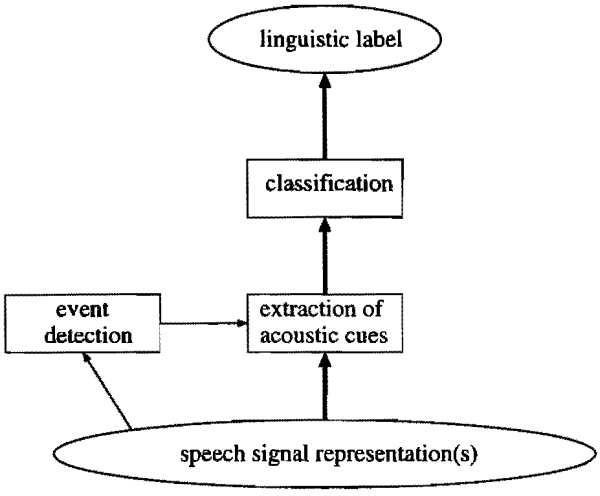

Figure 1.2: Flow diagram of the theory of acoustic invariance from the perspective of the general qualitative model of stop-consonant perception.

ondary" cues, and are only of decisive importance when the primary cues somehow cannot provide unequivocal information.

Due to criticism from various angles and new experimental findings, the theory of acoustic invariance has undergone revisions in certain aspects. First of all, the claim that the gross shape of the static onset spectrum provides the primary cue to place-perception in stop consonants has been abandoned. Instead, the *change* in the gross spectral characteristics during the first few tens of milliseconds after release has been put forward as the primary place-cue (Lahiri *et al.*, 1984). Secondly the claim that distinctive features are cued by the same acoustic properties in all languages has been weakened. Cross-linguistic acoustic analyses have shown that the acoustically invariant property associated with a certain feature is only maintained in a language if the feature actually plays a distinctive role (Jongman *et al.*, 1985).

The high specificity of the theory of acoustic invariance at the signal-analytic level allows it to be rather precisely translated in terms of our general model. In its most rigorous form, the theory can be viewed as the most basic instance of the general model, as illustrated in Figure 1.2.

The background-information module has disappeared, as well as the information flow from the classifier to the cue-measurement module and from the event detector to the classifier. Furthermore, the cue-measurement module is claimed to measure one integrative acoustic cue per distinctive feature. The classifier is relatively simple (the cue space has low dimensionality), and outputs distinctive features.

If we compare the theory of acoustic invariance with the motor theory from the perspective of our general model, two differences are striking. First of all, the type of cues that are hypothesized to be used by listeners are different. In the motor theory, relatively detailed, spectrographically explicit structures are viewed as the relevant cues, while the invariance theory emphasizes the role of gross, integrative acoustic properties. As the distinction is very relevant for the research presented in this thesis, it is useful to define the terms *detailed* and *gross* cues more formally here. We define a *detailed* spectro-temporal cue as the result of an acoustic measurement with high resolution in the time domain (in the order of 1 ms), or in the frequency domain (in the order of 50 Hz). Examples of detailed cues are formant frequencies, e.g. at voicing onset, formant transition rates, or certain release-burst characteristics such as the length of the burst. *Gross* spectro-temporal cues, on the other hand, are defined as the result of integrative acoustic measurements with resolutions in time and/or frequency above, say, 10 ms and 500 Hz, respectively. Typical gross cues are the global spectral tilt and compactness of a spectrum, and the evolution of these quantities over time. Furthermore, both types of cues can be of a static as well as dynamic nature.

A second obvious difference is the hypothesized complexity of the perception process. In the invariance theory, linguistic classification is assumed to take place on the basis of one cue per distinction, which is invariant across context, speaker, speaking rate and language. In the motor theory, linguistic classification takes place on the basis of multiple cues per distinction, all of which are highly variable with context, speaker, speaking rate and language.

Finally, it should be noted that, like in the motor theory, speech production plays an important role in the invariance theory. We have discussed that in the motor theory, speech perception is viewed as the "reverse" of speech production, in the sense that the articulatory gesture is recovered which could have produced the perceived speech sound. In the invariance theory, speech production theory is used to derive robust, invariant acoustic properties associated with a distinctive feature. Because the distinctive features have firm roots in articulation, the perceptual mechanism of extracting these robust acoustic properties is actually very similar to recovering articulatory information. In contrast with the motor theory, however, no claims are made on neurological modules or processes.

1.3.3 The fuzzy-logical model of speech perception

The fuzzy-logical model of speech perception (FLMSP, Oden and Massaro, 1978; Massaro and Oden, 1980) is an application of the general fuzzy-logical model of perception, which has been used to model various types of human classification behavior, in particular spoken and written text (e.g. Oden, 1979; Massaro and Friedman, 1990). In contrast with the two previously discussed theories, the FLMSP is not intended as a comprehensive linguistic or psychological theory of speech perception. Instead, the FLMSP provides a mathematical framework which is mainly used to test certain fundamental processing assumptions, such as whether or not acoustic

cues are measured independently on the speech signal. The model is discussed here because it is one of the very few mathematically fully developed categorization models which has been applied to speech perception. In chapter 3 of this thesis we will introduce a new general model of human categorization behavior which, like the FLMSP, is able to predict quantitative data, but differs from the FLMSP on a number of accounts.

Before we explain how the model works, we briefly discuss the general experimental procedure which is followed in Oden and Massaro (1978) and Massaro and Oden (1980). The two papers have large overlap. In both papers, the model is used to evaluate whether or not it is formally necessary to assume that two acoustic cues, which are relevant to both the place distinction and the voicing distinction, are measured *independently* on the speech signal. To this end, a two-dimensional synthetic stop-vowel continuum is created by varying VOT and the onset frequencies of F2 and F3 orthogonally. Listeners classify the stimuli at the four "corners" of the two-dimensional continuum reliably as /bae/, /dae/, /pae/, and /tae/, respectively. Next, the entire continuum is presented to listeners, who are instructed to classify the initial consonant as either B, D, P, or T. The results show that both acoustic cues (VOT and formant transitions) are used for perception of both distinctions, although VOT mainly cues voicing and the formant transitions mainly cue place of articulation.

In the FLMSP the basic assumption is made that each of the possible response classes is represented in long-term memory by prototypes, and that classification is based on the comparison of an incoming stimulus to each of the prototypes. The following steps are made in the model.

1. The levels of the acoustic parameters in the continuum are mapped onto an internal representation, which indicate the degree to which a particular acoustic property is present in the signal. With respect to this internal representation, it must be noted that the separation between *acoustic* features (cues) and *distinctive* features is somewhat obscured due to the terminology used. The fuzzy predicates representing the internal representation of the levels of VOT and formant onset frequencies are called respectively VOICED and LABIAL in Oden and Massaro (1978), while they are called SHORT VOT and HIGH F2-F3 TRANSITIONS in Massaro and Oden (1980). Secondly, it should be noted that each level of an acoustic parameter in the continuum receives a separate model parameter in the internal representation. Generally, after fitting the model, the resulting parameter values strongly suggest a sigmoid-like mapping of acoustic cue to internal representation (e.g. Massaro and Oden, 1980, Table V, p. 1006). Nevertheless, this rather obvious step is not explicitly modeled. In our opinion, this greatly reduces the generalizability of particular model fits, and leaves a gap in the complete trajectory from acoustic signal to perceived linguistic class.
2. The match of the stimulus to each of the prototypes is evaluated. Each response class is represented by a prototype, which is formalized in the form of

a fuzzy proposition. The fuzzy proposition is constituted of fuzzy predicates - representing the desired internal value of acoustic cues - linked by fuzzy-logical operators. For example, in Massaro and Oden (1980) the phoneme /t/ is represented by the fuzzy proposition

(HIGH F2-F3 TRANSITIONS) AND (NOT(SHORT VOT))

which can be translated into the mathematical expression

$$T = HFT \cdot (1 - SV)$$

where T represents the response strength for /t/, HFT is the internal representation of the onset frequencies of F2 and F3, and $(1 - SV)$ is the internal representation of the length of VOT (SV stands for "SHORT VOT", $(1 - SV)$ stands for "LONG VOT". T, HFT, and SV all have values in the interval $[0, 1]$.
Thus, for each stimulus, the response strength for all possible responses is determined by evaluating the match to each of the prototypes.

3. The response strength for each of the possible response classes is converted into response probabilities using the (unbiased) choice model of Luce (1963):

$$p_i = \frac{s_i}{\sum_j s_j}$$

where p_i represents the probability of choosing response class i, s_i is the response strength of class i, and the summation is made across all possible responses.

In terms of our general model, the assumptions of the FLMSP are almost completely confined to the classifier. Concerning the cue-measurement module, the only claim is that the output values represent the degree to which a particular property is present in the signal, which is expressed in a value between zero and one. In contrast with the previously discussed models, the classifier is mathematically fully developed in the FLMSP. The classification is based on the matching of a stimulus to category prototypes. Each prototype is essentially a collection of desirable properties. Like for the motor theory, the output labels of the FLMSP are phonemes.

Concerning the relation of the FLMSP to the categorization model used in this thesis we note the following. First of all, we use a model which is based on the multi-layer perceptron (MLP). Fundamentally, this model differs from the FLMSP in that it is based on the concept of linguistic distinctions rather than linguistic prototypes. Secondly, in contrast to the FLMSP, our model does account for the entire transformation of acoustic cues to linguistic classes. Furthermore, the FLMSP is mainly used to study the integration of 2 acoustic cues in the perception of place of articulation as well as voicing. We will confine ourselves to studying the integration of several acoustic cues in the perception of place of articulation alone. Finally, we remark that we will model the perception of (manipulated) natural utterances in several phonetic contexts. These stimuli have a far greater acoustic variability than

the highly stylized synthetic stimuli in Oden and Massaro (1978) and Massaro and Oden (1980). In their stimuli, only two acoustic parameters are varied, and only one vowel context is used.

1.4 Review of literature on the perception of stop consonants

In this section, a selection of the large number of research papers on the perception of place of articulation in initial prevocalic stop consonants. We will restrict ourselves to experimental studies which directly pertain to the cue-measurement and classification modules in our general model. In most of these studies, the influence of background information is minimized by fixing transmission conditions or using synthetic speech. We will review acoustic analysis studies in which the acoustic properties of stop consonants are described and in which potential perceptual cues are proposed. Furthermore, a number of perceptual studies will be discussed. Roughly, the perceptual studies can be divided into two groups. In one group the *necessity* or *sufficiency* of certain acoustic segments for the correct perception of place of articulation is established by respectively *deleting* or *isolating* them from natural utterances. In the other group, the influence of acoustic cues to the perception of place of articulation is investigated by systematically varying the values of the cues, for example by creating place-of-articulation continua and studying the resulting identification functions.

Rather than reviewing the literature in chronological order, we will discuss the relevant literature according to the type of cues under study. Specifically, the discussion is divided into two parts, one on detailed cues, and one on gross cues.

1.4.1 Detailed cues

Closure

In this thesis we will deal exclusively with initial prevocalic stops. The closure duration as a place cue is therefore of no concern. However, Dutch voiced plosives generally do have extensive voice bars, that is, low-frequency vocalic murmur prior to consonantal release (Slis and Cohen, 1969). Barry (1984) studied perception of place of articulation from voice bars excised from naturally uttered German stop-vowel syllables. Listeners identified place of articulation significantly above chance level. It has been shown by Van Wieringen (1995) that deletion of the voice bar from Dutch voiced stops strongly reduces correct perception of place of articulation.

Release burst

Release bursts carry considerable perceptual weight. This is demonstrated by experiments in which release bursts which are excised from natural speech were played back back to listeners, either presented in isolation, or spliced onto stationary vowels or formant transitions of conflicting utterances. Furthermore, deleting release bursts from stop consonants impairs the perception of place of articulation (Schatz, 1954; Fischer-Jørgensen, 1972; Cole and Scott, 1974; Dorman *et al.*, 1977; Ohde

and Sharf, 1977; Pols and Schouten, 1978; Pols, 1979; Schouten and Pols, 1983; Repp and Lin, 1988; Van Wieringen, 1995; see also chapter 5). A number of acoustic structures in the burst have been hypothesized to actually carry the place-of-articulation information.

First of all, various spectral properties of release bursts have been discussed. Acoustic analyses have shown that labial bursts have spectral peaks at low frequencies (below 1 kHz), or show a diffuse-falling shape. Dental or alveolar (henceforth *coronal*) bursts have high-frequency peaks (above 3 kHz), or a diffuse-rising shape. The spectral properties of both labial and coronal bursts are generally hardly dependent on the following vowel. Velar bursts, on the other hand, display a strong energy peak in the mid-frequency range (1 to 4 kHz), the position of which highly depends on the vowel context. In front-vowel context the energy peak is wide and lies in the F3-F4 region, while in non-front vowel contexts the peak is located at or slightly above the F2 at the onset of voicing (Fischer-Jørgensen, 1954; Halle *et al.*, 1957; Winitz *et al.*, 1971; Fant, 1973; Zue, 1976; Dorman *et al.*, 1977; Edwards, 1981; Repp and Lin, 1988; Keating and Lahiri, 1993; Keating *et al.*, 1994). Perception experiments with synthetic signals and burst-spliced natural utterances have shown that the spectral peaks of the burst are indeed perceptually relevant, and that they are evaluated in relation to the vowel context (Cooper *et al.*, 1952; Schatz, 1954; Hoffman, 1958; Ainsworth, 1968).

Beside the spectral properties, the length and energy of the release burst have been found to correlate with place of articulation. Labial bursts are generally weaker than coronal and velar bursts (Fischer-Jørgensen, 1954; Fant, 1973; Zue, 1976; Dorman *et al.*, 1977; Edwards, 1981; Repp and Lin, 1988). Ohde and Stevens (1983) have shown that this energy cue is indeed used in the perception of the labial-alveolar distinction.

The length of the release burst (excluding aspiration) has been found to increase with increasingly backward place of articulation, that is, velar bursts are longer than coronal bursts, which in turn are longer than labial bursts (Fischer-Jørgensen, 1954; Winitz *et al.*, 1971; Fant, 1973; Zue, 1976; Dorman *et al.*, 1977; Tekieli and Cullinan, 1979; Crystal and House, 1988). The perceptual importance of the burst length has been established by Ainsworth (1968). He found that, for synthetic stop-vowel syllables, the velar percept was enhanced by a longer burst. Notice that Dutch unvoiced stops generally have little or no aspiration, that is, the release burst is immediately followed by the onset of voicing (Slis and Cohen, 1969). Thus, burst length roughly coincides with VOT for Dutch unvoiced stops. Therefore we will not further discuss the literature on VOT as a place cue.

Formant transitions

Since the early days of speech-perception research, formant transitions have been considered to be important carriers of information regarding place of articulation (e.g. Cooper *et al.*, 1952). We define formant transitions here as changing vocal tract resonances, excited by either glottal vibration or aspiration noise, which are associated with the movement of articulators from the place of constriction to a

target position corresponding to the intended vowel. Perception of place of articulation is generally moderate to excellent when formant transitions are excised from natural speech and presented in isolation or combined with the subsequent (stationary) vowel. Furthermore, formant transitions which are preceded by a burst of a conflicting consonant will generally still carry considerable perceptual weight (Fischer-Jørgensen, 1972; Dorman *et al.*, 1977; Ohde and Sharf, 1977; Pols and Schouten, 1978; Pols, 1979; Schouten and Pols, 1983; Van Wieringen, 1995; see also chapter 5).

The place-of-articulation information is generally considered to reside in the frequencies of F2 and F3 at the onset of voicing (or aspiration), and in the direction of their initial change. Early research at Haskins laboratories demonstrated that perception of /b/, /d/, and /g/ could be induced by stylized synthetic two-formant stimuli in which the F2 "pointed at" invariant starting frequencies, or *loci*, of 720 Hz, 1800 Hz, and 3000 Hz, respectively. In terms of the initial direction of F2, labial, alveolar, and velar stops were found to need a rising F2, a falling F2 and a strongly falling F2, respectively (e.g. Cooper *et al.*, 1952; Liberman *et al.*, 1954; Delattre *et al.*, 1955). Subsequent work by Ainsworth (1968) showed that the locus concept, although adequate as a "minimal rule" in the synthesis of highly stylized stimuli (Liberman *et al.*, 1959), needed considerable refinement in more realistic stimuli. In particular, Ainsworth stressed the general importance of the release burst for the velar percept, and for other stop places in front vowel context. A number of acoustic studies showed that the F2-locus could hardly be found in natural speech signals. Frequencies of F2 and F3 at voicing onset or traced back to the instant of consonantal release showed high variability, particularly with vowel context (Fischer-Jørgensen, 1954; Halle *et al.*, 1957; Öhman, 1966; Fant, 1973; Kewley-Port, 1982). However, it was reported by several authors that the combination of various formant measures could lead to good clustering of stops according to place of articulation, e.g. F2 and F3 at voicing onset (Öhman, 1966; Fant, 1973), F2 at onset and in the vowel nucleus (Öhman, 1966; Sussman *et al.*, 1991), and F2 and F3 at voicing onset and F2 in the vowel nucleus (Öhman, 1966; Kewley-Port, 1982; Sussman, 1991). A discriminant analysis by Sussman (1991) on the basis of F2 and F3 at voicing onset and F2 in the vowel nucleus yielded an overall correct classification rate of 76% for voiced prevocalic stops (chance level at 33%). When separate analyses were performed for male and female voices and for front and back vowels, the average correct classification rose to 85%.

1.4.2 Gross cues

As discussed earlier, Blumstein and Stevens proposed a novel set of acoustic cues for place of articulation in stop consonants in a series of acoustic and perceptual studies (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980). Based on the acoustic theory of speech production (Fant, 1960), and previous observations by, amongst others, Halle *et al.* (1957) and Zue (1976), gross spectral characteristics of the initial 20-odd ms after consonantal release were hypothesized to con-

tain context-independent information for place of articulation of stop consonants. Labial, coronal, and velar stops were claimed to have diffuse falling, diffuse rising, and compact characteristics, respectively. An acoustic study showed that, on the basis of these gross spectral characteristics, naturally uttered stop-vowel syllables could be classified at a correct rate of 85% according to place of articulation, across speakers and vowel contexts (Blumstein and Stevens, 1979). Furthermore, perception experiments with synthetic stimuli demonstrated that the initial 25 ms of the speech signal after release was generally enough for perception of place of articulation, and that the subjects' responses were highly correlated with the gross spectral characteristics (Stevens and Blumstein, 1978; Blumstein and Stevens, 1980).

These results were criticized from three angles. Firstly, subsequent experiments demonstrated that formant transition information was perceptually more important than the gross onset characteristics. Synthetic stimuli in which formant transitions corresponded to one place of articulation and the gross onset spectrum corresponded to a conflicting place, are generally classified by listeners in accordance with the formant information (Blumstein *et al.*, 1982; Walley and Carrell, 1983). Secondly, Kewley-Port argued that dynamic, that is, time-dependent, information needed to be incorporated, and she added the dynamic cues "late onset of voicing" and "persistence of a mid-frequency peak over time" to the existing "static" spectral properties diffuse rising and diffuse falling. Acoustic classification significantly improved when the dynamic cues were used (Kewley-Port, 1983; Kewley-Port *et al.*, 1984), and perception of place of articulation was better for stimuli which had the intended dynamic properties than for stimuli which only had Blumstein and Stevens' static properties (Kewley-Port *et al.*, 1983). Thirdly, Suomi (1985, 1987) showed that, although the gross spectral characteristics of the first 20-odd ms of the speech signal after release were reasonably effective for vowel-independent classification of stop consonants, they were nevertheless strongly coarticulated. Suomi argued against the assumption, made in the invariance theory, that acoustic variability due to phonetic context is essentially "noise". The same general viewpoint was put forward by Elman and McClelland (1986). Suomi (1985) supported his claim by showing that automatic classification of stops based on context-dependent prototypes (essentially allophone prototypes) was more reliable than classification based on context-independent prototypes.

In perception experiments with synthetic stimuli in which the amplitude of the release burst was varied, Ohde and Stevens (1983) demonstrated that the evolution of high-frequency energy (above 2.5 kHz) after release was a cue to the labial-alveolar distinction. An increase in high-frequency energy from burst to voicing cues the labial place, a decrease cues the alveolar place. Subsequently, Lahiri *et al.* (1984) introduced an improved, dynamic, cue to the labial-coronal distinction. Roughly, the measure was based on the change in high-frequency energy (around 3.5 kHz), relative to the change in low-frequency energy (around 1.5 kHz) from burst to voicing onset. Based on this measure, dentals and alveolars could be separated from labials across various languages at a correct rate of 91%. Furthermore, a perception experiment employing synthetic stimuli in which the proposed measure cued one place of articulation and the formant transitions cued the conflicting place, showed

that listeners responded in accordance with the proposed gross measure.

Lindholm *et al.* (1988) constructed synthetic stimuli in which global spectral tilt and the abruptness of frequency change (roughly corresponding to the gross cues proposed by Kewley-Port, 1983) were pitted against formant transitions. Subjects classified the stimuli predominantly in accordance with the formant transitions.

Two large-scale automatic (machine) classification experiments have recently shown that it is indeed possible to perform an excellent speaker-independent and vowel-independent classification of prevocalic stops using only gross spectro-temporal information. Forrest *et al.* (1988) measured the first, third and fourth spectral moments (corresponding to spectral center of gravity, tilt and compactness) at four moments in the initial 50 ms after consonantal release. Using these gross cues, a correct stop consonant classification rate of 93% was obtained. Nossair and Zahorian (1991) explicitly tested the power of gross spectro-temporal information versus formant-transition information for the purpose of stop-consonant classification. The results demonstrated that classification was excellent for the gross cues (95% across speakers and vowel contexts), while classification based on formant transitions was significantly lower. Furthermore, it was shown that using dynamic spectral information improved recognition considerably compared to using static spectral information of the first 20 ms after release.

At present, the issue of the relative importance of detailed versus gross cues in speech perception is highly topical, which is demonstrated by the publication of a number of studies on this issue. Krull (1990) published a study of the perception of stop consonants in which acoustic cues, which were measured on short segments of naturally uttered stop-vowel syllables, were explicitly correlated with perceptual confusions of the same segments. The results showed that information of F2 and F3, combined with the length of the release burst, gave a good account of the observed perceptual confusions. Information based on spectral levels, which contained information on the gross spectral characteristics, correlated much less with the confusions. In a series of publications, Ter Keurs described the intelligibility of sentences and nonsense words in which spectral details were reduced by means of spectral smearing (Ter Keurs, 1992; Ter Keurs *et al.*, 1992; Ter Keurs *et al.*, 1993a, b). Her results showed that with a spectral smearing larger than 1/3 octave intelligibility was impaired. Confusions occurred mostly in place of articulation. These results suggest that in the F2 region (roughly 1 to 2kHz), spectral information at a level of detail of 500 Hz or coarser is not sufficient for correct perception of place of articulation. Drullman published a number of experimental studies on the perceptual effect of reducing rapid or slow fluctuations in the speech signal (Drullman *et al.* 1994a, b; Drullman, 1995a, b). His results show that intelligibility of sentences is not impaired when temporal fluctuations higher than 16 Hz are canceled. When slower fluctuations are deleted, intelligibility decreases. Decreasing slow fluctuations, while preserving rapid fluctuations does not affect sentence intelligibility for cut-off frequencies up to 4 Hz. Stop consonants appear to be mostly confused with either fricative consonants or glides. Place-of-articulation confusions also occur often. Thus, spectral changes with a time constant between roughly 50ms and 250ms

appear to be crucial for correct perception of place of articulation.

Before we proceed to motivate the purpose of the research presented in this thesis, we briefly summarize the findings and considerations discussed in this introductory chapter.

In accordance with the considerations made by other researchers, we have formulated the general problem of stop-consonant perception as a problem of variability versus invariance. Specifically, the acoustic variability of stop consonants with phonetic context is very high. A general qualitative model has been proposed, which covers the major findings and assumptions reported in the literature on stop-consonant perception. In the model several basic stages are discerned: input speech representation, background information, event detection, cue extraction, classification, and output linguistic representation.

In the light of the general model three theories of speech perception have been discussed. It was concluded that in the motor theory the perceptual importance of detailed cues is stressed. Furthermore, this theory emphasizes the high variability of these cues with phonetic context. In contrast, the invariance theory assumes that perception is principally guided by gross cues which are associated with distinctive features, and which are assumed to be invariant across phonetic contexts. The FLMSP was shown to be a mathematically well-developed model which is strongly focused on the classification problem in speech perception. The model assumes that the classification is based on a matching of stimuli to idealized category prototypes.

Finally, in a survey of the relevant experimental literature, a number of acoustic cues that have been investigated over the years have been discussed. The cues are summarized in Table 1.I, along with the typical values of each cue for the three major places of articulation. Automatic classification studies have shown that gross spectro-temporal properties are excellent potential perceptual cues, provided they contain dynamic information. Automatic classification based on detailed spectro-temporal properties has so far yielded less favorable results, although recent developments associated with locus equations are encouraging. Perception studies have shown that both detailed as well as gross cues are used in the perception of stop consonants. With respect to the gross cues, it seems fair to conclude that the hypothesis that place-perception of stops is principally guided by the gross characteristics of the static onset spectrum is disproved. However, dynamic gross information seems to provide strong perceptual cues. The issue whether perception of place of articulation is principally guided by detailed or gross information is far from settled, as studies employing synthetic, conflicting-cue stimuli have so far yielded unequivocal results. This issue remains therefore highly topical.

1.5 The purpose of this study

It is the purpose of the research presented in this thesis to investigate whether detailed or gross spectro-temporal properties are the primary cues for the perception of place of articulation of initial prevocalic stop consonants. In our study, we will emphasize two methodological aspects. First of all, like Krull (1990), we will use

Table 1.I: List of acoustic cues for the perception of place of articulation of stop consonants. The symbols −, ±, and + indicate that the acoustic cue either correlates with or induces the perception of the indicated place of articulation when it has a relatively low, medium, or high value, respectively. A + for spectral tilt means that the energy increases with increasing frequency. A + for change in tilt or HF energy means that the tilt or HF energy increases over time. A combination of symbols means that all indicated values commonly occur. Absence of a symbol means that the cue is thought not to be relevant for the indicated place of articulation.

detailed cue	typical value		
	labial	coronal	velar
freq. of burst peak	−	+	±+
strength of burst peak	−	−	+
energy of burst	−	±	+
burst length	−	±	+
$F2_o$	−±	±+	−±+
direction of $F2$	+	−±	−±
$F3_o$		+	−
direction of $F3$		−	+
gross cue	**typical value**		
	labial	coronal	velar
spectral tilt	−	+	
change of spectral tilt	+	−	
change of HF energy	+	−	
strength of MF peak	−	−	+
persistence of MF peak	−	−	+

(manipulated) natural utterances in our experiments in order to preserve the natural variability in the speech signal. Thus, we hope to avoid potential unnaturalness of the stimuli, which may have been partly responsible for the apparent contradiction in the results of previous studies using synthetic conflicting-cue stimuli (e.g. Lahiri *et al.*, 1984, versus Lindholm *et al.*, 1988). Secondly, we aim to model the entire process of stop-consonant perception as presented in our general model. That is, we will simulate the classification behavior of the subjects, including initial representation, event detection, extraction and integration of cues, and classification. Like many previous perception studies, the influence of background information will be minimized by fixing transmission conditions, and using a limited number of speakers.

The approach is as follows. We will create a set of stimuli by manipulating a number of natural stop-vowel utterances, which will be presented to listeners. Next we will simulate the listeners' classification behavior using a formal model. Before coming to the perception experiment and subsequent simulation, a number of methodological studies are presented. First, in chapter 2, a critical study of the

accuracy of the traditional quasi-stationary speech representations is made, which is especially relevant for the extraction of detailed cues, such as formant onset frequencies. The results show that accurate measurements of formant frequencies can be made using the spectrogram or linear prediction, provided the settings of the analysis parameters meet some basic conditions. Next, in chapters 3 and 4, a model of human classification behavior is introduced, which is based on the multi-layer perceptron (MLP). Chapter 3 focuses on the fundamental properties of the model, while in chapter 4 a number of practical issues are discussed which are important for the estimation and evaluation of the model. In chapter 5, the perception experiment is presented. First we discuss the procedures for manipulating the natural utterances. We encounter the difficulty that it is not possible to manipulate individual acoustic cues across a continuum in natural speech. Detailed and gross acoustic structures generally *covary* in natural speech, for example, a high F2-frequency at voicing onset will often be accompanied by a positive global spectral tilt. We will reduce this covariation by creating *deleted-cue* as well as *conflicting-cue* stimuli, by removing parts of the original utterances or exchanging information between utterances. To this end, we use the well-known "burst-splicing" technique (e.g. Fischer-Jørgensen, 1972). These stimuli are presented to listeners for classification of place of articulation. In chapter 6, the actual simulation of the classification behavior of the subjects is presented. First, the relevant acoustic events, burst onset and voicing onset, are detected and various detailed and gross cues are measured on the stimuli. Next, the proposed model is used to map the measured acoustic cues onto the observed perceptual data. The best-fitting models are interpreted with respect to (1) which cue set gives the better account of the perceptual data, (2) how are the cues actually integrated in the model. We remark that in our approach all cues are "treated equally", that is, we refrain from associating gross cues with invariance and detailed cues with variability. Finally, in chapter 7 the results of the study are discussed and some ideas for future research are presented.

Chapter 2

Accuracy of formant-frequency measurement in highly dynamic speech using spectrogram and linear prediction[1]

Abstract

In this chapter, the accuracy of the analysis of rapidly varying formants using spectrogram and Linear Prediction is assessed. Analysis of various dynamic signals shows that, when a long analysis window, like 25 ms, is used, the quality of the representation may be impoverished. Obvious unwanted effects are staircase-like formant tracks, flattening-off of formants close to voicing onset, and bending of the formant towards a strong energy concentration in the release burst. The parameters that have the largest influence on the quality of the representation are the length of the analysis window, the transition rate of the formant, the fundamental frequency, and the position and energy of the release burst. It is shown that the most accurate analysis using a quasi-stationary method is made when windows are positioned pitch-synchronously. Finally, a quantitative analysis of the influence of the mentioned parameters provides evidence that no deviations due to the quasi-stationarity assumption occur when the effective length of the analysis window is not larger than the pitch period. The wideband spectrogram is expected to be a reliable speech-analysis tool because it meets this condition for fundamental frequencies up to 370 Hz.

2.1 Introduction

When a stop consonant is produced, the acoustics of the vocal tract immediately prior to closure or after release change very rapidly. The resulting acoustic speech wave will generally be of a highly dynamic nature, that is, its spectral content will change rapidly over time (e.g. Fant, 1973). Among all classes of speech sounds, these essentially non-stationary sounds have proven to be the most difficult to define acoustically. In particular, it is nearly impossible to find any acoustically invariant structure in them.

In the search for acoustic invariance in stop consonants, roughly two approaches can be distinguished. One approach is to concentrate on gross spectral characteristics, either of a static nature (e.g. Blumstein and Stevens, 1979; Blumstein *et al.*, 1982) or of a dynamic nature (e.g. Kewley-Port, 1983; Lahiri *et al.*, 1984). The

[1]Based on: Smits, R. (**1994**), "Accuracy of quasi-stationary analysis of highly dynamic speech signals," J. Acoust. Soc. Am. **96**, 3401–3415.

other approach is to look in great detail for invariance in formant frequencies or formant transition rates (e.g. Kewley-Port, 1982; Sussman *et al.*, 1991). Two aspects of the latter approach are very important: first the definition of a formant frequency and transition rate, second the accuracy of the analysis tool that is used to measure these quantities.

Regarding the first aspect, explicit definitions of formant frequency and transition rate are, unfortunately, seldom found in these studies. Some form of spectral peak picking or linear predictive analysis is usually performed without strict definitions. Throughout this chapter we will use a definition which is related to these strategies. We define a formant frequency as a spectral peak which is associated with a vocal tract resonance. Ideally, one would require this measure to be more or less independent of the analysis technique. In practice, however, this is hardly ever the case. The dependence of the measurement result on the adopted analysis technique is one of the topics of this chapter.

Regarding the aspect of the accuracy of the analysis method, it is remarkable that, despite the dynamic nature of the signals, quasi-stationary analysis methods have nearly always been used, such as the Short Time Fourier Transform (STFT) or Linear Prediction (LP). In methods of this kind an analysis window is used within which the signal is assumed to be stationary. Despite the large amount of research effort invested, the problem of acoustic invariance for stop consonants has still not been solved (e.g. Perkell and Klatt, 1986). A reason for this fact may be that the type of analysis tools used in most of the studies is less than adequate. Indeed, one could argue that the analysis of highly dynamic signals, such as stop consonants, with essentially quasi-stationary methods will lead to inaccurate or even false estimations of important parameters such as formant frequencies and formant transition rates.

In fact, this line of reasoning has been followed by many researchers, particularly in the field of signal processing. During the last decade there has been an increasing interest in non-stationary analysis techniques. As a result, many new techniques have been developed and published.

In the next section an overview is presented of new non-stationary techniques that show promise, but still have to prove their merits for speech analysis. Next, it is argued that little is known about the suitability of the often-used quasi-stationary methods for the analysis of highly dynamic speech segments and in the remaining part of this chapter these techniques are subjected to a detailed investigation.

As a general analytic approach is impossible, an empirical approach is followed. In section 2.3, a qualitative survey of several problems with the quasi-stationary analysis of highly dynamic signals is presented by means of various representations of relevant examples. An inventorization is made of the signal and analysis parameters that have the largest influence on the accuracy. A quantitative description of the influence of these parameters on the accuracy is presented in Section 2.4. Finally, in section 2.5, the results are discussed and summarized.

2.2 New non-stationary analysis techniques

In this section, the class of "bilinear" time-frequency representations (TFRs) and time-varying LP-analysis will be discussed briefly. For extensive reviews see, for example, Boashash (1992), Cohen (1989), Hlawatsch and Boudreaux-Bartels (1992), or Loughlin *et al.* (1992).

The most thoroughly studied new type of TFRs is the class of bilinear TFRs. Based mostly on mathematically desirable properties, the general formulation of this class has been reduced to several interesting new TFRs (e.g. Choi and Williams, 1989; Zhao *et al.*, 1990). Many of these have been discussed in the context of speech analysis (Loughlin *et al.*, 1992; Cohen and Pickover, 1986), although few have actually been used for speech studies.

The Wigner Ville Distribution (WVD) is the best known non-stationary member of the class of bilinear TFRs. It has often been criticized for being uninterpretable when "real life" signals, such as speech, are analyzed, because of its strong interference terms and its negative values (Loughlin *et al.*, 1992; Atlas *et al.*, 1990; Riley, 1989). Still, besides the spectrogram, it is the only member of the bilinear class that has been seriously used for speech analysis (Dogil and Wokurek, 1989, 1991; Garudadri *et al.*, 1987; Velez and Absher, 1989; Velez and Garudadri, 1992). In all these studies a mild smoothing of the WVD is adopted in order to reduce interference. The resulting TFR has, strictly speaking, lost its non-stationarity, but still preserves a high resolving power simultaneously in time and frequency. Although the experience in using the WVD for speech analysis is limited, the results obtained so far are encouraging.

In LP-analysis usually a windowing technique similar to that in the Short Time Fourier analysis is applied. Within each windowed segment the signal is supposed to be stationary and the filter coefficients are constant. The traditional LP-model is therefore quasi-stationary. Recently, some time-varying LP-analysis methods have been presented (Grenier, 1983; Casacuberta and Vidal, 1987; Nathan *et al.*, 1991). Interestingly, in Nathan *et al.* (1991), the method was explicitly used to analyze highly dynamic final stop consonants. Analyses on the same material were conducted using pitch-synchronous quasi-stationary LP and results were compared. It appeared that in the very last part prior to closure the time-varying method sometimes showed strongly curving formant slopes, which highly deviated from the relatively flat tracks found by the quasi-stationary method. In Nathan and Silverman (1991) the F2-frequency and F2-slope prior to closure in vowel-stop syllables were calculated by quasi-stationary and time-varying LP and were presented in scatter diagrams. While the quasi-stationary data hardly showed any clustering, the time-varying data clustered almost completely into three linearly separable groups, corresponding to the consonants /p/, /t/ and /k/. These results suggest that (1) there might be some locus-related type of invariance (e.g. Delattre *et al.*, 1955) present in the signal, (2) the errors in measurements of formant frequencies by quasi-stationary analysis methods could be large enough to obscure the invariance that may be present in acoustical details.

2.2.1 Previous studies on the adequacy of quasi-stationary and non-stationary analysis techniques

The performance of new non-stationary methods for speech analysis is discussed among others by Atlas *et al.* (1990, 1991, 1992), Casacuberta and Vidal (1987), Dogil and Wokurek (1989, 1991), Loughlin *et al.* (1992), Velez and Absher (1989), and Velez and Garudadri (1992). In some of these studies, e.g. Casacuberta and Vidal (1987), Atlas *et al.* (1991) and Velez and Absher (1989), the validity of the traditional spectrographic or LP-analyses of dynamic speech signals is rather easily dispensed with because of its assumption of quasi-stationarity, and subsequently the proposed new method is claimed to be superior. However, it is generally accepted that the problem of the representation of time-varying signals by quasi-stationary methods is more complex than the well-known $\Delta f \Delta t$ trade-off (e.g. Cohen, 1989). Besides, this problem cannot be generally tackled by analytic means and must therefore be approached by empirical methods. Surprisingly, however, only a few attempts to describe when and how the quasi-stationary methods actually go wrong are reported in the literature.

First of all we mention two papers in which the accuracy of quasi-stationary techniques is discussed for analyzing *stationary* signals. In Lindblom (1961) the accuracy of the manual measurement of formant frequencies from sonagrams is described. Several signal parameters that may influence the measurement accuracy are discussed, especially the importance of the fundamental frequency (F_0). Next, the results of an experiment are briefly described in which five experienced phoneticians were instructed to estimate the formant frequencies of stationary synthetic vowels on wideband and narrowband sonagrams. The mean error in a fairly large number of measurements was reported to be 40 Hz for male voices. The error tended to increase with increasing F_0, but rarely exceeded $F_0/4$.

Monsen and Engebretson (1983) compare the accuracy of the spectrogram and LP-analysis in measuring the frequencies of the first three formants in stationary synthetic vowels. Four signal parameters were systematically varied: F_0, formant bandwidths, proximity of formants, and frequency location of formants. The resulting 90 different signals were subjected to automatic LP-analysis and to manual analysis by three experienced phoneticians using wideband and narrowband spectrograms. The results showed that the accuracy generally depended on all four signal parameters. The average accuracy of formant frequency measurement for the F1 and F2 was estimated at 60 Hz for both techniques. For the measurement of F3, this figure was the same for LP-analysis, but increased to roughly 110 Hz for the spectrogram.

In recent years some papers have been published on the accuracy of quasi-stationary techniques for *non-stationary* signals. Silverman and Lee (1987) claimed to have found an anomaly in the spectrographic representation of a rapid sinusoidal glide. A few years later, Wokurek (1991) showed that this result was based on an erroneous notion of instantaneous frequency. Wokurek subsequently claimed that the formant tracks of rapidly time-varying speech are displayed correctly by spectrograms. A similar claim was made by Riley (1989).

As discussed earlier, Nathan and Silverman (1991) and Nathan *et al.* (1991) describe that strong discrepancies are sometimes found in formant frequencies and transition rates measured by quasi-stationary and time-varying LP. However, as the speech material in which significant deviations occur is rather limited, it is hard to draw general conclusions concerning when and why quasi-stationary LP will give bad results.

Howitt (1991) presented a detailed quantitative comparison of the accuracy of naive subjects in measuring formant transition rates by hand from wideband spectrograms and unsmoothed WVDs. Synthetic single formant and multiple formant initial stop-like stimuli with various transition rates were used. The results show that the estimations of transition rates from spectrograms are generally too high (too rapid) for transitions up to 100 Hz/ms and too low (too shallow) for faster transitions. Deviations from the true transition rates were as large as 50%. The results for the WVD-estimations showed the same trend, but were slightly more accurate for the single formant situations.

In Loughlin *et al.* (1992) the accuracy of the estimation of transition rate from a wideband and a narrowband spectrogram was tested on one example of a single synthetic formant with a transition rate of 10 Hz/ms. It was concluded that the wideband and narrowband spectrogram underestimate the transition rate by 7% and 12%, respectively. Besides, it is stated that the WVD-representation of this signal is next to exact.

In summary, it remains unclear whether the traditional quasi-stationary methods are suited for the analysis of dynamic speech signals. Therefore, a detailed study of the performance of the quasi-stationary methods for dynamic speech signals still seems in order. Indeed, it is important for researchers in the area of speech analysis and perception to know if past (and future) studies of dynamic speech signals using quasi-stationary methods are valid and accurate. It is the aim of this chapter to provide qualitative as well as quantitative insight in the accuracy of measuring formant frequencies and formant transition rates in highly dynamic speech using spectrographic and LP techniques. It is emphasized that the issue of the perceptual relevance of various speech parameters, analysis techniques and measurement accuracies is not addressed in this chapter.

2.3 Qualitative experiments

The purpose of this section is (1) to give an overview of the type of problems that may occur when a dynamic signal is analyzed using a quasi-stationary technique, and (2) to discuss which signal and analysis parameters have the strongest influence on the quality of the representation. First, the methods for generation, analysis and display of the signals are presented, then the results are shown and discussed.

2.3.1 Method

Generation of signals

Tone glide of varying amplitude
First, we define the instantaneous frequency as the first time derivative of the instantaneous phase of the sinusoid. The first synthetic signal ($S1$) is a sinusoid of varying frequency. The instantaneous frequency increases linearly with time at a rate of 200 Hz/ms. The signal is multiplied by a Hanning (raised cosine) window with a total length of 10 ms.

Single formant signal
Signal $S2$ was created by exciting a time-varying second-order all-pole filter by a train of four impulses at time instances t = 9.9, 16.4, 22.9 and 29.4 ms (the fundamental period is 6.5 ms). The filter coefficients were adjusted at every sample instant in such a way that the resonance (formant) frequency increased linearly with time at a rate of 100 Hz/ms. This transition rate is extremely high but may still occur in natural speech (Howitt, 1991), although only during very short time intervals. The formant bandwidth was kept constant at 100 Hz.

Natural speech signal
The Dutch utterance /du/, spoken by a male talker, was recorded and filtered at 4.9 kHz and was sampled at a rate of 10 kHz. The release burst plus the first five glottal pulses were excised. The signal had a total duration of 51.2 ms. The signal was multiplied at onset by a ramp which rose linearly from zero to one in 2 ms. The time-reversed ramp was applied at signal offset.

Methods of analysis and display

Spectrogram
All time signals are transformed into several TFRs. First of all the STFT is performed. The STFT $\mathcal{F}(n, e^{j\omega})$ of the discrete-time signal $x(m)$ is defined as (e.g. Rabiner and Schafer, 1978):

$$\mathcal{F}(n, e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(nS - m)x(m)e^{-j\omega m} \tag{2.1}$$

where $w(nS - m)$ is the inverted time window at position nS, and S is the window shift. In all cases a Hanning window was used. As the purpose of the analyses is to study the inaccuracy of the methods due to the quasi-stationarity assumption, the shift S of the window is kept extremely small in all STFT calculations, viz. 0.2 ms (2 samples). In this way we achieve almost the highest accuracy possible, given a particular window length. A larger window shift may indeed introduce another type of inaccuracy, which is not essentially caused by the quasi-stationarity, but by a "subsampling" of the complete analysis.

The STFT $\mathcal{F}$ was separated into magnitude $|\mathcal{F}|$ and phase $\angle\mathcal{F}$. The phase was discarded and the magnitude was transformed from linear magnitude $|\mathcal{F}|$ into

logarithmic magnitude $|\tilde{\mathcal{F}}|$:

$$|\tilde{\mathcal{F}}| = 20 \log |\mathcal{F}| \tag{2.2}$$

$|\tilde{\mathcal{F}}|$, henceforth called the spectrogram, is displayed by means of contour plots. The option to use grey-level representations by means of small black dots was not used because the resulting representation is less detailed. The dynamic range of all contour plots, viz. the maximum displayed value minus the minimum displayed value, is 30 dB.

In many acoustic speech studies spectral peak picking is applied to arrive at formant frequencies. We have made the result of such peak picking method explicit by drawing an additional line through successive spectral maxima in the contour plots.

In some cases, viz. when the analysis window was larger than the pitch period, the resulting spectrograms were dominated by the harmonic structure of the signal. It was chosen to present some form of spectral envelope rather than the original spectrogram, because a harmonic structure hampers a precise determination of the spectral maxima. The spectral envelope was calculated by convolving each individual magnitude spectrum by a smoothing window. A normalized Hamming window with a total width of 332 Hz was used in all cases. it should be stressed that in all cases the resulting smoothed spectrum closely followed the original harmonic spectra. It was carefully checked that none of the observed effects described in the following sections are artifacts of the smoothing operation, but instead are all caused by the quasi-stationarity of the analysis method.

Wigner-Ville Distribution

The purpose of this research is to assess the accuracy of traditional quasi-stationary analysis methods, rather than to make a comparison between quasi-stationary and non-stationary techniques. Nevertheless, it is helpful to be able to make a comparison of the analysis results of quasi-stationary and non-stationary methods. Therefore, beside the STFT, the WVD was calculated for all time signals. The WVD $\mathcal{W}(n, e^{j\omega})$ of a discrete-time signal $x(n)$ is defined as:

$$\mathcal{W}(n, e^{j\omega}) = 2 \sum_{m=-\infty}^{\infty} \tilde{x}(n+m)\tilde{x}^*(n-m)e^{-2j\omega m} \tag{2.3}$$

where $\tilde{x}(n)$ is the analytical signal associated with the real signal $x(n)$ and * denotes complex conjugation (e.g. Velez and Absher, 1989). In fact, in order to keep calculations and data size within limits, the Pseudo-WVD $\mathcal{W}p$ was calculated:

$$\mathcal{W}p(n,k) = 2 \sum_{m=-L}^{L} w(m)w^*(-m)\tilde{x}(n+m)\tilde{x}^*(n-m)e^{-2jmk\frac{2\pi}{2L+1}} \tag{2.4}$$

where w is a time window of length L (e.g. Velez and Absher, 1989). In all our calculations w was chosen to be a 25.6 ms Hamming window. The resulting Pseudo-WVD equals a true WVD which is frequency-smoothed by the Fourier transform of the time window w.

It is well known that, when applied to non-trivial signals like speech, the WVD generally displays marked cross-terms which hamper a straightforward interpretation. As shown by e.g. Wokurek *et al.* (1987), light smoothing of the WVD along the time axis and the frequency axis effectively suppresses these cross-terms. In the case of signal $S3$ such smoothing appeared to be necessary. Following suggestions made by Wokurek *et al.* (1987), we used a Hanning window with a total width of 2.1 ms for the time smoothing and a Hamming window with a total width of 215 Hz for the frequency smoothing.

The negative parts of the smoothed Pseudo-WVDs are set to zero and the result it converted to logarithmic amplitude. Eq. 2.2 was used, with the factor 20 replaced by 10, because the WVD is a quadratic quantity, while the STFT magnitude is linear. The WVDs are displayed by means of contour plots, with a dynamic range of 30 dB.

Linear Prediction
Finally, LP-analysis is applied to signals $S2$ and $S3$. The LP-coefficients were estimated using the autocorrelation method (e.g. Rabiner and Schafer, 1978). As for the STFT, Hanning time windows of various lengths were used and the window shift was 0.2 ms. The order of the analysis varied between 2 and 14, depending on the signal. No preemphasis was used.

The result of the LP-analyses were visualized in two ways. First, the LP-coefficients were Fourier transformed for all window positions. The inverse of the resulting spectrum is the estimated transfer function of the LP-filter. The Fourier transform magnitude was converted to dBs using Eq. 2.2. The result can be viewed as an LP-spectrogram and is displayed in the same way as the STFT spectrograms. In order to get a better view of the exact pole frequencies estimated by the LP-analysis, the Fourier-transform phase spectrum of the LP-coefficients was converted to pole frequencies using the group delay method described by Yegnanarayana (1978). This method is used because the resolution of closely spaced formants is higher in group delay spectra than in magnitude spectra, which is relevant for our analyses. The pole frequencies are indicated by means of additional lines in the contour plots.

2.3.2 Results

Sinusoidal signals
It has been shown by Riley (1989) and Wokurek (1991) that the spectrogram gives a correct representation of a sinusoid of constant amplitude and linearly varying frequency. Although the quasi-stationarity of the analysis gives a broadening of the energy concentration in each individual spectrum, the instantaneous frequency of the signal can be accurately recovered from the series of successive spectra by picking the spectral peaks.

Kodera *et al.* (1978) discussed the problem of recovering the instantaneous frequency of short chirp-tones. They acknowledged that chirps of constant amplitude are generally represented accurately by the spectrogram, but subsequently showed

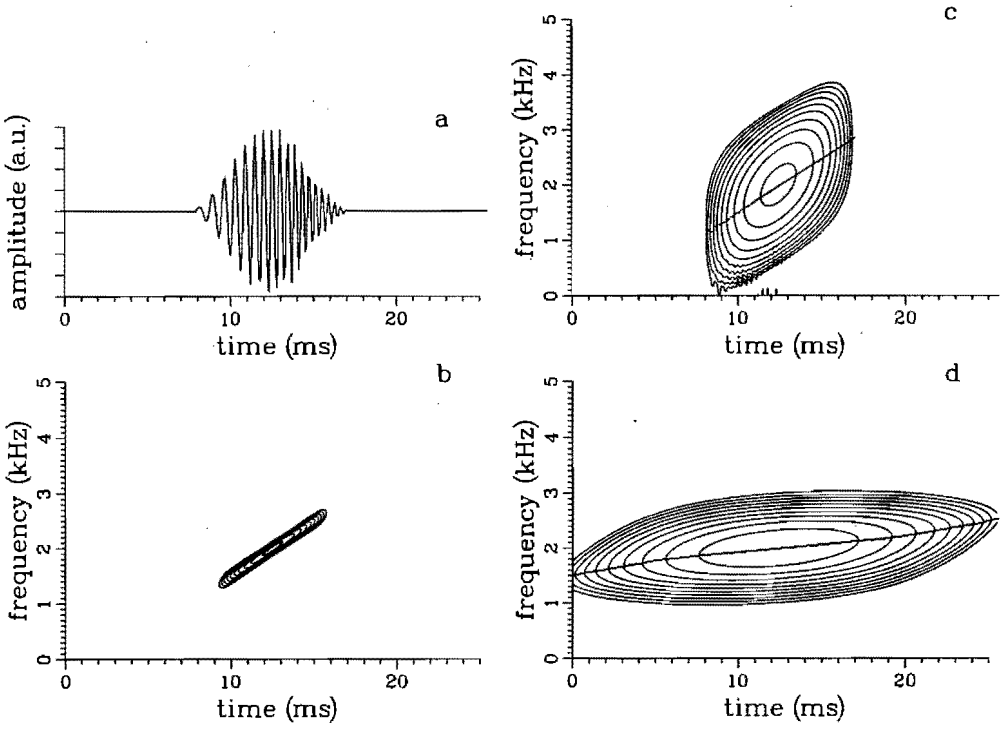

Figure 2.1: Hanning windowed linear chirp signal. a. Waveform. b. WVD. c. Spectrogram, analysis window length is 1.3 ms. d. Spectrogram, analysis window length is 25.6 ms.

that the representation of signals with both varying amplitude and varying frequency is ambiguous. The apparent transition rates, measured by picking spectral peaks, were found to be highly dependent on the length of the analysis window. Figure 2.1a shows signal $S1$. This signal is similar to those used by Kodera *et al.* The unsmoothed WVD is shown in Figure 2.1b. Figure 2.1c and 2.1d are spectrograms of the signal using an analysis window of 1.3 ms and 25.6 ms, respectively. The line crossing the contour lines indicates the location of the spectral maximum as a function of time.

It is well known that the WVD representation of mono-component signals closely follows the instantaneous frequency of the signal, which is defined as the time derivative of the instantaneous phase. This is also the case in Fig 2.1b. The spectrogram with the short window length (2.1c) shows a broadening in the frequency direction, but the instantaneous frequency can still be accurately recovered from the spectral maxima. The spectrogram for the long window (2.1d), however, gives a deceptive representation of the signal. The apparent transition rate, estimated from the line

indicating the spectral maximum, varies between 30 Hz/ms and 70 Hz/ms, which is clearly too low. This result can be understood in the following way. If the analysis window is larger than the time constant describing the amplitude modulation of the signal (in this case roughly 3 ms), then some "frequency weighing" due to the signal envelope will take place within the analysis window. Thus, the frequency at the amplitude maximum will, at many window positions, be weighted more strongly than the frequency at the skirts of the signal. As in Kodera *et al.* (1978), the result is a flattened structure, which is smeared in the time direction, and a highly underestimated transition rate. This finding is relevant for speech analysis because the speech-signal envelope varies markedly with time.

Single-formant signal

Signal $S2$ is shown in Figure 2.2. Figures 2.2a, b, c and d are the time signal, the WVD and the 1.3 ms and 25.6 ms spectrograms. Figures 2.2e, f and g are LP-analysis results. Figure 2.2e was calculated using a 1.3 ms analysis window, Figures 2.2f and g were calculated using a 25.6 ms window. A 2-pole model was fitted in Figure 2.2e and f, in Figure 2.2g a 10-pole model was fitted.

The short-window spectrogram and LP-analysis give accurate representations. The WVD shows marked interference products, but generally the instantaneous frequency and transition rate can still be measured correctly. The long-window spectrogram and LP-analyses show some interesting deviations from the actual formant track. First of all, the formant track seems to break up into separate horizontal structures in Figures 2.2d and g. At certain time instances, e.g. $t =$ 13 ms or $t = 20$ ms, individual spectra (vertical cross-sections) show two or three simultaneous formants. It should be noted that these horizontal structures do not simply reflect the harmonic structure of the signal. They are caused by the fact that successive impulses excite the formant at sufficiently distant frequencies. Interestingly, the 10-pole LP-analysis has fitted a single rapidly moving formant by several stationary formants, each of which is caused by another excitation. The 2-pole analysis gives a good impression of the formant movement in the center of the signal (better than the 10-pole analysis), but wrongly suggests a flattening-off behavior at the on- and offset of the signal. The long-window representations could be improved by using the clear definition of onset and offset of the signal offered by the oscillogram itself. This means ignoring the parts of the representations before $t = 10.0$ ms and after $t = 38.0$ ms. The transition rates at onset and offset of the signal would still be too low, as can be seen in Figure 2.2f.

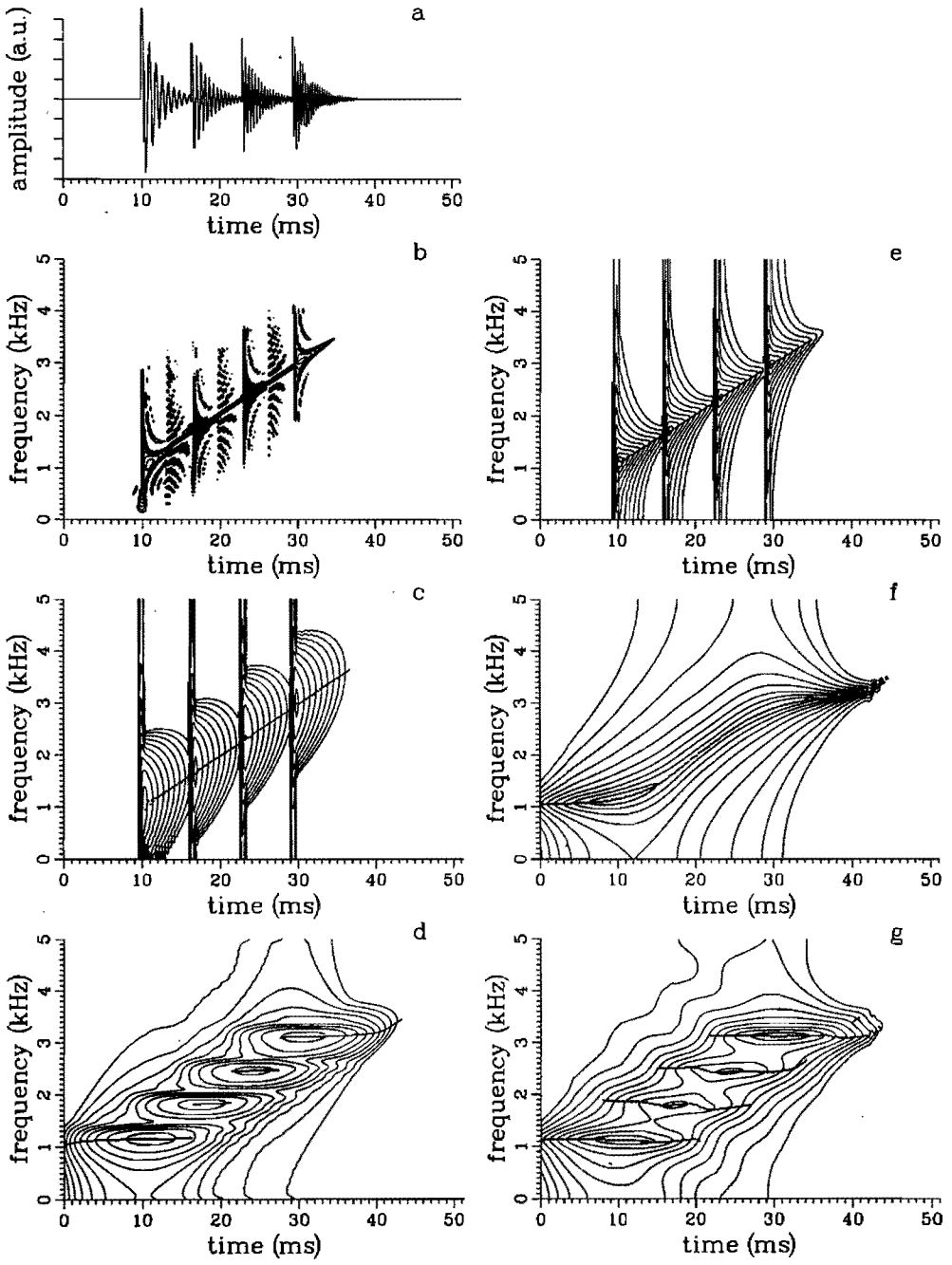
a
amplitude (a.u.)
time (ms)
b
frequency (kHz)
time (ms)
e
frequency (kHz)
time (ms)
c
frequency (kHz)
time (ms)
f
frequency (kHz)
time (ms)
d
frequency (kHz)
time (ms)
g
frequency (kHz)
time (ms)

Figure 2.2: *(previous page)* Single gliding formant, excited by 4 impulses. a. Waveform. b. WVD. c. Spectrogram, analysis window length is 1.3 ms. d. Spectrogram, analysis window length is 25.6 ms. e. Second order LP-spectrogram, analysis window length is 1.3 ms. f. Second order LP-spectrogram, analysis window length is 25.6 ms. g. Tenth order LP-spectrogram, analysis window length is 25.6 ms.

From the above examples it would seem that the use of an extremely short analysis window, e.g. 1.3 ms, is optimal. However, when moving from sinusoidal and single-formant signals to complex multiple-formant signals, the frequency blurring by these windows would be too large. According to Harris (1978), the relation between the total length D of a Hanning window, and the associated resolving power in frequency ΔF, defined as the -6dB bandwidth, is described by

$$D \cdot \Delta F = 2.0 \tag{2.5}$$

A compromise is therefore necessary. The short-window analyses of the next signal were made using a window length of 6 ms, corresponding to a resolution of 333 Hz. This is close to the resolution of 300 Hz of the traditional wideband spectrogram (e.g. Flanagan, 1972, p. 151).

In the remaining part of the chapter, the "effective length" of an analysis window will often be used. The effective length of a window is here defined as the interval between the two -3dB points of the window, which is half the total length in case of a Hanning window.

Natural speech signal

A Dutch prevocalic voiced stop will generally consist of the following phases: (1) a silent or prevoicing phase, (2) the release burst, followed more or less closely by (3) formant transitions. Thus, the very first part of the formant transitions, containing the most rapid movements, is flanked on the left side by the noise burst. If the noise burst is weak or absent, a situation similar to Figure 2d, f and g may be expected to occur, i.e. the spectrum at the onset of the formant transitions will be strongly smeared in the time direction and will thus dominate the region to the left of the onset. In this way a flattening-off effect may be expected to take place at the onset of the formant transitions. If, on the other hand, the energy in the noise burst is comparable to the energy at the onset of the formant transitions, we may expect the formant transitions at onset to be displayed correctly because they will not dominate the region to the left of the onset.

Figure 2.3 shows the oscillogram (2.3a), the smoothed WVD (2.3b), the 6 ms window and 25.6 ms window spectrograms (2.3c and d), and the 6 ms window and 25.6 ms window LP spectra (2.3e and f) of a relevant part of the utterance /du/. The LP-analyses are of order 14. Figure 2.3g and h show the pole frequencies of the 6 ms and 25.6 ms window LP analysis.

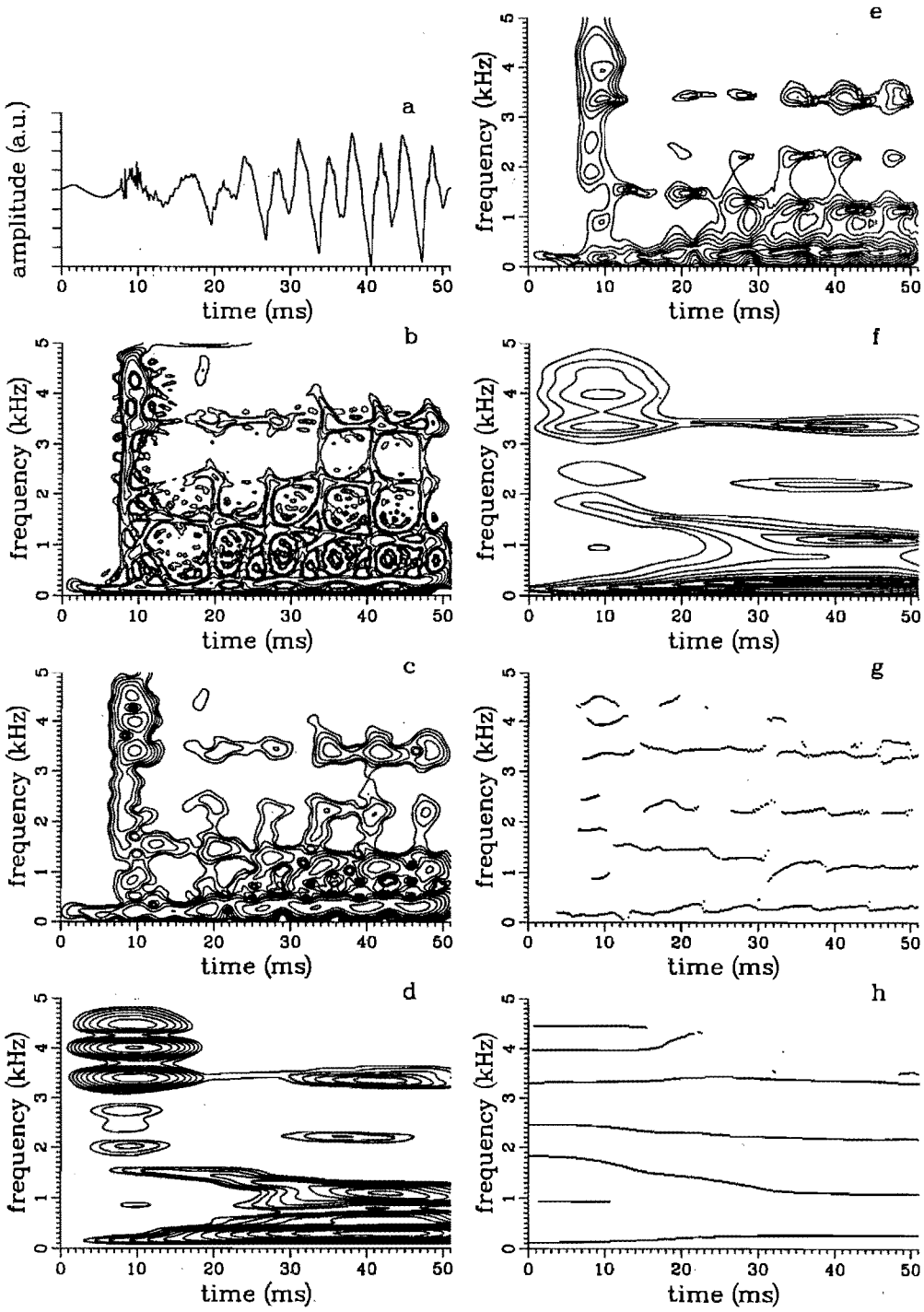
a
amplitude (a.u.)
time (ms)
b
frequency (kHz)
time (ms)
c
frequency (kHz)
time (ms)
d
frequency (kHz)
time (ms)
e
frequency (kHz)
time (ms)
f
frequency (kHz)
time (ms)
g
frequency (kHz)
time (ms)
h
frequency (kHz)
time (ms)

Figure 2.3: *(previous page)* Part of a recorded utterance /du/. a. Waveform. b. Smoothed WVD. c. Spectrogram, analysis window length is 6.0 ms. d. Spectrogram, analysis window length is 25.6 ms. e. LP-spectrogram, analysis window length is 6.0 ms. f. LP-spectrogram, analysis window length is 25.6 ms. g. LP pole frequencies, analysis window length is 6.0 ms. h. LP pole frequencies, analysis window length is 25.6 ms.

Before we describe the results, we want to stress the distinction between the *representation* of a formant and the *interpretation* of this representation. The reason for this is that the measurement of formant onset/offset frequencies and transition rates in natural speech signals is not a clear-cut problem and needs some additional definitions. This point will be addressed more extensively in the next sections.

In the WVD (Figure 2.3b), the parts of the formant between excitations do not appear to be reliable because they tend to split into several lines. The WVD representation therefore does not seem to have a clear advantage over the short-window spectrogram in Figure 2.3c. The most reliable interpretation of the WVD and short-window spectrogram would then be to use only the frequency values close to the instants of excitation. In this way we would obtain a sampled version of the formant track, where the sampling instants are dictated by the glottal pulses. It appears that a straight line can be drawn through the excitations at $t = 20$, 27 and 34 ms, and the resulting transition rate is 25 Hz/ms. If, on the other hand, the energy concentration at 12 ms and 1.5 kHz is assumed to be part of the voiced transition rather than of the burst, and if this point and the excitation at 20 ms are used, a transition rate of 11 Hz/ms is found. These observations indicate that, even if the representation is excellent, still an interpretation step has to be taken and the results depend greatly on the exact method of measurement.

The long-window spectrogram in Figure 2.3d shows some of the staircase-like behavior we have already found in Figure 2.2. In contrast with Figure 2.2, however, the formant track does not break up into separate horizontal structures. A clear flattening-off tendency can be observed at the initial part of the F2. Insufficiently careful measurement of the transition rate may therefore yield too shallow a slope. However, the formant frequencies at the excitation instants are rather close to the values of Figures 2.3b and c. Thus, a transition-rate estimation based on the frequencies at excitation instants will give the most accurate interpretation of the long-window representation.

The 6 ms window LP representation in Figure 2.3e looks very much like the 6 ms window spectrogram (2.3c). In the 25.6 ms window LP spectrogram it can be seen that, in contrast to Figure 2.3d, the second formant has merged with a frequency peak in the release burst. The respective part (from 8 to 17 ms) suggests a transition rate of some 40 Hz/ms. From $t = 17$ ms on, the second formant track closely matches the formant frequencies in Figures 2.3b and c. Thus it seems that, even in the case of prevocalic stops with a strong burst portion and no aspiration, the formant tracks to the left of the first excitation in long-window representations cannot be trusted, because they may be just the connecting line between an energy concentration in the release burst and a formant at the first glottal pulse.

2.3.3 Discussion

We have seen that, when a relatively long analysis window, like 25 ms, is used, the quality of the quasi-stationary representation of highly dynamic speech signals can become rather poor. Formants may seem to have staircase-like tracks or even split up into separate horizontal structures. If no burst is present close to voicing onset, the formant tracks will flatten off and they will seem to have very low initial transition rates. If, on the other hand, the formant is closely flanked by a noise burst with a strong energy concentration, the formant track close to the onset of voicing may be strongly bent.

Generally, we find that "short-time" energy concentrations, like glottal excitations or noise excitations of formants, dominate their immediate surroundings in time and frequency in the quasi-stationary representation. The exact range of influence in time and frequency of a certain energy concentration is dictated by the analysis window and the $\Delta f \Delta t$ trade-off. A long window will of course cause the range to be extensive in time and narrow in frequency, while for a short window the reverse will hold. The time-frequency region in which a certain energy concentration *dominates* the representation, however, also depends on the location and energy of neighboring energy concentrations. Thus, it is clear that, when a long window is used, the most reliable measurement is made close in time to the actual energy concentration, that is, pitch-synchronously.

Along this line of reasoning, we hypothesize that the following signal and analysis parameters have a dominant influence on the quality of a representation and the accuracy of a measurement on this representation. First, the window length, because it determines the range of influence of energy concentrations. A second parameter is of course the formant transition rate. For low transition rates the deviations will be smaller than for high transition rates. Another signal parameter is the fundamental frequency of the voiced speech signal, because an energy concentration due to a glottal excitation of a formant will dominate only within a region as large as a pitch period. Finally, for the same reason, the location in time and frequency and the energy of a noise burst will be important.

2.4 Quantitative experiments

The goal of this section is to give a quantitative description of the deviations due to the quasi-stationarity assumption when measuring formant onset frequencies and formant transition rates. Two aspects will be emphasized: (1) When do deviations occur? (2) If deviations occur, what is their magnitude? The estimations will be based on measurements on synthetic single-formant signals. As stated before, it is not the purpose of this study to make a quantitative comparison between quasi-stationary and non-stationary analysis techniques. Because the most important qualitative differences between these two types of methods, when applied to speech, have been made sufficiently clear in the previous section, the WVD analysis was not used in the quantitative experiments.

Experiment 1 is intended to give a general quantitative description of the devia-

tions for a wide range of the most relevant signal and analysis parameters suggested in the previous section. Experiment 2 zooms in on a subset of the parameter settings used for experiment 1. An analytical model is devised and is fitted on the data of experiment 2. This model predicts when deviations occur and how large these deviations are.

2.4.1 Experiment 1

Method

Generation of signals

In the previous section it was suggested that the analysis and signal parameters that have the largest influence on the accuracy of the measurements are (1) the length (wl) of the analysis window, (2) the fundamental frequency (F_0) of the signal, (3) the transition rate (R) of the formant, (4) the position in time and frequency of the main energy concentration in the burst relative to the first glottal pulse and (5) the burst energy. It is not realistic to aim at an exhaustive description of the influence of these parameters, because the number of necessary signals and measurements would be too large. We have therefore chosen to limit the number of parameter values for signal generation and analysis, while ensuring that we may still make general predictions on when deviations occur and how large they are.

For the purpose of experiment 1, 45 synthetic signals were created. All signals were similar to signal $S2$. The signals are created by filtering an excitation signal by a time-varying second-order IIR-filter. The excitation signal consists of an impulse at time $t = 14.9$ ms, followed by 3 or more impulses at periodic intervals. Three fundamental periods are used: 12.5, 6.7, and 4.0 ms, corresponding to an F_0 of 80.0, 149.3, and 250.0 Hz, respectively. The IIR-filter (formant) is stationary from the beginning of the signal to the instant of the first impulse, and has a frequency of 1.0 kHz. At the instant of the first impulse ($t = 14.9$ ms), the formant frequency starts increasing linearly with time at various rates. The used rates are 0, 10 and 100 Hz/ms. As for signal $S2$, the filter coefficients are adjusted at every sample instant. Throughout the signal, the formant bandwidth is 150 Hz.

To a subset of the signals, a noise burst is added. The noise burst consists of a burst of white noise, which starts at $t = 4.9$ ms and ends at $t = 8.8$ ms, and which is filtered by a stationary second-order IIR-filter with a resonance frequency of either 1.0 kHz or 1.3 kHz and a bandwidth of 150 Hz. The burst energy was adjusted in the following way. Using an analysis window of 6 ms, an STFT was made of the signal consisting of a burst and a quasi-periodic part. The burst energy was adjusted in such a way that the maximum STFT amplitude due to the burst was either 3 dB below or 3 dB above the maximum STFT amplitude due to the first excitation. The cases where the burst energy is 3 dB lower or higher then the first glottal pulse energy will be referred to as the soft burst and loud burst conditions, respectively.

Thus, we created 5 burst conditions (no burst, soft burst at 1.0 kHz, soft burst at 1.3 kHz, loud burst at 1.0 kHz and at 1.3 kHz), 3 fundamental frequencies (80.0 Hz, 149.3 Hz, 250.0 Hz) and 3 transition rates (0 Hz/ms, 10 Hz/ms and 100 Hz/ms),

for a total of 45 signals.

Method of analysis
The signals were analyzed using STFT and LP. In both cases, Hanning windows were used of 4 different lengths: 3.0, 6.0, 12.5, 25.0 ms. In order to obtain the maximum accuracy, given the signal and the window length, a pitch-synchronous analysis was used. Note that many possible errors discussed in the previous section, e.g. deviations caused by asynchronous window-positioning, are avoided in this way.

Because we use impulses as glottal excitations, the discontinuities in the signal at the instants of excitation are large. These discontinuities cause the pole frequencies in the immediate surroundings of the instant of excitation to deviate from the general smooth track, especially for short analysis windows (e.g. Figure 2.2c). For all analyses we therefore centered the analysis windows not at the instants of excitation, but 1.0 ms after the instant of excitation, where the tracks of the pole frequencies were again regular. The measurements were corrected for this time shift. For example, an exact measurement of the formant onset frequency of a formant starting at a frequency of 1000 Hz and increasing at a rate of 100 Hz/ms should yield a value of 1100 Hz.

Two measurements were made on all signals: the formant onset frequency $\hat{F}_{ons}$ and the formant transition rate $\hat{R}$. In this notation, $\hat{X}$ refers to a measured value as opposed to the true value X. For these 2 measurements, only two windows were necessary, one for each of the first two excitations. As for the qualitative experiments discussed in the previous section, the windowed segments were converted into Fourier amplitude spectra for the spectrographic measurements and into inverted group delay spectra of the LP-coefficients for the LP-measurements. For all LP-analyses a 2-pole model was used. Again, the formant frequencies were measured by picking the peaks from these spectra.

Because the amplitude spectra for the spectrographic measurements contained harmonics for many of the parameter settings, a spectral envelope needed to be calculated before picking the peaks. As in the case of signals $S2$ and $S3$, a simple smoothing was applied by convolving the amplitude spectrum with a Hamming window. The effective bandwidth of the Hamming window was 98 Hz, 234 Hz, and 470 Hz for signals with an F_0 of 80.0 Hz, 149.3 Hz and 250.0 Hz, respectively. These figures were arrived at by carefully inspecting the effects of smoothing on many signals. The smoothing was applied to all analyses made with a 25.0 ms window, to analyses of signals with an F_0 of 149.3 Hz and 250.0 Hz, made with a 12.5 ms window, and to analyses of signals with an F_0 of 250.0 Hz, made with a 6.0 or 3.0 ms window.

$\hat{F}_{ons}$ is defined as the formant frequency for the first window. $\hat{R}$ is defined as the difference between the formant frequencies for the second and the first window, divided by the distance between the centers of the two windows (the fundamental period).

The peak picking was carried out by selecting the frequency point which had the maximum amplitude. Because in all our cases the frequency axis from 0 Hz to 5 kHz is represented by 257 points, the between-sample distance equals 19.53 Hz.

Thus, the basic measurement error in the $\hat{F}_{ons}$ lies within the range of plus or minus half the between-sample distance, that is, within [-9.8 Hz, 9.8 Hz]. The basic error in $\hat{R}$ depends on the time interval between the 2 formant frequency measurements, which is the fundamental period of the signal. The maximum positive or negative measurement error equals the between-sample distance on the frequency axis, divided by the fundamental period. This yields maximal basic measurement errors of 1.6 Hz/ms, 2.9 Hz/ms, and 4.9 Hz/ms for $\hat{F}_{ons}$'s of 80.0 Hz, 149.3 Hz and 250.0 Hz, respectively. Of course, the basic error can be made arbitrarily small by using some kind of interpolation technique, e.g. increasing the FFT-size by using zero padding. However, our basic errors were small enough compared to the observed effects.

Results

The deviations in $\hat{F}_{ons}$ and $\hat{R}$ for the no-burst, soft-burst and loud-burst conditions are shown in Table 2.I, 2.II, and 2.III, respectively. In all tables, the left column shows the parameter settings. In table 2.I, the middle and right columns show the measurement deviations for the LP-analysis and the spectrographic analysis, respectively. In tables 2.II and 2.III, columns 2 and 3 show the measurement deviations for the LP-analysis and the spectrographic analysis for the burst frequency of 1.0 kHz, columns 4 and 5 show the deviations for the burst frequency of 1.3 kHz.

The LP-results for the rate of 100 Hz/ms are shown graphically in Figure 2.4. As the data are very similar for spectrogram and LP, we chose to display only the LP-results, because the LP-data are generally slightly more regular than the spectrogram data. All figures on the left display the deviation in $\hat{R}$ as a function of wl, the figures on the right show the deviation in $\hat{F}_{ons}$ as a function of wl. The top, middle and bottom 2 figures contain data for the signals with F_0's of 80.0 Hz, 149.3 Hz and 250.0 Hz, respectively.

Let us first discuss the stationary condition, that is, $R = 0$ Hz/ms and no burst. The measurements are very accurate and the basic measurement error is exceeded only in very few cases. This compares favorably with the results of Monsen and Engebretson (1983), discussed earlier. When R is not zero, in many cases the measurement is still accurate. For F_0's up to 149.3 Hz, only the largest window length (25 ms) gives a deviation which exceeds the basic error. For the F_0 of 250.0 Hz, however, the accuracy starts deviating from zero earlier, that is, for a window of 12.5 ms. Therefore, it seems that both wl and F_0 determine whether deviations occur or not.

The data suggest that when significant deviations occur, they increase with increasing R, increasing F_0 and increasing wl. The dependence on F_0 and wl is evident in Figure 2.4.

The noise burst only has a significant influence on the measurements in the loud-burst condition. Note that in this condition large deviations may occur even for the stationary formant ($R = 0$ Hz/ms) when the burst frequency deviates from the formant frequency (see Table 2.III). For $wl = 25$ ms in the loud-burst condition, Figure 2.4 shows a regular pattern. When the burst frequency is 1.3 kHz (symbol $\triangle$), $\hat{F}_{ons}$ is higher than for the no-burst condition (symbol o). On the other

hand, when the burst frequency is 1.0 kHz (symbol ∇), $\hat{F}_{ons}$ is lower than for the no-burst condition. So like in Figure 2.3, we find that, when the burst is stronger than the first excitation, the formant measurement of the first excitation is biased towards the burst frequency.

Under maximally unfavorable conditions, deviations are as large as 200 Hz in $\hat{F}_{ons}$ and 50% in $\hat{R}$.

Table 2.I: Deviations in measured formant onset frequency and transition rate in the burstless condition. For further explanation, see text.

			LP		spectrogram	
rate	$F0$	wl	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate
0	80	3.0	-4	0.0	-4	0.0
		6.0	-4	0.0	-4	0.0
		12.5	-4	0.0	-4	0.0
		25.0	-4	0.0	-4	0.0
	150	3.0	-4	0.0	-4	0.0
		6.0	-4	0.0	-4	0.0
		12.5	-4	0.0	-4	0.0
		25.0	-4	0.0	-4	2.9
	250	3.0	-4	0.0	-4	0.0
		6.0	-4	0.0	-23	4.9
		12.5	-4	0.0	-23	0.0
		25.0	-4	0.0	-23	0.0
10	80	3.0	6	-0.6	-4	0.9
		6.0	-14	0.9	-4	0.9
		12.5	-14	0.9	6	-0.6
		25.0	-14	0.9	-4	0.9
	150	3.0	6	-1.3	-14	1.7
		6.0	-14	1.7	-14	1.7
		12.5	-14	1.7	-14	1.7
		25.0	6	-1.3	6	-1.3
	250	3.0	6	-0.2	-14	-0.2
		6.0	-14	-0.2	-14	-0.2
		12.5	-14	-0.2	-14	-0.2
		25.0	6	-5.1	-14	-5.1
100	80	3.0	-6	0.0	-6	1.6
		6.0	-26	0.0	33	0.0
		12.5	-26	1.6	72	0.0
		25.0	-6	0.0	72	1.6
	150	3.0	-6	-0.9	-6	2.0
		6.0	-26	2.0	33	-0.9
		12.5	-6	-0.9	33	-0.9
		25.0	72	-24.2	52	-0.9
	250	3.0	-6	2.5	-6	2.5
		6.0	-26	2.5	-6	2.5
		12.5	33	-12.1	13	-2.3
		25.0	150	-46.3	91	-41.4

2.4.2 Experiment 2

The goal of experiment 2 is to focus on the no-burst condition that was part of experiment 1 and to derive an analytical expression for when deviations occur and how large they are.

Table 2.II: Deviations in measured formant onset frequency and transition rate, soft-burst condition. For further explanation, see text.

			soft burst, 1.0 kHz				soft burst, 1.3 kHz			
			LP		spectrogram		LP		spectrogram	
rate	$F0$	wl	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate
0	80	3.0	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		6.0	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		12.5	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		25.0	-4	0.0	35	-3.1	-4	0.0	-4	0.0
	150	3.0	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		6.0	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		12.5	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		25.0	-4	0.0	-4	2.9	-4	0.0	-4	2.9
	250	3.0	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		6.0	-4	0.0	-23	4.9	-4	0.0	-23	4.9
		12.5	-4	0.0	-23	0.0	-4	0.0	-23	0.0
		25.0	-4	0.0	-23	0.0	-4	0.0	-4	-4.9
10	80	3.0	6	-0.6	-14	0.9	6	-0.6	-14	0.9
		6.0	-14	0.9	-14	0.9	-14	0.9	-14	0.9
		12.5	-14	0.9	6	-0.6	-14	0.9	6	-0.6
		25.0	-14	0.9	25	-0.6	6	-0.6	6	0.9
	150	3.0	6	-1.3	-14	1.7	6	-1.3	-14	1.7
		6.0	-14	1.7	-14	1.7	-14	1.7	-14	1.7
		12.5	-14	1.7	-14	1.7	-14	1.7	-14	1.7
		25.0	6	-1.3	6	-1.3	25	-4.2	6	-1.3
	250	3.0	6	-0.2	-14	-0.2	6	-0.2	-14	-0.2
		6.0	-14	-0.2	-14	-0.2	-14	-0.2	-14	-0.2
		12.5	-14	-0.2	-14	-0.2	-14	-0.2	-14	-0.2
		25.0	6	-5.1	-14	-5.1	6	-5.1	6	-10.0
100	80	3.0	-6	0.0	-6	1.6	-6	0.0	-6	1.6
		6.0	-26	0.0	33	0.0	-26	0.0	33	0.0
		12.5	-26	1.6	52	1.6	-26	1.6	52	1.6
		25.0	-26	1.6	-26	9.4	-6	0.0	72	1.6
	150	3.0	-6	-0.9	-6	2.0	-6	-0.9	-6	2.0
		6.0	-26	2.0	33	-0.9	-26	2.0	33	-0.9
		12.5	-6	-0.9	33	-0.9	-6	-0.9	33	-0.9
		25.0	72	-24.2	52	-0.9	91	-27.1	52	-0.9
	250	3.0	-6	2.5	-6	2.5	-6	2.5	-6	2.5
		6.0	-26	2.5	-6	2.5	-26	2.5	-6	2.5
		12.5	33	-12.1	13	-2.3	33	-12.1	13	-2.3
		25.0	130	-41.4	72	-36.5	150	-46.3	91	-41.4

Method

Generation of signals

In total, 25 burstless single-formant signals were used in this experiment, 9 of which were already used in experiment 1. In order to determine the influence of the signal parameters F_0 and R more precisely, more values for these parameters were used. The fundamental period had values of 12.5 ms, 9.3 ms, 6.7 ms, 5.2 ms and 4.0 ms, corresponding to F_0's of 80.0 Hz, 107.5 Hz, 149.3 Hz, 192.3 Hz and 250.0

Table 2.III: Deviations in measured formant onset frequency and transition rate, loud-burst condition. For further explanation, see text.

			loud burst, 1.0 kHz				loud burst, 1.3 kHz			
			LP		spectrogram		LP		spectrogram	
rate	$F0$	wl	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate	ΔF_{ons}	Δ rate
0	80	3.0	-4	0.0	-4	0.0	16	-1.6	-4	0.0
		6.0	-4	0.0	16	-1.6	-4	0.0	-4	0.0
		12.5	-4	0.0	55	-4.7	-4	0.0	-4	0.0
		25.0	-4	0.0	35	-3.1	113	-9.4	-4	0.0
	150	3.0	-4	0.0	-4	0.0	16	-2.9	-4	0.0
		6.0	-4	0.0	16	-2.9	-4	0.0	-4	0.0
		12.5	-4	0.0	16	-2.9	-4	0.0	-23	2.9
		25.0	-4	0.0	-4	2.9	94	-14.6	16	0.0
	250	3.0	-4	0.0	-4	0.0	16	-4.9	-4	0.0
		6.0	-4	0.0	-4	0.0	-4	0.0	-4	0.0
		12.5	-4	0.0	-4	0.0	-4	0.0	-23	4.9
		25.0	-4	0.0	-23	0.0	55	-14.7	55	-14.7
10	80	3.0	6	-0.6	-14	0.9	6	-0.6	6	-0.6
		6.0	6	-0.6	6	-0.6	-14	0.9	-14	0.9
		12.5	6	-0.6	45	-3.8	6	-0.6	6	-0.6
		25.0	-14	0.9	25	-0.6	103	-8.4	6	0.9
	150	3.0	6	-1.3	-14	1.7	6	-1.3	6	-1.3
		6.0	6	-1.3	6	-1.3	-14	1.7	-14	1.7
		12.5	6	-1.3	25	-4.2	6	-1.3	-14	1.7
		25.0	-14	1.7	6	-1.3	103	-15.8	25	-4.2
	250	3.0	6	-0.2	-14	4.7	6	-0.2	6	-0.2
		6.0	6	-5.1	-14	-0.2	6	-5.1	-14	-0.2
		12.5	6	-5.1	-14	-0.2	6	-5.1	-14	-0.2
		25.0	-14	0.2	33	-0.2	84	-24.7	64	-19.8
100	80	3.0	13	-1.6	13	0.0	-6	0.0	-6	1.6
		6.0	-6	-1.6	52	-1.6	-26	0.0	33	0.0
		12.5	-6	0.0	33	3.1	-6	0.0	13	4.7
		25.0	-65	4.7	-45	10.9	72	-6.3	209	-9.4
	150	3.0	13	-3.8	13	-0.9	-6	-0.9	-6	2.0
		6.0	-6	-3.8	52	-3.8	-26	-0.9	33	-0.9
		12.5	-6	-0.9	52	-3.8	-6	-0.9	33	-0.9
		25.0	-6	-9.6	-84	19.5	130	-33.0	50	-15.5
	250	3.0	13	-2.3	-6	2.5	-6	2.5	-6	2.5
		6.0	6	2.5	13	-2.3	-26	2.5	-6	2.5
		12.5	52	-12.1	52	-12.1	52	-17.0	33	-12.1
		25.0	52	-26.8	-26	-12.1	170	-56.1	130	-51.2

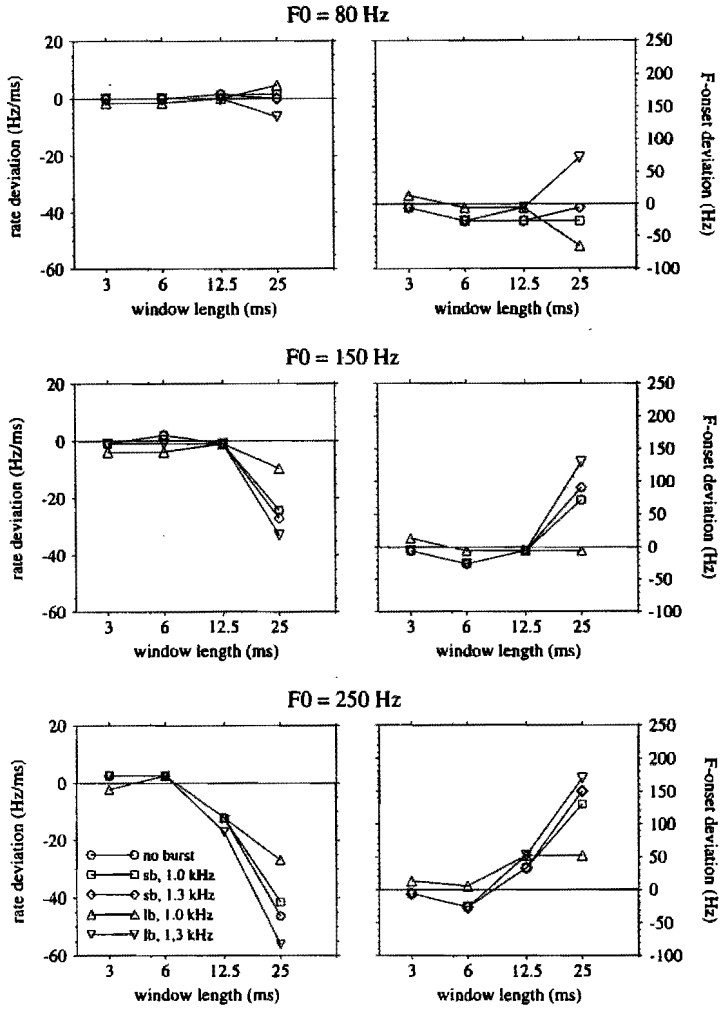

Figure 2.4: Results of experiment 1 for $R = 100$ Hz/ms. The deviations in $\hat{R}$ are displayed in the figures on the left, the deviations in $\hat{F}_{ons}$ on the right. Deviations for $F_0 = 80.0$, 149.3 and 250.0 Hz are presented in the figures at the top, middle and bottom, respectively.

Hz, respectively. For R, the values 0 Hz/ms, 10 Hz/ms, 25 Hz/ms, 50 Hz/ms and 100 Hz/ms were used.

Method of analysis

For experiment 2, only LP-analysis was applied. The window lengths and other details of the analysis method were identical to experiment 1.

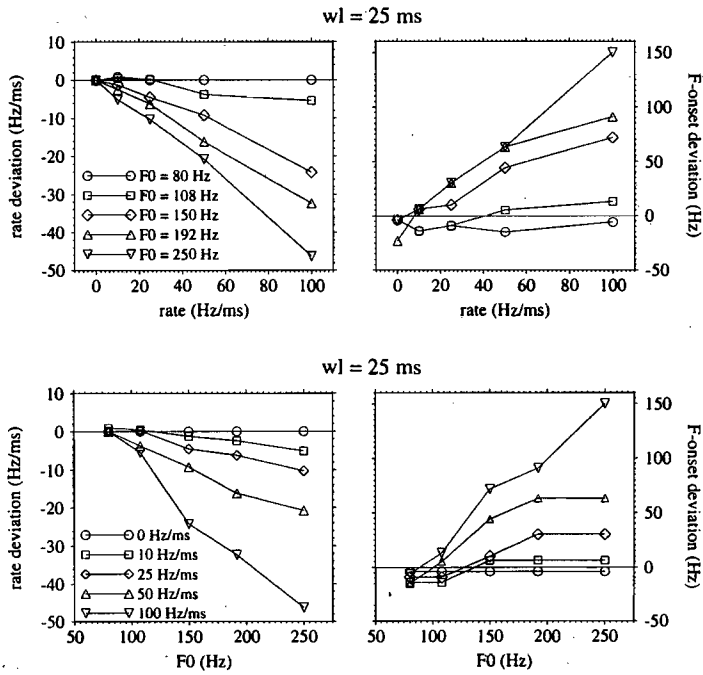

Figure 2.5: Results of experiment 2 for $wl = 25$ ms. Again, the deviations in $\hat{R}$ and $\hat{F}_{ons}$ are displayed in the figures on the left and right, respectively. In the top figures, the deviations are plotted as a function of R, in the bottom figures the deviations are plotted as a function of F_0.

Results

The deviations for the largest window length (25 ms) are displayed in Figure 2.5. The figures on the left and right again show the deviation in $\hat{R}$ and $\hat{F}_{ons}$, respectively. In order to make the dependence of the deviations on R and F_0 clear, the top figures give the deviations as a function of R, the bottom figures contain the same data, now plotted as a function of F_0.

Figure 2.5 suggests that the deviations in $\hat{R}$ and $\hat{F}_{ons}$ depend almost linearly on F_0 and R when the other parameters are held constant. Furthermore, the lines depicting the deviations as a function of R go through the origin (y-intercept 0). Different values of the fundamental frequency only have an influence on the slope of the lines, not on the y-intercept.

The lines depicting the deviations as a function of F_0 clearly have an y-intercept which is not zero. Different rate-values clearly influence the slope of the lines, but whether the y-intercepts are also different is not evident.

We will now construct a simple analytical model which predicts when deviations occur and how large they are. First of all, the numerical data and Figures 2.4 and

2.5 show that for a certain subset of realistic parameter values, there is no error in the measurement due to the quasi-stationarity assumption. Arguing along the lines of the discussion of the qualitative experiments (section 2.3.3), it seems reasonable to suggest that deviations due to quasi-stationarity only occur when there are more than a certain number of periods present within the analysis window, each of which excites the gliding formant at a different frequency. The number of periods within the analysis window is equal to $F_0 \cdot wl$, with F_0 expressed in [Hz] and wl in [s]. The critical number of periods within the window will be denoted by C_0. If the number of periods within the window exceeds C_0, we hypothesize that (1) the deviation will be proportional to R, and (2) the deviation has a linear relation with the number of periods within the window $F_0 \cdot wl$. Thus, the following model arises:
If $(wl \cdot F_0) > C_0$ then

$$|\Delta\hat{R}| = C_1 \cdot |R| \cdot (wl \cdot F_0 - C_0), \tag{2.6}$$

and

$$|\Delta\hat{F}_{ons}| = C_2 \cdot |R| \cdot (wl \cdot F_0 - C_0), \tag{2.7}$$

else

$$|\Delta\hat{R}| = |\Delta\hat{F}_{ons}| = 0, \tag{2.8}$$

where C_0, C_1 and C_2 are parameters, the values of which must be determined by fitting the model to the data of experiment 2. If we express R and $\Delta\hat{R}$ in [Hz/ms], wl in [s], and F_0 and $\Delta\hat{F}_{ons}$ in [Hz], a least-squares model fit yields the following parameter values:
$C_0 = 2.2$,
$C_1 = 0.12$,
$C_2 = 0.37$ ms.

In order to describe the goodness-of-fit of the model, correlation coefficients between observed deviations and predicted deviations were calculated for the data points for which $(wl \cdot F_0) > 2.2$, that is, for 30 out of 100 data points for R and for 30 out of 100 for F_{ons}. The correlation coefficients are 0.99 for R and 0.97 for F_{ons}. It may be hypothesized that this extremely high goodness-of-fit is caused by the fact that the density of data points in the region where the deviations are small is higher than the density in the region where the deviations are large. This hypothesis was checked by recalculating the correlation coefficients for the data points for which $(wl \cdot F_0) > 3.0$. This restriction excludes 12 of 30 data points at the high-density end of the line both for R and for F_{ons}. The new correlation coefficients were even slightly higher: 0.99 for R and 0.98 for F_{ons}. Thus, we see that the high correlation coefficients are not caused by an artifact, and the model fit is truly very good.

A value of 2.2 for the parameter C_0 means that there are no deviations due to the quasi-stationarity assumption as long as the analysis window is not wider than 2.2 pitch periods. This value only holds for the Hanning window. The effective length of a window was earlier defined as the interval between the two -3dB points of the window, which is half the total length in case of a Hanning window. Therefore, in

general, it is concluded that effectively there may not be more than 1.1 period in the window. Thus, as a general rule of thumb, one should use an analysis window that is effectively not longer than one pitch period, when analyzing highly dynamic speech signals. This value agrees with the intuitive notion that a good measurement of formant frequency is made when one glottal excitation dominates the windowed speech segment.

Using the parameter value of 2.2, we can make a general statement about the adequacy of the wideband spectrogram for analyzing highly dynamic speech. As stated earlier, the wideband spectrogram has an analysis window of roughly 6 ms, in case a Hanning window is used. Substitution of this value in the expression

$$wl \cdot F_0 \leq 2.2 \tag{2.9}$$

leads to

$$F_0 \leq 370\text{Hz} \tag{2.10}$$

Thus, our study indicates that the accurate measurements can be made from the wideband spectrogram as long as F_0 is smaller than 370 Hz. This is always the case for male voices, and very often for female voices.

2.5 Conclusions and recommendations

Short energy concentrations, like glottal excitations of formants, dominate their immediate surrounding area in a quasi-stationary TFR. If a relatively long analysis window is used, the influence of each of these concentrations is spread along the time dimension, which may corrupt the representation. The most important unwanted effects caused by this mechanism are staircase-like formant tracks, flattening-off of formants close to voicing onset, and bending of the formant towards a strong energy concentration in the release burst.

The parameters that have the largest influence on the quality of the representation are the length of the analysis window, the transition rate of the formant, the fundamental frequency, the position in time and frequency of the burst and the energy in the burst. The most accurate analysis using a quasi-stationary method is made when windows are positioned pitch-synchronously.

A quantitative analysis of the influence of the mentioned parameters shows that no deviations due to the quasi-stationarity assumption are to be expected when the inequality $wl \cdot F_0 \leq 2.2$ is met, in case a Hanning window is used. This inequality can be generalized to the statement that no deviations due to the quasi-stationarity assumption are to be expected when the effective length of the analysis window is not larger than one pitch period. Expressions predicting the deviations in $\Delta\hat{R}$ and $\Delta\hat{F}_{ons}$ are derived for situations where the inequality is not met. Based on the results and the derived inequality, it is expected that the wideband spectrogram is a reliable tool for making measurements on rapidly moving formants for signals with an F_0 that is lower than 370 Hz. A word of caution is however in order here. The general conclusions are derived from experiments with very simple synthetic single-formant signals. These signals may differ significantly from natural speech signals

in a number of respects, such as glottal source signal, and variations in bandwidth within and across fundamental period.

The following practical recommendations are suggested. First of all, it is important to be aware of the maximal F_0 in the signals that are to be analyzed. In the case of stop consonants, extra care is in order, because just after voicing onset, where the formants are most dynamic, the F_0 may for a short time be much higher than in the following less dynamic parts. Ohde (1984) found that the ratio between F_0 just after voicing onset and in the vowel target may be as high as 1.4.
Next, an analysis window must be chosen with an effective length of about the shortest pitch period in the signals to be analyzed.
Finally, pitch-synchronous window positioning is recommended.

Chapter 3

The multi-layer perceptron as a model of human categorization behavior. I. Theory[1]

Abstract

A model of human categorization behavior is presented in which the multi-layer perceptron (MLP) is the central part. First the modeling behavior of the single-layer perceptron (SLP) is studied through an analysis of the mathematical expressions and a discussion of a number of theoretical examples. A number of similarities and differences between the SLP and the well-known similarity-choice model (SCM) are discussed and it is shown that the SLP and SCM coincide in a certain limit case. Finally, the theory for the SLP is extended to the two-layer perceptron (TLP). It is shown that the TLP has substantial modeling power, but it can become hard to interpret. A linearization of the sigmoid functions in the hidden nodes is introduced, which facilitates interpretation.

3.1 Introduction

Categorization plays an important role in everyday processes of perception and cognition, such as the recognition of spoken and written language. A number of formal models for the categorization process have been developed, such as the similarity-choice model (SCM, Shepard, 1958; Luce, 1963), multi-dimensional scaling (MDS, Kruskal, 1964), the fuzzy logical model of perception (Oden and Massaro, 1978), multiple-exemplar models (Medin and Schaffer, 1978; Nosofsky, 1986), and general recognition theory (GRT, Ashby and Perrin, 1988).

During the last decade, connectionist models have become very popular, not in the least for the modeling of perceptual and cognitive processes (e.g. McClelland, Rumelhart and the PDP research group, 1986; Quinlan, 1991). The multi-layer perceptron (MLP) is probably the mathematically best-developed connectionist model (e.g. Lippman, 1987; Hertz *et al.*, 1991; Haykin, 1994). Nevertheless, the MLP is still met with considerable suspicion as a model for human categorization behavior (e.g. Massaro, 1988), predominantly due to lack of understanding of its modeling capabilities and because the MLP is often considered to be a "black box" which would not allow for the extraction of knowledge from its parameters.

[1]Based on: Smits, R., and Ten Bosch, L. (**1994a**), "The multi-layer perceptron as a model of human categorization behavior. I. Theory," submitted to J. Math. Psych.

It is the purpose of this chapter to give a formal description of the MLP as a model of human categorization behavior. We will provide insight in the modeling behavior and capabilities of the MLP by studying the relevant mathematical expressions and a number of examples. Furthermore, we will compare the MLP with some alternative models, in particular the SCM. In the next chapter, a number of important practical issues are addressed, such as how to estimate the model (*training*), and how to evaluate the performance and generalizability of the model (*testing*), and a practical example is elaborated.

The structure of this chapter is as follows. In the next section the general structure of the model is set up. Next, the basic mathematics of the simplest MLP, that is, the *single-layer perceptron* (SLP), are described. The SCM is recapitulated in section 3.4 and the modeling behavior of the SLP is compared to that of the SCM in section 3.5. Section 3.6 presents an intermediate discussion of the prototype concept which is relevant for the comparison of the SLP and the SCM. In section 3.7, the theory is extended to MLPs with one hidden layer, and in the final section the results are discussed and summarized.

3.2 General model structure

On each trial in a categorization experiment a subject is presented with one of N_s stimuli and is required to assign one of N_r predefined labels to this stimulus. Essentially, in a categorization or classification[2] task $N_r < N_s$, while for identification $N_r = N_s$. For the sake of simplicity, we assume that each of the N_s stimuli is presented N_p times to the subject. Furthermore, we assume that on each trial the categorization of the presented stimulus does not depend on previous trials. Based on this assumption, the results of the experiment can, without loss of information, be summarized in a stimulus-response matrix consisting of N_s rows and N_r columns. Each entry R_{ij} in the stimulus-response matrix denotes the number of times the stimulus S_i has been labeled as belonging to category C_j. Note that $\forall_i : \sum_{j=1}^{N_r} R_{ij} = N_p$ and $\sum_{i=1}^{N_s} \sum_{j=1}^{N_r} R_{ij} = N_s N_p$.

We will now propose a model which simulates the mapping of a set of stimuli onto a set of categorical responses. The proposed model consists of 3 steps:

1. extraction of stimulus features,
2. evaluation of class probabilities on the basis of stimulus features,
3. actual choice of a single response class on the basis of class probabilities.

These steps can be described as a cascade, as is shown in Figure 3.1.

The model fits into the general framework described by Ashby (1992), which consists of 3 stages: the *representation stage*, the *retrieval stage*, and the *response selection stage*. Throughout the rest of the chapter we will use this terminology introduced by Ashby. Note also that the model corresponds to the central branch of the general model for stop-consonant perception presented in chapter 1 (Figure 1.1).

[2]Throughout this thesis the terms classification and categorization are used to indicate one and the same paradigm.

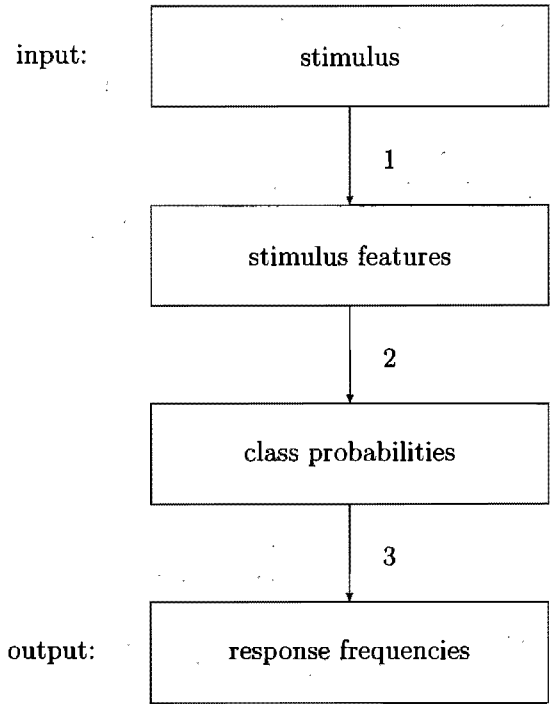

Figure 3.1: Schematic representation of the model for perceptual classification. 1, 2, and 3 indicate the representation stage, retrieval stage, and response selection stage, respectively.

In the *representation stage* the stimulus S_i is transformed from the physical domain into an internal representation, which is a vector containing N_F stimulus features. Thus, in the representation stage, each stimulus is mapped to a point in an N_F-dimensional *feature space* $\mathcal{F}$. Feature vectors in $\mathcal{F}$ will be denoted by $\mathbf{F}$, the *value* of $\mathbf{F}$ for stimulus S_i is denoted by $\mathbf{F_i}$, and the kth component of $\mathbf{F_i}$, that is, the value of feature k of stimulus S_i, is denoted by F_{ik}. Generally, the specific choice of features in a model will be based on either knowledge of the potential perceptual relevance of various stimulus features, or on knowledge of statistics of the stimulus set. Often, in the preparation of the experimental stimuli, a number of stimulus features is explicitly manipulated in order to test their perceptual relevance. The feature extraction is assumed to be deterministic, which means that each presentation of a particular stimulus will lead to the same feature vector, that is, this stage is noiseless.

In the *retrieval stage* the feature vector $\mathbf{F_i}$ of each stimulus S_i is mapped to a vector $\mathbf{p_i}$ of length N_r containing the *a posteriori* probabilities of choosing each of

the response categories. The jth component of the probability vector $\mathbf{p_i}$ is denoted by p_{ij}. The feature-to-class-probability mapping is assumed to be deterministic. Here, it is modeled by an MLP.

In the *response selection stage* the actual labeling takes place. This labeling is assumed to be probabilistic, and here it is modeled by a multinomial function. Suppose that after N_p presentations of stimulus S_i, a subject has generated an output vector $\mathbf{R_i}$ of length N_r. Each component R_{ij} of $\mathbf{R_i}$ denotes the number of times the subject has assigned stimulus S_i to class C_j. The multinomial model states that the probability $p(\mathbf{R_i}|\mathbf{p_i})$ of generating the response vector $\mathbf{R_i}$ after N_p presentations of stimulus S_i, given class probabilities $\mathbf{p_i}$, equals

$$p(\mathbf{R_i}|\mathbf{p_i}) = N_p! \prod_{j=1}^{N_r} \frac{p_{ij}^{R_{ij}}}{R_{ij}!} \tag{3.1}$$

The 3-stage mapping can be summarized as follows:

$$S_i \longrightarrow \mathbf{F_i} \longrightarrow \mathbf{p_i} \longrightarrow \mathbf{R_i} \tag{3.2}$$

We will briefly compare the general assumptions for each of the 3 stages described above to two major existing models of human categorization behavior, that is the *prototype model*, e.g. multi-dimensional scaling (MDS, e.g. Kruskal, 1964), or the classical Shepard-Luce similarity-choice model (SCM, e.g. Shepard, 1958; Luce, 1963), and *general recognition theory* (GRT, e.g. Ashby and Perrin, 1988; Ashby and Maddox, 1993).

The central assumption in *prototype models* is that a "prototype" or ideal reference exemplar exists for each response category, and that the categorization of each stimulus is based on a calculation of similarity of this stimulus to each of the prototypes. Generally, the representation and retrieval stages are assumed to be deterministic, while the response selection stage is assumed to be probabilistic.

In *GRT* it is assumed that the subject divides the feature space $\mathcal{F}$ into disjunct subspaces or "response regions", each of which is associated with a distinct category label. On each experimental trial the presented stimulus is mapped to a point in the feature space and the stimulus receives the label of the corresponding response region. The representation stage in GRT is assumed to be essentially probabilistic, while the retrieval and response selection stages are generally assumed to be deterministic.

As stated earlier, the representation and retrieval stages in the model proposed in this chapter are assumed to be deterministic, and the response selection stage is assumed to be probabilistic. In this sense, our model is equivalent to the prototype-based models and differs conceptually from GRT.

3.3 The multi-layer perceptron

The MLP is the core of our categorization model. It is used to model the retrieval stage, that is, the mapping of stimulus features to class probabilities. In this section, we will develop the theory for a "single-layer perceptron" (SLP), which is a

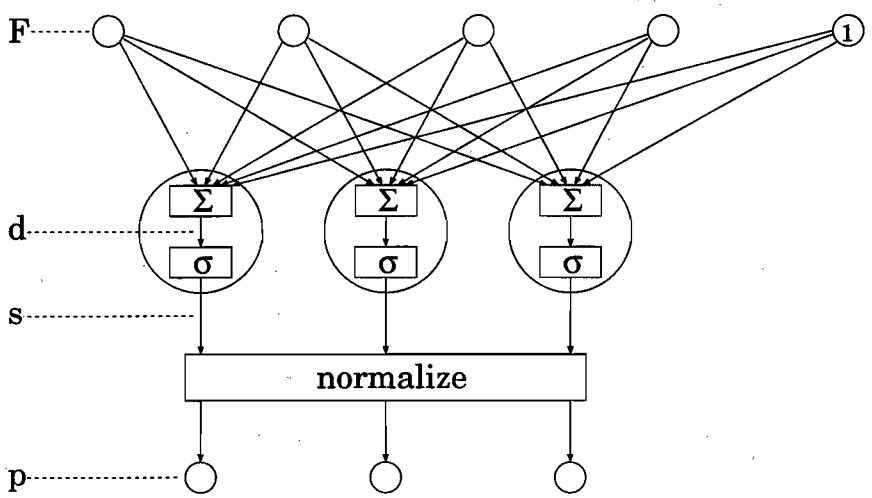

Figure 3.2: Schematic representation of the SLP as it is used in our model. The top row of small circles represents the input layer including the bias. The middle row of large circles represents the output layer. The symbols Σ and σ represent summation and sigmoid transformation, respectively.

perceptron consisting an input layer and an output layer, and no hidden layers.[3] In a later section we will extend this theory to an MLP with one hidden layer. The structure of the SLP used in our model is shown schematically in Figure 3.2.

The stimulus features are clamped to the input nodes, represented as the top row of circles in Figure 3.2. The input nodes pass the features unchanged. One input node is assigned to each stimulus feature. The SLP in Figure 3.2 has 4 input nodes. The top right circle with the number "1" is the *bias node*. Instead of transferring a feature value, this node simply outputs a fixed value 1 (e.g. Haykin, 1994).

All input nodes, including the bias node are connected to all output nodes. A weight is associated with each connection. The feature value traveling through the connection is multiplied by the respective weight before reaching the output node. The weights are the parameters of the model. The number of weights N_w, including biases, in the SLP equals

$$N_w = (N_F + 1)N_r \tag{3.3}$$

In the output nodes two processing steps are made. First, the weighted feature

[3] A "two-layer perceptron" may seem to be a more appropriate name for perceptron consisting of an input layer and an output layer. Nevertheless, in accordance with other authors (e.g. Haykin, 1994; Lippman, 1987), we have chosen to use the term single-layer perceptron because the output layer is the only "real" layer in the sense that it consists of neurons which perform the summation and nonlinear activation transfer.

values of stimulus S_i are summed, yielding a quantity d_{ij}:

$$d_{ij} = b_j + \sum_{k=1}^{N_F} w_{kj} F_{ik} \tag{3.4}$$

where b_j is the *bias*, which is the weight between the bias node and output node j, and w_{kj} is the weight between input node k and output node j. All weights and biases are real numbers and can assume negative values. Thus, d_{ij} also is a real number and can assume negative values.

Next, d_{ij} is passed through a sigmoid activation function yielding a quantity s_{ij} defined by

$$s_{ij} = \frac{1}{1+\exp(-d_{ij})} = \frac{1}{1+\exp(-b_j - \sum_{k=1}^{N_F} w_{kj} F_{ik})} \tag{3.5}$$

Note that $\forall_{i,j} : s_{ij} \in \langle 0,1 \rangle$, and $\lim_{d_{ij}\to-\infty} s_{ij}(d_{ij}) = 0$ and $\lim_{d_{ij}\to\infty} s_{ij}(d_{ij}) = 1$. Throughout the chapter, s_{ij} will be interpreted as the *similarity* of stimulus S_i to category C_j.

Finally, the outputs s_{ij} of the output nodes are normalized yielding the quantity p_{ij}:

$$\begin{aligned} p_{ij} &= \frac{s_{ij}}{\sum_{l=1}^{N_r} s_{il}} = \frac{(1+\exp(-d_{ij}))^{-1}}{\sum_{l=1}^{N_r}(1+\exp(-d_{il}))^{-1}} \\ &= \frac{(1+\exp(-b_j - \sum_{k=1}^{N_F} w_{kj} F_{ik}))^{-1}}{\sum_{l=1}^{N_r}(1+\exp(-b_l - \sum_{k=1}^{N_F} w_{kl} F_{ik}))^{-1}} \end{aligned} \tag{3.6}$$

This normalization step, which is not part of the actual SLP, is similar to the normalization in the (unbiased) SCM (Luce, 1963). The p_{ij} can now be interpreted as probabilities because $\forall_{i,j} : p_{ij} \in \langle 0,1 \rangle$ and $\forall_i : \sum_{j=1}^{N_r} p_{ij} = 1$. The probability p_{ij} is the *a posteriori* probability of class C_j, that is, the probability that the model responds with class C_j when it is presented with stimulus S_i.

Generally, before being clamped to the input node, the set of values for each feature is normalized over all stimuli using

$$\tilde{F}_{ik} = \frac{F_{ik} - \mu_k}{\sigma_k} \tag{3.7}$$

where F_{ik} and $\tilde{F}_{ik}$ are the original and normalized value of feature k of stimulus S_i, and $\mu_k = \frac{1}{N_s}\sum_{i=1}^{N_s} F_{ik}$ and $\sigma_k = \sqrt{\frac{1}{N_s-1}\sum_{i=1}^{N_s}(\mu_k - F_{ik})^2}$.

An important concept in categorization models is the location of decision boundaries. The *equal-probability boundary* B_{mn} between classes C_m and C_n is defined as the subspace of $\mathcal{F}$ where the ratio L of the probability $p_m(\mathbf{F})$ of responding class C_m and the probability $p_n(\mathbf{F})$ of responding class C_n is one:

$$B_{mn} = \{\mathbf{F} \in \mathcal{F} | L(\mathbf{F}) = \frac{p_m(\mathbf{F})}{p_n(\mathbf{F})} = 1\} \tag{3.8}$$

For the SLP the boundary B_{mn} is defined by

$$B_{mn} = \{\mathbf{F} \in \mathcal{F} \left| \frac{p_m(\mathbf{F})}{p_n(\mathbf{F})} = \frac{1 + \exp\left(-b_n - \sum_{k=1}^{N_F} w_{kn} F_k\right)}{1 + \exp\left(-b_m - \sum_{k=1}^{N_F} w_{km} F_k\right)} = 1\right\}$$

which reduces to

$$B_{mn} = \{\mathbf{F} \in \mathcal{F} | (b_n - b_m) + \sum_{k=1}^{N_F} F_k(w_{kn} - w_{km}) = 0\} \tag{3.9}$$

Equation (3.9) states that the equal-probability boundary between any two classes is linear in the SLP model (see also Haykin, 1994; Lippman, 1987).

3.4 The similarity-choice model

In order to make a detailed comparison between the SLP and the SCM possible in the next section, we will briefly recapitulate the well-known SCM in this section.

Like the SLP, the SCM actually only models the retrieval stage in the categorization process. For the purpose of the comparison we will use a popular instance of the class of distance-based SCMs, that is, the weighted SCM with Euclidean distance and exponential decay function. This model will be briefly recapitulated below.

Each response class C_j has one prototype $\mathbf{P_j}$ which is a vector containing N_F components P_{jk}. The weighted Euclidean distance d_{ij} of a stimulus S_i to prototype $\mathbf{P_j}$ is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{N_F} w_k (F_{ik} - P_{jk})^2} \tag{3.10}$$

where w_k is a non-negative parameter representing the *attention* allocated to feature dimension k. For the sake of generality we will not impose the restriction $\sum_{k=1}^{N_F} w_k = 1$.

It is assumed that the similarity s_{ij} of stimulus S_i to category C_j is related to the psychological distance d_{ij} of stimulus S_i to prototype $\mathbf{P_j}$ via the exponential decay function (e.g. Shepard, 1958):

$$s_{ij} = \exp(-d_{ij}) = \exp\left(-\sqrt{\sum_{k=1}^{N_F} w_k (F_{ik} - P_{jk})^2}\right) \tag{3.11}$$

Note that s_{ij} lies within the range $\langle 0, 1]$, with $s_{ij}(0) = 1$ ("self-similarity"), and $\lim_{d_{ij} \to \infty} s_{ij}(d_{ij}) = 0$.

Finally, the probability p_{ij} of responding class C_j, given stimulus S_i is defined as (Luce, 1963):

$$\begin{aligned} p_{ij} &= \frac{b_j s_{ij}}{\sum_{l=1}^{N_r} b_l s_{il}} = \frac{b_j \exp(-d_{ij})}{\sum_{l=1}^{N_r} b_l \exp(-d_{il})} \\ &= \frac{b_j \exp(-\sqrt{\sum_{k=1}^{N_F} w_k (F_{ik} - P_{jk})^2})}{\sum_{l=1}^{N_r} b_l \exp(-\sqrt{\sum_{k=1}^{N_F} w_k (F_{ik} - P_{lk})^2})} \end{aligned} \tag{3.12}$$

where $b_j \in \mathbb{R}^+$ is the *response bias* for category C_j. Note that this response bias is different from the MLP-bias.

The above defined SCM contains $N_F N_r$ prototype parameters P_{jk}, N_F attention weights w_k, and $N_r - 1$ response biases[4] b_j. Thus, the model has $(N_F+1)(N_r+1)-1$ free parameters in total.

The equal-probability boundary B_{mn} between classes C_m and C_n in the SCM is defined by

$$B_{mn} = \{\mathbf{F} \in \mathcal{F} \left| \frac{p_m(\mathbf{F})}{p_n(\mathbf{F})} = \frac{b_m \exp\left(-\sqrt{\sum_{k=1}^{N_F} w_k (F_k - P_{mk})^2}\right)}{b_n \exp\left(-\sqrt{\sum_{k=1}^{N_F} w_k (F_k - P_{nk})^2}\right)} = 1\right.\}$$

which leads to

$$B_{mn} = \{\mathbf{F} \in \mathcal{F} \left| \sqrt{\sum_{k=1}^{N_F} w_k (F_k - P_{nk})^2} - \sqrt{\sum_{k=1}^{N_F} w_k (F_k - P_{mk})^2} + \ln \frac{b_n}{b_m} = 0\right.\} \quad (3.13)$$

If $b_m = b_n$ this reduces to

$$B_{mn} = \{\mathbf{F} \in \mathcal{F} \left| \sum_{k=1}^{N_F} w_k ((F_k - P_{nk})^2 - (F_k - P_{mk})^2) = 0\right.\} \quad (3.14)$$

which is a hyperplane passing through the midpoint between prototypes P_m and P_n (the quadratic term in F_k vanishes). If all w_ks are equal, Eq. (3.14) defines the mid-sagittal plane between P_m and P_n.

3.5 Comparison of SLP and SCM

The purpose of this section is to gain insight in the modeling capabilities and implicit assumptions of the SLP through a comparison of the SLP with the SCM.

In both the SLP and the SCM we can distinguish 3 processing steps, that is, the transformations of $\mathbf{F}$ to $\mathbf{d}$, of $\mathbf{d}$ to $\mathbf{s}$, and of $\mathbf{s}$ to $\mathbf{p}$:

$$\mathbf{F} \longrightarrow \mathbf{d} \longrightarrow \mathbf{s} \longrightarrow \mathbf{p} \quad (3.15)$$

N.B., these steps take place *within* the retrieval stage, and should not be confused with the 3 stages described in section 1. The transformations are summarized in Table 3.I.

In order to get a notion of the differences between the processing steps of the two models and the influence of the various parameters, we will discuss some examples. In section 3.5.1, a number of one-dimensional examples are presented which illustrate the basic situations. The observations will be generalized in section 3.5.2 using the analytic expressions. In section 3.5.3, the general model properties derived from the analytic forms are further illustrated with a number of two-dimensional examples.

[4]Without loss of generality we may impose the restriction $\sum_{j=1}^{N_r} b_j = 1$, which brings the number of free bias parameters down from N_r to $N_r - 1$.

3.5.1 Examples: 1 feature, 2 classes

First we will consider the simplest case of interest, namely that of one stimulus feature and two response classes. Figure 3.3 shows an artificial example (not fitted on real data) for the SCM. $\mathbf{d_j}$ (symbols $\times$ and $+$), $\mathbf{s_j}$ (symbols — and $- \cdot -$), and $\mathbf{p_j}$ (symbols $--$ and $\cdots$) are displayed as a function of the stimulus feature F for each of the two categories.

In all three subfigures the prototypes are located at $F_1 = -1$ and $F_2 = 2$. In Figure 3.3a the biases for both response classes, as well as the attention weight for the single stimulus feature are 1. Due to the Luce-normalization, in the regions to the left as well as to the right of both prototypes the probabilities of responding either class become constant, although both similarities quickly drop to zero. This is related to the subject's having to make a forced choice. Although a particular stimulus may have very low similarity to both prototypes the subject has to choose one of the two. In Figure 3.3b all parameters are the same as in Figure 3.3a, only the attention weight is changed to 0.1. The transitions in $\mathbf{p}$ become shallower and different "saturation probabilities" in the regions to the left and right of both prototypes are reached. In Figure 3.3c all parameters are the same as in Figure 3.3a, except the bias for the right category, which is increased to 8. Note that the equal-probability class boundary has shifted to the left. Furthermore, the left and right saturation probabilities are now different.

In Figure 3.4, a number of basic (artificial) situations are shown for the SLP, again with one stimulus feature and two response classes. $\mathbf{d_j}$, $\mathbf{s_j}$, and $\mathbf{p_j}$ are indicated by the same symbols as in Figure 3.3.

In Figure 3.4a both biases are 0 and the weights connecting the input node to output nodes 1 and 2 are $w_1 = -1$ and $w_2 = 0.5$, respectively. A first clear difference with the SCM is that the function $\mathbf{d}$ in the SLP cannot be interpreted as a distance because it can become negative. Secondly, if we define a prototype as a point in the feature space where the similarity to the associated category reaches its maximum value 1, the SLP's prototypes are located at infinity. Stated differently,

Table 3.I: The three levels of processing in the SLP and SCM. For further explanation see text.

model	SLP	SCM
"distance"	$d_{ij} = b_j + \sum_{k=1}^{N_F} w_{kj} F_{ik}$	$d_{ij} = \sqrt{\sum_{k=1}^{N_F} w_k (F_{ik} - P_{jk})^2}$
similarity	$s_{ij} = \frac{1}{1+\exp(-d_{ij})}$	$s_{ij} = \exp(-d_{ij})$
probability	$p_{ij} = \frac{s_{ij}}{\sum_{l=1}^{N_r} s_{il}}$	$p_{ij} = \frac{b_j s_{ij}}{\sum_{l=1}^{N_r} b_l s_{il}}$

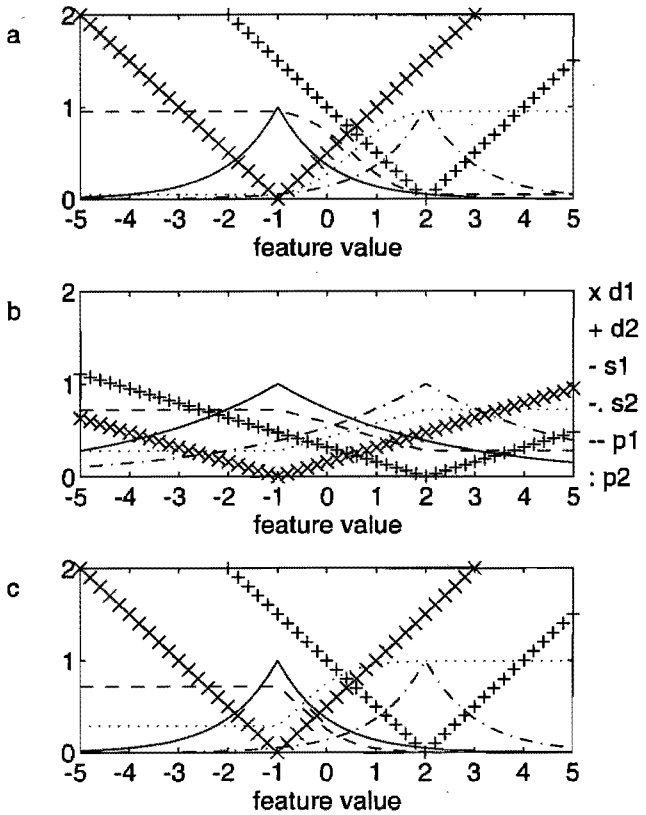

Figure 3.3: The functions d_1, d_2, s_1, s_2, p_1 and p_2 as function of the stimulus feature for three examples of the SCM. The parameter values are: Figure 3.3a: $P_1 = -1, P_2 = 2, b_1 = b_2 = w = 1$. Figure 3.3b: $P_1 = -1, P_2 = 2, b_1 = b_2 = w = 0.1$. Figure 3.3c: $P_1 = -1, P_2 = 2, b_1 = 8, b_2 = w = 1$.

the SLP model essentially supports prototypical *directions*, rather than prototypes. This issue will be extensively discussed in the next section.

Interestingly, although the shapes of the similarity curves are very different for the SCM and the SLP, the shapes of the probabilities are very similar. The clearest difference lies in the asymptotic behavior. While the SLP-probabilities approach 0 or 1 far away from $F = 0$, the SCM-probabilities are constant outside both prototypes. Finally, we note that the equal-probability boundary in Figure 3.4a is located at $F = 0$, as can be easily verified from Eq. (3.9).

The effect of increasing the absolute value of the weights is illustrated in Figure 3.4b. Here, $w_1 = -4$ and $w_2 = 2$, while $b_1 = b_2 = 0$. The transitions between the two classes become steeper.

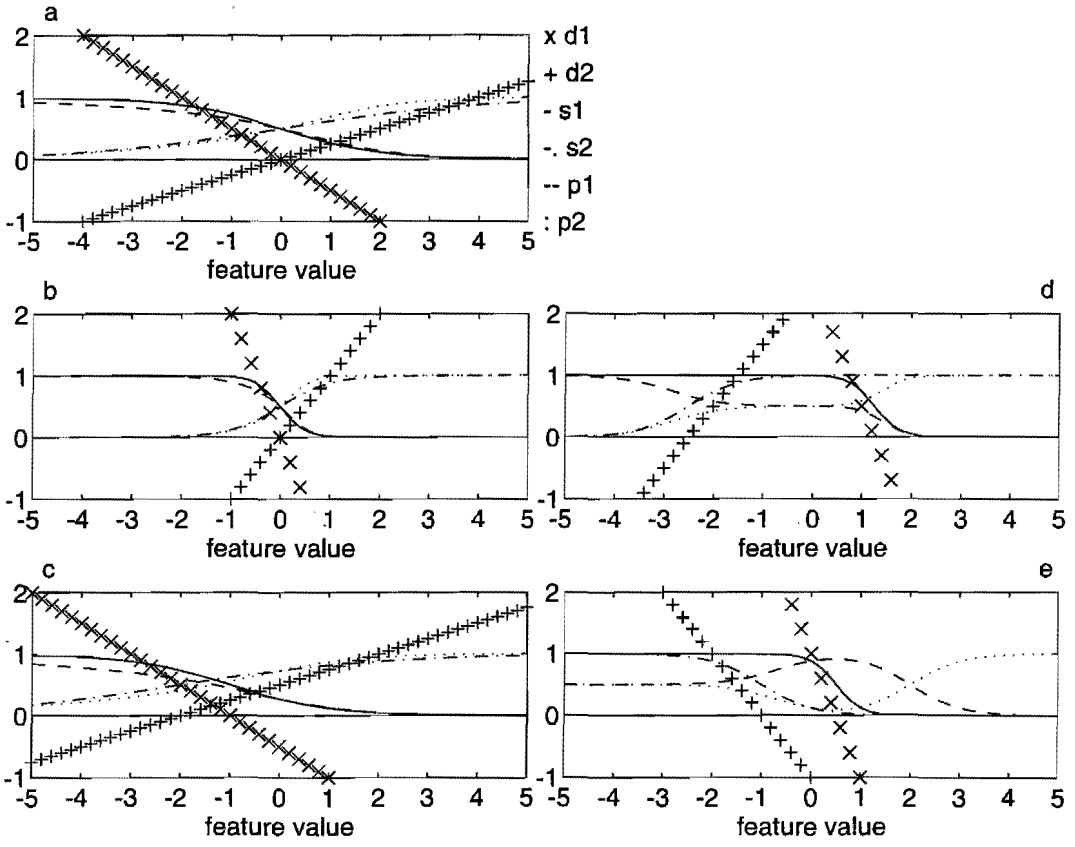

Figure 3.4: The functions d_1, d_2, s_1, s_2, p_1 and p_2 as function of the stimulus feature for five examples of the SLP. The parameter values are: Figure 3.4a: $w_1 = -1, w_2 = 0.5, b_1 = b_2 = 0$. Figure 3.4b: $w_1 = -4, w_2 = 2, b_1 = b_2 = 0$. Figure 3.4c: $w_1 = -1, w_2 = 0.5, b_1 = -1, b_2 = 1$. Figure 3.4d: $w_1 = -4, w_2 = 2, b_1 = b_2 = 5$. Figure 3.4e: $w_1 = -4, w_2 = -2, b_1 = 2, b_2 = -2$.

Figure 3.4c illustrates the effect of having different biases for different classes. Here, $w_1 = -1$ and $w_2 = 0.5$, like in Figure 3.4a, while now $b_1 = -1$ and $b_2 = 1$. Clearly, the equal-probability shifts toward class C_1, as can be verified from Eq. (3.9).

In Figures 3.4d and e, two situations are presented which have no equivalent in the SCM. In both cases there is a substantial area in the feature space where both classes have high similarities, which results in an "ambiguous area" where both class probabilities are close to 0.5. In Figure 3.4d, $w_1 = -4$ and $w_2 = 2$, and $b_1 = b_2 = 5$. Like in Figure 3.4b, the probability of responding class C_1 is high for low feature values and the probability of responding class C_2 is high for high feature values. In the middle region, however, both similarities are high due to the high biases,

resulting in ambiguousness.

The parameter values for Figure 3.4e are $w_1 = -4, w_2 = -2, b_1 = 2, b_2 = -2$. Both weights have the same sign, causing the similarities to the two categories to be high in the same general area, in this case for negative feature values. After normalization we find an ambiguous area for p at low feature values, high p_2 for high feature values, and high p_1 for feature values close to 0.

In summary, the 1-dimensional examples have shown that, although the similarity functions of the SLP and SCM are very different, the probability curves may be very similar. The most striking difference between the probability functions of the SLP and the SCM is that for the SLP the probability functions may show extensive ambiguous regions where both classes have equal probability. In this sense, the SLP is more general than the SCM.

3.5.2 Generalization of the examples

In order to generalize these observations we go back to the general analytic forms. In general the probability ratio $L(\mathbf{F})$ of responding class C_m and class C_n in the SLP equals

$$L(\mathbf{F}) = \frac{1 + \exp\left(-b_n - \sum_{k=1}^{N_F} w_{kn} F_k\right)}{1 + \exp\left(-b_m - \sum_{k=1}^{N_F} w_{km} F_k\right)} \tag{3.16}$$

Let us now decompose the weights w_{km} and w_{kn} and the biases b_m and b_n into an average component $\overline{w}_k$ and $\overline{b}$, and a variable component Δw_k and Δb, respectively:

$$w_{km} = \overline{w}_k + \Delta w_k \tag{3.17}$$

$$w_{kn} = \overline{w}_k - \Delta w_k \tag{3.18}$$

$$b_m = \overline{b} + \Delta b \tag{3.19}$$

$$b_n = \overline{b} - \Delta b \tag{3.20}$$

Note that $\overline{w}_k = \frac{1}{2}(w_{km} + w_{kn})$, $\Delta w_k = \frac{1}{2}(w_{km} - w_{kn})$, and $\overline{b} = \frac{1}{2}(b_m + b_n)$, $\Delta b = \frac{1}{2}(b_m - b_n)$.
Now Eq. (3.16) can be rewritten as

$$L(\mathbf{F}) = \frac{1 + \exp\left(-\overline{b}\right) \exp\left(-\sum_{k=1}^{N_F} \overline{w}_k F_k\right) \exp\left(\Delta b + \sum_{k=1}^{N_F} \Delta w_k F_k\right)}{1 + \exp\left(-\overline{b}\right) \exp\left(-\sum_{k=1}^{N_F} \overline{w}_k F_k\right) \exp\left(-\Delta b - \sum_{k=1}^{N_F} \Delta w_k F_k\right)} \tag{3.21}$$

Each of the 3 exponential factors in the numerator and denominator has a distinct interpretation.

The factor $\exp\left(-\overline{b}\right)$ can be interpreted as a general "attenuator" of the influence of the stimulus features on the probability ratio. If $\exp\left(-\overline{b}\right)$ is small compared to the other exponential factors (that is, when $\overline{b}$ is large), we can write

$$L(\mathbf{F}) = \frac{1 + \epsilon_1(\mathbf{F})}{1 + \epsilon_2(\mathbf{F})} \approx 1 + \epsilon_1(\mathbf{F}) - \epsilon_2(\mathbf{F}) \approx 1 \tag{3.22}$$

with $0 < \epsilon_1(\mathbf{F}) \ll 1$ and $0 < \epsilon_2(\mathbf{F}) \ll 1$.
This means that both probabilities hardly depend on $\mathbf{F}$, and they are about equal. This is the case in Figure 3.4d for feature values close to 0.
If, on the other hand, $\exp(-\bar{b})$ is large compared to the other exponential factors ($\bar{b}$ is very negative), we can omit the term 1 and obtain

$$\begin{aligned} L(\mathbf{F}) &\approx \frac{\exp(-\bar{b})\exp(-\sum_{k=1}^{N_F}\overline{w}_k F_k)\exp(\Delta b + \sum_{k=1}^{N_F}\Delta w_k F_k)}{\exp(-\bar{b})\exp(-\sum_{k=1}^{N_F}\overline{w}_k F_k)\exp(-\Delta b - \sum_{k=1}^{N_F}\Delta w_k F_k)} \\ &= \exp(2\Delta b + 2\sum_{k=1}^{N_F}\Delta w_k F_k) \end{aligned} \tag{3.23}$$

In this case the SLP coincides with a special case of the SCM, as will be shown in the next section.

The second exponential function in Eq. (3.21), that is, $\exp(-\sum_{k=1}^{N_F}\overline{w}_k F_k)$, can be interpreted as an attenuator which, in contrast with $\exp(-\bar{b})$, depends on F_k. In regions of $\mathcal{F}$ where $\exp(-\sum_{k=1}^{N_F}\overline{w}_k F_k)$ is large compared to the other factors, L approximates 1. This is the case for very negative values of F in Figure 3.4e, where $\overline{w}_k = -3$ and $\Delta w_k = -1$.

The third exponential function in Eq. (3.21), that is, $\exp(\Delta b + \sum_{k=1}^{N_F}\Delta w_k F_k)$, exclusively determines the location of the equal-probability boundary B_{mn}. Recall from Eq. (3.9) that B_{mn} is defined by

$$(b_n - b_m) + \sum_{k=1}^{N_F} F_k(w_{kn} - w_{km}) = 0$$

which is equivalent to

$$\Delta b + \sum_{k=1}^{N_F}\Delta w_k F_k = 0 \tag{3.24}$$

Thus, we see that, when a constant is added to all outgoing weights of an input node, or when a constant is added to all biases, the shape of the probability "landscape" will change, but the classification boundaries will be unaffected.

3.5.3 Examples: 2 features, 3 classes

Categorization experiments where only one stimulus feature is varied by the experimenter, cq. used by the subject, are rare. Often two or more features have to be taken into account in the model. We will briefly study two examples with $N_F = 2$ and $N_r = 3$.

Figure 3.5 represents a basic example of the similarity functions $\mathbf{s_j}$ and probability functions $\mathbf{p_j}$ of the SCM and the SLP. Figure 3.5a shows the SCM's $\mathbf{s_1}, \mathbf{s_2}$ and $\mathbf{s_3}$ simultaneously, and Figure 3.5b shows the SCM's $\mathbf{p_1}, \mathbf{p_2}$ and $\mathbf{p_3}$. Figures 3.5c and d show the SLP's $\mathbf{s_1}, \mathbf{s_2}, \mathbf{s_3}$ and $\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}$, respectively. The two feature dimensions are indicated by x and y. For the SCM, the 3 prototypes are located at (-2,0), (1,1), and (1,-1). The 3 biases and 2 attention weights are set to 1. The SLP-weights[5]

[5]Recall that w_{kj} indicates the weight between input node k and output node j in the SLP.

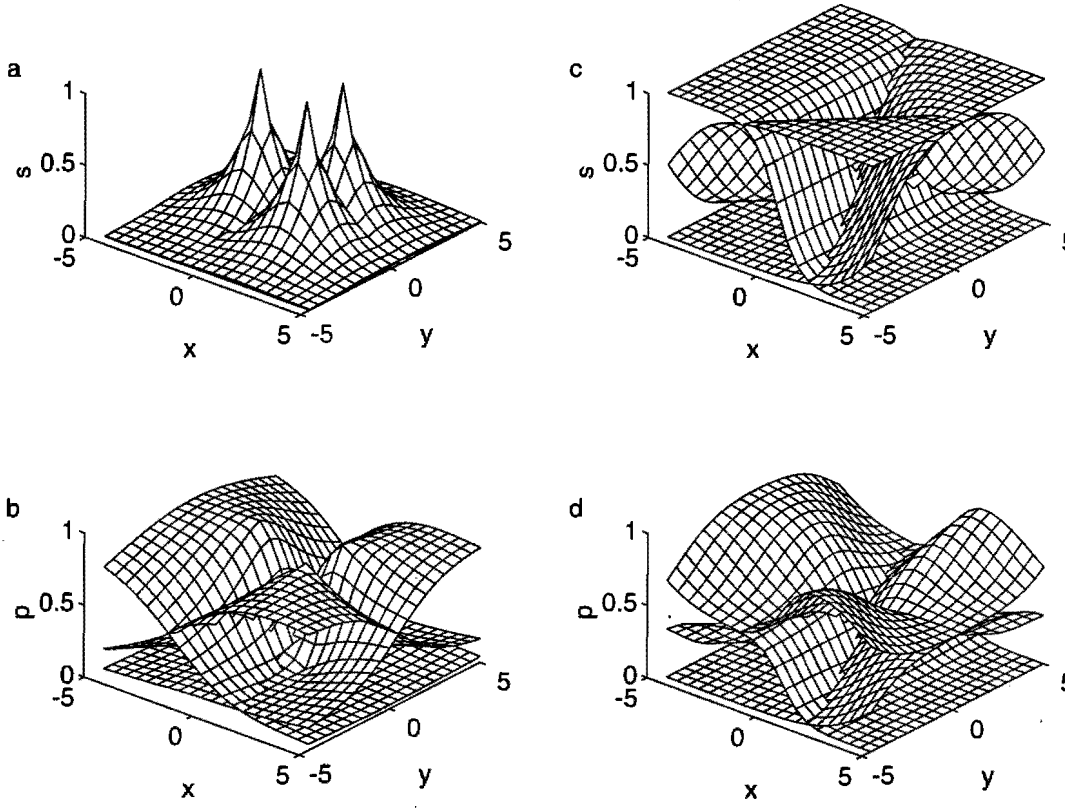

Figure 3.5: Similarities s_1, s_2, s_3 (Figures 3.5a and c), and class probabilities p_1, p_2, p_3 (Figures 3.5b and d), for the SCM (Figures 3.5a and b), and the SLP (Figures 3.5c and d). The parameter values are: Figure 3.5a and b: $\mathbf{P_1} = (-2,0), \mathbf{P_2} = (1,1), \mathbf{P_3} = (1,-1), b_1 = b_2 = b_3 = w_1 = w_2 = 1$. Figure c and d: $w_{11} = -2, w_{21} = 0, w_{12} = 1, w_{22} = 1, w_{13} = 1, w_{23} = -1, b_1 = b_2 = b_3 = 0$. x and y are the stimulus features.

are set to $w_{11} = -2$, $w_{21} = 0$, $w_{12} = 1$, $w_{22} = 1$, $w_{13} = 1$ and $w_{23} = -1$, and all biases are set to 0.

Like in the one-dimensional case, we find that the shapes of the similarity functions of the SCM and the SLP are very different. In the SCM, the similarities to each of the three prototypes strongly peak at the prototype locations, while the SLP-similarities reflect the sigmoid shape. The class-probability functions of the two models are, however, rather similar. Note that, in contrast with the SCM, the SLP probabilities flatten off in non-prototypical directions, e.g. at feature values (5,0). As shown earlier with Eq. (3.21), the amount of flattening-off between two categories is controlled by the average bias of the two categories.

Figure 3.6 shows two basic situations for the SLP which, like Figures 3.4d and

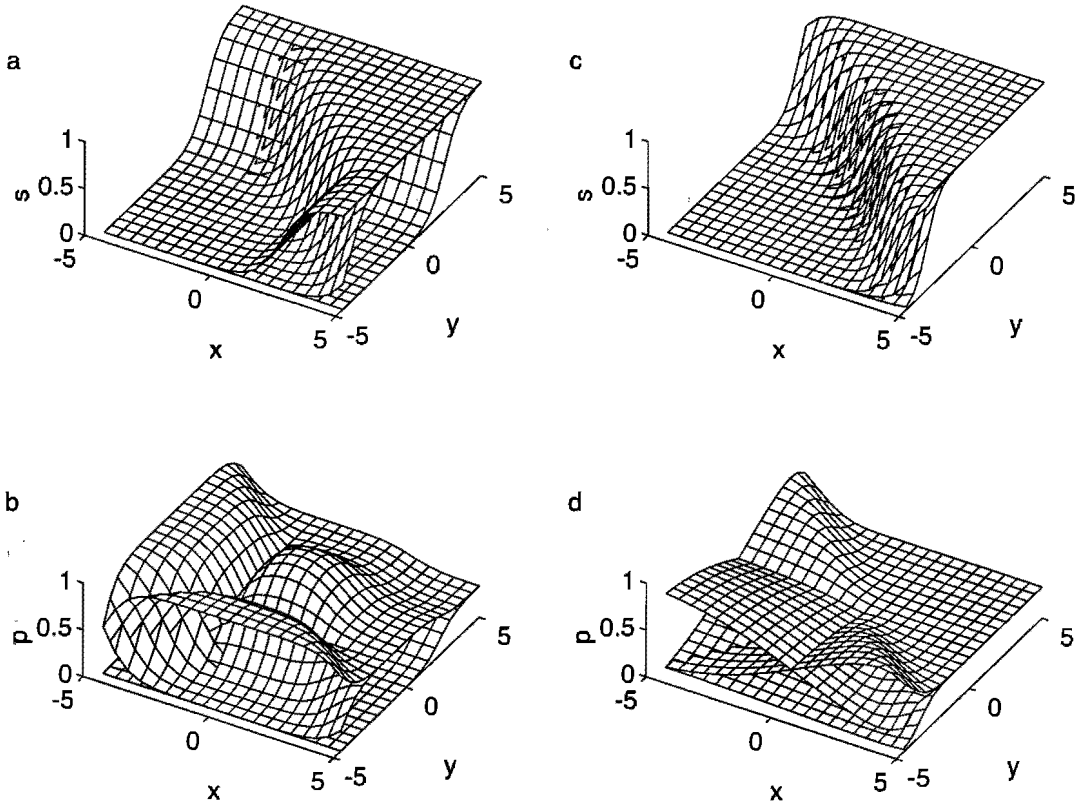

Figure 3.6: Similarities s_1, s_2, s_3 (figures 3.6a and c), and class probabilities p_1, p_2, p_3 (figures 3.6b and d), for two examples of the SLP. The parameter values are: Figure 3.6a and b: $w_{11} = 0, w_{21} = 2, w_{12} = 2, w_{22} = 2, w_{13} = 2, w_{23} = 0, b_1 = -4, b_2 = 0, b_3 = -4$. Figure c and d: $w_{11} = 2, w_{21} = 2.5, w_{12} = 2, w_{22} = 2, w_{13} = 2.5, w_{23} = 2, b_1 = b_2 = b_3 = 0$. x and y are the stimulus features.

e, have no equivalent in the SCM. The Figures on the left side show the 3 similarity functions (Figure 3.6a) and probability functions (Figure 3.6b) for parameter values $b_1 = -4$, $b_2 = 0$, $b_3 = -4$ and $w_{11} = 0$, $w_{21} = 2$, $w_{12} = 2$, $w_{22} = 2$, $w_{13} = 2$ and $w_{23} = 0$. Figures 3.6c and d show similarities and probabilities for parameter values $b_1 = b_2 = b_3 = 0$ and $w_{11} = 2$, $w_{21} = 2.5$, $w_{12} = w_{22} = 2$, $w_{13} = 2.5$ and $w_{23} = 2$. Note that for both cases the $\overline{w}$s are relatively large.

In Figure 3.6b, we find that categories 1 and 3 show asymptotic behavior, and category C_2 peaks for feature values close to the origin. We will calculate the class boundaries for this example by substituting the parameter values in Eq. (3.9), which leads to

$$
\begin{aligned}
B_{12} &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | x + 2 = 0\} \\
B_{13} &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | x - y = 0\} \\
B_{23} &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | y + 2 = 0\}
\end{aligned}
$$

In accordance with Ashby and Perrin (1988), we will use the term *response region* to indicate the subspace $\mathcal{F}_j$ of $\mathcal{F}$ where class C_j is the most probable response class:

$$\mathcal{F}_j = \{\mathbf{F} \in \mathcal{F} | \forall_{k \neq j} : p_k < p_j\} \tag{3.25}$$

We can establish the response region for each response class by taking the cross section of the subspaces defined by the 2 relevant boundaries. For the example of Figure 3.6b, this leads to

$$
\begin{aligned}
\mathcal{F}_1 &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | x + 2 < 0 \wedge x - y < 0\} \\
\mathcal{F}_2 &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | x + 2 > 0 \wedge y + 2 > 0\} \\
\mathcal{F}_3 &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | y + 2 < 0 \wedge x - y > 0\}
\end{aligned}
$$

In Figure 3.6d, we find a large ambiguous region for positive x and y, because the similarities to all 3 categories are close to unity. Note that the counter-intuitive situation occurs that p_3 increases with decreasing s_3. This means that, although the likeness to category C_3 decreases, the likeness to the other categories decreases even more, so the subject is more likely to respond with category C_3 in a forced choice task. A comparison of similarity (or "typicality") ratings to categorization data for the same stimuli may indicate whether or not this is a psychologically realistic effect. Prototype-based models can only display this behavior when class-dependent *rate decreasing parameters* are used, as proposed by Ashby and Maddox (1993).[6]

The classes in Figure 3.6d, as described by their boundaries, are

$$
\begin{aligned}
\mathcal{F}_1 &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | y > 0 \wedge x < y\} \\
\mathcal{F}_2 &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | x < 0 \wedge y < 0\} \\
\mathcal{F}_3 &= \{(x,y) \in \mathbb{R} \times \mathbb{R} | x > 0 \wedge x > y\}
\end{aligned}
$$

3.6 The prototype concept in relation to the natural range of feature values

The concept of prototype, as it is used in prototype models of categorization, is based on at least 3 assumptions.

1. The prototype of a category is the most typical member of the category.
2. The prototype contains the only information concerning the category which is memorized.
3. Prototypes are used as reference exemplars during categorization, that is, responses are based on a computation of similarity of the stimulus to all relevant prototypes.

[6] Distance measure E5, pages 382-383.

The general validity of these assumptions has been subject to criticism (e.g. Ashby, 1992). Rather than further criticizing these assumptions, we want to discuss an assumption which is often made more or less tacitly in prototype models. This assumption is that all prototypes must lie within the natural range of feature values, that is, stimuli exist - or can be synthesized - which coincide with the prototypes.

Let us look at the following example. Suppose that in an experiment subjects are asked to classify adult men into two categories, namely category C_1 of tall men and category C_2 of short men. Fundamentally, the prototype concept is not suited to model this problem. The taller the man, the more typical he is of category C_1, the shorter the man, the more typical of category C_2. However, in a categorization task, each stimulus feature has a certain limited natural range. In our example no stimuli will be presented with a height below 1m or above 2.5m. If we now allow the prototypes to lie outside this natural range, e.g. we allow the prototypes of the short and long category to be men of 0.1m and 10m, respectively, the observed categorization data may be well modeled by a prototype model.

More generally, we want to state that

1. In some situations it is fundamentally more appropriate to use a concept of *prototypical direction* than of prototype.
2. These situations may still be accurately modeled with prototype models when the prototypes are allowed to lie outside the natural feature range.

As the SCM is a prototype-based model, these statements should apply. As stated in an earlier section, the SLP is a conceptually different model which is based on the notion of prototypical direction. In support of the second statement, however, it is shown in Appendix 1 that the SCM with prototypes at infinite distance from the origin[7] coincides with the SLP with biases at minus infinity. Using symbolic notation, we may write

$$\lim_{b_j \to -\infty} \text{SLP} = \lim_{\|P_j\| \to \infty} \text{SCM} \tag{3.26}$$

Furthermore, in this situation we can express the SLP-parameters in terms of the SCM-parameters as is listed in Table 3.II.

Figure 3.7 illustrates the close correspondence of the SLP to the SCM in this limit case. Figures 3.7a and b represent the similarities and class probabilities for the SCM with $b_1 = b_2 = b_3 = w_1 = w_2 = 1$ and the prototypes far away from the origin: $\mathbf{P_1} = (-20, 0), \mathbf{P_2} = (18, 10), \mathbf{P_3} = (18, -10)$. Figures 3.7c and d represent the similarities and class probabilities for the SLP with the parameter values calculated by substituting the SCM-parameters in the equations listed in Table 3.II, yielding $b_1 = -20, b2 = b3 = -20.6, w_{11} = -1, w_{21} = 0, w_{12} = 0.874, w_{22} = 0.486, w_{13} = 0.874$ and $w_{23} = -0.486$.

Clearly, in line with Eq. (3.26), the similarities as well as the probabilities of the two models are almost identical.

[7]Recall that usually the stimulus features are normalized to Z-scores using Eq. (3.7), so the feature vectors for all stimuli are grouped around the origin. Here the origin has no perceptual interpretation.

Table 3.II: The correspondence between SLP-parameters and SCM-parameters in the limit case $\lim_{b_j \to -\infty} \text{SLP} = \lim_{\|P_j\| \to \infty} \text{SCM}$.

SLP-parameter	Corresponding function in terms of SCM-parameters
b_j	$\ln(b_j) - \sqrt{\sum_{k=1}^{N_F} w_k P_{kj}^2}$
w_{kj}	$\frac{w_k P_{kj}}{\sqrt{\sum_{l=1}^{N_F} w_l P_{lj}^2}}$

We conclude this section with a remark on the GRT. Ashby and Perrin (1988) define the unbiased probability p_{ij} of responding with category C_j, when presented with stimulus S_i, as

$$p_{ij} = \int_{R_j} \Phi_i(\mathbf{F}) d\mathbf{F} \tag{3.27}$$

where R_j is the response region of category C_j, and $\Phi_i(\mathbf{F})$ is the multivariate normal probability density function (pdf) which is associated with the *perceptual effect* of stimulus S_i. Furthermore, the similarity of stimulus S_i to response class C_j has the same definition. When a GRT-model does not have enclosed (finite) response regions, which will often occur when N_r is not much larger than N_F, similarities as well as class probabilities are cumulative normal pdfs (Ashby and Perrin, 1988). Therefore, as for the SLP and the SCM with remote prototypes, the GRT-similarities as well as probabilities will reach their extreme values at infinity.

3.7 Extension to one hidden layer

In this section, the theory which was developed for the SLP is extended to the MLP. We will restrict ourselves to the case of the two-layer perceptron (TLP), that is, the MLP with one hidden layer.

3.7.1 Definitions

In the TLP, a layer of N_H hidden nodes plus one bias node is situated between the input layer and the output layer. All input nodes (including the bias node) are connected to each hidden node, and all hidden nodes (including the "hidden" bias node) are connected to each output node (e.g. Haykin, 1994; Lippman, 1987). In total the number of parameters N_w in the TLP equals

$$N_w = (N_F + 1)N_H + (N_H + 1)N_r \tag{3.28}$$

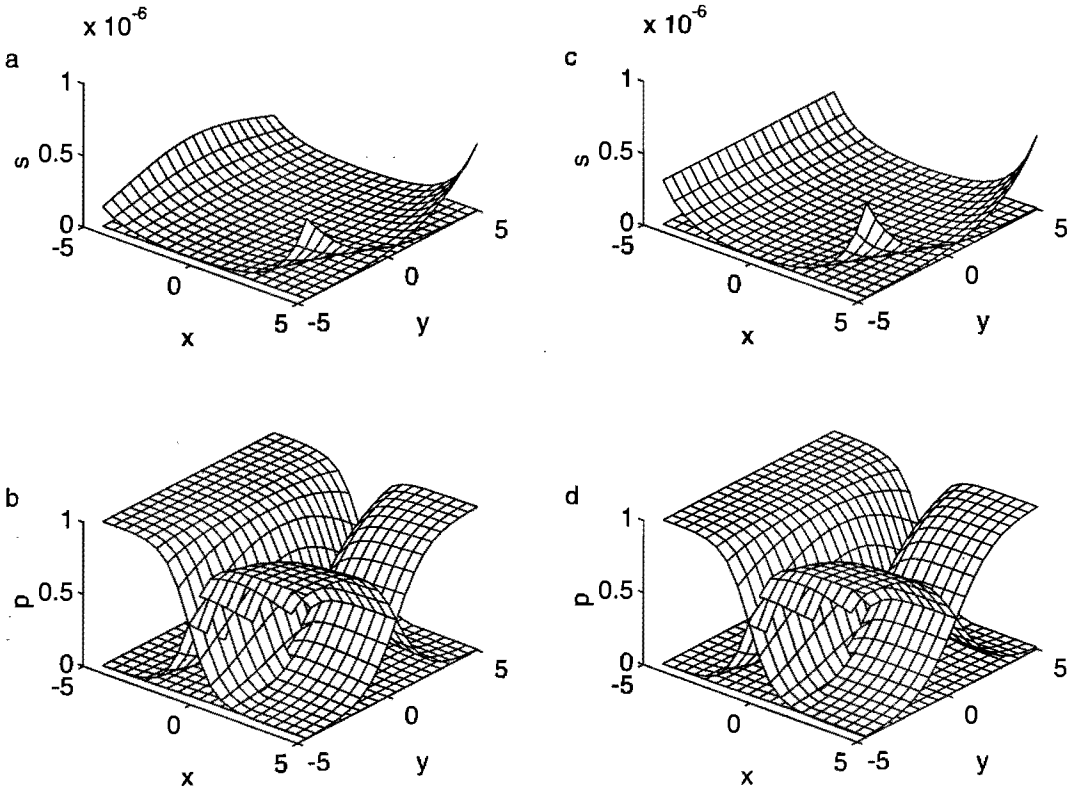

Figure 3.7: Similarities s_1, s_2, s_3 (Figures 3.7a and c), and class probabilities p_1, p_2, p_3 (Figures 3.7b and d), for the SCM (Figures 3.7a and b), and the SLP (Figures 3.7c and d). The parameter values are: Figure 3.7a and b: $\mathbf{P_1} = (-20, 0), \mathbf{P_2} = (18, 10), \mathbf{P_3} = (18, -10), b_1 = b_2 = b_3 = w_1 = w_2 = 1$. Figure c and d: $w_{11} = -1, w_{21} = 0, w_{12} = 0.874, w_{22} = 0.486, w_{13} = 0.874, w_{23} = -0.486, b_1 = -20, b_2 = b_3 = -20.6$. x and y are the stimulus features.

The hidden nodes perform the same processing - summation and sigmoid transformation - as the output nodes.

The notation is adjusted as follows. The weight connecting node k in layer l to node j in layer $l+1$ is indicated by w_{lkj}. The input, hidden and output layers have indices $l = 0$, 1 and 2, respectively. In hidden node j, a weighted sum of input feature values F_{ik} of stimulus S_i is made, yielding a quantity d_{0ij}:

$$d_{0ij} = b_{0j} + \sum_{k=1}^{N_F} w_{0kj} F_{ik} \tag{3.29}$$

Hidden node j will output a quantity s_{0ij} defined by

$$s_{0ij} = (1 + \exp(-d_{0ij}))^{-1} \tag{3.30}$$

In output node l, a weighted sum of values s_{0ij} of stimulus S_i is made, yielding a quantity d_{1il}:

$$d_{1il} = b_{1l} + \sum_{j=1}^{N_H} w_{1jl} s_{0ij} \tag{3.31}$$

The output s_{1il} of output node l is defined by

$$s_{1il} = (1 + \exp(-d_{1il}))^{-1} \tag{3.32}$$

Finally, the outputs of the output layer are again normalized yielding the probability p_{il} of assigning stimulus S_i to response class C_l:

$$p_{il} = \frac{s_{1il}}{\sum_{m=1}^{N_r} s_{1im}} \tag{3.33}$$

Depending on the number of hidden nodes, the modeling power of the TLP is much larger than that of the SLP, of course at the expense of a number of additional parameters. It has been established by several authors (e.g. Weenink, 1991) that the TLP is capable of modeling any nonlinear boundary between convex or non-convex subspaces. We will refrain from studying the general analytical form which expresses the relationship between input features and response probabilities, because it is far from translucent. We will confine ourselves to discussing one illustrative example, and we will introduce an approximation method which allows us to interpret the TLP-model in the same way as the SLP-model.

3.7.2 Example

It is the purpose of this example to illustrate the processing steps within the TLP and to give an impression of the extra modeling power provided by the hidden layer. Like in a number of previous examples, $N_F = 2$ and $N_r = 3$. The number of hidden nodes is set to $N_H = 3$. For this example, the weights between the input layer and the hidden layer are similar to those in the example of Figure 5c: $w_{011} = -4, w_{021} = 0, w_{012} = w_{022} = 2, w_{013} = 2, w_{023} = -2$. The biases between input and hidden layer are $b_{01} = b_{02} = b_{03} = -5$.

The outputs s_{0j} of the 3 hidden nodes are shown in Figure 3.8a as a function of the 2 input features x and y. Note that the shapes are similar to those in Figure 5c. The weights and biases between the hidden layer and the output layer are chosen such that the resulting classes are more or less concentric, that is, class C_2 surrounds class C_3, and class C_1 surrounds class C_2. The parameter values are $w_{111} = w_{121} = w_{131} = 4.5, w_{112} = w_{122} = w_{132} = 0, w_{113} = w_{123} = w_{133} = -4.5$, and $b_{11} = -7, b_{12} = -3, b_{13} = -2$. The outputs s_{1j} of the three output nodes as a function of the 2 input features x and y are shown simultaneously in Figure 3.8b, and are shown separately in Figures 3.8d (s_{11}), 3.8e (s_{12}) and 3.8f (s_{13}). Note that output node 1 is positively connected to all hidden nodes, output node 3 is negatively connected to all hidden nodes, and output node 2 is not connected (weights 0) to any hidden node. Thus the functions s_{11}, s_{12} and s_{13} are high in the outside region, constant, and high in the inside region, respectively. After

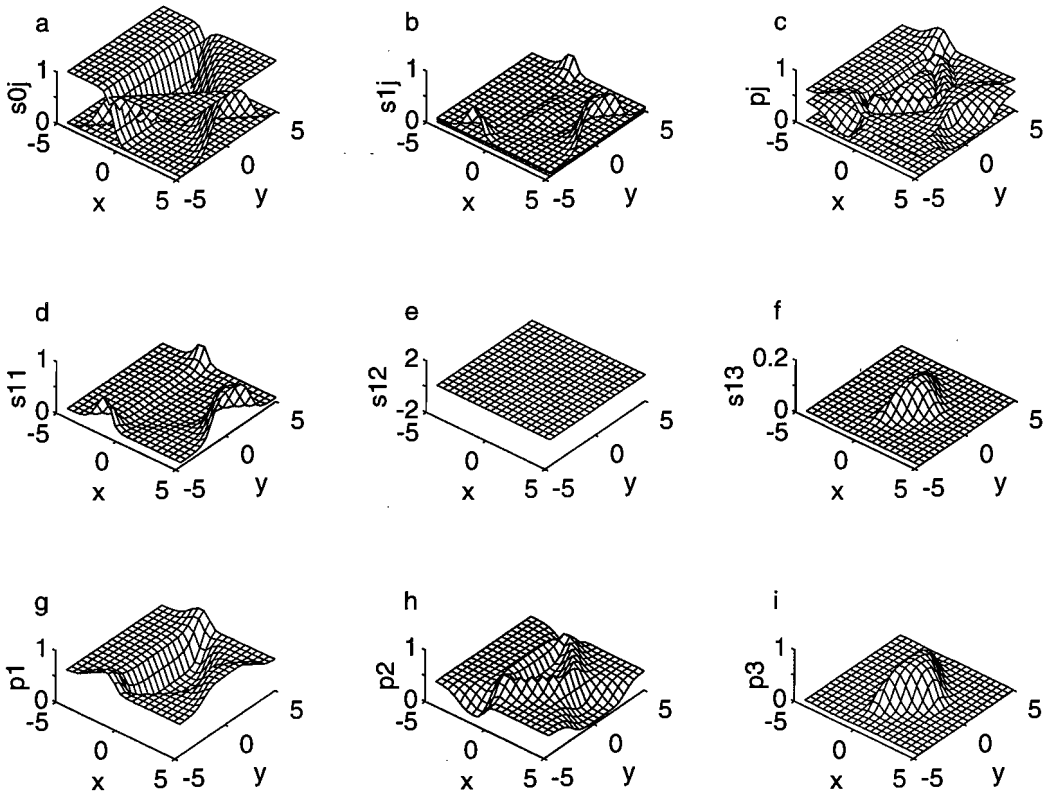

Figure 3.8: Various processing levels within the TLP. Parameter values are: $w_{011} = -4, w_{021} = 0, w_{012} = w_{022} = 2, w_{013} = 2, w_{023} = -2$, $b_{01} = b_{02} = b_{03} = -5$; $w_{111} = w_{121} = w_{131} = 4.5, w_{112} = w_{122} = w_{132} = 0, w_{113} = w_{123} = w_{133} = -4.5$, $b_{11} = -7, b_{12} = -3, b_{13} = -2$. Figures 3.8a shows the functions s_{01}, s_{02}, s_{03}, Figure 3.8b shows the functions s_{11}, s_{12}, s_{13}, Figure 3.8c shows the functions p_1, p_2, p_3. The functions s_{11}, s_{12} and s_{13} are shown separately in Figures 3.8d, e and f, respectively. The functions p_1, p_2 and p_3 are shown separately in Figures 3.8g, h and i, respectively. x and y are the stimulus features.

normalization this leads to the "concentric" response regions: p_1 is high in the outside region (Figure 3.8g), p_3 is high in the inside region (Figure 3.8i), and p_2 is high in a band in between (Figure 3.8h). p_1, p_2 and p_3 are displayed simultaneously in Figure 3.8c.

Figure 3.9 shows the equal-probability boundaries between class C_1 and class C_2 (outer curve) and between class C_2 and C_3 (inner curve) for the above example. The boundary between class C_1 and C_3 lies between the two other boundaries, but it is not displayed.

The example shows that the TLP is capable of modeling non-convex classes

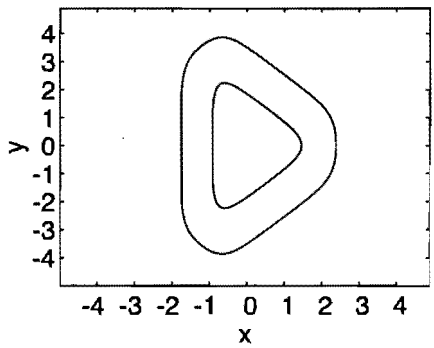

Figure 3.9: The equal-probability boundaries B_{12} (outer curve) and B_{23} (inner curve), for the example of Figure 3.8. x and y are the stimulus features.

with nonlinear boundaries. The hidden layer linearly separates a number of convex "auxiliary" classes, which are combined in the output layer to get the output classes.

3.7.3 Linearization of hidden nodes

In general, the interpretation of the categorization behavior of the TLP is much more difficult than that of the SLP. The equal-probability boundaries in the TLP are nonlinear, and their analytical expressions are - although easy to derive - hard to interpret. In this section we will propose a method for the interpretation of the TLP which is based on replacing the sigmoid function in the hidden nodes by a piecewise linear approximation. The simplest piecewise linear approximation to the sigmoid is the well-known *Heaviside* function or *hard-limiter* (e.g. Lippman, 1987; Minsky and Papert, 1969). The Heaviside function equals zero for negative input and equals one for positive output. We consider this approximation too crude for our purposes because the resulting similarity and probability functions are discontinuous and show no transitional regions. We choose the simplest-but-one piecewise linear approximation to the sigmoid, that is, a continuous function consisting of three line segments. We will use it as follows.

After a TLP has been trained on experimental data[8], the sigmoid $\sigma(x)$ in all hidden nodes is replaced by the piecewise linear "pseudo-sigmoid"[9] $\hat{\sigma}(x)$, defined by

$$\begin{aligned}
\hat{\sigma}(x) &= 1-b, & x &\in \langle \leftarrow, -a], \\
\hat{\sigma}(x) &= x(\tfrac{b-0.5}{a}) + 0.5, & x &\in \langle -a, a\rangle, \\
\hat{\sigma}(x) &= b, & x &\in [a, \rightarrow\rangle
\end{aligned} \tag{3.34}$$

[8]Methods for training and testing of the models are presented in the next chapter.

[9]The pseudo-sigmoid cannot be used during training because all its derivatives are zero everywhere apart from a finite point set.

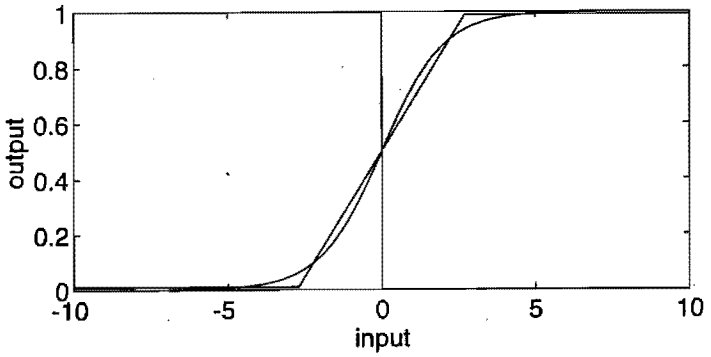

Figure 3.10: The sigmoid function $\sigma(x)$ and the pseudo-sigmoid function $\hat{\sigma}(x)$. Parameter values are $a = 2.71, b = 0.99$.

where a and b are parameters to be determined. Note that $\hat{\sigma}(x)$ is continuous and always passes through (0,0.5).

An example of the pseudo-sigmoid $\hat{\sigma}(x)$ is displayed in Figure 3.10, together with the regular sigmoid $\sigma(x)$. For this example, the parameter b was set to 0.99, and the value of a was calculated such that the L_∞-norm $|\hat{\sigma}(x) - \sigma(x)|$ is minimized. This leads to the value $a = 2.71$.

The pseudo-sigmoid in each hidden node j divides the range of incoming values d_{0j} into three subranges, namely $d_{0j} \in \langle \leftarrow, -a], d_{0j} \in \langle -a, a \rangle$, and $d_{0j} \in [a, \rightarrow \rangle$. In each of these subranges, a different linear relation exists between input and output of the pseudo-sigmoid. The combined effect of all N_H linearized hidden nodes is that the feature space $\mathcal{F}$ is divided into (at most) 3^{N_H} subspaces. In each of these subspaces a different linear relationship exists between the stimulus features $\mathbf{F}$ and the input of the output nodes $\mathbf{d_1}$. Thus, we can substitute the TLP by 3^{N_H} SLPs, each of which processes a different subset of stimuli. For each of the SLPs we can use the interpretation method described earlier, that is, define the appropriate linear equal-probability boundaries. A few remarks are in order, however.

First, let us define a hidden node to be *saturated* when its input x is in a region where $\frac{d\hat{\sigma}(x)}{dx} = 0$, that is, when $x \in \langle \leftarrow, -a] \cup [a, \rightarrow \rangle$. We define a hidden node to be *operational* when $\frac{d\hat{\sigma}(x)}{dx} \neq 0$, that is, when $x \in \langle -a, a \rangle$. In Appendix 2 it is shown that, when 3^{N_H} subspaces exist, the number of subspaces N_k where k hidden nodes are operational equals

$$N_k = \binom{N_H}{k} 2^{N_H - k} \tag{3.35}$$

Thus, we see that in 2^{N_H} subspaces, all hidden nodes are saturated ($k = 0$). This means that in all of these subspaces $\mathbf{s_0}, \mathbf{d_1}, \mathbf{s_1}$ and $\mathbf{p}$ are independent of the stimulus features. N.B. although $\mathbf{s_0}, \mathbf{d_1}, \mathbf{s_1}, \mathbf{p}$ are *constant* on each saturated subspace, they are generally *different* in different saturated subspaces. Class boundaries may only

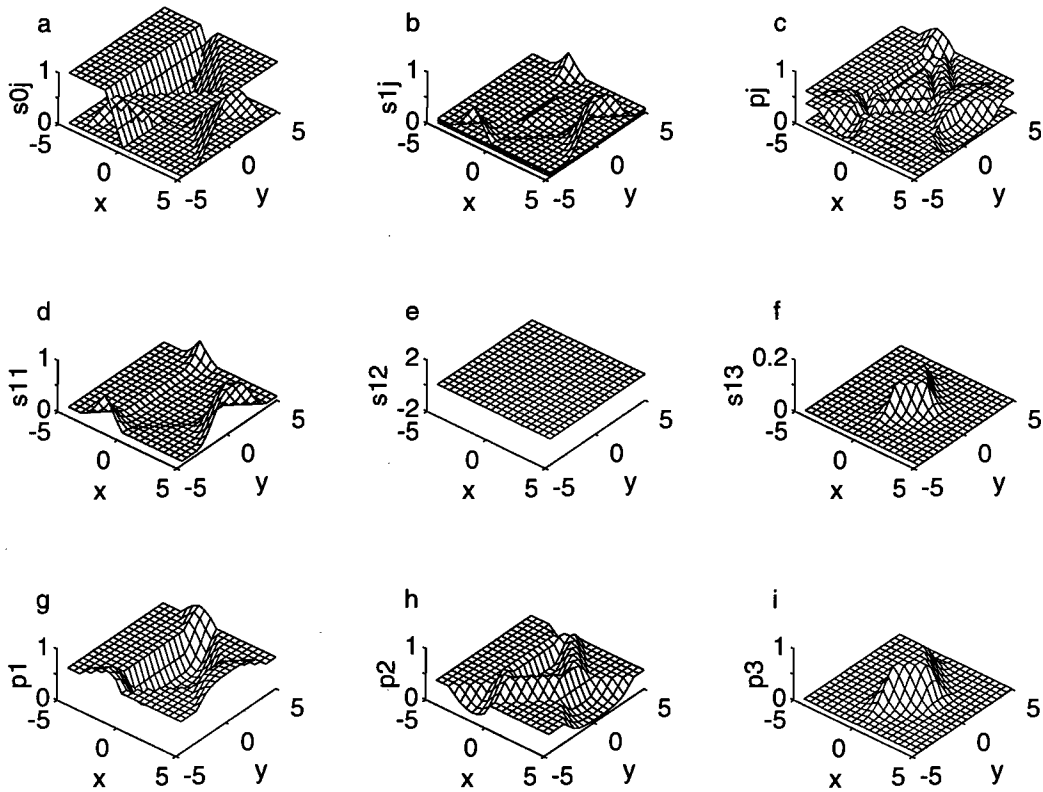

Figure 3.11: Various processing levels of the linearized TLP. Parameters and subfigures are the same as in Figure 8.

exist in the subspaces where the number of operational hidden nodes is larger than zero.

Let us go back to the example of section 6.2. We replace the sigmoids in the hidden nodes by the pseudo-sigmoid defined for Figure 3.10 ($a = 2.71$, $b = 0.99$). Figure 3.11 shows the various functions in the linearized TLP. All subfigures correspond to the subfigures of Figure 3.8.

Although all functions are clearly less smooth, the general patterns are very similar to those in Figure 3.8.

Figure 3.12 shows the equal-probability boundaries for the linearized TLP (thick lines), and the boundaries between the various subspaces (thin lines).

The piecewise linear boundaries are good approximators to the nonlinear boundaries displayed earlier in Figure 3.9. Note that, because the 3 functions d_{0j} are dependent, only 19 subspaces are present, which is less than the maximum of 27 for 3 hidden nodes. Some combinations are not possible here, like all 3 pseudo-sigmoids

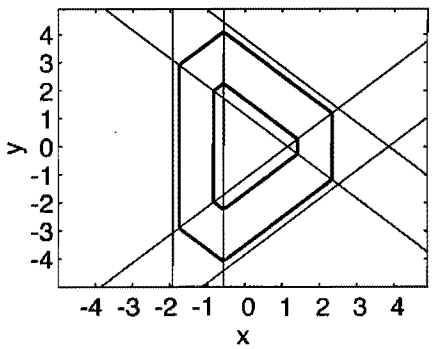

Figure 3.12: The equal-probability boundaries B_{12} (outer thick curve) and B_{23} (inner thick curve), for the example of Figure 11. The thin lines indicate the boundaries between the subspaces as defined by the linearized hidden nodes. x and y are the stimulus features.

being operational simultaneously, or all 3 being right-hand saturated (see Figure 3.11a).

The piecewise linear approximation to the nonlinear boundaries of the TLP can be arbitrarily precise, by choosing an appropriate number of linear segments in the pseudo-sigmoid. If we would have used the Heaviside function in our example, the boundaries of Figure 3.9 would have been approximated by triangles instead of hexagons (that is, Figure 3.12 without the "clipped tips"). The more line segments used in replacing the sigmoid, the closer the approximation to the nonlinear boundaries of Figure 3.9 are.

3.8 General discussion and summary

In the present study, a model for human categorization behavior has been developed, of which the MLP is the central part. The model has been embedded in the general 3-stage framework put forward by Ashby (1992). Similar to the SCM, the representation and retrieval stages are assumed to be deterministic, while the response selection stage is assumed to be probabilistic.

For the SLP, the analytic expression for the ratio of the probabilities of responding class C_m and C_n can be rewritten in a basic form (Eq. 3.21) which allows for the direct interpretation of the behavior of the model. In this expression, three exponential factors appear. The first factor, $\exp(-\overline{b})$, is interpreted as a "global attenuator" of the influence of the stimulus features on the class C_m vs. class C_n probability ratio. The second factor, $\exp(-\sum_{k=1}^{N_F} \overline{w}_k F_k)$, is interpreted as a "feature-dependent attenuator". The third factor, $\exp(\Delta b + \sum_{k=1}^{N_F} \Delta w_k F_k)$, solely determines the position of the equal-probability boundaries and sets the basic shape of the probability space.

It is established that, also within our model, the SLP applies linear equal-probability boundaries between each pair of response classes. The response region for each of the classes is demarcated by piecewise linear boundaries.

We have discussed the fact that the MLP is essentially not a prototype-based model: the concept of distance to a prototypical exemplar does not play a role. It is shown, however, that, the MLP is similar to a prototype model which has its prototypes located *outside* the stimulus-feature range. In particular, it is proved that for a certain limit case the SLP and SCM coincide, which is symbolically expressed as

$$\lim_{b_j \to -\infty} \text{SLP} = \lim_{\|P_j\| \to \infty} \text{SCM}$$

The psychologically important concept of similarity appears explicitly in the model as the output of the output nodes of the MLP. Like in the unbiased SCM the class probabilities are derived from the similarities simply by normalization.

The theory set up for the SLP is extended to the TLP (the MLP with one hidden layer). It is indicated that the modeling power of the TLP can be very large, as it is capable of applying nonlinear equal-probability boundaries, and of modeling non-convex response regions. Because the equations for the nonlinear class boundaries are hard to interpret, a method is introduced in which the sigmoid function in the hidden nodes is approximated by a piecewise linear "pseudo-sigmoid". The combined effect of all pseudo-sigmoids is that the entire feature space is subdivided into a number of non-overlapping subspaces, in each of which the similarities are linear functions of the stimulus features, like in the SLP. The resulting class boundaries are piecewise linear approximations to the nonlinear boundaries in the original TLP.

Appendix 3.A
Proof of limit case

In this appendix it is shown that the SLP and the SCM coincide in the limit case, when the SLP-biases tend to $-\infty$ and the distance of all prototypes to the origin approach infinity.

We assume that all stimulus features are normalized using Eq. (3.7), so that all values are grouped around the origin. First it is to be shown that the distance between $\mathbf{F}$ and $\mathbf{P_j}$ is linear in $\mathbf{F}$ when $\|\mathbf{P_j}\|$ tends to ∞. Let $\mathbf{\Pi_F}$ represent the orthogonal projection of the feature vector $\mathbf{F}$ on the prototype vector $\mathbf{P_j}$ of class C_j, in the vector space with dotproduct

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t W \mathbf{y}$$

W denoting the diagonal matrix of attention weights.
Since $\langle \mathbf{\Pi_F} - \mathbf{F}, \mathbf{P_j} \rangle = 0$ by definition, it follows that the distance d_j of $\mathbf{F}$ to prototype $\mathbf{P_j}$ equals

$$d_j = d(\mathbf{F}, \mathbf{P_j}) = \sqrt{d^2(\mathbf{F}, \mathbf{\Pi_F}) + d^2(\mathbf{\Pi_F}, \mathbf{P_j})} \tag{A–1}$$

Since

$$\lim_{x \to \infty} \{\sqrt{a + x} - \sqrt{x}\} = 0$$

it follows that

$$\lim_{\|\mathbf{P_j}\| \to \infty} \{\sqrt{d^2(\mathbf{F_i}, \mathbf{\Pi_F}) + d^2(\mathbf{\Pi_F}, \mathbf{P_j})} - \sqrt{d^2(\mathbf{\Pi_F}, \mathbf{P_j})}\} = 0$$

Thus, when $\|\mathbf{P_j}\|$ is large, we find

$$d(\mathbf{F}, \mathbf{P_j}) \approx \sqrt{d^2(\mathbf{\Pi_F}, \mathbf{P_j})} = d(\mathbf{\Pi_F}, \mathbf{P_j}) \tag{A–2}$$

Because $\mathbf{\Pi_F}$ and $\mathbf{P_j}$ have the same direction

$$d(\mathbf{\Pi_F}, \mathbf{P_j}) \quad = \quad \|\mathbf{P_j}\| - \|\mathbf{\Pi_F}\|$$

After some calculation, it follows that

$$d(\mathbf{\Pi_F}, \mathbf{P_j}) = \sqrt{\sum_{k=1}^{N_F} w_k P_{jk}^2} - \frac{\sum_{k=1}^{N_F} w_k P_{jk} F_k}{\sqrt{\sum_{l=1}^{N_F} w_l P_{jl}^2}}$$

Hence, $d(\mathbf{\Pi_F}, \mathbf{P_j})$ is a linear function of $\mathbf{F}$.

If we substitute

$$\omega_{kj} = \frac{w_k P_{jk}}{\sqrt{\sum_{l=1}^{N_F} w_l P_{jl}^2}} \tag{A–3}$$

Eq. (A–3) simplifies to

$$d(\mathbf{\Pi_F}, \mathbf{P_j}) = \sqrt{\sum_{k=1}^{N_F} w_k P_{jk}^2} - \sum_{k=1}^{N_F} \omega_{kj} F_k$$

According to Shepard (1958) and Luce (1963), the biased similarity $b_j s_j$ of $\mathbf{F}$ to $\mathbf{P_j}$ is defined by

$$b_j s_j = b_j \exp(-d(\mathbf{F}, \mathbf{P_j})) \tag{A–4}$$

Using Eq. A–2 when $\|\mathbf{P_j}\|$ is large we find

$$\begin{aligned} b_j s_j &\approx b_j \exp(-d(\mathbf{\Pi_F}, \mathbf{P_j})) \\ &= \exp(\ln b_j - d(\mathbf{\Pi_F}, \mathbf{P_j})) \\ &= \exp\left(\ln b_j - \sqrt{\sum_{k=1}^{N_F} w_k P_{jk}^2} + \sum_{k=1}^{N_F} \omega_{kj} F_k\right) \end{aligned} \tag{A–5}$$

If we now substitute

$$\beta_j = \ln b_j - \sqrt{\sum_{k=1}^{N_F} w_k P_{jk}^2} \tag{A–6}$$

Eq. (A–5) simplifies to

$$b_j s_j \approx \exp(\beta_j + \sum_{k=1}^{N_F} \omega_{kj} F_k) \tag{A–7}$$

Let us now turn to the SLP. As stated in Eq. (3.5), the similarity s_j of $\mathbf{F}$ to class C_j is defined as

$$s_j = (1 + \exp(-d_j))^{-1} \tag{A–8}$$

$$= (1 + \exp(-b_j - \sum_{k=1}^{N_F} w_{kj} F_k))^{-1} \tag{A–9}$$

Using

$$\lim_{d_j \to -\infty} \{\frac{(1 + \exp(-d_j))^{-1}}{\exp(d_j)}\} = 1$$

we find that

$$\begin{aligned} \lim_{b_j \to -\infty} \{\frac{s_j}{\exp(b_j + \sum_{k=1}^{N_F} w_{kj} F_k)}\} &= \lim_{b_j \to -\infty} \{\frac{(1 + \exp(-b_j - \sum_{k=1}^{N_F} w_{kj} F_k))^{-1}}{\exp(b_j + \sum_{k=1}^{N_F} w_{kj} F_k)}\} \\ &= 1 \end{aligned} \tag{A–10}$$

Thus, for $b_j < 0$ and $|b_j|$ large, Eq. A–9 simplifies to

$$s_j \approx \exp(b_j + \sum_{k=1}^{N_F} w_{kj} F_k) \tag{A–11}$$

Expression (A–11) is equivalent to Eq. (A–7) for the similarity of the SCM with the prototypes $\mathbf{P_j}$ at infinite distance from the origin. Thus we find that, in this limit case, the SLP-biases b_j are equivalent to the SCM parameters β_j, which stand for $\ln b_j - \sqrt{\sum_{k=1}^{N_F} w_k P_{jk}^2}$, and the SLP-weights w_{kj} are equivalent to the SCM parameters ω_{kj}, which stand for $\frac{w_k P_{jk}}{\sqrt{\sum_{l=1}^{N_F} w_l P_{jl}^2}}$ (see Table 3.II).

Appendix 3.B
Number of subspaces

In this appendix it is shown that the number of subspaces N_k where k hidden nodes are operational is maximally equal to

$$N_k = \binom{N_H}{k} 2^{N_H - k}$$

Suppose that in a certain subspace k hidden nodes are operational. Then there are $\binom{N_H}{k}$ combinations possible where k out of N_H hidden nodes are operational.

When k hidden nodes are operational, $N_H - k$ hidden nodes are saturated. Suppose that i out of these $N_H - k$ hidden nodes are saturated on the left-hand side ($x \leq -a$). Then there are $\binom{N_H - k}{i}$ possible combinations that i out of $N_H - k$ hidden nodes are left-saturated. Thus, there are $\binom{N_H}{k}\binom{N_H - k}{i}$ possible combinations where k hidden nodes are operational and i hidden nodes are left-saturated.

In order to calculate the total number of combinations N_k where k hidden nodes are operational, irrespective of the number of left-saturated hidden nodes, we sum over all possibilities of having i left-saturated hidden nodes:

$$N_k = \binom{N_H}{k} \sum_{i=0}^{N_H - k} \binom{N_H - k}{i} \tag{B–1}$$

Using

$$\sum_{n=0}^{m} \binom{m}{n} = 2^m$$

Eq. (B–1) simplifies to

$$N_k = \binom{N_H}{k} 2^{N_H - k} \tag{B–2}$$

Note that, when the d_{0j} are dependent, that is, when $\exists_{(\lambda_0, \lambda_1, \ldots)} : d_{0j} = \lambda_0 + \sum_{k \neq j} \lambda_k d_{0k}$, the number of subspaces is smaller than 3^{N_H} because not all combinations are possible. This will always be the case when $N_H > N_F$.

Chapter 4

The multi-layer perceptron as a model of human categorization behavior. II. Practical aspects[1]

Abstract

In this chapter practical methods are presented for estimating the model parameters and the goodness-of-fit (GOF) of the MLP as a model of human categorization behavior. A measure of GOF is defined which is interpreted as a generalized "percent correct score". The "leaving-one-out method", a cross-validation technique which is commonly used in the field of statistical pattern recognition, is adopted to estimate the generalizability of a model estimation. Finally, the methodology is illustrated by a practical example which deals with the issue of the perception of stop consonants. The danger of overfitting is demonstrated, and the best fitting model is interpreted using the theory presented in the previous chapter.

4.1 Introduction

In the previous chapter we presented a model for human categorization behavior which is centered around the multi-layer perceptron (MLP). The general purpose of using a formal categorization model, like the MLP, for the analysis of categorization data, is the extraction of knowledge about the perceptual process under study. In order to achieve this goal, three more or less distinct steps have to be made:

1. The model parameters have to be estimated in such a way that the best possible account is given of the observed behavior. We will call this step the *model estimation.*
2. The performance of the model in accounting for the observed behavior has to be estimated. This step will be called *model evaluation.*
3. The model has to be interpreted in order to gain insight into the relevant perceptual processes. This step will be called *model interpretation.*

The third step of the analysis, the model interpretation, is very specific for the model that is used. The model will be interpreted in terms of its assumptions and parameters. For example, a similarity-choice model (SCM, Shepard, 1958; Luce, 1963) will be interpreted in terms of where prototypes lie, what the attention weights

[1]Based on: Smits, R., and Ten Bosch, L. (**1994b**), "The multi-layer perceptron as a model of human categorization behavior. II. Practical aspects," submitted to J. Math. Psych.

are, and possibly, through a comparison of the performance of various instances of the model, whether or not the response classes should be described by multiple prototypes. The specific way of interpreting the MLP is extensively described in the previous chapter.

The present chapter will focus on the first two steps of the analysis, the model estimation and evaluation. These two steps have a practical nature and are sometimes rather underexposed or treated in an offhand manner, which may reduce the validity of the model interpretation. Therefore we considered it necessary to describe useful estimation and evaluation methods in sufficient detail to allow for implementation by interested readers. Although the methods presented in this chapter are developed particularly for the MLP, they may also be useful for the estimation and evaluation of other categorization models, such as the SCM.

The chapter is organized as follows. In the next section we will present a method for estimating the MLP-parameters, which is based on a function minimization technique. Next, we will discuss the evaluation of the MLP-model. A *cross-validation* technique which stems from the field of statistical pattern recognition will be adapted to test the generalizability of our model. Next, an example is given where the MLP is estimated, evaluated and interpreted for categorization data that are generated in a speech perception experiment. In the final section, the findings are discussed and summarized.

4.2 Model estimation

Given a particular model architecture, e.g. the topology of the MLP, the parameters of the model have to be estimated such that the "best" possible account is given of the observed data by maximizing the level of "goodness-of-fit" (GOF). In practice, often the "badness-of-fit" (BOF), a monotonously decreasing function of the GOF, is minimized. Our specific choice of GOF and BOF are defined later.

The estimation of MLP-parameters is often called *training* of the MLP. We define the term *training* here as the iterative adjustment of the weights and biases of the MLP, with the purpose of minimizing the BOF. It is important to notice that no direct method exists for the computation of the MLP weights and biases from the observed data.[2] For a general understanding of the methods used for the determination of the optimal MLP-parameter values, it is useful to think in terms of a *search* in a *model-parameter space* (e.g. Haykin, 1994; Hertz *et al.*, 1991). The model-parameter space is spanned by the model parameters, and a *cost function* defines the BOF of the model at each point in this space. Thus, the purpose of the training is to search for the absolute minimum in the model-parameter space. For the training procedure we need (1) to define a cost function, and (2) to choose a search method.

[2]Some alternative models, like general recognition theory (GRT, Ashby and Perrin, 1988) or discriminant analysis (DA, e.g. Fukunaga, 1972), do allow for direct computation of their model parameters.

4.2.1 Cost function and goodness-of-fit

In section one of the previous chapter, it was established that, given a vector of stimulus features $\mathbf{F_i}$ for stimulus S_i, the normalized output of the MLP is a vector $\mathbf{p_i}$, of which each component p_{ij} represents the probability of assigning stimulus S_i to class C_j. Furthermore, it was assumed that the probability $P(\mathbf{R_i}|\mathbf{p_i})$ that a model with class probabilities $\mathbf{p_i}$ generates the observed response vector $\mathbf{R_i}$ after N_p presentations of stimulus S_i is given by the multinomial distribution:

$$P(\mathbf{R_i}|\mathbf{p_i}) = N_p! \prod_{j=1}^{N_r} \frac{p_{ij}^{R_{ij}}}{R_{ij}!} \tag{-1}$$

We define the *optimal model* as the model for which the probability of generating the observed response $\mathbf{R_i}$ is maximal. As is well-known, the class probability vector $\mathbf{p_i^o}$ of this model is given by

$$\mathbf{p_i^o} = \frac{1}{N_p}\mathbf{R_i}$$

Note that

$$P(\mathbf{R_i}|\mathbf{p_i}) \leq P(\mathbf{R_i}|\mathbf{p_i^o}) \tag{-2}$$

The ratio L_i of the probability of the actual model generating $\mathbf{R_i}$ and the probability of the optimal model generating $\mathbf{R_i}$, given stimulus S_i, equals

$$L_i = \frac{P(\mathbf{R_i}|\mathbf{p_i})}{P(\mathbf{R_i}|\mathbf{p_i^o})} = \prod_{j=1}^{N_r} \left(\frac{N_p p_{ij}}{R_{ij}}\right)^{R_{ij}}$$

Because of Eq. (-2) $L_i \leq 1$. L_i is a probability ratio defined for N_p presentations of stimulus S_i. The average value of L_i *per single presentation* of stimulus S_i is given by

$$\overline{L_i} = L_i^{\frac{1}{N_p}} = \prod_{j=1}^{N_r} \left(\frac{N_p p_{ij}}{R_{ij}}\right)^{\frac{R_{ij}}{N_p}} \tag{-3}$$

Since $L_i \leq 1$ also $\overline{L_i} \leq 1$. Note that, if we define $p_{ij}^{\text{obs}} = \frac{R_{ij}}{N_p}$, then Eq. (-3) can be rewritten as

$$\overline{L_i} = \prod_{j=1}^{N_r} \left(\frac{p_{ij}}{p_{ij}^{\text{obs}}}\right)^{p_{ij}^{\text{obs}}}$$

We now define the GOF P_c of a model as the average of $\overline{L_i}$ over all stimuli S_i:

$$P_c = \frac{1}{N_s}\sum_{i=1}^{N_s} \overline{L_i} = \frac{1}{N_s}\sum_{i=1}^{N_s}\prod_{j=1}^{N_r} \left(\frac{N_p p_{ij}}{R_{ij}}\right)^{\frac{R_{ij}}{N_p}} \tag{-4}$$

Clearly, $0 \leq P_c \leq 1$ since $\overline{L_i} \leq 1$.
P_c is interpreted as the average probability of the model's generating the observed response on a single presentation of a stimulus. For the sake of simplicity, we will interpret P_c as the *probability of a correct response*, hence the subscript "c". During training, instead of maximizing P_c, we minimized the cost function $1 - P_c$. This cost function can be interpreted as the probability of an *incorrect* response.

A few remarks are in order here.

Firstly, P_c is derived in such a way that it is a generalization of the probability of correct classification (or *percentage correct*) which is widely used in the field of statistical pattern recognition (e.g. Fukunaga, 1972). In this field, classification is usually "crisp", that is, a classification is either correct or incorrect. This translates into the p_{ij} always being either 0 or 1, the R_{ij} always being either 0 or N_p, and $P(\mathbf{R_i}|\mathbf{p_i^o}) = 1$.

Secondly, we chose to average the ratio $\overline{L_i}$ over all stimuli S_i. Instead of optimizing the average probability of a correct response to a presentation of a *single* stimulus, we could have chosen to optimize the probability of generating the *entire* response set. For this GOF-measure the summation sign in Eq. (–4) would be replaced by a multiplication sign. Although the latter measure is equally interpretable, it makes the model fit particularly sensitive to "outliers". While the summation measure optimizes the average fit, the multiplication measure biases the fit to the worst cases.

Finally, we remark that our cost function $1 - P_c$ differs substantially from the sum-of-squared-errors (SSE) cost function used in many studies (e.g. Nosofsky, 1986; Massaro and Friedman, 1990; Ashby and Lee, 1991). We define the sum-of-squared-errors cost function SSE here as

$$SSE = \sum_{i=1}^{N_s} \sum_{j=1}^{N_r} (N_p p_{ij} - R_{ij})^2 \qquad (-5)$$

It is our opinion that our measure $1 - P_c$ has a statistical interpretation which is related to the multinomial distribution, while the cost function SSE has a statistical interpretation only if the differences between observation and prediction for all cells of the confusion matrix can be assumed to be independent and Gaussian. A simple example clearly demonstrates an important conceptual difference between the function $1 - P_c$ and the SSE cost functions. Suppose that the observed response frequencies for a 3-class labeling task are $(98, 1, 1)$ for stimulus S_1 and $(40, 30, 30)$ for stimulus S_2, and suppose that a model predicts class probabilities $(0.90, 0.05, 0.05)$, and $(0.32, 0.34, 0.34)$ for stimulus S_1 and S_2, respectively. SSE is equal for both stimuli, namely $SSE = (98-90)^2 + (1-5)^2 + (1-5)^2 = (40-32)^2 + (30-34)^2 + (30-34)^2 = 96$. $1 - P_c$ for stimulus S_1 and S_2 is $1 - \left(\frac{90}{98}\right)^{0.98} \cdot \left(\frac{5}{1}\right)^{0.01} \cdot \left(\frac{5}{1}\right)^{0.01} = 0.05$, and $1 - \left(\frac{32}{40}\right)^{0.40} \cdot \left(\frac{34}{30}\right)^{0.30} \cdot \left(\frac{34}{30}\right)^{0.30} = 0.14$, respectively. In other words, when the cost function $1 - P_c$ is used, the closer the response fractions are to 0 or 1, the more accurately they must be modeled; the mid-range can be modeled in a sloppier sense.

For SSE the differences count equally. An elaborate discussion of these issues is given in Ten Bosch and Smits (1994).

4.2.2 Search Method

A *search method* is a numerical technique for finding a minimum in an model-parameter space. The three most widely used search methods are (e.g. Hertz *et al.*, 1991; Haykin, 1994):

1. Back-propagation, a technique which is developed particularly for MLPs with SSE-type cost functions;
2. Conjugate-gradient method, a function minimization technique which is especially efficient for quadratic functions;
3. Quasi-Newton method, a general-purpose function minimization technique.

Because of our non-quadratic cost function we used a quasi-Newton method (algorithm by Gill and Murray, 1976) in all minimizations carried out for the examples presented in a later section.

As with each minimization procedure, we here encounter the problem of ending up in local instead of global minima (Hertz *et al.*, 1991; Haykin, 1994). Each local minimum represents an instantiation of the model that locally performs best. The number of local minima in the model-parameter space can be very high and, since we do not have additional information on the shape of the cost function, it is generally difficult, if not impossible, to decide whether or not the global minimum has been reached. A number of practical solutions to this problem are presented in the literature, e.g. Aarts and Korst (1989), Haykin (1994). One of the simplest and most commonly adapted practical procedures is to randomly choose several initial positions in the model-parameter space, train the MLP for each of these initial positions and finally choose the model which yields the lowest final cost. It is not easy, however, to estimate beforehand how many initial positions have to be tried before one can be reasonably confident of having obtained the global minimum, because the number of local minima depends on the complexity of the data and on the MLP topology.

In order to overcome this problem, we used a method which is related to the *bootstrapping technique* (Efron, 1982). The essence of the proposed method is that several minimizations with different initial positions are carried out, and after each minimization it is evaluated whether or not a good estimate of the distribution of final costs has been reached. To this end, all final costs obtained so far are binned, yielding a cost histogram. This histogram can, after normalization, be viewed as giving an estimate of the probability-density function (PDF) of final costs associated with all local minima in the model-parameter space. If, after a number of initial positions has been tried, the shape of the histogram is stable, it is concluded that the probability of finding a new minimum which is lower than the lowest so far is acceptably low, and no new initial positions are tried. Otherwise, additional training runs are initiated.

The decision whether or not the distribution is stable can be made as follows. After each training run, the final cost $1 - P_c$ is binned. In the examples discussed in a later section, the interval for $1 - P_c$ that was spanned by the bins was chosen to be [0, 0.5]. A bin size of 0.05 was used, which makes the total number of bins $N_{\text{bins}} = 10$. Initially, $2N_i$ training runs are carried out, before making the comparison between the cost distribution of the first N_i runs and the last N_i runs. The value of N_i was chosen in such a way that the average number of costs per bin was large enough to use a χ^2-test. Therefore, N_i was chosen to be 6 times the number of bins, which makes $2N_i$ equal to 120 for the bin size of 0.05.

After the $2N_i$ initial training runs, the two cost distributions are compared. The likelihood ratio test (LRT, Wilks, 1935) is used to test the null-hypothesis stating that the two distributions are sampled from the same underlying distribution. Suppose that the number of costs in bin i of distribution 1 and bin j of distribution 2 are n_i and m_j, respectively. The maximum probability P_s of finding the two distributions under the restriction that they are generated by the same underlying distribution is

$$P_s = N_c!^2 \prod_i \frac{\left(\frac{n_i+m_i}{2N_c}\right)^{n_i+m_i}}{n_i!m_i!}$$

where N_c is the total number of costs per distribution.
The maximum probability P_d of finding the two distributions without the restriction is

$$P_d = N_c!^2 \prod_i \frac{\left(\frac{n_i}{N_c}\right)^{n_i} \left(\frac{m_i}{N_c}\right)^{m_i}}{n_i!m_i!}$$

Taking minus twice the logarithm of the ratio $\frac{P_s}{P_d}$ yields

$$-2 \ln \frac{P_s}{P_d} = 2 \sum_i \{ n_i \ln \left(\frac{2n_i}{n_i + m_i} \right) + m_i \ln \left(\frac{2m_i}{n_i + m_i} \right) \} \tag{-6}$$

which follows a χ^2-distribution with $(N_{\text{bins}} - 1)$ degrees of freedom (Wilks, 1935). If the probability of a type II error, that is, the probability of falsely accepting the null-hypothesis, is larger than a threshold value χ^2_{thr}, 2 new training runs are initiated and the test is performed again. Otherwise, the process is terminated and the initialization which yields the best fit is chosen. For all model fits in the examples, a threshold β for the probability of falsely accepting the null-hypothesis of $\beta = 0.10$ was used. In almost all cases, after 120 training runs the test value of χ^2 was below χ^2_{thr}, so the procedure was automatically terminated.

With an increasing number of training runs, the two distributions will become more and more similar. When a small MLP is used, the model-parameter space will contain relatively few local minima. Consequently, the cost distribution is simple and the training process is terminated early. When, on the other hand, the MLP is large, the model-parameter space is complex, and many runs are needed before the histograms are stable, as will be shown in some examples in a later section.

4.2.3 Treatment of empty cells

It may occur that a number of cells in the observed response matrix contain zeros. This is problematic for two reasons. Firstly, although a zero is highly informative on the observed response behavior, the cell does not play a role in the actual model estimation due to our particular choice of cost function (Eq. -4). This is caused by the fact that an empty cell, $R_{ij} = 0$, gives a factor 1 in our cost function, irrespective of the value of p_{ij}. Secondly, we have observed in practice that the search in the model-parameter space is more troublesome when many zeros are present in the confusion matrix than when it contains few zeros.

A practical remedy to these difficulties is to perform a slight smoothing of the data before the models are trained (e.g. Agresti, 1990, pp. 249, 250). For the calculations for our examples we decided to follow the suggestions made by Agresti (1990). If, for a particular stimulus S_i, a zero occurred in $\mathbf{R_i}$, a small constant was subtracted from the component of $\mathbf{R_i}$ which has the highest value, and this constant was put in place of the zero, thus leaving the total number of responses unchanged. In our case, the total number of responses N_p for each stimulus is equal to 120. Various small constants, ranging from 1 to 10^{-4}, were tried out. A value of 0.1 (about 0.08% of N_p) appeared to give the best performance in terms of goodness-of-fit and was therefore chosen for all calculations.

4.3 Model evaluation

In this section we will deal with the question how the performance of a model can be estimated. Special emphasis is put on the generalizability of models, and an evaluation technique which is commonly used in the field of pattern classification is adapted to suit our categorization model. Finally, a note is made on chance-level performance of models.

4.3.1 Generalizability

In a perception experiment, for practical reasons only a limited set of stimuli can be used. Nevertheless, one wants to make claims about the general validity of the model. In Fukunaga and Kessell (1971) and Fukunaga (1972) a statistical method is presented for estimating the generalizable GOF of *crisp* categorization models. A crisp categorization model is defined as a categorization model which has deterministic representation and retrieval stages and which generates class probability vectors which always contain $N_r - 1$ components equal to zero and one component equal to one. *Fuzzy* categorization models (like the model presented in the previous chapter), on the other hand, are defined as models which generate output probabilities that can take on any value between zero and one. We will briefly review the evaluation method for the crisp case below and extend it to fuzzy classification models.

First we need to introduce some new notation. Let us indicate a categorization model by $\boldsymbol{\theta}$, which stands for a vector of model parameters. Furthermore, we define

a *data set* $\mathcal{D}$ as a set of vector pairs $\{\mathbf{F_i}, \mathbf{R_i}\}, i = 1, 2, ...$, where $\mathbf{F_i}$ and $\mathbf{R_i}$ are the feature vector and response vector for a stimulus S_i. The term *testing* is defined as determining the GOF of a trained model on a (possibly new) set of data. Finally, let $P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}])$ indicate the probability that the model $\boldsymbol{\theta}$, which is trained on a data set $\mathcal{D}_{\text{train}}$, makes an incorrect categorization of a datum from the test set $\mathcal{D}_{\text{test}}$.

The experimenter aims to estimate the performance of a certain categorization model on a general data set $\mathcal{D}_{\text{gen}}$. In other words, ideally, the experimenter wants to measure $P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{gen}}, \mathcal{D}_{\text{gen}}])$ (the entire set $\mathcal{D}_{\text{gen}}$ is used as training set as well as test set). In general, however, only a representative subset $\mathcal{D}_{\text{sub}}$ of $\mathcal{D}_{\text{gen}}$ is available. In Fukunaga and Kessell (1971) and Fukunaga (1972) it is shown that lower and upper bounds for $P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}])$ can be estimated from $\mathcal{D}_{\text{sub}}$. This is expressed in the inequality

$$E\{P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{sub}}, \mathcal{D}_{\text{sub}}])\} \leq P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{gen}}, \mathcal{D}_{\text{gen}}]) \leq E\{P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{sub}}, \mathcal{D}_{\text{gen}}])\} \qquad (-7)$$

where $E\{x\}$ denotes the expected value of a quantity x. In Eq. –7 the expected values of P_ϵ are calculated based on the distribution of P_ϵ for all possible subsets $\mathcal{D}_{\text{sub}} \subset \mathcal{D}_{\text{gen}}$.

The lower bound can be simply estimated by training and testing the model on the entire data set $\mathcal{D}_{\text{exp}}$ produced in an experiment. Thus we replace $E\{P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{sub}}, \mathcal{D}_{\text{sub}}])\}$ by $P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{exp}}, \mathcal{D}_{\text{exp}}])$. The upper bound can be estimated using a *cross validation* technique. In a cross validation technique the training set $\mathcal{D}_{\text{train}} \subset \mathcal{D}_{\text{exp}}$ and test set $\mathcal{D}_{\text{test}} \subset \mathcal{D}_{\text{exp}}$ are disjunct: $\mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset$.

The two best-known cross-validation methods are the 'sample partitioning' method and the 'Leaving-One-Out' (LOO) method. In the 'sample partitioning' method, the N available data are subdivided into two or more distinct subsets, and the model is trained on all but one subsets and tested on the remaining subset. In order to get accurate estimates of both the model and the GOF of the model, the training sets as well as the test set must be sufficiently large. As this is often not the case for perception data, this method may not lead to accurate estimates of GOF. In the case that N is not large, the LOO-method can be used, which is computationally more expensive but gives more accurate GOF estimates. In this method, the N data are subdivided into 2 subsets, one containing $N-1$ data and the other subset containing the single remaining datum. The model is trained on the $N-1$ data and tested on the remaining datum. Next, the N data are again subdivided into 2 subsets, one containing $N-1$ data and the other subset containing a different remaining datum. Again, a model is trained and tested as before, and the process is carried out N times in total, leaving out each of the N data once in the process. The resulting test error is defined as the average of the N test errors. As each datum is effectively used as training as well as test item, the method can be shown to make optimal use of the available samples, that is, it gives the closest possible upper bound approximation of $P_\epsilon(\boldsymbol{\theta}[\mathcal{D}_{\text{gen}}, \mathcal{D}_{\text{gen}}])$, given a set of N data (Fukunaga, 1972).

4.3.2 The LOO-method for fuzzy classification

The method for estimating $P_\epsilon(\theta[\mathcal{D}_{\text{gen}}, \mathcal{D}_{\text{gen}}])$ described above was developed for crisp categorization models. It can, however, be easily generalized to suit fuzzy categorization models, like our MLP-based model. The probability of a correct response P_c was defined earlier as

$$P_c = \frac{1}{N_s} \sum_{i=1}^{N_s} \prod_{j=1}^{N_r} \left(\frac{N_p p_{ij}}{R_{ij}} \right)^{\frac{R_{ij}}{N_p}} \tag{-8}$$

and the probability of an incorrect response P_ϵ was simply defined as

$$P_\epsilon = 1 - P_c \tag{-9}$$

In the LOO-method the model is repeatedly tested on a single datum. The probability of a correct response $P_{c,i}^{\text{test}}$ when the model is tested on datum i, that is on stimulus-response pair $\{\mathbf{F_i}, \mathbf{R_i}\}$, equals

$$P_{c,i}^{\text{test}} = \prod_{j=1}^{N_r} \left(\frac{N_p p_{ij}}{R_{ij}} \right)^{\frac{R_{ij}}{N_p}} \tag{-10}$$

Note that the p_{ij} are calculated by training the model on the set $\mathcal{D}_{\text{gen}} \backslash \mathcal{D}_i$, where $\mathcal{D}_i$ represents the left-out datum $\{\mathbf{F_i}, \mathbf{R_i}\}$.
The average probability of a correct response P_c^{test} after the entire LOO-cycle is given by

$$\begin{aligned} P_c^{\text{test}} &= \frac{1}{N_s} \sum_{i=1}^{N_s} P_{c,i}^{\text{test}} \\ &= \frac{1}{N_s} \sum_{i=1}^{N_s} \prod_{j=1}^{N_r} \left(\frac{N_p p_{ij}}{R_{ij}} \right)^{\frac{R_{ij}}{N_p}} \end{aligned} \tag{-11}$$

Note that Eq. (-11) is identical to Eq. (-8).

We summarize the process of estimating $P_c(\theta[\mathcal{D}_{\text{gen}}, \mathcal{D}_{\text{gen}}])$ by the inequality

$$P_c(\theta[\mathcal{D}_{\text{exp}}, \mathcal{D}_{\text{exp}}]) \geq P_c(\theta[\mathcal{D}_{\text{gen}}, \mathcal{D}_{\text{gen}}]) \geq \frac{1}{N_s} \sum_{i=1}^{N_s} P_{c,i}(\theta[\mathcal{D}_{\text{exp}} \backslash \mathcal{D}_i, \mathcal{D}_i]) \tag{-12}$$

Note that, compared to Eq. -7, the inequality signs have changed because Eq. -7 is expressed in P_ϵ, while Eq. -12 is expressed in P_c.

4.3.3 Chance-level performance

When the performance of a model is evaluated, it is important to be aware of the chance-level performance of the model. The chance-level performance is here defined as the highest possible goodness-of-fit that can be obtained without any knowledge

of the stimulus features. This means that, at the output of the chance-level model, we find a fixed class probability vector **p** which does not depend on the stimulus. In order to determine the chance level P_{ch}, we must find the vector **p** for which the probability of finding the observed responses reaches its maximum value:

$$P_{ch} = \text{argmax}_{p_j} \left\{ \frac{1}{N_s} \sum_{i=1}^{N_s} \prod_{j=1}^{N_r} \left(\frac{N_p p_j}{R_{ij}} \right)^{\frac{R_{ij}}{N_p}} \right\} \tag{-13}$$

In our examples we have calculated P_{ch} by minimizing $1 - P_c$. In general, GOF-train will always be larger than P_{ch}, but situations may occur where GOF-test is actually lower than P_{ch}.

Note that the chance-level model can be represented by a single-layer perceptron (SLP) in which all weights connecting the stimulus features to the output nodes are zero and only the biases are nonzero.

4.4 Example

In this section, the developed methodology is illustrated by a practical example. The data in this example are part of the data set presented in chapter 5.

4.4.1 Perception experiment

A set of stimuli used for this example consisted the release bursts, which were isolated from natural utterances consisting of an unvoiced stop consonant (/p/, /t/ or /k/) followed by a vowel (/a/, /i/, /y/ or /u/). These stimuli were presented to subjects who responded to each presentation with either P, T, or K ($N_r = 3$). In total 24 stimuli (2 tokens × 3 consonants × 4 vowels) were used ($N_s = 24$). Each stimulus was presented 6 times to each of 20 subjects. The responses of all subjects were summed, yielding a total of 120 responses per stimulus ($N_p = 120$). For a more extensive presentation of the experimental procedure, see chapter 5. The resulting response fractions p_{ij}^{obs} are shown in the matrix in Table 4.I.

4.4.2 Stimulus features

On the basis of a number of phonetic studies (e.g. Blumstein and Stevens, 1979) it was decided to measure the following 5 stimulus features on each of the 24 stimuli:

1. Energy of the burst (E);
2. Length of the burst (L);
3. Global spectral tilt of the burst (T);
4. Frequency of a broad mid-frequency peak of the burst (Fr);
5. Height of the broad mid-frequency peak of the burst (H).

The specific methods for measuring these features are described in chapter 6. The stimulus features were converted to Z-scores using Eq. 3.7.

Table 4.I: Matrix containing observed response fractions. The fractions are derived from the actual responses frequencies by dividing by $N_p = 120$.

stim.no.	P	T	K
1	0.8250	0.1166	0.0583
2	0.8333	0.0750	0.0916
3	0.4333	0.2250	0.3416
4	0.6083	0.2833	0.1083
5	0.0083	0.0166	0.9750
6	0.0083	0.0083	0.9833
7	0.7916	0.1500	0.0583
8	0.9000	0.0750	0.0250
9	0.1916	0.4833	0.3250
10	0.2916	0.4416	0.2666
11	0.0083	0.0416	0.9500
12	0.0008	0.2083	0.7908
13	0.7166	0.0666	0.2166
14	0.7750	0.1500	0.0750
15	0.0250	0.8333	0.1416
16	0.0166	0.8916	0.0916
17	0.0166	0.0166	0.9666
18	0.0083	0.0250	0.9666
19	0.9083	0.0583	0.0333
20	0.8916	0.0916	0.0166
21	0.4416	0.3916	0.1666
22	0.5083	0.4500	0.0416
23	0.0250	0.0083	0.9666
24	0.0416	0.0083	0.9500

4.4.3 Training and testing various topologies

In order to demonstrate the influence of the model topology on the GOF, various topologies were trained and tested on the data: SLPs with 1, 2, 3, or 4 input nodes and two-layer perceptrons (TLPs) with 2 hidden nodes and 2, 3, or 4 input nodes. For all topologies the number of output nodes was 3.

The earlier described automatic procedure for terminating the training process was carried out many times for estimating the GOF intervals (see Eq. –12), namely once to determine the upper bound ("GOF-train") and N_s times to determine the lower bound ("GOF-test"). Ideally, we would have trained and tested each model on all possible subsets of stimulus features to assess which set gives the best generalizable account of the data. However, as the computing cost of the LOO-method is very high we adopted the following less expensive method. Each model having N_F input nodes ($N_F = 1, 2, 3, 4$) was trained on all possible subsets of N_F features. The 3 subsets that gave the best GOF-train were then used for cross validation using the LOO-method. Finally, the feature subset which resulted in the best GOF-test was selected as the overall best subset of N_F features, given the model topology.

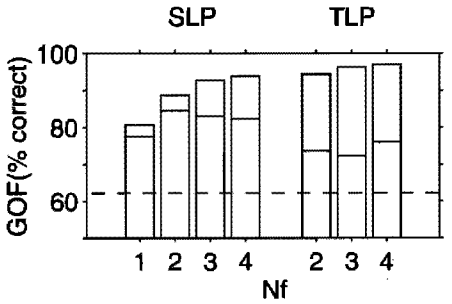

Figure 4.1: Goodness of fit on training and testing for various model fits, expressed in %. In each bar, the upper value indicates GOF-train, and the lower value indicates GOF-test. The dashed line represents chance level.

4.4.4 Results

Chance level for the observed stimulus-response matrix, after the treatment of empty cells, was 62.1%. While the marginal distribution of the stimulus-response matrix is $(0.387, 0.213, 0.400)$ for the response classes P, T and K, respectively, the chance-level model has fixed output probabilities $(0.481, 0.225, 0.294)$.

The GOF-levels for training and testing of the various model topologies are listed in Table 4.II and are shown graphically in Figure 4.1. In Table 4.II also the number of parameters is given for each topology.

Clearly, with increasing number of parameters, GOF-train keeps increasing. GOF-test on the other hand, quickly reaches a maximum with increasing number of parameters, and then decreases. This is a typical example of *overfitting* (e.g. Haykin, 1994). In general, overfitting, or non-generalizability, occurs when the number of model parameters is in the order of - or larger than - the number of data. For our example the number of degrees of freedom of the data is $(N_r - 1)N_s = 48$. Apparently, we need to keep the number of model parameters in our example roughly below $\frac{1}{4}$ of the number of degrees of freedom of the data in

Table 4.II: Goodness of fit on training and testing for various model fits, expressed in %.

	SLP				TLP		
N_F	1	2	3	4	2	3	4
N_w	6	9	12	15	15	17	19
GOF train	80.7	88.7	92.8	93.8	94.4	96.4	97.0
GOF test	77.5	84.6	83.0	82.3	73.6	72.3	76.0
Chance level	62.1	62.1	62.1	62.1	62.1	62.1	62.1

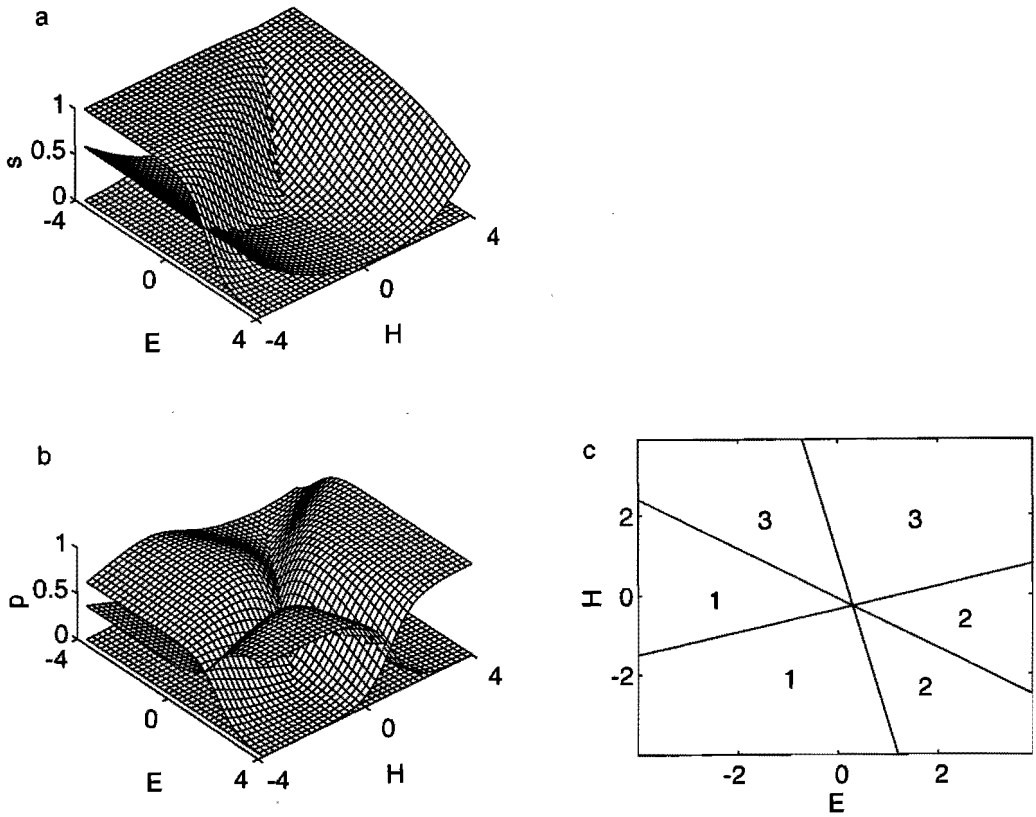

Figure 4.2: 4.2a. The functions s_1, s_2, s_3 as function of the stimulus features E and H for the SLP with the highest GOF-test. 4.2b. The functions p_1, p_2, p_3 as function of the stimulus features E and H for the same model. 4.2c. The equal-probability boundaries between classes 1 ("P"), 2 ("T") and 3 ("K") for the same model.

order to make a generalizable model estimation.

Let us look more closely at the model fit with the highest GOF-test, that is, the SLP with 2 input nodes. GOF-train and GOF-test are 88.7 and 84.6, respectively. The optimal stimulus features for this model are the burst energy E and the height of the mid-frequency peak H. The model parameters are $w_{11} = -2.447$, $w_{12} = -0.016$, $w_{13} = -0.670$, $w_{21} = -1.483$, $w_{22} = -0.893$, $w_{23} = 1.373$, $b_1 = -2.721$, $b_2 = -3.284$, $b_3 = -2.488$.

Figure 4.2 shows the similarities s_1, s_2, s_3 (Figure 4.2a), the class probabilities p_1, p_2, p_3 (Figure 4.2b), and the equal-probability class boundaries (Figure 4.2c). Roughly, we find that subjects tend to respond "P" (class 1) to stop-consonant release bursts when they have low energy and a weak mid-frequency peak. Bursts with high energy and a weak mid-frequency peak are labeled "T" (class 2), and

Table 4.III: Matrix containing class probabilities predicted by the SLP with two input nodes, using E and H as stimulus features.

stim.no.	P	T	K
1	0.8506	0.1293	0.0200
2	0.7866	0.0864	0.1269
3	0.6605	0.1006	0.2387
4	0.3499	0.0746	0.5753
5	0.0640	0.0818	0.8541
6	0.1205	0.0590	0.8203
7	0.8480	0.0694	0.0824
8	0.7137	0.2060	0.0802
9	0.1505	0.5717	0.2776
10	0.4372	0.4011	0.1615
11	0.0940	0.1516	0.7542
12	0.0319	0.0880	0.8800
13	0.6618	0.1241	0.2140
14	0.8165	0.1654	0.0180
15	0.0675	0.8443	0.0881
16	0.0318	0.5575	0.4106
17	0.0084	0.0724	0.9191
18	0.0017	0.1000	0.8981
19	0.8612	0.0821	0.0565
20	0.8671	0.0870	0.0458
21	0.2975	0.4544	0.2479
22	0.0097	0.0808	0.9093
23	0.0043	0.0384	0.9572
24	0.0059	0.0099	0.9840

bursts with a strong mid-frequency peak are generally labeled "K" (class 3).[3] These findings confirm the results of earlier phonetic studies in which acoustic classification experiments were carried out, like Halle *et al.* (1957) and Blumstein and Stevens (1979), as well as the results of phonetic perception studies where synthetic stimuli were used, like Blumstein and Stevens (1980) and Kewley-Port *et al.* (1983).

The class probabilities predicted by the model with the highest GOF-test are shown in Table 4.III. For a comparison with the observed response fractions, see Table 4.I.

4.5 General discussion and summary

The general process of analyzing categorization data using a formal categorization model can be subdivided into 3 steps: model estimation, model evaluation, and model interpretation. In this chapter, methods were described for carrying out the first two steps of this process when an MLP is used as categorization model.

[3]Recall that all stimulus features were transformed to Z-scores, so almost all feature values lie within the range $[-2, 2]$.

With respect to model *estimation*, a measure of goodness-of-fit has been introduced which naturally suits categorization models which have class probabilities as output. The GOF-measure is based on the multinomial function and is interpreted as a generalized "percent correct score" for fuzzy categorization models. A cost function is associated with the GOF-measure, which is minimized during training of the MLP. A heuristic method is proposed for automatically deciding if sufficient initial estimates of the weight vector have been tried.

With respect to model *evaluation*, the importance of generalizability of the model is stressed. The leaving-one-out method, a cross-validation technique which is regularly used in the field of statistical pattern recognition, is used to estimate the lower bound of the GOF of our model.

Finally, the methodology is demonstrated by an example dealing with the issue of the perception of stop consonants. A comparison of GOF-levels for training and testing MLPs of various complexities on a very limited data set clearly demonstrates the danger of overfitting. The model with the highest GOF-test is chosen as the "best" model in the sense that the best possible generalizable account is given of the observed data. The interpretation of the model suggests that the listeners' categorization of stop-consonant release bursts is in accordance with findings of earlier phonetic studies.

Chapter 5

The perception of burst-spliced prevocalic stop consonants[1]

Abstract

The experiments presented in this chapter and chapter 6 address the basic research question formulated in chapter 1, that is, to evaluate the relative importance of detailed and gross acoustic structures for the perception of place of articulation in prevocalic stop consonants. To this end, first a perception experiment is carried out with "burst-spliced stimuli". This experiment is described in this chapter. From a number of stop-vowel utterances, burst-only, burst-less and cross-spliced stimuli were created and presented to listeners. The results of the experiment show that the relative perceptual importance of burst and transitions highly depends on the stop consonant, the vowel context and whether the stop is voiced or unvoiced. Velar bursts are generally much stronger in cueing place of articulation than other bursts. The dental transitions appear to be weaker than labial or velar transitions. In front-vowel contexts the release burst dominates the perception of place of articulation, while in non-front vowel contexts the formant transitions are generally dominant. The bursts of unvoiced stops are perceptually more important than the bursts of voiced stops.

5.1 Introduction

During the last decades various types of acoustic cues to the perception of stop consonants have been proposed. Initially, signal structures were investigated which are explicitly visible in waveforms and spectrograms of stop-consonant signals. Release burst and formant transitions were treated as essentially separate signal portions, and the perceptual relevance of acoustic structures such as burst length, burst frequency and formant onset frequencies were studied (e.g. Cooper *et al.*, 1952; Liberman *et al.*, 1954; Schatz, 1954; Hoffman, 1958; Ainsworth, 1968). As indicated in the introductory chapter, we will call this type of acoustic properties *detailed cues.*

More recently, a number of cues have been proposed which are less clearly visible in the spectrogram and are of a more gross nature. In this approach, no explicit distinction is made between burst and formant transitions, but instead integrative structures are proposed as being the main cues for perception. Initially, the importance of the gross shape of the static spectrum of the first 20-odd ms of the signal

[1]Based on: Smits, R., Ten Bosch, L., and Collier, R. (**1995a**), "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants: I. Perception experiment," submitted to J. Acoust. Soc. Am.

after burst release was stressed (e.g. Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980). After a number of critical reappraisals (e.g. Blumstein *et al.*, 1982; Walley and Carrell, 1983), emphasis shifted towards the *dynamic* gross spectral shape of the first few tens of ms after release (Kewley-Port, 1983; Kewley-Port *et al.*, 1983; Kewley-Port and Luce, 1984; Lahiri *et al.*, 1984). We will call this type of acoustic properties *gross cues.*

Two studies have explicitly addressed the question whether detailed cues or (dynamic) gross cues are more important for perception of place of articulation in stop consonants. Both in Lahiri *et al.* (1984) and Lindholm *et al.* (1988), stylized synthetic stop-vowel stimuli were used, in which the formants cued one place of articulation, while the gross cues cued another place. Listeners classified the stimuli of Lahiri *et al.* mainly in accordance with the gross cues, while the stimuli of Lindhom *et al.* were mainly classified according to formant transitions. Thus, the issue whether detailed or gross cues are perceptually more important remains unsettled.

Recently, a number of acoustic studies have addressed the question whether it is possible to correctly classify a great number of naturally uttered stop-vowel signals using only detailed formant information or only gross spectro-temporal information. Amongst others, Forrest *et al.* (1988) and Nossair and Zahorian (1991) have shown that it is possible to make an excellent speaker-independent classification based on gross spectro-temporal information. In Sussman *et al.* (1991) and Sussman (1991), on the other hand, it was shown that a good classification is possible based on formant information only , namely F2-frequency at voicing onset and in the vowel, especially when combined with the F3 onset frequency.

It is the aim of this chapter to evaluate the importance of various cues for the perception of initial prevocalic stops. In particular we want to test the perceptual importance of detailed cues versus gross cues. Two methodological aspects will be emphasized. First of all, we will use (manipulated) natural utterances in our experiments in order to preserve the natural variability in the speech signal. Thus, we hope to avoid potential unnaturalness of the stimuli, which may have been partly responsible for the apparent contradiction in the results of previous studies using synthetic conflicting-cue stimuli (e.g. Lahiri *et al.*, 1984, versus Lindholm *et al.*, 1988). Secondly, we will make a complete simulation of the response behavior of the subjects using a formal model of categorization behavior. The simulation will enable us to establish (1) which set of cues gives the better account of the observed response behavior, and (2) how are the cues integrated by the listeners in order to arrive at their responses.

The approach is as follows. First we will create a set of stimuli by manipulating a number of natural stop-vowel utterances. However, here we encounter the difficulty that it is not possible to manipulate individual acoustic cues across a continuum. Detailed and gross acoustic structures generally *covary* in natural speech, for example, a high F2-frequency at voicing onset will often be accompanied by a positive global spectral tilt. We will reduce this covariation by creating *deleted-cue* as well as *conflicting-cue* stimuli, by removing parts of the original utterances or

exchanging information between utterances. To this end, we use the well-known "burst-splicing" technique (e.g. Fischer-Jørgensen, 1972). These stimuli will be presented to listeners for classification of place of articulation. Next, the relevant acoustic events, burst onset and voicing onset, are detected and various detailed and gross cues are measured on the stimuli. Before actually measuring the detailed cues, a critical study of the accuracy of the traditional quasi-stationary speech representations is made, which is especially relevant for the extraction of detailed cues, such as formant onset frequencies. Next, the measured cue values are mapped onto the observed responses of the listeners using a formal classification model.

In this chapter the perception experiment is described in which the burst-spliced utterances were used. The next chapter presents the model simulations, that is, the cue-measurement procedures, and the estimation and interpretation of the categorization models. It is the primary purpose of the burst-splicing experiment to provide the stimulus-response matrices on which the model fits can be performed for the evaluation of the cue sets. We have decided, however, to present and discuss the burst-splicing experiment independently, because we think that the perceptual data themselves (without the subsequent model fits) deserve special attention. As will be discussed later, there is only one previous study in which genuine *conflicting-cue* stimuli are used which are created through burst-splicing on initial stops, namely Fischer-Jørgensen (1972). However, this study is not widely accessible. Moreover, the "tape-cutting" technique that was used by Fischer-Jørgensen (1972) to create the stimuli, is possibly less precise than the digital signal processing techniques which are currently available.

The chapter is organized as follows. In the next section, results of previous studies using the burst-splicing technique are discussed. The procedure for our experiments is discussed in section 5.3. In section 5.4, the results of the experiments are presented. The chapter will be concluded with a discussion.

5.2 Previous studies using the burst-splicing technique

Before our specific implementation of the burst-splicing technique is described in detail, some results of earlier experiments using burst-spliced utterances of prevocalic stops are briefly discussed. For the purpose of our discussion, we define *deleted-cue* stimuli as stimuli from which one or more cues are removed. *Conflicting-cue* stimuli are defined as stimuli in which cues are present which point to different (conflicting) responses. Logically, perception experiments with deleted-cue stimuli generally measure the *necessity* of the deleted cues and the *sufficiency* of the remaining cues. Perception experiments with conflicting-cue stimuli measure the *relative importance* of the conflicting cues.

5.2.1 Experiments with deleted-cue stimuli

Experiments in which the formant transitions have been replaced by silence or a steady-state vowel have been conducted by Schatz (1954), Halle *et al.* (1957), Fischer-Jørgensen (1972), Winitz *et al.* (1972), Cole and Scott (1974), LaRiviere *et*

al. (1975), Dorman *et al.* (1977), Ohde and Sharf (1977), and Schouten and Pols (1983). The results and conclusions of these experiments were at variance with each other, as will be shown below.

Winitz *et al.* (1972), Cole and Scott (1974) and LaRiviere *et al.* (1975) found almost perfect place of articulation perception from burst-only stimuli. However, in all three studies the aspiration noise following the burst was included in the stimuli. Halle *et al.* (1957), Fischer-Jørgensen (1972), Ohde and Sharf (1977), and Schouten and Pols (1983) generally found a good (70% - 90% correct) place of articulation perception from bursts excised from voiceless stops and a somewhat worse score (50% - 80%) for bursts of voiced stops. Dorman *et al.* (1977) found a rather low performance (chance level to 50%) for bursts isolated from voiced stops.

Schatz (1954), Cole and Scott (1974) and Dorman *et al.* (1977) performed experiments in which release bursts spoken in various vowel contexts were spliced onto other (steady state) vowels. For velar bursts, which show a strongly vowel-dependent spectral peak, the resulting percept appeared to depend highly on the following vowel. Schatz showed that, for example, a release burst excised from /ski/ and spliced onto the steady state vowels /i,a,u/ results in clear perception of /ki,ta,pu/, respectively.

Experiments in which the release bursts have been replaced by silence have been conducted by Fischer-Jørgensen (1972), LaRiviere *et al.* (1975), Dorman *et al.* (1977), Ohde and Sharf (1977, 1981), Pols and Schouten (1978, 1981), Pols (1979) and Schouten and Pols (1983). Correct recognition of place of articulation from transitions isolated from voiced stops ranged between moderate and rather high (40% to 80% correct). In all cases, when results were not pooled over consonants, the results were clearly lower for velars than for labials and alveolars. Performance was always close to chance for voiced formant transitions from English unvoiced stops, because the formant transitions are generally completed within the aspiration phase. Dutch unvoiced stops, on the other hand, have no aspiration and the correct perception of place of articulation from voiced formants may be as high as 66% (Pols and Schouten, 1978).

5.2.2 Experiments with conflicting-cue stimuli

To our knowledge, Fischer-Jørgensen (1972) was the only one to perform conflicting-cue experiments to determine the relative importance of burst and transitions for the perception of place of articulation. She used stimuli in which bursts and transitions excised from real speech were combined in such a way that the release burst cued one place of articulation and the formant transitions another. Bursts were cross-spliced only within the same vowel contexts. One of the most important results of her study was that the relative importance of burst and formant transitions was highly dependent on the vowel context. This does not appear to be a widely acknowledged phenomenon. The perceptual data showed that the burst determines the percept in /i/-context, while the transitions determine the percept in /a/-context. In /u/-context the /g/-burst and /d/-transitions appeared to be strong cues. For unvoiced stops the situation was somewhat different: /t/-transitions were very ro-

bust, /k/ always needed a /k/-burst and /p/-transitions could only be overridden by /k/-bursts. Furthermore, Fischer-Jørgensen found that the perceived place of articulation may be different from the place of articulation of burst as well as transitions in their original contexts: a /tu/-burst spliced onto /ku/-transitions gave a clear /pu/-percept.

In our experiments, deleted-cue stimuli (burst-only and no-burst stimuli) as well as conflicting-cue stimuli (mixed-burst stimuli) will be used, together with original utterances.

5.3 Method

5.3.1 Stimuli

Two male Dutch talkers each spoke 2 tokens of all possible CV-combinations consisting of the Dutch stops /b,d,p,t,k/ followed by the Dutch vowels /a,i,y,u/. Dutch voiced stops show extensive prevoicing and unvoiced stops are not aspirated. The phoneme /g/ does not exist in the Dutch language. The talkers were seated in a sound-treated room and spoke into a microphone which was placed at approximately 30 cm from their mouths. All utterances were spoken moderately emphatically and in isolation. The material was low-pass filtered at 4.9 kHz and converted into 12-bit numbers at a sampling rate of 10 kHz. Thus, 80 original utterances were obtained, which were scaled to equal maximum instantaneous amplitude.

Next, for all utterances, the instant of burst onset was determined by hand with the aid of the waveform and the wideband spectrogram. Also the instant of burst offset was determined from the waveform and the wideband spectrogram.

For every utterance the release burst was separated from the rest of the signal. Pols and Schouten (1981) have shown that simply cutting a waveform in two may cause abrupt onsets and clicks, which can significantly bias the perceptual responses. They showed that this can be avoided by using a smooth time window to make a cut. For this study it was decided to follow this strategy. The procedure is illustrated for the utterances /ta/ and /ka/ in Figure 5.1.

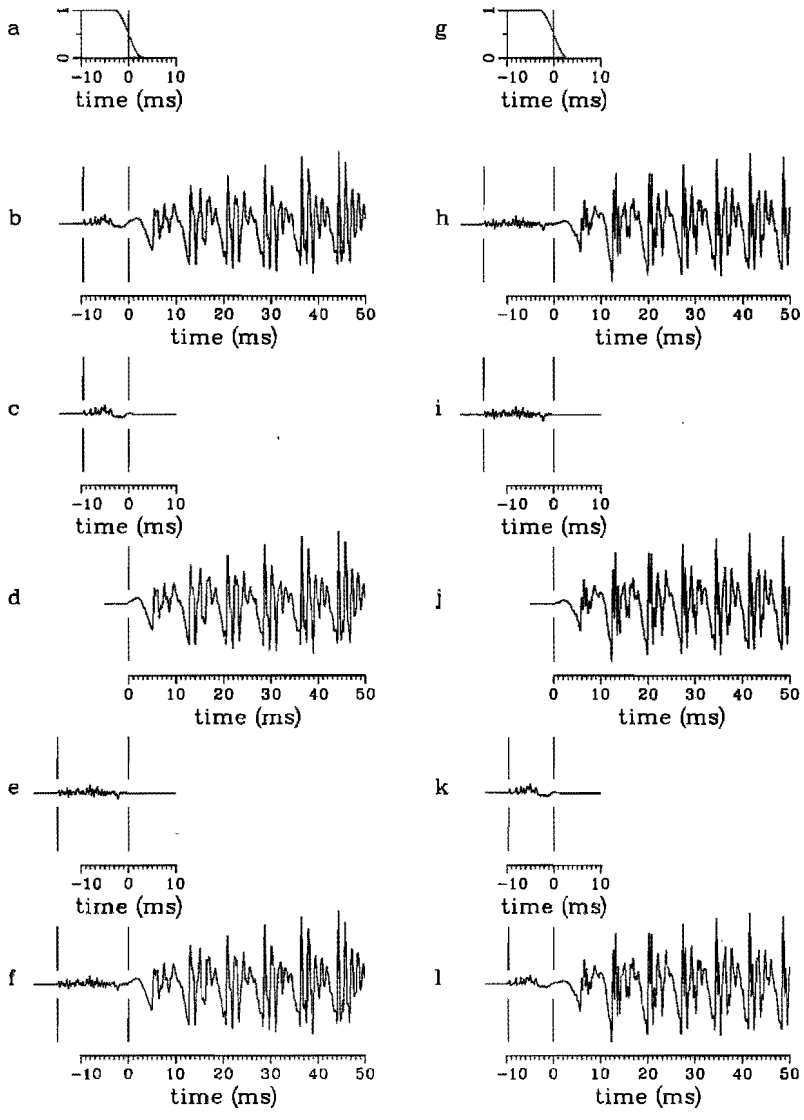
a
1
0
-10 0 10
time (ms)
b
-10 0 10 20 30 40 50
time (ms)
c
-10 0 10
time (ms)
d
0 10 20 30 40 50
time (ms)
e
-10 0 10
time (ms)
f
-10 0 10 20 30 40 50
time (ms)
g
1
0
-10 0 10
time (ms)
h
-10 0 10 20 30 40 50
time (ms)
i
-10 0 10
time (ms)
j
0 10 20 30 40 50
time (ms)
k
-10 0 10
time (ms)
l
-10 0 10 20 30 40 50
time (ms)

Figure 5.1: (*previous page*) Burst-splicing procedure for the unvoiced stops, illustrated for the utterances /ta/ and /ka/. The relevant parts of the original utterances /ta/ and /ka/ are shown in Figures 5.1b and h. The vertical lines are the burst-onset markers and burst-offset markers. Figures 5.1a and g show the cutting windows centered at the burst-offset markers. Applying these windows to the signals yields the burst-only signals, shown in Figures 5.1c and i, and the burst-less signals, shown in Figures 5.1d and j. Next, the /ta/-burst is aligned at its offset marker with the burst-offset marker of the burst-less portion of /ka/ (Figure 5.1k), and likewise for the burst of /ka/ and the burst-less portion of /ta/. The mixed-burst signals, displayed in Figures 5.1f and l, are obtained by adding the signals in 5.1d and e and adding the signals in 5.1j and k, respectively.

The right half of a Hanning window was centered at the voice-onset marker. The total length of the half window was 6.0 ms. The window is shown in Figures 5.1a and 5.1g. By multiplying the original signal (Figures 5.1b, 5.1h) with the window the burst was isolated (Figures 5.1c, 5.1i). Subtracting the isolated burst from the original signal yielded the "no-burst"-signal (Figures 5.1d, 5.1j). The conflicting-cue stimuli were created by aligning a burst-only signal from one utterance (Figures 5.1e, 5.1k) with a no-burst signal of another utterance at their voice-onset markers and adding the two signals (Figures 5.1f, 5.1l).

Conflicting-cue signals were created only within the same vowel-context, speaker and token number. For instance, the burst-only portion from the first /pa/-token by speaker 1 was only spliced onto the no-burst portion of the first /ta/- and /ka/-token by speaker 1. For all signals, except the burst-only signals, the duration of the voiced part was limited to 400 ms. This was achieved by multiplying the signal from 380 ms to 400 ms after the voice-onset marker with a linearly falling window. As the burst-only stimuli were rather soft, their levels were scaled up by 3.8 dB. Thus, from the 48 original utterances containing a voiceless stop, 48 burst-only stimuli, 48 no-burst stimuli, and 96 conflicting-cue stimuli were created.

The procedure for the voiced stops was more complicated because of the clear voice bars on which the release bursts are superimposed. The option of using the same windowing technique again was discarded, because the voice bar may contain place-of-articulation information (Barry, 1984). Moreover, cross-splicing a voice bar onto a different utterance may introduce a distracting lack of continuity in the signal, even when a smooth time window is used. Therefore, we resorted to a time-frequency filtering method which can be viewed as a specific instance of the general technique described by Saleh and Subotic (1985). The time-frequency filtering technique is similar to the separation of time-frequency "tiles", which has recently been used by Ghitza (1993). The procedure is illustrated for the utterances /by/ and /dy/in Figure 5.2.

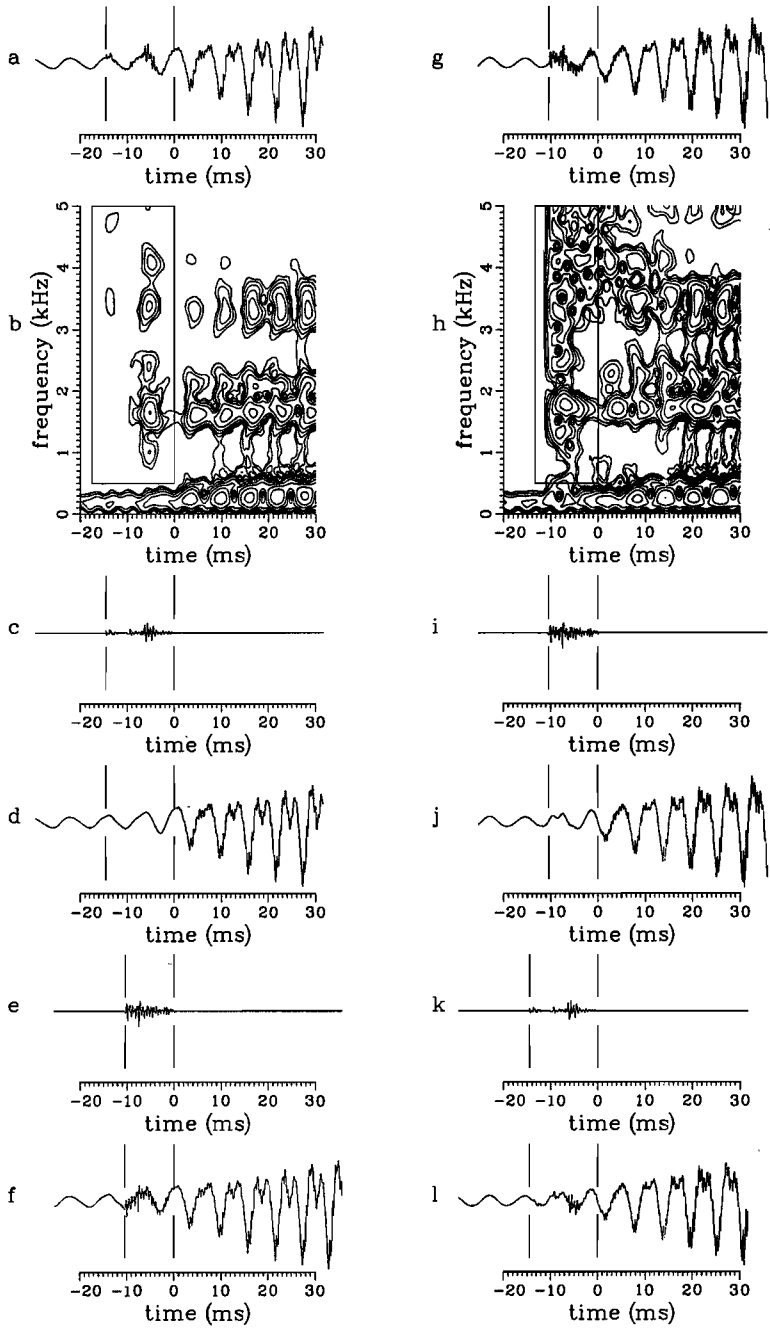
a
b
c
d
e
f
g
h
i
j
k
l
frequency (kHz)
0
1
2
3
4
5
-20 -10 0 10 20 30
time (ms)

Figure 5.2: (*previous page*) Burst-splicing procedure for the voiced stops, illustrated for the utterances /by/ and /dy/. The relevant parts of the original utterances /by/ and /dy/ are shown in Figures 5.2a and g, together with their burst-onset and burst-offset markers. Figures 5.2b and h show contour plots of the wideband spectrograms of these signals. The rectangles in the plots indicate the time-frequency areas which define the burst-signals. The right side of the area coincides with the burst-offset marker, the left side is positioned 3 ms to the left of the burst-onset marker. The lower boundary is located at 500 Hz. Setting the amplitudes outside the rectangles to zero and applying the resynthesis procedure yields the burst signals, which are shown in Figures 5.2c and i. The burst-less signals, shown in Figures 5.2d and 5.2j, are obtained by subtracting the burst signals from the original signals. The procedure for creating the mixed-burst signals, shown in Figures 5.2e, f, k and l, is identical to the procedure used for the unvoiced stops.

A Short-Time Fourier Transform (STFT, Rabiner and Schafer, 1978) was made of every utterance. A Hanning window with a total length of 6.0 ms and a window shift of 0.2 ms (2 samples) was used. Figures 5.2b and h show contour-plot spectrograms of the waveforms shown in Figures 5.2a and g, respectively. Next, all numbers in the STFT-amplitude which were either more than 3 ms left to the burst-onset marker or right to the voicing-onset marker or below 500 Hz were set to zero. This is the area outside the boxes in Figures 5.2b and h. The resulting STFT amplitude was combined with the unchanged STFT phase and the weighted overlap addition technique described by Griffin and Lim (1984) was used to synthesize the burst-only signal.[2] As in the case of the unvoiced stops, subtraction of the burst-only signal from the original signal yielded the no-burst signal, and conflicting-cue stimuli were created by adding burst-only signals to no-burst signals originating from an utterance with the same vowel but a different consonant, after alignment at the voicing-onset marker. Again, all signals, except the burst-only signals, were multiplied with a linearly falling window from 380 ms to 400 ms after the voice-onset marker. The burst-only signals were even softer than the burst-only signals of the unvoiced stops. Their levels were therefore scaled up by 7.0 dB. Thus, from the 32 originals, we created 32 burst-only stimuli, 32 no-burst stimuli, and 32 conflicting-cue stimuli.

5.3.2 Subjects and procedure

Twenty Dutch university students served as subjects. All subjects reported no history of any hearing loss. The 20 subjects were divided into 4 groups of five. Each group participated in 4 sessions, each of which consisted of 2 parts. The stimuli were divided into 8 blocks, according to speaker 1 vs. speaker 2, voiced vs. unvoiced, and burst-only vs. original, no-burst and mixed-burst. These blocks were distributed over the 4 sessions according to Table 5.I.

[2]The time-windowing method used for the unvoiced stops can be shown to be identical to the time-frequency technique used for the voiced stops, with a window shift of 0.1 ms instead of 0.2 ms, and no lower frequency boundary, instead of the lower frequency boundary of 500 Hz.

Table 5.I: Order of presentation of the stimuli over the 4 sessions for the 4 groups of subjects. "V" = voiced stops, "UV" = unvoiced stops.

	group 1		group 2		group 3		group 4	
session 1	speaker 1	V	speaker 1	UV	speaker 2	V	speaker 2	UV
		UV		V		UV		V
session 2	speaker 2	V	speaker 2	UV	speaker 1	V	speaker 1	UV
		UV		V		UV		V
session 3	speaker 1	V	speaker 1	UV	speaker 2	V	speaker 2	UV
(bursts)		UV		V		UV		V
session 4	speaker 2	V	speaker 2	UV	speaker 1	V	speaker 1	UV
(bursts)		UV		V		UV		V

The burst-only stimuli were presented separately from the other stimuli because of their special character. Each half of a session consisted of seven randomized blocks of stimuli. The first block was intended for familiarization and the results were not used for further analysis. Thus, each subject received each of the 368 stimuli 6 times.

Subjects were seated in a silent room 3 to 5 at a time. They received the stimuli through Sennheiser HD425 headphones. The presentation level was set to a comfortable position before the first experiment and was kept constant thereafter. In order to alert the subjects, each stimulus was preceded by a soft 500 Hz beep with a smooth onset and offset and with a duration of 300 ms. The silent interval between the offset of the beep and the onset of the stimulus was 800 ms. After the offset of the stimulus the subjects had 3.5 s time to indicate their response on an answer sheet, before the next beep was sounded. The subjects could choose between B and D for the voiced stops, and P, T, and K for the unvoiced stops. The sessions were interrupted every 20 minutes for a 5-minute break.

5.4 Results

5.4.1 Unvoiced stops

In Table 5.II, the stimulus-response matrix is presented for the experiments with the syllables containing the unvoiced stops /p,t,k/. Responses are pooled across subjects and presented in percentages, with 100% equal to 120 judgements. As subjects had 3 response alternatives, chance level corresponds to 33% correct.

The first eight blocks of four rows show the results for the original, no-burst, and burst-spliced stimuli. The first block of four rows contains the results for the four tokens of the original utterances /pa/ (first three columns), /pi/ (second three columns), /py/ (third three columns) and /pu/ (fourth three columns). The first and second token are spoken by speaker 1, the third and fourth token are spoken by speaker 2. The second group of four rows contains the results for the no-burst

stimuli from /pa/, /pi/, /py/ and /pu/. In the third group of four rows, the results are presented for the mixed-burst stimuli containing the bursts of /ta, ti, ty, tu/ and the transitions of /pa, pi, py, pu/, respectively. The same holds for the fourth group of four rows, except that the bursts are now from /ka, ki, ky, ku/. The bottom 12 rows give the results for the burst-only stimuli.

All *original stimuli* are recognized correctly at a rate higher than 95%, except token 1, 3 and 4 of /py/ (86.7%, 59.2% and 81.7%, respectively), token 3 and 4 of /ta/ (53.3% and 73.3%, respectively), and token 3 and 4 of /ti/ (85.0% and 93.3%, respectively). So, all but one of the less well recognized original utterances were produced by speaker 2. Listening analytically to these tokens revealed that the lower recognition rates were caused by a less clear pronunciation and not by clicks or other disturbances.

The average rate of correct recognition of the *burst-less stimuli* is 68.8% (94.8% for /p/, 64.2% for /t/, 47.4% for /k/), which is in agreement with results found by Fischer-Jørgensen (1972), LaRiviere *et al.* (1975), Ohde and Sharf (1977), Pols and Schouten (1978, 1981) and Schouten and Pols (1983). In accordance with other studies in which burst-less stimuli were presented to subjects, we find a strong preference for /p/-responses for the burst-less stimuli. This phenomenon is not well understood (e.g. Fischer-Jørgensen, 1972). On the one hand, it can be viewed as a response bias, which means that when insufficient or ambiguous information is present in the stimuli, the decision mechanism has a preference for a response /p/. On the other hand, it may be a genuine perceptual effect in the sense that the absence of a burst gives the stimuli a /p/-like quality, e.g. because labial bursts are often somewhat weak. It is also in agreement with previous studies, like Fischer-Jørgensen (1972) and Pols (1979), that the velars are poorly recognized when the burst is absent.

The average rate of correct recognition of the *burst-only stimuli* is 73.6% (80.0% for /p/, 49.6% for /t/, 91.1% for /k/). Although these levels are in agreement with those reported by Halle *et al.* (1957), Fischer-Jørgensen (1972), Ohde and Sharf (1977), and Schouten and Pols (1983), the variation in the recognition levels for bursts of /p/, /t/ and /k/ are so large that presenting only averages, as in some of these studies, seems unwarranted. The performance for /k/-bursts is strikingly high. Note that the recognition of the bursts from /ka/, /ky/ and /ku/ are generally above 95%.

Overall, the responses to the *mixed-burst stimuli* corresponded to the burst identity in 49% of the cases, and to the transitions in 42.8% of the cases. 7.8% of the responses corresponded with neither the burst nor the transitions. We performed a number of statistical analyses on the data for the mixed-burst stimuli. It was tested whether the number of responses which correspond with the burst-identity was sig-

BU	TR	vowel /a/ p	vowel /a/ t	vowel /a/ k	vowel /i/ p	vowel /i/ t	vowel /i/ k	vowel /y/ p	vowel /y/ t	vowel /y/ k	vowel /u/ p	vowel /u/ t	vowel /u/ k
p	p	99.2	0.8	0.0	100.0	0.0	0.0	86.7	0.0	13.3	100.0	0.0	0.0
		100.0	0.0	0.0	100.0	0.0	0.0	97.5	0.8	1.7	99.2	0.0	0.8
		100.0	0.0	0.0	99.2	0.8	0.0	59.2	5.8	35.0	98.3	0.0	1.7
		100.0	0.0	0.0	99.2	0.8	0.0	81.7	1.7	16.7	100.0	0.0	0.0
-	p	100.0	0.0	0.0	93.3	0.0	6.7	82.5	5.8	11.7	92.5	0.8	6.7
		99.2	0.0	0.8	96.7	1.7	1.7	80.0	5.8	14.2	90.8	2.5	6.7
		100.0	0.0	0.0	99.2	0.0	0.8	89.2	6.7	4.2	100.0	0.0	0.0
		99.2	0.8	0.0	100.0	0.0	0.0	94.2	4.2	1.7	100.0	0.0	0.0
t	p	97.5	2.5	0.0	40.0	60.0	0.0	9.2	90.8	0.0	98.3	1.7	0.0
		100.0	0.0	0.0	52.5	47.5	0.0	0.0	100.0	0.0	100.0	0.0	0.0
		100.0	0.0	0.0	6.7	93.3	0.0	0.0	100.0	0.0	96.7	3.3	0.0
		100.0	0.0	0.0	3.3	85.0	11.7	0.8	99.2	0.0	98.3	1.7	0.0
k	p	93.3	5.8	0.8	12.5	13.3	74.2	7.5	0.8	91.7	0.0	0.0	100.0
		95.0	4.2	0.8	64.2	35.8	0.0	39.2	6.7	54.2	0.8	0.8	98.3
		97.5	2.5	0.0	4.2	9.2	86.7	0.8	0.0	99.2	0.0	0.8	99.2
		99.2	0.8	0.0	2.5	9.2	88.3	0.0	0.0	100.0	0.0	0.0	100.0
t	t	0.8	98.3	0.8	0.0	100.0	0.0	0.8	99.2	0.0	0.0	100.0	0.0
		2.5	97.5	0.0	0.0	100.0	0.0	0.0	100.0	0.0	3.3	96.7	0.0
		35.0	53.3	11.7	0.0	85.0	15.0	0.0	100.0	0.0	0.0	100.0	0.0
		23.3	73.3	3.3	0.0	93.3	6.7	0.0	98.3	1.7	0.0	100.0	0.0
-	t	26.7	64.2	9.2	52.5	43.3	4.2	38.3	57.5	4.2	15.0	80.8	4.2
		20.8	67.5	11.7	45.0	41.7	13.3	29.2	65.8	5.0	18.3	79.2	2.5
		55.8	32.5	11.7	21.7	75.0	3.3	41.7	55.8	2.5	0.0	100.0	0.0
		36.7	51.7	11.7	16.7	83.3	0.0	68.3	28.3	3.3	0.0	100.0	0.0
p	t	6.7	92.5	0.8	65.8	34.2	0.0	29.2	35.0	35.8	50.0	49.2	0.8
		10.8	88.3	0.8	52.5	45.8	1.7	51.7	46.7	1.7	41.7	58.3	0.0
		60.8	30.0	9.2	54.2	40.8	5.0	29.2	45.0	25.8	10.8	25.0	64.2
		12.5	72.5	15.0	34.2	65.8	0.0	80.0	5.8	14.2	10.8	81.7	7.5
k	t	0.8	25.8	73.3	0.0	15.0	85.0	0.0	29.2	70.8	0.8	37.5	61.7
		0.0	74.2	25.8	0.0	37.5	62.5	4.2	59.2	36.7	0.8	38.3	60.8
		10.8	28.3	60.8	0.8	3.3	95.8	0.0	3.3	96.7	0.8	10.8	88.3
		4.2	56.7	39.2	1.7	11.7	86.7	0.0	0.8	99.2	0.0	13.3	86.7
k	k	4.2	0.0	95.8	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0
		0.8	0.0	99.2	0.0	0.0	100.0	0.0	0.0	100.0	0.8	0.0	99.2
		0.8	0.0	99.2	0.0	0.8	99.2	0.8	0.0	99.2	0.0	0.0	100.0
		0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0
-	k	25.8	4.2	70.0	43.3	14.2	42.5	13.3	0.0	86.7	90.0	0.8	9.2
		10.0	5.8	84.2	16.7	10.8	72.5	23.3	1.7	75.0	71.7	2.5	25.8
		0.0	0.8	99.2	39.2	30.8	30.0	44.2	37.5	18.3	95.0	0.0	5.0
		0.8	1.7	97.5	33.3	39.2	27.5	81.7	5.0	13.3	99.2	0.0	0.8
p	k	31.7	3.3	65.0	15.8	4.2	80.0	20.0	0.0	80.0	99.2	0.8	0.0
		1.7	1.7	96.7	12.5	1.7	85.8	49.2	0.0	50.8	100.0	0.0	0.0
		0.8	0.8	98.3	45.0	14.2	40.8	29.2	5.8	65.0	96.7	0.0	3.3
		0.0	0.8	99.2	48.3	31.7	20.0	60.0	3.3	36.7	95.8	0.8	3.3
t	k	8.3	20.8	70.8	0.8	40.0	59.2	4.2	77.5	18.3	95.8	4.2	0.0
		4.2	7.5	88.3	0.0	24.2	75.8	0.0	100.0	0.0	98.3	1.7	0.0
		0.0	3.3	96.7	0.0	65.0	35.0	0.8	97.5	1.7	72.5	27.5	0.0
		0.0	0.8	99.2	0.0	46.7	53.3	0.0	95.0	5.0	87.5	12.5	0.0
p	-	82.5	11.7	5.8	79.2	15.0	5.8	71.7	6.7	21.7	90.8	5.8	3.3
		83.3	7.5	9.2	90.0	7.5	2.5	77.5	15.0	7.5	89.2	9.2	1.7
		93.3	5.8	0.8	96.7	1.7	1.7	36.7	19.2	44.2	55.0	3.3	41.7
		86.7	9.2	4.2	91.7	5.8	2.5	73.3	15.0	11.7	81.7	5.8	12.5
t		43.3	22.5	34.2	19.2	48.3	32.5	2.5	83.3	14.2	44.2	39.2	16.7
		60.8	28.3	10.8	29.2	44.2	26.7	1.7	89.2	9.2	50.8	45.0	4.2
		72.5	5.8	21.7	16.7	60.8	22.5	0.8	74.2	25.0	10.8	60.0	29.2
		77.5	10.8	11.7	10.0	54.2	35.8	0.8	65.8	33.3	7.5	61.7	30.8
k		0.8	1.7	97.5	0.8	4.2	95.0	1.7	1.7	96.7	2.5	0.8	96.7
		0.8	0.8	98.3	0.0	20.8	79.2	0.8	2.5	96.7	4.2	0.8	95.0
		1.7	2.5	95.8	5.0	25.8	69.2	0.0	4.2	95.8	3.3	0.0	96.7
		1.7	5.8	92.5	3.3	36.7	60.0	0.0	4.2	95.8	0.0	3.3	96.7

Table 5.II: (*previous page*) Stimulus-response matrix for the unvoiced stops. Results are pooled across subjects and are given in percentages. 100% is equal to 120 judgements. The identities of the burst portion and transition portion are indicated in the columns marked "BU" and "TR", respectively. The first eight groups of four rows give the results for the original, no-burst, and burst-spliced stimuli. The bottom three groups of four rows give the results for the burst-only stimuli.

nificantly dependent on various factors. To this end we used generalized linear modeling (GLIM, e.g. Aitkin *et al.*, 1989)[3] with factors BUR (consonant place-of-articulation of the burst), TRA (consonant place-of-articulation of the transitions), VOW (vowel identity), SPK (speaker), and LST (listener). The only interactions which were included in the model were those which seemed phonetically relevant, which are BUR×TRA, BUR×VOW, TRA×VOW, and BUR×SPK, TRA×SPK, VOW×SPK. Separate GLIM-analyses were carried out for dependent variables NB and NT, which indicate the number of responses corresponding to the burst and to the transitions, respectively. As the listener was used as a factor, the analyses were of course not carried out on the pooled data in Table 5.II but on the raw data which were not pooled across subjects.

The results of the analyses showed that all main effects and interactions were highly significant ($p < 0.001$), except the main effect TRA ($p = 0.35$) and the interaction TRA×SPK ($p = 0.55$) for the dependent variable NB, and the interaction TRA×SPK ($p = 0.31$) for the dependent variable NT. The interaction BUR×SPK for dependent NT was only marginally significant ($p = 0.05$). The Bonferroni[4] test for multiple post-hoc comparisons of means gave the following results.

1. The /k/-burst is significantly stronger in cueing place of articulation than the /p/ and /t/-burst. The /t/-transitions are significantly weaker in cueing place of articulation than the /p/ and /k/-transitions.
2. The burst is most effective before vowel /y/, less before vowels /i/ and /u/, and least effective before vowel /a/. The effectiveness of the transitions is highest before /a/, second highest before /i/, then /u/, and lowest before /y/.
3. Speaker 2 has significantly more effective bursts and less effective transitions than speaker 1.
4. Although LST was a highly significant factor in both the analysis for NB and for NT, the post-hoc grouping of subjects resulted in 3 highly overlapping groups. As there was no clear clustering of subjects into more or less disjunct groups, we judged that pooling across subjects before performing the model fits in the next chapter was reasonable.

The mixed-burst data are recapitulated in Table 5.III. The left column in Table 5.III describes the stimulus, and the three columns on the right-hand side present

[3] A regular ANOVA could not be used because the mixed-burst data contained empty cells for certain factor combinations, namely for equal burst and transition identity.

[4] The Bonferroni criterion was used because for some factors, like LST, a large number of comparisons was necessary (e.g. Hays, 1994).

the percentage of responses corresponding to either the burst, the transitions, or the remaining category (neither in accordance with the burst nor the transitions). This type of response was predominantly given for the stimuli with the burst of /tu/ and the voiced part of /ku/ (/p/-response). Part of the data of Table 5.III are presented visually in Figure 5.3. The bars in Figure 5.3a represent the proportion of responses corresponding with the burst, given the consonant identity of the transitions (first 3 bars), the consonant identity of the burst (second 3 bars) or the vowel identity (last 4 bars). The dotted line indicates the average proportion of responses corresponding to the burst (49%). Non-significant differences, as determined by the post-hoc Bonferroni tests, are indicated by the "staples" above the bars. The same holds for Figure 5.3b, but here for the proportion of responses corresponding with the transitions. Note that the /k/-burst is relatively strong in cueing place of articulation, and that in the /a/-context transitions are much more effective than in the other contexts.

Table 5.III: Average probabilities, presented in percentages, of responding according to the burst, transitions or remaining category for the mixed-burst stimuli of the unvoiced stops. For a further explanation see text.

stimulus	response		
	burst	trans	other
trans = /p/	52.5	44.4	3.1
trans = /t/	54.1	39.4	6.5
trans = /k/	41.5	44.6	13.9
burst = /p/	40.8	51.3	7.9
burst = /t/	40.9	47.1	12.0
burst = /k/	66.4	30.0	3.6
vowel = /a/	15.0	81.9	3.1
vowel = /i/	57.0	37.1	5.9
vowel = /y/	73.2	22.5	4.3
vowel = /u/	52.1	29.8	18.1
total	49.4	42.8	7.8

5.4.2 Voiced stops

In Table 5.IV, the stimulus-response matrix is presented for the experiments with the syllables containing the voiced stops /b,d/. Again, the responses are pooled across subjects. The results are presented in percentages, with 100% equal to 120 judgements. As subjects had 2 response alternatives, chance level corresponds to 50% correct. The upper two groups of four rows give the results for the original and no-burst stimuli created from the utterances starting with /b/. The results for the stimuli consisting of a /d/-burst spliced onto the syllable starting with /b/ are presented in the third group of four rows. The lower eight rows give the results for the burst-only stimuli.

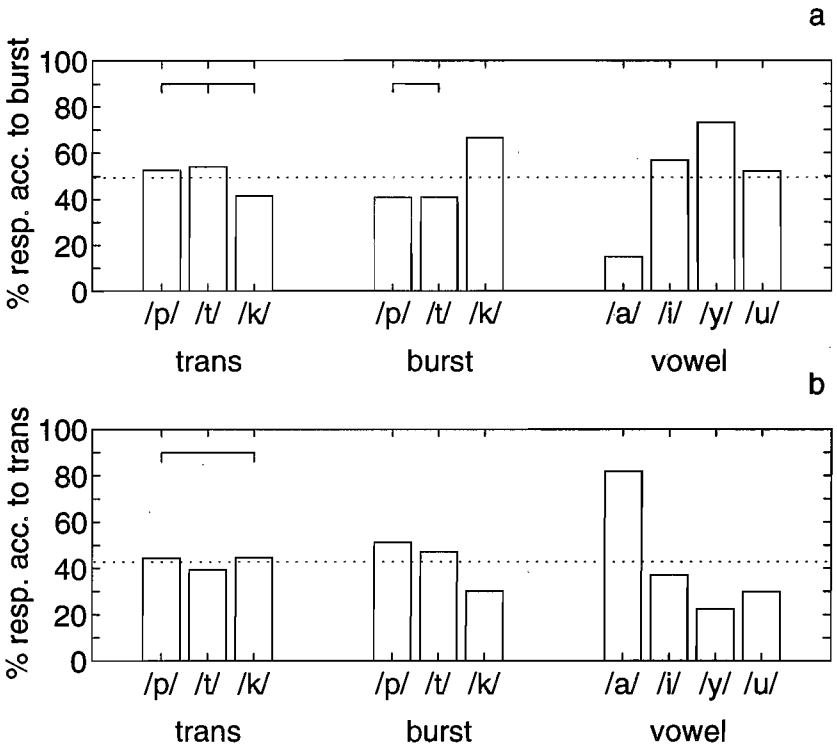

Figure 5.3: Figure 5.3a presents the percentage of responses to the (unvoiced) mixed-burst stimuli corresponding to the burst, given the stop from which the transitions originate (first bar: transitions from /p/, second bar: transitions from /t/, third bar, transitions from /k/), the identity of the burst (second group of 3 bars) or the vowel identity (last group of 4 bars). The dotted line indicates the average proportion of responses corresponding to the burst (49%). Percentages which are not significantly different, as determined by the post-hoc Bonferroni tests, are indicated by the "staples" above the bars. Figure 5.3b presents the percentage of responses corresponding with the transitions for the same parameters.

All original stimuli are correctly recognized at a rate higher than 95%, except the third token of /bi/ (90.8%), the fourth token of /dy/ (84.2%) and the second token of /du/ (90.8%). Again, these less than perfect scores could be traced to a slightly less clear pronunciation.

The average rate of correct recognition of the *burst-less stimuli* is 91.9% (96.4% for /b/, 87.4% for /d/), which is higher than results reported in the literature. This may be the result of our particular method of creating the burst-less stimuli, which leaves the part of the burst below 500 Hz largely intact. In agreement with other studies, recognition of place of articulation from burst-less stimuli seems to

be better for voiced stops than for unvoiced stops. As for the unvoiced stops, we find the preference for labial responses.

The average rate of correct recognition of the *burst-only* stimuli is 61.1%, which is only slightly above chance level (50%). This is much lower than the rates for the unvoiced stops. It is an often-reported finding that bursts isolated from unvoiced stops are more powerful in cueing place of articulation than bursts isolated from voiced stops. Note that this figure is based only on the perception of labial and dental bursts, while for the unvoiced stops also velars are included. As velar bursts generally cue place of articulation more strongly than labial or dental bursts, the figure for the voiced stops is relatively deflated. However, the particular rate of 61.1% is still lower than the usually reported percent correct rates for voiced stops

Table 5.IV: Stimulus-response matrix for the voiced stops. Results are pooled across subjects and are given in percentages. 100% is equal to 120 judgements. The identities of the burst portion and transition portion are indicated in the columns marked "BU" and "TR", respectively. The first six groups of four rows give the results for the original, no-burst, and burst-spliced stimuli. The bottom two rows give the results for the burst-only stimuli.

		vowel /a/		vowel /i/		vowel /y/		vowel /u/	
BU	TR	b	d	b	d	b	d	b	d
b	b	100.0	0.0	98.3	1.7	100.0	0.0	100.0	0.0
		99.2	0.8	100.0	0.0	97.5	2.5	100.0	0.0
		100.0	0.0	90.8	9.2	98.3	1.7	100.0	0.0
		100.0	0.0	96.7	3.3	100.0	0.0	100.0	0.0
-	b	100.0	0.0	96.7	3.3	100.0	0.0	100.0	0.0
		100.0	0.0	100.0	0.0	98.3	1.7	100.0	0.0
		99.2	0.8	80.0	20.0	96.7	3.3	100.0	0.0
		100.0	0.0	0.0	100.0	94.2	5.8	100.0	0.0
d	b	100.0	0.0	0.0	100.0	94.2	5.8	100.0	0.0
		87.5	12.5	7.5	92.5	41.7	58.3	100.0	0.0
		100.0	0.0	0.8	99.2	12.5	87.5	99.2	0.8
		98.3	1.7	8.3	91.7	27.5	72.5	100.0	0.0
d	d	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
		0.0	100.0	0.0	100.0	1.7	98.3	9.2	90.8
		0.0	100.0	0.0	100.0	0.8	99.2	0.0	100.0
		0.0	100.0	0.8	99.2	15.8	84.2	0.8	99.2
-	d	0.0	100.0	0.8	99.2	15.8	84.2	0.8	99.2
		10.8	89.2	50.8	49.2	36.7	63.3	58.3	41.7
		0.0	100.0	4.2	95.8	9.2	90.8	7.5	92.5
		0.0	100.0	27.5	72.5	1.7	98.3	3.3	96.7
b	d	0.0	100.0	27.5	72.5	1.7	98.3	3.3	96.7
		0.8	99.2	58.3	41.7	10.0	90.0	40.8	59.2
		0.8	99.2	39.2	60.8	13.3	86.7	5.8	94.2
		0.0	100.0	2.5	97.5	16.7	83.3	1.7	98.3
b	-	59.2	40.8	65.0	35.0	55.0	45.0	62.5	37.5
		50.0	50.0	50.8	49.2	35.0	65.0	69.2	30.8
		70.8	29.2	53.3	46.7	65.8	34.2	61.7	38.3
		75.0	25.0	38.3	61.7	76.7	23.3	70.8	29.2
d		35.8	64.2	44.2	55.8	34.2	65.8	32.5	67.5
		37.5	62.5	31.7	68.3	27.5	72.5	35.8	64.2
		60.0	40.0	29.2	70.8	30.0	70.0	32.5	67.5
		63.3	36.7	35.0	65.0	40.8	59.2	34.2	65.8

(e.g. Schouten and Pols, 1983). A rather obvious reason for this discrepancy is that, due to our procedure for isolating the bursts from the voiced stops, the bursts contain no energy below 500 Hz. This may bring about a certain unnaturalness in the bursts when presented in isolation, which may negatively affect recognition. The burst-only data should therefore be treated with caution.

For the *mixed-burst* stimuli, the overall proportion of responses corresponding to the burst was 26%, and with the transitions 74%. Like for the unvoiced stops, we performed a GLIM analysis on the response data for the mixed-burst stimuli. The analysis was slightly different, however. Because for the voiced stops only two responses were possible ("B" or "D"), the number of responses corresponding to either burst or transitions summed to a constant value. Moreover, the effect of the /b/-burst is indistinguishable from the effect of the /d/-transitions - and vice versa - because they always co-occur. Therefore we only made a GLIM-analysis for the dependent variable NB, and TRA was not used as a factor. Thus, the following main effects and interactions were used in the model: BUR, VOW, SPK, LST, and BUR×VOW, BUR×SPK, VOW×SPK.

The results of the analysis showed that all main effects and interactions were highly significant ($p < 0.001$), except the main effects SPK ($p = 0.47$) and LST ($p = 0.79$). The Bonferroni test for post-hoc comparison of means revealed that

1. The /d/-burst and the /b/-transitions were significantly stronger in cueing place of articulation than the /b/-burst and /d/-transitions.
2. The burst is most effective before vowel /i/, less effective before /y/, and least effective before /a/ and /u/. The reverse holds for the transitions.

The mixed-burst data are recapitulated in Table 5.V. The left column in Table 5.V again describes the stimulus, and the two columns on the right-hand side present the percentage of responses corresponding to either the burst or the transitions. Part of the data of Table 5.V are presented visually in Figure 5.4. The bars in Figure 5.4 represent the proportion of responses corresponding with the burst, given the consonant identity of the transitions (first 2 bars), or the vowel identity

Table 5.V: Average probabilities, presented in percentages, of responding according to the burst or transitions for the mixed-burst stimuli of the voiced stops. For a further explanation see text.

stimulus	response	
	burst	trans
trans = /b/ & burst = /d/	38.9	61.1
trans = /d/ & burst = /b/	13.9	86.1
vowel = /a/	2.0	98.0
vowel = /i/	63.9	36.1
vowel = /y/	33.2	66.8
vowel = /u/	6.6	93.4
total	26.4	73.6

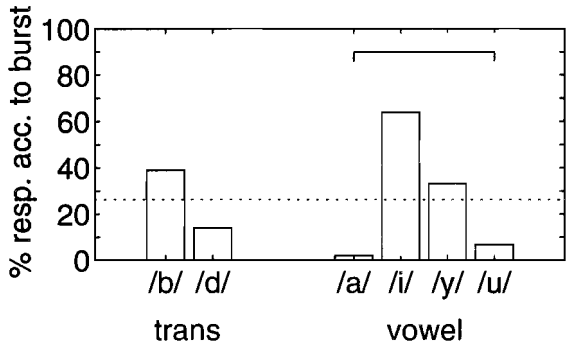

Figure 5.4: Percentage of responses to the (voiced) mixed-burst stimuli corresponding to the burst, given the identity of the transitions (first 2 bars), or the vowel identity (last 4 bars). The dotted line indicates the average proportion of responses corresponding to the burst (26%). Non-significant differences, as determined by the post-hoc Bonferroni tests, are indicated by the "staples" above the bars.

(last 4 bars). The dotted line again indicates the average proportion of responses corresponding to the burst (26%). The proportion of responses in accordance with the transitions are not plotted, as they are simply equal to 100% minus the percentage of responses in accordance with the burst. Note the exceptionally large vowel effect: in the front vowel contexts the burst is much more effective than in the other contexts. Clearly this is not an effect which is restricted to the velar category.

5.5 General discussion and conclusions

The results of the perception experiment show a complex picture of the relative importance of release burst and formant transitions for the perception of place of articulation of initial prevocalic stop consonants. The perceptual importance of burst and transitions highly depends on (1) whether the burst is labial, dental or velar, (2) whether the transitions are labial, dental or velar, (3) the vowel context, and, for the unvoiced stops, (4) the speaker. Furthermore, the results show that the bursts of unvoiced stops are more effective in cueing place of articulation than the bursts of their voiced counterparts. We will discuss these dependencies in more detail below.

As described in an earlier section, it has been reported in previous studies that the perceptual relevance of the release burst depends on the place of articulation of the stop consonant. In particular, it has been found that the velar burst is a more effective cue than either a labial or a dental/alveolar one. Our results for the burst-only as well as burst-spliced stimuli are in complete agreement with these findings. Indeed, we found, first of all, that /k/-bursts, when presented in isolation, are al-

most perfectly recognized, while bursts of /p/ and /t/ were much less accurately recognized. Secondly, the perceptual data for the mixed-burst stimuli showed that the /k/-bursts were significantly more effective than /p/ and /t/-bursts.

Also in accordance with earlier studies, we found that, for burst-less stimuli, the /k/-transitions are the least effective cues, while /p/-transitions give rise to almost perfect recognition. This observation may prompt the conclusion that /p/-transitions are strong or robust place cues, while /k/-transitions are weak place cues. This conclusion would, however, be at variance with the mixed-burst data, for which the /p/ and /k/-transitions are not significantly different in their effectiveness in cueing place of articulation, while both are significantly more effective than /t/-transitions. We offer the following explanation for this seeming contradiction. The physical removal of the burst from prevocalic stops does not bring about the intended deletion of burst *cues*, but rather suggests a very weak release burst. In terms of the gross cues of Stevens and Blumstein (1978), a strongly positive spectral tilt at signal onset, is suggested, which will cue a /p/-response, unless other information is present which indicates a different place of articulation. In other words, the removal of the burst does not bring about a deletion but rather a *modification* of the "burst cues", as was suggested earlier by Pols (1979). The strategy of replacing the release burst with a 300ms burst of noise, as employed by Pols and Schouten (1978), Pols (1979), and Van Wieringen (1995), seems an adequate method to avoid this problem. Their data show that, when the noise burst is used, the proportion of correct /p/-responses decreases while the proportion of correct /t/ and /k/-responses increases dramatically, compared to the condition where the burst is replaced by silence.
For similar reasons, we argue that splicing the release burst onto the stationary formants, as performed by Dorman *et al.* (1977), does not result in an actual deletion of transition cues. Instead, the new cue "straight formant transitions" is introduced, which in a back-vowel context can be strong evidence for a labial consonant. This may partially explain the strong bias in favor of formant transitions which is observed by some authors, like Dorman *et al.* (1977). As a consequence, we like to argue that the conflicting-cue paradigm is more reliable for measuring the relative contributions of acoustic cues than the cue-deletion paradigm.

Our study shows that the relative perceptual importance of burst and transitions is highly dependent on the vowel context. In general the release burst dominates place perception in front vowel contexts, while for non-front vowels the transitions are most important, although exceptions exist, like the /pu/-/ku/ contrast, which is almost entirely established by the burst. These findings are in agreement with Fischer-Jørgensen (1972) and Pols (1979), and in partial agreement with Dorman *et al.* (1977) and Van Wieringen (1995). Although the vowel-dependence seems to be very robust and important, it appears to be rather underexposed in studies investigating the relative importance of burst and transitions. In a large number of such papers, either a single vowel context is tested (e.g. Hoffman, 1958; Ohde and Stevens, 1983; Lindholm *et al.*, 1988) or results are pooled across vowels (e.g.

Pols and Schouten, 1978; Ohde and Scharf, 1981; Schouten and Pols, 1983), thus ignoring or obliterating this interesting phenomenon.

A qualitative explanation for the influence of the vowel context on the perceptual importance of burst and transitions is suggested by Fischer-Jørgensen (1972). She reasoned that, because the production of stop consonants differs for different vowel contexts, the acoustics of the stops show that in one vowel context the formant tracks for the 3 places of articulation are very different and the bursts are similar, like in the /a/-context, while in an other context, like /i/, the formant tracks are similar for the 3 places of articulation and the bursts are very distinct. Fischer-Jørgensen hypothesizes that the perceptual system has learned to focus on the most distinctive properties of the signal and therefore weighs the burst and transition cues differently depending on the vowel context.

The results of our experiments have shown that the bursts of unvoiced stops are more effective in cueing place of articulation than the bursts of voiced stops. One may argue that this finding is an artifact, brought about by the difference in the signal manipulations for the voiced and unvoiced stops. As our signal manipulation procedure for deleting the burst only removes the burst energy above 500 Hz, a substantial part is left intact, which may result in an apparent reduction of the influence of the release burst. We think that this is not the reason, because the most relevant burst properties are generally thought to reside in frequencies above 500 Hz (e.g. Halle *et al.*, 1957; Jongman and Miller, 1991; Kewley-Port *et al.*, 1983). In fact our findings are in accordance with a number of earlier studies which show that the perceptual system more heavily depends on burst information for unvoiced stops than for voiced stops (e.g. Schouten and Pols, 1983; Fischer-Jørgensen, 1972).

In conclusion, we have found that the relative perceptual importance of burst and transitions highly depends on the stop consonant, the vowel context and whether the stop is voiced or unvoiced. Velar bursts are generally much stronger in cueing place of articulation than other bursts. The dental transitions appear to be weaker than labial or velar transitions. In front-vowel contexts the release burst dominates the perception of place of articulation, while in non-front vowel contexts the formant transitions are generally dominant. Finally, we have found that the bursts of unvoiced stops are perceptually more important than the bursts of voiced stops.

Chapter 6

Detailed versus gross cues for the perception of prevocalic stop consonants: Modeling and evaluation[1]

Abstract

The purpose of the study presented in this chapter is to evaluate whether detailed or gross time-frequency structures are more relevant for the perception of prevocalic stop consonants. To this end, first a perception experiment is carried out with "burst-spliced" stop-vowel utterances. This experiment is described in the previous chapter. The present chapter describes the second part of the investigation, that is, the simulation of the behavior of the listeners in the perception experiment. First, a number of detailed and gross cues are measured on the stimuli. Next, these cues are mapped onto the observed perceptual data using a formal model of human classification behavior. The results show that the detailed cues, such as formant transitions, give a better account of the perceptual data than the gross cues in all cases. The best-performing models are interpreted in terms of the acoustic boundaries which are associated with the perceived linguistic contrasts. These boundaries are highly interpretable linear functions of 5 or 6 acoustic cues, which give a quantitative description of the often-discussed "trade-off" relation between the various cues for perception of place of articulation in stop consonants.

6.1 Introduction

In this chapter we present the actual evaluation of the relevance of various sets of acoustic cues for the perception of initial prevocalic stop consonants. In particular we will compare the perceptual relevance of *detailed* versus *gross* spectro-temporal cues, using manipulated natural utterances. As indicated in the previous chapter, a three-step procedure is adopted. First, a perception experiment is carried out in which manipulated natural stop-vowel utterances are presented to listeners for classification. This experiment is described in the chapter 5. The second step is the measurement of a number of detailed and gross cues on the stimuli. The third step is the mapping of the acoustic cues onto the observed perceptual responses. The second and third step, which are intended as a formal simulation of the behavior of the subjects in our perception experiment, are presented in this chapter.

[1] Based on: Smits, R., Ten Bosch, L., and Collier, R. (**1995b**), "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants: II. Modeling and evaluation," submitted to J. Acoust. Soc. Am.

The chapter is structured as follows. In the next section, the methods are described that were used for measuring the various cues on the signals. In section 6.3, the classification model is discussed, along with the methods for training, testing and interpreting the model. Section 6.4 presents the results of the various model fits and an interpretation of the best-fitting models. Finally, the results are summarized and discussed in the last section.

6.2 Extraction of acoustic cues

In this section, our specific choices of the detailed and gross acoustic cues are motivated, and the details of the measurement procedures are given.

6.2.1 Detailed cues

Preparations

As described in the previous chapter, the original utterances used for the purpose of the perception experiment were all possible CV-combinations consisting of the Dutch stops /b,d,p,t,k/ followed by the Dutch vowels /a,i,y,u/, spoken by two male speakers. For all utterances, the instants of burst onset and offset were used, which were determined manually for the purpose of the burst-splicing, see chapter 5.

In this study, frequencies of spectral peaks and formants are transformed to an 'auditory' frequency axis. For this purpose, the 'Equivalent Rectangular Bandwidth' or ERB-scale was chosen (Glasberg and Moore, 1990), which is similar to the Bark scale. The ERB-scale is defined in such a way that each ERB corresponds to a constant distance along the basilar membrane. The ERB-rate e can be calculated from the frequency f using the following equation (Glasberg and Moore, 1990):

$$e = 21.4 \log(4370f + 1) \qquad (-1)$$

with f expressed in Hz.

Burst cues

Based on the results of previous studies, it was decided to measure the following 4 burst cues: the burst length, the frequency of the highest burst peak, the level of this peak and the total level of the burst.

The length of the release burst has been found to increase with increasingly backward place of articulation, that is, velar bursts are longer than coronal bursts, which in turn are longer than labial bursts (Fischer-Jørgensen, 1954; Winitz *et al.*, 1971; Fant, 1973; Zue, 1976; Dorman *et al.*, 1977; Tekieli and Cullinan, 1979; Crystal and House, 1988). We define the burst length l_b here as the interval between the instants of burst onset and offset. As described in chapter 5, it was carefully checked, by looking as well as listening to the speech signals, that none of the stop-vowel utterances contained aspiration, as is common in the Dutch language.

Acoustic analyses have shown that labial bursts have spectral peaks at low frequencies (below 1 kHz), while dental or alveolar bursts have high-frequency peaks

(above 3 kHz). Velar bursts, on the other hand, display a strong energy peak in the mid-frequency range (1 to 4 kHz), the position of which highly depends on the vowel context. In front-vowel context the energy peak is wide and lies in the F3-F4 region, while in non-front vowel contexts the peak is located near or slightly above the F2 at the onset of voicing (Fischer-Jørgensen, 1954; Halle *et al.*, 1957; Winitz *et al.*, 1971; Fant, 1973; Zue, 1976; Dorman *et al.*, 1977; Edwards, 1981; Repp and Lin, 1988; Keating and Lahiri, 1993; Keating *et al.*, 1994). Perception experiments with synthetic signals and burst-spliced natural utterances have shown that the spectral peaks of the burst are indeed perceptually relevant, and that they are evaluated in relation to the vowel context (Cooper *et al.*, 1952; Schatz, 1954; Hoffman, 1958; Ainsworth, 1968). For our measurement of the burst peak frequency F_{bp}, an FFT was calculated of the entire burst. From the resulting amplitude spectrum, the highest peak in the interval of 700 Hz to 5 kHz was picked. The frequency of the peak was converted to the ERB-scale, using Eq. –1.

The level of the burst peak L_{bp} was defined as

$$L_{bp} = 20\log(|\mathcal{F}(F_{bp})|) \tag{-2}$$

where $|\mathcal{F}(F_{bp})|$ is the spectral amplitude at frequency F_{bp}.

Finally, the total level of the burst L_b was measured. Labial bursts are generally weaker than coronal and velar bursts (Fischer-Jørgensen, 1954; Fant, 1973; Zue, 1976; Dorman *et al.*, 1977; Edwards, 1981; Repp and Lin, 1988). Ohde and Stevens (1983) have shown that this energy cue is indeed used in the perception of the labial-alveolar distinction. Our measurement procedure was as follows. In order to eliminate the contribution of artificial low-frequency components to the burst energy, only the energy above 100 Hz was used. The burst level was defined as:

$$L_b = 10\log(\sum_i |\mathcal{F}(f_i)|^2) \tag{-3}$$

where f_i is the discrete frequency, expressed in Hz, $|\mathcal{F}(f_i)|$ is the spectral amplitude at frequency f_i, and the summation is carried out across frequencies ranging from 100 Hz to 5 kHz.

Formant frequencies

Since the early days of speech perception research, formant transitions, especially F2 and F3, have been considered to be important carriers of information regarding place of articulation. Throughout this chapter, the second and third formant will be indicated by F2 and F3, respectively. The *frequencies* of F2 and F3 at onset and in the stationary part of the utterance will be indicated by $F2_o$, $F3_o$, and $F2_{st}$, $F3_{st}$, respectively. A number of perception studies showed that the frequencies of F2 and F3 are strong cues to place of articulation (e.g. Cooper *et al.*, 1952; Liberman *et al.*, 1954; Delattre *et al.*, 1955; Ainsworth, 1968). Beside the perception studies, a number of acoustic studies showed that the *individual* frequencies of F2 and F3 at voicing onset (or traced back to the instant of consonantal release) showed high variability, particularly with vowel context (Fischer-Jørgensen, 1954; Halle *et al.*, 1957; Öhman, 1966; Fant, 1973; Kewley-Port, 1982). It was, however,

reported by several authors that a *combination* of various formant measures could give reasonable to good clustering of stops according to place of articulation, e.g. F2 and F3 at voicing onset (Öhman, 1966; Fant, 1973), F2 at onset and in the vowel nucleus (Öhman, 1966; Sussman *et al.*, 1991), and F2 and F3 at voicing onset and F2 in the vowel nucleus (Öhman, 1966; Kewley-Port, 1982; Sussman, 1991).

In our study, the frequency of the second and third formant were measured at voicing onset and in the stationary part at the end of the utterance. The onset frequencies $F2_o$ and $F3_o$ were measured in accordance with the suggestions made in chapter 2. From all utterances, the first glottal period after the release was excised using a rectangular window. The first glottal period was defined as the interval starting at the last negative zero crossing before the first glottal pulse and ending at the last negative zero crossing before the second glottal pulse. From the windowed portion, an FFT was calculated. From the resulting amplitude spectrum, the F2 and F3 peaks were measured manually. In cases where multiple peaks were present, the peak was chosen that was part of a continuous formant track in the corresponding wideband spectrogram. In a few cases, the F3 was not significantly excited by the first glottal pulse. In these cases, $F3_o$ was measured from the second glottal period.

The formant frequencies in the stationary part of the utterance, $F2_{st}$ and $F3_{st}$, were measured manually from wideband spectrograms. Because only stop-vowel utterances were used in this study, all vowels ended in a clear stationary portion, and the measurement of the formant frequencies was unambiguous.

The formant frequencies at onset as well as in the stationary part were converted from Hz to ERB using Eq. -1.

Locus equations

In Sussman (1991) and Sussman *et al.* (1991) it was shown that prevocalic stop consonants can be very well classified using the concept of locus equations for F2. A locus equation for F2 is defined as

$$F2_o = k \cdot F2_{st} + c \qquad (-4)$$

where the slope k indicates the linear variation of F2 with the F2 of the following vowel, and c is a constant. A distinct pair (k, c) is associated with each stop consonant.

Sussman hypothesized that locus equations may play an important role in the perception of stops. According to his hypothesis, the listener has a locus equation for each of the stop consonants stored in memory. These locus equations will generally be speaker-dependent. From each new occurrence of a stop consonant, the onset frequency and frequency in the 'vowel nucleus' of F2 is estimated by the listener. The distance of the resulting point $(F2_{st}, F2_o)$ to the locus equations of each of the stop consonants is calculated. The perceived consonant will generally correspond to the locus equation which is closest to the point $(F2_{st}, F2_o)$ of the new utterance. For certain vowel contexts, especially high front vowels, the locus equations for the different places of articulation are generally very close together. In these cases $F3_o$ as well as certain burst properties are used as additional cues.

In order to test the role of locus equations in the perception of stops, locus equations were determined from our stimuli and for all stimuli the distance to each of the locus equations was calculated. The calculations were made as follows. First locus equations were computed separately for each speaker and separately for voiced stops and unvoiced stops. That is, for each of the stops, the pair (k, c) of Eq. –4 is determined from the measured values of $F2_o$ and $F2_{st}$, using linear regression. Separate locus equations were calculated for velars in front-vowel contexts /i,y/ and velars in non-front vowel contexts /a,u/, as suggested by Sussman *et al.* (1991). The resulting locus equations are listed in Table 6.XI of Appendix 6.A.

The distance of each utterance to each of the relevant locus equations is determined as follows. Suppose the F2 of an utterance has an onset and vowel frequency $F2_o$ and $F2_{st}$, respectively. For each locus equation, the point that is closest to $(F2_{st}, F2_o)$ is obtained by orthogonal projection of $(F2_{st}, F2_o)$ on the locus equation. Next, all coordinates are converted to ERB-rate, using Eq. –1, and the Euclidean distance in ERB between the points is calculated. Thus, for the unvoiced stops 3 distances are measured: the distance D_l to the labial locus equation, the distance D_d to the dental locus equation, the distance D_v to the velar locus equation. For the voiced stops, only D_l and D_d were determined.

6.2.2 Gross cues

In a series of acoustic and perceptual studies, Blumstein and Stevens hypothesized that gross spectral characteristics of the initial 20-odd ms after consonantal release contain context-independent information for place of articulation of stop consonants. Labial, coronal, and velar stops were claimed to have diffuse falling, diffuse rising, and compact characteristics (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980). Kewley-Port argued that dynamic, that is, time-dependent, information needed to be incorporated, and she added the dynamic cues "late onset of voicing" and "persistence of a mid-frequency peak over time" to the existing static spectral properties (Kewley-Port, 1983; Kewley-Port and Luce, 1984; Kewley-Port *et al.*, 1983). Lahiri *et al.* (1984) introduced an improved, dynamic, cue to the labial-coronal distinction. Roughly, the measure was based on the change in high-frequency energy (around 3.5 kHz), relative to the change in low-frequency energy (around 1.5 kHz) from burst to voicing onset.

Based on these studies, we decided to measure the following gross cues. In accordance with Blumstein and Stevens, we measured the global spectral tilt and compactness after release. Additionally, we measured the change of global spectral tilt and compactness during the first 50 ms after release. The change of spectral tilt was intended to capture the measure proposed by Lahiri *et al.* (1984), while the change of compactness was intended to indicate the persistence of the mid-frequency peak, as proposed in Kewley-Port (1983) and in Kewley-Port and Luce (1984). In addition we decided to measure the global spectral tilt and compactness in the stationary part of the utterance, as well as the frequency location of the mid-frequency peak immediately after release and in the stationary part.

Preparations

In contrast to some of the detailed cues, all gross cues were measured automatically. For the measurement of the gross cues, a Short-Time Fourier Transform (STFT, Rabiner and Schafer, 1978) was made of all stimuli, using a Hanning window with a total length of 20.0 ms. In total, 12 windows were used. The first window was centered at the instant of burst onset. For the no-burst stimuli, the first window was centered at the voice-onset marker, which was used as the cutting position for the separation of the burst and the rest of the utterance. The next 10 windows were positioned to the right of the first window, using a window shift of 5.0 ms. Thus, approximately 50 ms of the signal after burst onset was analyzed by the first 11 windows. The last window was centered at 200 ms after burst onset. This window was intended to capture gross spectral properties in the stationary vowel.

All windowed segments were padded with 312 zeros to a sequence of 512 samples. From each of these sequences an FFT was calculated, resulting in a Fourier spectrum containing 257 frequency points. Like in many previous studies on acoustic cues for the perception of stops (e.g. Kewley-Port and Luce, 1984, Krull, 1990), the spectra were converted to "auditory spectra". This was done in two steps. First the frequency axis was warped to an ERB-axis, second the spectral amplitudes were weighed for the local bandwidth.

The amplitude spectrum for each of the windows was converted from Hz to ERB using Eq. –1. In order to obtain ERB-spectra containing 257 equidistant points, linear interpolation was used. In effect, the Hz-spectrum was upsampled at the low-frequency end, and downsampled at the high-frequency end. The interval between the last-but-one and the last sample on the ERB-axis corresponds to approximately 63.6 Hz, which corresponds to a downsampling with a factor of 3.26. Generally, if no smoothing of the spectrum is applied beforehand, downsampling may cause aliasing. In our case, however, aliasing was avoided by zero-padding. As a Hanning window of 20 ms has an effective spectral resolving power of approximately 72 Hz (Harris, 1978), which is larger than the maximum frequency step of 63.6 Hz, aliasing was negligible.

From a spectrum that is obtained by warping of the frequency axis of a Fourier amplitude spectrum, a valid energy measurement cannot be made. A correction must be made by multiplying the individual spectral amplitudes by the square root of the local auditory filter width. This is shown in the following derivation.

Generally, the spectral energy E between frequency f_1 and f_2 is calculated as follows:

$$E = \int_{f_1}^{f_2} |\mathcal{F}(f)|^2 df \qquad (-5)$$

where $|\mathcal{F}(f)|$ is the amplitude of the Fourier spectrum at frequency f. Suppose $A(e)$ is a function that transforms ERB-rate e into linear frequency f ($A(e)$ is the inverse of Eq. –1) and define e_1 and e_2 by $f_1 = A(e_1)$ and $f_2 = A(e_2)$. Then:

$$E = \int_{A(e_1)}^{A(e_2)} |\mathcal{F}(A(e))|^2 dA(e) \qquad (-6)$$

$$= \int_{e_1}^{e_2} |(\mathcal{F} \cdot A)(e)|^2 \frac{dA(e)}{de} de \tag{-7}$$

$$= \int_{e_1}^{e_2} \{|(\mathcal{F} \cdot A)(e)| \sqrt{\frac{dA(e)}{de}}\}^2 de \tag{-8}$$

where $|(\mathcal{F} \cdot A)(e)|$ is the frequency-warped amplitude spectrum, and $dA(e)/de$ is the first derivative of $A(e)$ with respect to e, which is always positive. $dA(e)/de$ is equal to $df/dA^{inv}(f)$, which is equal to the reciprocal of the first derivative of Eq. –1 with respect to f, which is the auditory bandwidth as a function of f.

Thus we find that a warped spectrum $|\mathcal{F}^w(e)|$, which is valid for energy measurements, can be obtained by multiplying the warped spectrum $|(\mathcal{F} \cdot A)(e)|$ by the square root of the local bandwidth:

$$|\mathcal{F}^w(e)| = |(\mathcal{F} \cdot A)(e)| \cdot \sqrt{24.7(0.00437f + 1)} \tag{-9}$$

where f is the frequency in Hz corresponding with the ERB-rate e.

Figure 6.1 shows the 2-step transformation of the frequency spectrum into the ERB-spectrum for a spectrum of one of the utterances, plus the effect of smoothing of the ERB-spectrum (which is used for the measurement of the mid-frequency peak).

Global spectral tilt

For the purpose of this study, the spectral tilt at onset, the change of spectral tilt over time, and the spectral tilt in the stationary part were measured. As the measure of spectral tilt used by Lahiri *et al.* (1984) was very successful for separating labials from dentals and alveolars, our strategy for measuring the global spectral tilt was chosen to be similar to this measure. In essence, the spectral tilt in Lahiri *et al.* (1984) was obtained by drawing a straight line by hand through the F2 and F4 peaks in an LPC spectrum with a logarithmic amplitude axis. Our automatic procedure works as follows. A straight line is fitted through the spectra with an ERB-axis and a logarithmic amplitude axis. In order to exclude the F1 from contributing to the spectral tilt, while ensuring that the F2 is always used, only the spectrum points in the range of 650 Hz to 5000 Hz were used. Linear regression was used for the fit, and spectral peaks were emphasized by weighing each squared local error by the local linear amplitude value, which is always larger than zero. The squared error E^2 that was minimized in the linear regression was:

$$E^2 = \sum_i w_i E_i^2 = \sum_i |\mathcal{F}^w(e_i)| \cdot (20 \log |\mathcal{F}^w(e_i)| - a_0 - a_1 e_i)^2 \tag{-10}$$

where, w_i and E_i are the weight and the error at ERB-rate e_i, a_0 and a_1 are the y-intercept and slope of the regression line, and the summation is over the ERB-rate interval corresponding to the frequency interval of 650 Hz to 5000 Hz.

The results of the fitting procedure for the spectra of the 1st, 3rd, 5th and 7th frames of an utterance /pa/ are shown in Figure 6.2.

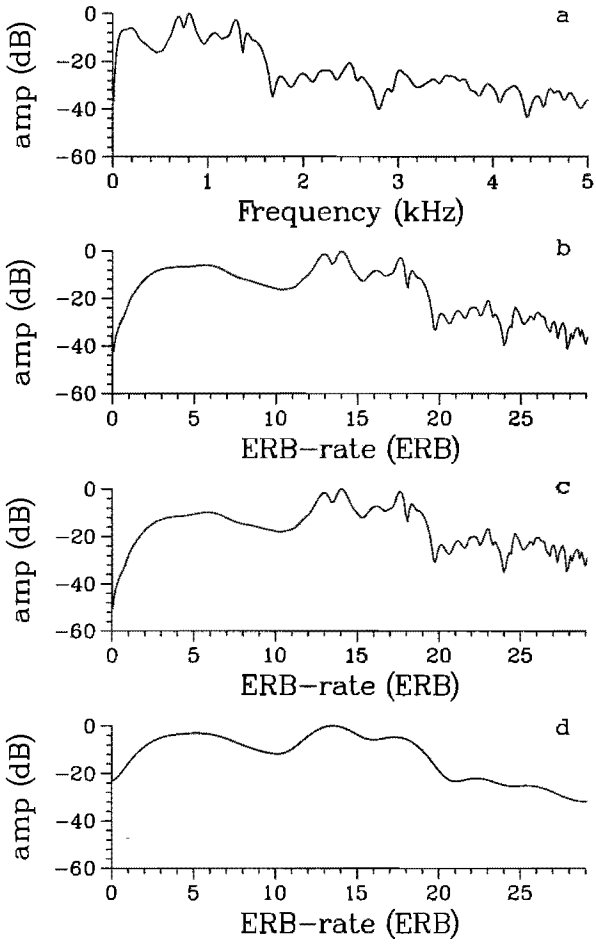

Figure 6.1: The transformation of the frequency spectrum into the ERB-spectrum, illustrated for a vowel spectrum of an utterance /ba/. Figures 6.1a, b, c and d show the original amplitude spectrum $|\mathcal{F}(f)|$, the frequency-warped amplitude spectrum $(|\mathcal{F} \cdot A)(e)|$, the final ERB-spectrum $|\mathcal{F}^w(e)|$, and $|\mathcal{F}^w(e)|$ after smoothing with a Hamming window with a total width of 4 ERB, respectively.

Each regression fit resulted in two numbers: the y-intercept, expressed in dB, and the slope, which is the actual spectral tilt, expressed in dB/ERB. Only the spectral tilt was used for further analysis.

The 12 values of spectral tilt for the 12 window positions were reduced to 3 numbers: the tilt for the first window (at burst onset), the tilt for the last window (in the stationary part) and the change of the tilt over the first 50 ms after the burst onset. For this change-of-tilt measure, a linear regression line was fitted through

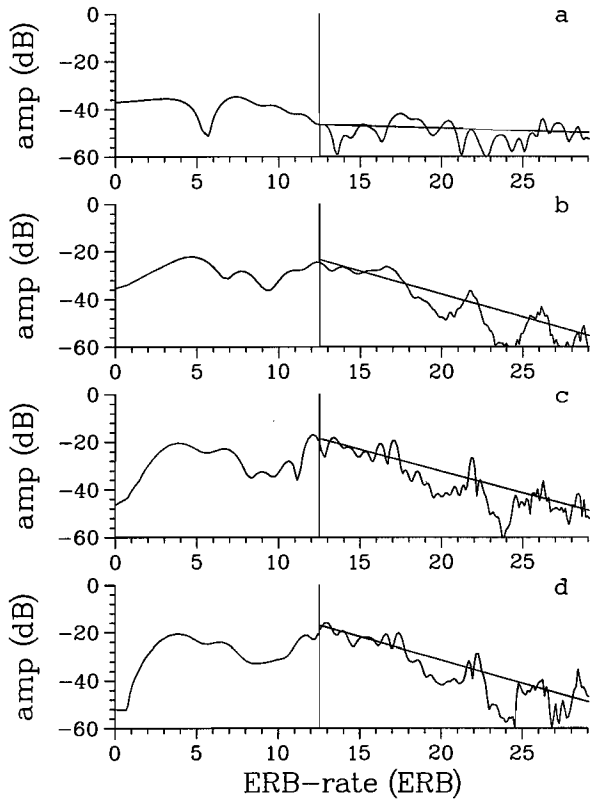

Figure 6.2: Illustration of the measurement procedure of the spectral tilt. Figures 6.2a, b, c, and d show the ERB-spectra of the 1st, 3rd, 5th and 7th frame of an utterance /pa/, together with the straight line segments that are fitted through the part of the ERB-spectrum between 650 Hz and 5 kHz. The vertical line indicates the ERB-rate corresponding to 650 Hz. Note the effect of emphasizing the spectral peaks in the linear regression.

the tilt values of the first 11 windows. The regression line was forced to contain the onset value of the tilt, so only one parameter value, the change of tilt, was determined. The squared error E^2 that was minimized in the linear regression was:

$$E^2 = \sum_i E_i^2 = \sum_{i=2}^{11} (T_i - (t_i \Delta T_o + T_o))^2 \qquad (-11)$$

Where t_i is the time interval between the center of window i and the instant of burst onset ($t_1 = 0$ ms, $t_2 = 5$ ms, etc.), T_i is the spectral tilt at time t_i, and $T_o = T_1 =$ the tilt for the first window. ΔT_o is the change of spectral tilt, expressed in dB/(ERB·ms), which is estimated in the minimization.

Thus, for each stimulus, we obtain 3 measures related to the global spectral tilt: the tilt at onset T_o, the change of spectral tilt immediately after onset ΔT_o, and the tilt T_{st} in the stationary part of the utterance.

Mid-frequency peak

For the purpose of this study, the intensity of the mid-frequency peak at onset L_o, the change of this intensity after onset ΔL_o, and the intensity of the mid-frequency peak in the stationary part L_{st} were measured. Also the frequency location of the mid-frequency peak at onset F_o and in the stationary part of the utterance F_{st} were determined. The procedure for locating the mid-frequency peak and measuring its intensity was aimed to be an automated version of the procedure followed by the judges inspecting the auditory running spectra in Kewley-Port and Luce (1984).

First, the auditory spectra $|\mathcal{F}^w(e_i)|$ were smoothed as follows. Each auditory spectrum was mirrored around $e = 0$ ERB and $e = 29.08$ ERB (5 kHz) and subsequently convolved with a Hamming window with a total width of 4 ERB. The figure of 4 ERB was arrived at through a manual optimization. Next, in the smoothed spectrum of the first window, the highest peak was picked in the mid-frequency range, that is, the range of ERB-rates corresponding to the frequency interval of 650 Hz to 3.5 kHz. Only true local maxima were considered, that is, spectral amplitudes that were locally maximal in their surroundings to the left as well as to the right. Thus, spectral amplitudes at the boundary of the mid-frequency region were never eligible for mid-frequency peak. The ERB-rate position F^{mfp} of the peak was determined and the intensity of the peak L^{mfp} was defined as the logarithm of the peak energy divided by the average mid-frequency energy:

$$L^{\mathrm{mfp}} = 10\log\left(\frac{|\mathcal{F}^w(F^{\mathrm{mfp}})|^2}{\frac{1}{c}\sum_i |\mathcal{F}^w(e_i)|^2}\right) \qquad (-12)$$

where the summation $\sum_i$ is made over the mid-frequency region, and c is the number of terms in the summation.

In order to ensure that only peaks were tracked that varied slowly in frequency, the onset peak was tracked within a narrow frequency window in subsequent spectra. Therefore, in the 2nd to 11th smoothed auditory spectrum, the maximum spectral amplitude was picked in a frequency region of width 2 ERB, centered around the peak position in the onset spectrum. For these spectra, the mid-frequency peak was allowed to be a spectral amplitude at the boundary of the 2-ERB region. For the 2nd to 11th spectrum, only the peak intensities were kept, the peak locations were discarded. Like for the onset spectrum, the peak intensities were calculated using Eq. –12. Finally, in the smoothed auditory spectrum for the stationary part, the peak position and peak intensity were determined in the same way as in the onset spectrum. So, the onset peak and the stationary peak could be farther apart than 1 ERB.

The results of the measurement procedure for the mid-frequency peak are shown in Figure 6.3 for the 1st, 3rd, 5th and 7th smoothed ERB-spectrum of an utterance /ky/.

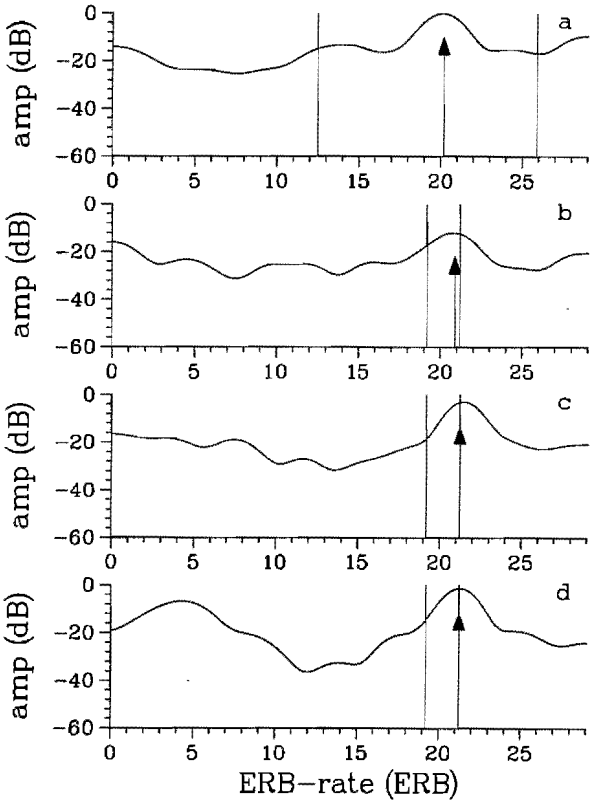

Figure 6.3: Illustration of the measurement procedure of the mid-frequency peak. Figures 6.3a, b, c, and d show the 1st, 3rd, 5th and 7th ERB-spectrum of an utterance /ky/, after smoothing. The vertical lines in Figure 6.3a show the mid-frequency region on the ERB-axis. The arrow indicates the position of the mid-frequency peak. The pair of straight lines in Figures 6.3b, c, and d indicate the 2-ERB region centered around the ERB-rate of the initial peak, and the arrows indicate the maximum value of the spectrum within this region. Note that the mid-frequency peak shifts slowly upward.

The measurements of the mid-frequency peak yield 12 peak intensities and 2 peak positions per stimulus. Using the 'clamped' linear regression which was identical to the calculation of the change of spectral tilt, the 12 intensity values were reduced to 3 numbers: the peak intensity for the first window, the peak intensity for the last window, and the change of the peak intensity over the first 50 ms after the burst onset.

6.2.3 Cues for the various stimuli

In summary, the following 19 cues were measured on the stimuli.
Detailed cues:

L_b : total level of the burst;
F_{bp} : frequency of the spectral peak of the burst;
L_{bp} : spectral level of the burst peak;
l_b : burst length;
$F2_o$: frequency of F2 at voicing onset;
$F2_{st}$: frequency of F2 in the stationary part of the utterance;
$F3_o$: frequency of F3 at voicing onset;
$F3_{st}$: frequency of F3 in the stationary part of the utterance;
D_l : distance to the labial locus equation;
D_d : distance to the dental locus equation;
D_v : distance to the velar locus equation.

Gross cues:

T_o : spectral tilt at onset;
ΔT_o : change of spectral tilt after onset;
T_{st} : spectral tilt in the stationary part of the utterance;
L_o : level of the mid-frequency peak at onset;
ΔL_o : change of the level of the mid-frequency peak after onset;
L_{st} : level of the mid-frequency peak in the stationary part of the utterance;
F_o^{mfp} : ERB-rate of mid-frequency peak at onset;
F_{st}^{mfp} : ERB-rate of mid-frequency peak in the stationary part of the utterance.

Four types of stimuli were used in the perception experiments: original utterances, no-burst stimuli, burst-only stimuli and mixed-burst stimuli. As the burst-only signals by definition do not contain voiced parts, only the detailed cues l_b, F_{bp}, L_{bp} and L_b, and the gross cues T_o, F_o and L_o were measured. For evident reasons, no burst cues were measured on the no-burst signals. On the original and mixed-burst stimuli, all cues are measured. For the voiced stops no velar locus equations were measured, so D_v was not used as a cue. The measured cue values for the stimuli containing the unvoiced stops are listed in Tables 6.XII to 6.XIX in Appendix 6.B. The measured cue values for the stimuli containing the voiced stops are listed in Tables 6.XX to 6.XXIII in Appendix 6.C.

6.3 The classification model

6.3.1 Introduction

The cue values that were measured on the stimuli were mapped onto the observed perceptual responses using the classification model described in chapters 3 and 4. In this section we will briefly recapitulate a number of basic issues concerning this model which are necessary to understand the modeling results.

In accordance with the general theory put forward by Ashby (1992), it is assumed that the underlying identification process which is performed by the subject can be subdivided into three intermediate steps: (1) extraction of stimulus features, (2) evaluation of class probabilities on the basis of stimulus features, (3) actual choice of response class on the basis of class probabilities. For a detailed discussion of these steps, see section 3.2.

6.3.2 The single-layer perceptron

The SLP is the core of the classification model. It performs a mapping of the stimulus cues onto the class probabilities. The number of input nodes of the SLP is equal to the number of stimulus cues. The number of output nodes equals the number of response classes. The SLP contains no hidden nodes.

We define the *class boundary* between two response classes C_i and C_j as the subspace of the cue space where the probability of choosing C_i equals the probability of choosing C_j. The *response region* of a response class C_i is defined as the subspace of the cue space where the most probable response class is C_i. As described in chapter 3, a basic assumption in the SLP is that the class boundaries are linear. No assumptions are made concerning the existence of category prototypes or underlying distributions of the acoustic cues.

Addition of a "hidden layer" to the perceptron may enhance the classification power of the model by enabling the use of non-linear boundaries. Beside the SLP-model fits we tried this enhanced model, but significant improvements in goodness of fit were seldom found. Therefore, only the results for the SLP-model are presented in this chapter.

6.3.3 Model estimation (training)

The term *training* or *estimation* of the model is defined here as the estimation of the model parameters, that is, the SLP's weights and biases, for which the goodness of fit (GOF) of the model is optimized. As measure of GOF we will use the "fuzzy" percent-correct score, henceforth GOF-fuzzy, which is derived from the multinomial function (see chapter 4, Eq. –4). In some cases GOF-fuzzy will be transformed into a *winner-takes-all* (WTA) score. For each stimulus it is determined whether the model's most likely response is equal to the actually most-occurring response, and the WTA-score expresses the overall proportion of stimuli for which this is the case.

In all model estimations it is important to be aware of the *chance level performance* of the model. Chance level is defined as the maximum GOF which can be obtained by a model with a fixed, stimulus-independent output (see chapter 4). Chance level for GOF-fuzzy increases with increasing response bias and with increasing confusion in the responses. GOF-WTA only increases with increasing response bias.

6.3.4 Model evaluation (testing).

In order to avoid over-fitting of the model, we tested the generalizability of all model fits using a formal cross-validation method. Although the perception experiments were extensive, the number of data that are available for the model fits is necessarily rather limited compared to e.g. the data bases used in automatic classification experiments, such as Forrest *et al.* (1988) and Nossair and Zahorian (1991). Therefore, we chose to employ the *leaving-one-out* (LOO) method for the cross-validation (see chapter 4). Although the LOO-method is computationally expensive, it makes maximally efficient use of the available data by using, in effect, each datum in training as well as in testing, while maintaining the independence of training set and test set.

6.3.5 Model interpretation

By means of interpreting the SLP's weights and biases in an appropriate way (see chapter 3) the class boundaries can be derived. The class boundaries are linear functions of the stimulus cues. The absolute value of the coefficient of each of the cues indicates the perceptual relevance of the cue, the sign of the coefficient gives the direction in which the cue works (e.g. whether a high cue value favors a /p/ or a /t/-response).

Apart from being relevant for a class distinction, cues may have the function of indicating troublesome areas in the cue space, that is, areas where either the SLP is unable to model the observed stimulus dependent response behavior, or the responses themselves have been rather stimulus-independent. The SLP-model is capable of "squeezing" the output probabilities to fixed levels in a certain cue subspace, by attenuating the influence of the cues on the output probabilities to zero (see chapter 4).

6.3.6 Modeling program

The comprehensive set of 19 cues was subdivided into 4 cue sets, which will be indicated by **fo** (formant plus burst cues), **le** (locus-equation plus burst cues), **gr** (gross cues), and **su** ("super set"):

fo: Detailed cues, comprising burst cues L_b, F_{bp}, L_{bp}, l_b, plus formant cues $F2_o$, $F2_{st}$, $F3_o$, $F3_{st}$.

le: Detailed cues, comprising burst cues, L_b, F_{bp}, L_{bp}, l_b, formant cue $F3_o$, and locus equation distances D_l, D_d, D_v. $F3_o$ is incorporated in accordance with suggestions by Sussman (1991).

gr: Gross cues T_o, ΔT_o, T_{st}, L_o, ΔL_o, L_{st}, F_o^{mfp}, F_{st}^{mfp}.

su: "Super set", containing a selection of burst cues, formant cues and gross cues which, after a number of preliminary model fits, gave the most promising results: L_b, l_b, $F2_o$, $F2_{st}$, $F3_o$, $F3_{st}$, T_o, ΔT_o, L_o, F_o^{mfp}.

For the purpose of the model fits, the data of the perception experiment were pooled across subjects, thus obtaining 120 responses per stimulus. Pooling was

allowed because data analysis showed that, although subject was a significant factor in the perception of the unvoiced stops, the subjects could only be clustered in 3 largely overlapping groups, as described in the previous chapter.

The first type of model fits that were performed were speaker-independent model fits on the original utterances, separately for unvoiced and voiced stops. The purpose of the fits was to make a comparison with previous automatic classification experiments. Next, the actual perception experiments were simulated. As indicated in Table 5.I in the previous chapter, the complete perception experiment consisted of 8 subsessions. In each of the first subsessions, subjects were presented with the original plus no-burst plus mixed-burst stimuli in random order, but separately per speaker and separately for voiced versus unvoiced stops. In the last 4 subsessions, subjects were presented with the burst-only stimuli, again separately per speaker and separately for voiced and unvoiced stops. In accordance with these experimental procedures, the perceptual data were subdivided into 8 stimulus response matrices, one for each subsession. For practical reasons, the set of stimuli consisting of the original plus no-burst plus mixed-burst stimuli will henceforth be indicated by the *combined* stimuli. Calculation of chance levels for GOF-fuzzy showed that the burst-only stimuli for the voiced stops were essentially labeled randomly by the subjects: chance levels were 97% and 94% for speaker 1 and 2, respectively. No model fits were therefore made on these data. On each of the remaining 6 stimulus-response matrices separate model fits were made.

The procedure for the model fits was as follows. On each stimulus-response matrix, except the burst-only data, models were trained for each possible combination of 4, 5, 6, and 7 cues within each of the cue sets **fo**, **le**, **gr**, and **su**. Next, for each number of cues, the three cue combinations with the highest GOF-fuzzy on training were tested using the LOO-procedure.[2] Subsequently, the cue set with the highest GOF on testing was chosen as the best cue subset, and the corresponding model obtained during training was chosen as the best model. Thus, for each of the stimulus-response matrices and for each of the cue sets **fo**, **le**, **gr**, and **su**, we obtained a best cue subset consisting of 4, 5, 6, and 7 cues. In all cases, the highest GOF-fuzzy on testing occurred for either 5 or 6 cues. These cue subsets plus corresponding model were finally selected as the overall best-performing cues and model.

The procedure for the burst-only data of the unvoiced stops was slightly different. First of all, instead of the cue sets **fo**, **le**, **gr**, and **su**, the sets **de**, **gr**, and **su** were constructed, which consisted of the following cues:

de: Detailed cues L_b, F_{bp}, L_{bp}, l_b.

gr: Gross cues T_o, L_o, F_o^{mfp}.

su: "Super set", containing all cues of **de** and **gr** combined.

The subsets trained and tested consisted of 1, 2, 3, or 4 cues. In all cases the subset of 2 cues provided the highest GOF-fuzzy on testing.

[2]The LOO-method could not be applied to all possible cue subsets because of computation costs.

Before the model fits were initiated, the measured values for each cue were normalized across all stimuli using Eq. 3.7. For the burst cues L_b, F_{bp}, L_{bp}, and l_b of the burst-less stimuli the value 0 was inserted. Although it seems that the burst cues thus receive the *average* value rather than *no* value for the no-burst stimuli, the actual effect of this manipulation is that, during training, the weights emanating from the burst-cue nodes are free to assume any value for the burst-less stimuli, because they are multiplied by zero and thus have no contribution to the GOF.

Fitting all the models according to the procedures described above will finally result in the set of "best" models and cue subsets listed in Table 6.I. The results of these 32 model fits will be discussed in the next section.

6.4 Results

6.4.1 Levels of goodness-of-fit

Original utterances

In order to assess the agreement with previous automatic classification studies, speaker-independent model fits were obtained for the original utterances. Note that, although the fuzzy GOF-measure was optimized, the WTA-levels are most interesting for the purpose of comparison.

Unvoiced stops

The subsets of each of the cue types that produced the highest GOF-fuzzy on testing are listed in Table 6.II, together with the respective GOF-levels on training and testing. The GOF-levels on testing are shown graphically in Figure 6.4. For each pair of bars for a particular cue set, the left bar represents GOF-fuzzy, the right bar represents GOF-WTA, respectively. The dashed and dotted lines indicate chance level for GOF-fuzzy and GOF-WTA. Chance levels are 38.8% for GOF-fuzzy

Table 6.I: Modeling program: The model simulations will yield "best" models for the listed sets of stimuli and cues.

stimulus set		speaker	cue sets
original utterances	unvoiced	1+2	**fo, le, gr, su**
	voiced	1+2	**fo, le, gr, su**
combined stimuli	unvoiced	1	**fo, le, gr, su**
	unvoiced	2	**fo, le, gr, su**
	voiced	1	**fo, le, gr, su**
	voiced	2	**fo, le, gr, su**
burst-only stimuli	unvoiced	1	**de, gr, su**
	unvoiced	2	**de, gr, su**

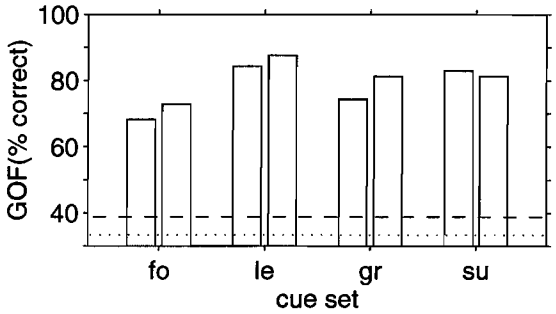

Figure 6.4: The GOF-levels on testing for each of the cue types **fo**, **le**, **gr**, and **su**, for the speaker-independent model fits on the *original* utterances containing the *unvoiced stops*. For each pair of bars for a particular cue set, the left and right bar represent GOF-fuzzy and GOF-WTA, respectively. The dashed and dotted lines indicate chance level for GOF-fuzzy (38.8%) and GOF-WTA (33.3%).

and 33.3% for GOF-WTA. The subscripts **fo**, **le**, **gr**, and **su** indicate the various cue sets.

Comparing the GOF-test for the **fo**, **le**, and **gr** sets, we find both for GOF-fuzzy and for GOF-WTA that the **le**-set provides the best fit (84% and 88% on testing for GOF-fuzzy and GOF-WTA, respectively), followed by **gr** (74.3% and 81.3% on testing), while **fo** gives the worst fit (68% and 73%, respectively). The **su**-set provides a GOF which is intermediate between **le** and **gr** (83% and 81% on testing). Note that for the best subset for **su**, the formant onset frequencies $F2_o$ and $F3_o$ are combined with the "Blumstein-and-Stevens" cues T_o and L_o and the burst level.

Although the WTA-levels are reasonable, they are far from perfect and clearly lower than some of the results reported in the literature (e.g. Forrest *et al.*, 1988; Nossair and Zahorian, 1991). The ordering of GOF-levels for the 3 cue types is,

Table 6.II: The best subsets of each of the cue types **fo**, **le**, **gr**, and **su**, for the speaker-independent model fits on the *original* utterances containing the *unvoiced stops*. The two columns on the right present the levels of GOF-fuzzy and GOF-WTA on training and testing for each of the cue subsets. The chance levels are 38.8% for GOF-fuzzy, and 33.3% for GOF-WTA.

cue type	cues used					GOF(fuzzy) train	GOF(fuzzy) test	GOF(WTA) train	GOF(WTA) test
fo	L_b	L_{bp}	l_b	$F2_o$	$F3_o$	85.1	68.2	87.5	72.9
le	F_{bp}	L_{bp}	$F3_o$	D_d	D_v	90.3	84.3	89.6	87.5
gr	T_o	F_o^{mfp}	F_{st}^{mfp}	L_o	ΔL_o	89.4	74.3	93.8	81.3
su	L_b	$F2_o$	$F3_o$	T_o	L_o	88.7	82.9	87.5	81.3

however, in agreement with previous findings. Both locus equations and gross cues have been reported to provide higher classification scores than raw formant data (e.g. Sussman *et al.*, 1991; Nossair and Zahorian, 1991). Apparently for our data the locus-equation data, in turn, perform slightly better than the gross cues.

The discrepancy of our WTA-levels with the data of e.g. Forrest *et al.* (1988) and Nossair and Zahorian (1991) can be attributed to the small size of our data set for the model fits on the original utterances: only 48 utterances were available. Both in Forrest *et al.* (1988) and Nossair and Zahorian (1991), large numbers of data were available, thus permitting a more precise model estimation and the use of a larger number of cues. Forrest *et al.* (1988) and Nossair and Zahorian (1991) used 12 and 20 cues, respectively, in the best-performing conditions, while we could use only 5 cues. It is important to remember that the purpose of our study is to study the *perceptual* rather than *machine* classification of stops, and the fits on the original utterances are only intended to allow for a crude validation.

Voiced stops

The cue sets that produce the highest GOF-levels for the voiced stops are listed in Table 6.III, together with the respective GOF-levels for training and testing. The GOF-levels on testing are plotted in Figure 6.5. Chance levels are 52.8% for GOF-fuzzy and 50.0% for GOF-WTA.

All 3 subsets of cues have a high GOF-level on testing. The **fo**-set produces the highest levels on testing (97% and 100% for GOF-fuzzy and GOF-WTA, respectively), followed by **gr** (91% and 97%), while **le** produces the worst (but still good) fit (88% and 97%). The best subset of the cue set **su** appeared to be identical to the best subset of **fo**. Finally, we note that the number of utterances used for these classifications is again low, namely 32.

Table 6.III: The best subsets of each of the cue types **fo**, **le**, **gr**, and **su**, plus the respective GOF-levels, for the speaker-independent model fits on the *original* utterances containing the *voiced stops*. The chance levels are 52.8% for GOF-fuzzy, and 50.0% for GOF-WTA.

cue type	cues used	GOF(fuzzy) train	GOF(fuzzy) test	GOF(WTA) train	GOF(WTA) test
fo	l_b $F2_o$ $F2_{st}$ $F3_o$ $F3_{st}$	99.5	96.9	100.0	100.0
le	L_b L_{bp} $F3_o$ D_l D_d	98.3	88.3	100.0	96.9
gr	T_o ΔT_o F_{st}^{mfp} L_o ΔL_o	98.9	91.3	100.0	96.9
su	l_b $F2_o$ $F2_{st}$ $F3_o$ $F3_{st}$	99.5	96.9	100.0	100.0

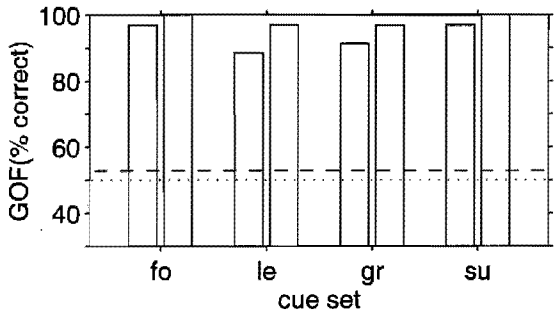

Figure 6.5: The GOF-levels on testing for the cue sets **fo**, **le**, **gr**, and **su**, for the speaker-independent model fits on the *original* utterances containing the *voiced stops*. Interpretation of the bars is as in Figure 6.4. Chance levels are 52.8% for GOF-fuzzy and 50.0% for GOF-WTA.

Combined stimuli

Unvoiced stops

The best cue subsets and their respective levels of GOF-fuzzy for the fits on the combined stimuli (original plus no-burst plus mixed-burst) are listed separately for speaker 1 and 2 in Table 6.IV. The GOF-levels on testing for both speakers are shown graphically in Figure 6.6. The left and right bar in each pair now represent GOF-fuzzy for speaker 1 and 2, respectively. As the levels of GOF-WTA are less relevant here, they are omitted. The dashed and dotted lines indicate chance level for GOF-fuzzy for the perception data for speaker 1 (chance level 52.2%) and

Table 6.IV: The best subsets of each of the cue types **fo**, **le**, **gr**, and **su**, plus the respective levels of GOF-fuzzy, for the speaker-dependent model fits on the *combined* stimuli containing the *unvoiced stops*. The chance levels are 52.2% for speaker 1, and 50.8% for speaker 2.

cue type	speaker	cues used						GOF train	GOF test
fo	1	L_b	L_{bp}	$F2_o$	$F2_{st}$	$F3_o$		75.4	68.4
fo	2	L_{bp}	$F2_o$	$F2_{st}$	$F3_o$	$F3_{st}$		77.5	70.2
le	1	L_b	L_{bp}	l_b	$F3_o$	D_l	D_d	82.0	75.2
le	2	F_{bp}	L_{bp}	D_l	D_d	D_v		77.9	70.7
gr	1	T_o	ΔT_o	T_{st}	F_{st}^{mfp}	L_o	L_{st}	73.3	61.4
gr	2	T_o	ΔT_o	T_{st}	F_{st}^{mfp}	L_o		74.5	66.7
su	1	L_b	$F2_o$	$F3_{st}$	T_o	F_o^{mfp}		74.4	69.5
su	2	l_b	$F2_o$	$F3_o$	F_o^{mfp}	L_o		81.6	75.0

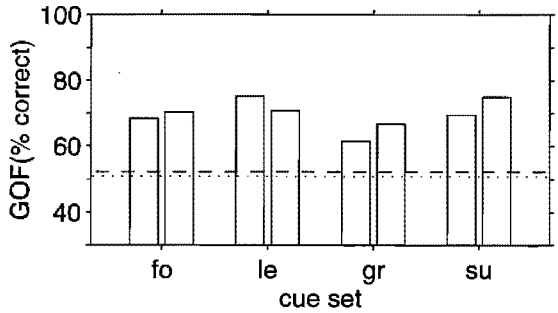

Figure 6.6: The levels of GOF-fuzzy on testing for the cue sets **fo**, **le**, **gr**, and **su**, for the speaker-dependent model fits on the *combined* stimuli containing the *unvoiced stops*. The left and right bar in each pair represent the GOF for speaker 1 and 2, respectively. The dashed and dotted lines indicate chance levels for GOF-fuzzy for speaker 1 (52.2%) and speaker 2 (50.8%).

speaker 2 (chance level 50.8%).

First of all, we note that, compared to the model fits for the original utterances only, the GOF-levels have generally decreased considerably, even while the chance levels have gone up. Secondly, it is evident that the pattern for the two speakers is very similar. Comparing the various cue sets, we find that for both speakers, the **le**-set provides the best account of the data (75% and 71% on testing for speaker 1 and 2, respectively), followed by the **fo**-set (68% and 70%). The gross cues (**gr**) give the worst account of the data (61% and 66%). Adding gross cues to the detailed cues based on raw formant information (**su**) only gives an improvement for speaker 2 (GOF-fuzzy is 75%).

The optimal cue subsets are very similar for the two speakers. For both speakers the subsets for **fo** contain the cues L_{bp}, $F2_o$, $F2_{st}$, and $F3_o$. The subsets for **le** both contain L_{bp}, D_l, and D_d; the subsets for **gr** contain T_o, ΔT_o, T_{st}, F_{st}^{mfp}, and L_o. The subsets for **su** are less similar, as they only both contain $F2_o$ and F_o^{mfp}.

Voiced stops

The best cue subsets and their respective levels of GOF-fuzzy are listed separately for speaker 1 and 2 in Table 6.V. The GOF-levels are shown graphically in Figure 6.7. Chance levels for speaker 1 and speaker 2 are 60.9% and 61.1%, respectively.

As was the case for the unvoiced stops, the GOF-levels are lower here than for the fits for the original utterances only. For both speakers GOF on testing is highest for the **fo**-set (89% for both speakers), closely followed by the **le**-set (88% for both speakers), and clearly lowest for the **gr**-set (79% and 78% for speaker 1 and 2, respectively). Only for speaker 1 did the **su**-set produce a GOF higher than the GOF for the **fo**-set, namely 91%.

The optimal subsets of cues are again rather similar for the two speakers. For

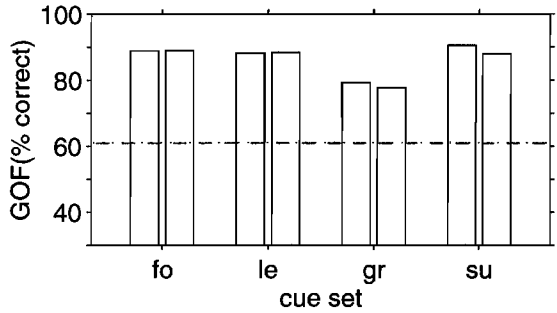

Figure 6.7: The levels of GOF-fuzzy on testing for the cue sets **fo**, **le**, **gr**, and **su**, for the speaker-dependent model fits on the *combined stimuli* containing the *voiced stops*. The left and right bar in each pair again represent the GOF for speaker 1 and 2, respectively. The dashed and dotted lines indicate chance levels for speaker 1 (60.9%) and speaker 2 (61.1%).

both speakers the optimal subsets for **fo** contain L_{bp}, $F2_o$, $F2_{st}$, and $F3_{st}$, and the subsets for **le** both contain L_b, L_{bp}, $F3_o$, and D_d. The optimal subsets for the **gr** and **su** sets are somewhat less similar.

Burst-only stimuli

As explained earlier, model fits on the burst-only data were only made for the unvoiced stops. The best cue subsets and the respective levels of GOF-fuzzy are listed separately for speaker 1 and 2 in Table 6.VI. The GOF-levels are shown graphically in Figure 6.8. Chance levels are 62.0% for speaker 1 and 62.1% for

Table 6.V: The best subsets of each of the cue types **fo**, **le**, **gr**, and **su**, plus the respective levels of GOF-fuzzy, for the speaker-dependent model fits on the *combined* stimuli containing the *voiced stops*. The chance levels are 60.9% for speaker 1, and 61.1% for speaker 2.

cue type	speaker	cues used						GOF train	GOF test
fo	1	L_b	L_{bp}	$F2_o$	$F2_{st}$	$F3_{st}$		93.2	88.9
fo	2	F_{bp}	L_{bp}	$F2_o$	$F2_{st}$	$F3_o$	$F3_{st}$	94.9	88.9
le	1	L_b	L_{bp}	$F3_o$	D_l	D_d		95.2	88.1
le	2	L_b	F_{bp}	L_{bp}	$F3_o$	D_d		95.2	88.3
gr	1	T_o	ΔT_o	F_o^{mfp}	ΔL_o	L_{st}		86.5	79.3
gr	2	T_o	ΔT_o	T_{st}	F_{st}^{mfp}	L_o		88.1	77.8
su	1	L_b	$F2_o$	$F2_{st}$	T_o	ΔT_o		95.2	90.5
su	2	$F2_o$	$F3_o$	$F3_{st}$	T_o	ΔT_o		95.0	88.0

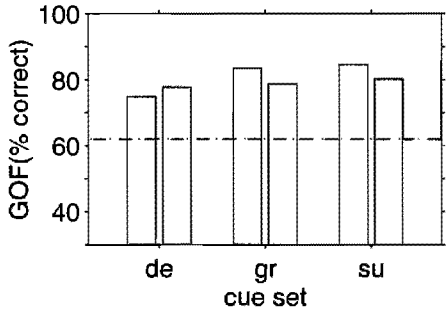

Figure 6.8: The levels of GOF-fuzzy on testing for the cue sets **de**, **gr**, and **su**, for the speaker-dependent model fits on the *burst-only* stimuli of the *unvoiced stops*. The left and right bar in each pair again represent the GOF for speaker 1 and 2, respectively. The dashed and dotted lines indicate chance levels for speaker 1 (62.0%) and speaker 2 (62.1%).

speaker 2.

In contrast with the previous fits, here the gross cues perform slightly better than the detailed cues for both speakers. GOF for the **de**-set is 75% and 78%, respectively, for speaker 1 and 2, and 84% and 79% for the **gr**-set. The combination of a detailed and a gross cue (**su**-set), however, produces the highest GOF-levels for both speakers (85% and 80%).

6.4.2 Interpretation of the classification models

In this section we will look in detail at some of the models that were trained on the data for the combined stimuli. We emphasize again that in our model no assump-

Table 6.VI: The best subsets of each of the cue types **fo**, **le**, **gr**, and **su**, plus the respective levels of GOF-fuzzy, for the speaker-dependent model fits on the *burst-only* stimuli of the *unvoiced stops*. The chance levels are 62.0% for speaker 1, and 62.1% for speaker 2.

cue type	speaker	cues used		GOF train	GOF test
de	1	L_b	l_b	85.3	74.8
de	2	L_{bp}	F_{bp}	86.1	77.7
gr	1	F_o^{mfp}	L_o	87.3	83.5
gr	2	T_o	L_o	84.5	78.7
su	1	L_b	L_o	88.7	84.6
su	2	L_{bp}	F_o^{mfp}	88.7	80.2

tion is made that listeners arrive at linguistic classification by comparing the incoming "idealized" category prototypes. Instead, our model can be viewed as being "boundary-based". Using a boundary-based classification model is a classification-theoretic expression of the fundamental axiom that linguistic communication is achieved by transmitting *distinctions* rather than *idealized symbols* (e.g. Jakobson *et al.*, 1952). Thus, the model's class boundaries can be interpreted as the acoustic correlates of linguistic distinctions. In the model interpretations, we will therefore restrict ourselves to describing the linear boundaries in the multi-dimensional cue space for each place-of-articulation contrast, and we will discuss the importance of the various cues for each of these contrasts.

Unvoiced stops

Speaker 1

For speaker 1, the highest GOF on testing was obtained for the le-set (75%). GOF on training for this model was 82% (see Table 6.IV). Interpretation of the weights and biases of the SLP leads to the following boundaries B_{p-t}, B_{p-k}, and B_{t-k} between the response regions for /p/ versus /t/, /p/ versus /k/, and /t/ versus /k/:

$$B_{p-t}: \quad -1.6L_b + 0.7L_{bp} - 0.6l_b - 1.8F3_o - 0.6D_l + 1.5D_d + 0.2 = 0 \qquad (-13)$$

$$B_{p-k}: \quad 2.2L_b - 3.2L_{bp} + 0.5l_b + 0.5F3_o - 1.5D_l + 0.1D_d + 0.6 = 0 \qquad (-14)$$

$$B_{t-k}: \quad 3.9L_b - 3.9L_{bp} + 1.1l_b + 2.3F3_o - 0.9D_l - 1.4D_d + 0.4 = 0 \qquad (-15)$$

Note that these equations are in fact mathematical expressions of the often (qualitatively) discussed cue trading relations (e.g. Dorman *et al.*, 1977). The cues in Eqs. –13—15 have been normalized using Eq. 3.7. Note that the "~" symbols have been left out for readability. The actual means and standard deviations used in this normalization of the cues in Eqs. –13—15 are listed in Table 6.VII. Using Table 6.VII, Eqs. –13—15 can be transformed into linear combinations of the true (unnormalized) cues, and can thus be tested on new stimuli.

For the *labial-dental distinction* (Eq. –13), the most important cues are the burst level L_b, the frequency of F3 at voicing onset $F3_o$, and the distance to the dental locus equation D_d. We find that a low burst level, combined with a low frequency of F3 at voicing onset and distance to the dental locus equation cue the labial place of articulation versus the dental place.

Table 6.VII: Means μ and standard deviations σ of the cues used in the best model for the combined stimuli with the *unvoiced stops* of *speaker 1*.

	L_b (dB)	L_{bp} (dB)	l_b (ms)	$F3_o$ (ERB)	D_l (ERB)	D_d (ERB)
μ	90.5	71.9	13.9	22.3	0.99	1.37
σ	4.5	7.2	5.3	1.0	1.10	1.29

For the *labial-velar distinction* (Eq. –14), the most important cues are the burst level L_b, the level of the spectral peak in the burst L_{bp}, and the distance to the labial locus equation D_l. For the sake of interpretation, we rewrite the terms $2.2L_b - 3.2L_{bp}$ as $-2.2\Delta L_{bp} - 1.0L_{bp}$, where $\Delta L_{bp} = L_{bp} - L_b$, which is the level of the spectral peak in the burst relative to the total level of the burst. A large value of ΔL_{bp} indicates that almost all the burst energy is concentrated in the spectral peak, while a small value indicates that the burst peak is relatively weak. Thus, we find that the "pronouncedness" of the spectral peak in the burst is important for the labial-velar distinction: the labial class is cued by a "diffuse" burst spectrum, while the velar class is cued by a strong spectral peak. Furthermore, the labial class is distinguished from the velar class by a proximity to the labial locus equation.

The most important cues for the *dental-velar distinction* (Eq. –15) are the burst level L_b, the level of the spectral peak in the burst L_{bp}, the frequency of F3 at voicing onset $F3_o$, and the distance to the dental locus equation D_d. Again we substitute $\Delta L_{bp} = L_{bp} - L_b$, which transforms the terms $3.9L_b - 3.9L_{bp}$ into the single term $-3.9\Delta L_{bp}$. Clearly, like in the labial-velar distinction, here the velar class is cued by a pronounced spectral peak, while the dental class is cued by a diffuse burst spectrum. Furthermore, we find that a high $F3_o$ and proximity to the dental locus equation are cues to the dental class versus the velar class.

If we inspect the coefficients separately per cue, we find that some of the cues are important for only two distinctions, and are close to zero for the remaining distinction. This indicates that the cue mainly triggers one particular class (e.g. dental) rather than cueing a particular distinction (e.g. labial versus dental). Thus, we find that a high $F3_o$ indicates the dental category while a low $F3_o$ is not a particular cue to the labial or velar place, but simply indicates "not dental". Similarly, the distance to the dental locus equation is only relevant for contrasts involving the dental class (Eqs. –15 and –13).

Speaker 2

The highest GOF on testing for speaker 2 was obtained for the **su**-set (75%), where GOF on training was 82% (see Table 6.IV). The class boundaries are given by:

$$B_{p-t}: \quad 0.2l_b - 2.0F2_o + 0.05F3_o - 3.6F_o^{\mathrm{mfp}} + 1.0L_o + 1.9 = 0 \qquad (-16)$$

$$B_{p-k}: \quad -3.8l_b - 4.9F2_o + 0.4F3_o + 4.8F_o^{\mathrm{mfp}} + 1.6L_o - 0.4 = 0 \qquad (-17)$$

$$B_{t-k}: \quad -4.1l_b - 2.9F2_o + 0.3F3_o + 8.4F_o^{\mathrm{mfp}} + 0.6L_o - 2.3 = 0 \qquad (-18)$$

The means and standard deviations used in the normalization of the cues in Eqs. –16—18 are listed in Table 6.VIII.

The *labial-dental distinction* (Eq. –16) is mainly determined by the frequency of F2 at voicing onset $F2_o$, and the location of the mid-frequency peak at release F_o^{mfp}. A low mid-frequency peak at release, as well as a low frequency of F2 at voicing onset cue the labial class versus the dental class.

The most important cues for the *labial-velar distinction* (Eq. –17) are the burst length l_b, the frequency of F2 at voicing onset $F2_o$, and the frequency of the mid-

Table 6.VIII: Means μ and standard deviations σ of the cues used in the best model for the combined stimuli with the *unvoiced stops* of *speaker 2*.

	l_b (ms)	$F2_o$ (ERB)	$F3_o$ (ERB)	F_o^{mfp} (ERB)	L_o (dB)
μ	19.8	18.7	21.6	19.9	5.2
σ	11.4	2.4	1.4	3.9	2.1

frequency peak at consonantal release F_o^{mfp}. We rewrite the terms $-4.9F2_o+4.8F_o^{\mathrm{mfp}}$ as $-4.9\Delta F_o+0.1F_o^{\mathrm{mfp}}$, where $\Delta F_o = F2_o - F_o^{\mathrm{mfp}}$. The purpose of this manipulation is that the new measure ΔF_o seems to be a good indicator of the continuity of the burst peak into the formant transitions, that is, it is a "spectral" alternative to ΔL_o. Inspection of the acoustic data reveals that F_o^{mfp} is generally close to, or larger than $F2_o$. Thus, the maximum value of ΔF_o occurs when $F_o^{\mathrm{mfp}} \approx F2_o$, that is, when the spectral peak at release is spectrally continuous into F2. A small value of ΔF_o, on the other hand, occurs when F_o^{mfp} is much higher than $F2_o$, in which case the transition of the mid-frequency peak into F2 is discontinuous. Obviously, a large value of ΔF_o cues /k/, a small value cues /p/. Additionally, we find that a long burst cues /k/ versus /p/.

The *dental-velar distinction* (Eq. –18) is mainly determined by the burst length l_b, The F2-frequency at onset $F2_o$, and the frequency of the mid-frequency peak at release F_o^{mfp}. A similar argument as before leads to the transformation of $-2.9F2_o+8.4F_o^{\mathrm{mfp}}$ into $-2.9\Delta F_o + 5.5F_o^{\mathrm{mfp}}$. Thus, both the continuity of the mid-frequency peak, as well as a low frequency of the mid-frequency peak cues /k/ versus /t/. Additionally, a long burst cues /k/ versus /t/.

As discussed earlier, a number of acoustic analysis studies have demonstrated that the burst length is an acoustic correlate of place of articulation, and thus is a potential perceptual cue. The coefficients of l_b for the 3 boundaries indicate that the burst length is mainly used as a cue for "velar versus not velar" by the listeners.

Because the coefficients of $F3_o$ are close to zero for all 3 boundaries, it may seem that this cue is irrelevant and can be deleted from the cue set without damage. Inspection of the SLP-weights shows, however, that $F3_o$ does play an important role in the model as indicator of "troublesome areas", as discussed in section 6.3.5. Apparently, for high $F3_o$, which generally occurs for the vowel context /i/, the model is not capable of reproducing the observed response behavior based on the cue set in question. Calculation of GOF as well as chance level separately per vowel context, shows that the fit for the vowel context /i/ is by far the worst. Chance levels for the vowels /a, i, y, u/ are 51%, 58%, 53% and 49%, respectively (highest for /i/), while GOF-fuzzy on testing was 75%, 64%, 89%, and 71%, respectively (lowest for /i/).

Voiced stops

Speaker 1

For the data of the voiced stops of speaker 1 the highest GOF on testing was obtained for the **su**-set (19%), with a GOF on training of 95% (see Table 6.V). As there is only a labial-dental contrast for Dutch voiced stops, the model generates only one boundary B_{b-d}:

$$B_{b-d} : 0.8L_b - 9.4F2_o - 1.0F2_{st} + 1.1T_o + 5.7\Delta T_o + 0.3 = 0 \qquad (-19)$$

The means and standard deviations used in the normalization of the cues in Eq. -19 are listed in Table 6.IX.

The most relevant cues for the /b/-/d/ distinction are the F2-frequency at voicing onset $F2_o$, and the change of spectral tilt after release ΔT_o. In accordance with past perception studies, the labial place is perceptually distinguished from the dental place by a low $F2_o$ and an increasing spectral tilt, that is, a spectral tilt which becomes less falling when going from burst to voiced transitions. The roles of the burst level L_b and the spectral tilt at onset T_o are somewhat mystifying, as they appear to have the wrong sign. Their influence on the /b/-/d/ distinction is, however, small, and their main contribution is in the overall attenuation of the cue influence, like we have encountered before.

Table 6.IX: Means μ and standard deviations σ of the cues used in the best model for the combined stimuli with the *voiced stops* of *speaker 1*.

	L_b (dB)	$F2_o$ (ERB)	$F2_{st}$ (ERB)	T_o (dB·ERB^{-1})	ΔT_o (dB·ERB^{-1}·ms^{-1})
μ	88.1	18.8	18.2	-0.142	-0.100
σ	6.5	2.1	3.0	0.570	0.262

Speaker 2

The highest GOF on testing for speaker 2 was obtained for the **fo**-set (89%), with a GOF on training of 95%. The labial-dental boundary B_{b-d} is given by

$$B_{b-d} : 0.04F_{bp} - 1.2L_{bp} - 6.2F2_o + 3.1F2_{st} - 0.7F3_o - 1.4F3_{st} - 1.5 = 0 \qquad (-20)$$

The means and standard deviations used in the normalization of the cues in Eq. -20 are listed in Table 6.X.

We rewrite the terms $-6.2F2_o + 3.1F2_{st}$ as $-3.1F2_o - 3.1\Delta F2$, where $\Delta F2 = F2_o - F2_{st}$, which is the frequency change of F2 from onset to the stationary position. Thus, we find that the /b/-/d/ distinction is mainly cued by the onset

Table 6.X: Means μ and standard deviations σ of the cues used in the best model for the combined stimuli with the *voiced stops* of *speaker 2*.

	F_{bp} (ERB)	L_{bp} (dB)	$F2_o$ (ERB)	$F2_{st}$ (ERB)	$F3_o$ (ERB)	$F3_{st}$ (ERB)
μ	19.9	72.8	18.6	18.3	22.3	22.8
σ	3.4	9.1	2.2	3.1	1.0	1.9

frequency of the F2 and its frequency change. In accordance with previous studies we find that perception of the labial place versus the dental place benefits from a low $F2_o$ as well as a rising F2. The role of the $F3_{st}$ needs some clarification. We surmise that $F3_{st}$ basically has a trading relation with the dominant F2-cues. $F3_{st}$ for speaker 2's vowels /a, i, y, u/ are approximately 2350 Hz, 3400 Hz, 2050 Hz, and 2050 Hz, respectively. When $F3_{st}$ is relatively low, that is, in vowel contexts /a, i, y/, the /b/-/d/ boundary for the cues $F2_o$ and $\Delta F2$ lies relatively close to /b/. This means that a /d/-percept needs a relatively high $F2_o$ and a strongly falling F2. When $F3_{st}$ is high (vowel context /i/), on the other hand, the /b/-/d/ boundary shifts toward /d/, which means that the demands for the /d/-percept are relaxed, and now the /b/-percept needs a relatively low $F2_o$ and a distinctly rising F2. Note that these trading relations, which have also been reported in a number of previous studies, are here compactly represented in a single linear expression.

6.5 General discussion and conclusions

In this chapter and the previous chapter, we have described an experiment which consists of three major steps. The first step is a perception experiment in which original and manipulated ("burst-spliced") natural stop-vowel utterances have been presented to listeners for classification. The second step is the measurement of various detailed and gross cues on the stimuli. The third step is the mapping of the measured acoustic cues onto the observed perceptual responses. With the second and third step we have intended to simulate the behavior of the listeners with two purposes: (1) to establish whether detailed or gross cues give a better account of the perceptual data, and (2) to model how the listeners have integrated the cues in their perception of the stimuli.

6.5.1 Detailed versus gross cues

Concerning the first question, whether detailed or gross cues give a better account of the perceptual data, we have found the following. For the original utterances only, the levels of goodness-of-fit did not provide a clear basis for preferring either detailed or gross cues. However, as discussed earlier, detailed and gross cues will generally *covary* in natural speech. Therefore, the model fits on the manipulated utterances with reduced covariance and redundancy provide the real test of which

type of cues has been used by the listeners. For these stimuli - the combined stimuli - the performance of detailed and gross cues clearly differs: in all cases the detailed cues give a better account of the perceptual data than the gross cues. Often, however, the "super"-set - a number of detailed cues combined with one or two gross cues - gave the overall best performance.

From these observations we conclude that it is likely that detailed spectro-temporal properties, such as formants, have been the primary cues for the listeners' perception of place of articulation in our experiments. We formulate this conclusion in somewhat cautious terms because the procedure adopted in our study is based on a number of assumptions which, naturally, are liable to criticism.

First of all, we want to remark that part of the detailed cues, namely the formant frequencies and the burst length, have been measured manually, while all gross cues were measured automatically. This may give the detailed cues an advantage, as phonetic knowledge has been used in the complex process of the manual measurement of these cues, e.g. knowledge on the continuity of formants into the vowel (not every spectral peak is a formant).

Secondly, while there has been some consensus on the identity of the detailed cues in the existing literature (frequencies of the second and third formant plus certain burst parameters), the gross cues that have been proposed over the years seem to vary from author to author. We have attempted to capture the major gross spectral properties that have been proposed, namely the global spectral tilt and compactness, and their evolution throughout the utterance. Nevertheless, our cues may fail to capture certain potentially important gross spectral properties, such as the change of high-frequency energy over time (Ohde and Stevens, 1983).

Thirdly, we remark that our categorization model makes certain assumptions on the nature of the categorization process. The most basic assumptions are (1) linear class boundaries, and (2) one convex subspace per response class. In a first investigation of this kind, where one wants to start simple, these assumptions seem warranted. Moreover, as discussed earlier, it is our opinion that these assumptions are superior to, for instance, the assumption of category prototypes combined with an isotropic distance functions, as is often used in other studies (e.g. Oden and Massaro, 1978; Suomi, 1985). Nevertheless, these assumptions may substantially deviate from reality. For instance, the velar category, for which most cues are extremely context-dependent, may be associated with a non-convex (e.g. strongly curved) cue-subspace, or even several disjunct subspaces associated with allophonic variants, for instance for front versus back vowel contexts.

Finally, we want to look in more detail at the differences in the results for the unvoiced stops and the voiced stops. Let us define the *badness-of-fit* (BOF) as 1 minus the goodness-of-fit (GOF), and let us judge the modeling success in terms of the reduction of BOF compared to chance level. The best-performing model fit for the unvoiced stops provides a reduction in BOF of roughly 50% (BOF shifts from 49% at chance level to 25%). Clearly, the model cannot account for a substantial part of the observed behavior. Note, however, that these figures are based upon

cross-validation on independent test data.[3] The best-performing model for the voiced stops provides a higher reduction in BOF, namely 75% (BOF shifts from 39% at chance level to 10%). We attribute the difference in the BOF-reduction between the voiced and the unvoiced stops to the fact that the voiced stops lack the velar category, rather than to the possibility that the cues (or model) used in this study would be more suitable for voiced stops than for unvoiced stops. It seems to be the case that the perception of the labial-dental distinction is modeled more "easily" than the distinctions involving the velar category. This may be caused by the fact that the velar cues vary more strongly with phonetic context than the labial or dental cues, as mentioned earlier.

6.5.2 Cue integration

We have interpreted the best-fitting models in terms of the boundaries between each pair of response classes. In our classification model, the class boundaries are linear combinations of 5 or 6 cues. Although we have discussed the resulting class boundaries individually per cue, we stress that it is the *combination* of cues in the boundary equations that is important here. As noted earlier, the boundary equations can actually be interpreted as quantitative expressions of the often-discussed "trading relations" between cues (e.g. Dorman *et al.*, 1977). The stimulus is labeled on the basis of a set of cues, rather than just one, and the value of one cue can be "compensated for" by the value of another cue.

In a sense, our boundary equations are similar to locus equations. Compare, for instance, Eq. –4 to Eq. –20. These equations are very similar, Eq. –20 just contains additional terms which incorporate information of F3 and the release burst. It is, however, important to note two basic differences. Firstly, locus equations are considered to be category *prototypes*, while our equations are category *boundaries*. Secondly, locus equations (so far) have always been derived from speech *production* data, while our equations are derived from speech *perception* data.

6.5.3 Comparison with Krull (1990)

Krull (1990) investigated whether perceptual confusions of stop consonants could be predicted from acoustic distances between the speech signals. In a number of respects Krull's study is similar to ours. Firstly, she presented truncated natural stop-vowel utterances to listeners for classification. Next, two types of acoustic properties were measured on the stimuli: formant frequencies plus the length of the release burst, and running spectral levels. Finally, the acoustic distances between each utterance and prototypical utterances of each of the response classes were mapped onto the observed perceptual confusions: large acoustic distance should lead to low confusability, small acoustic distance should lead to high confusability. Krull found that the distances based on formant frequencies plus burst length had a

[3] We employed this severe test of model-generalizability in order to avoid over-fitting. To our knowledge, this technique is seldom used in perceptual modeling, with the result that the presented GOF-levels - obtained from the training set only - may often be over-optimistic.

high correlation with the perceptual confusions. The running spectral levels, on the other hand, correlated much less with the perceptual confusions. If we assume that the running spectral levels capture gross spectro-temporal properties, these results are in good agreement with our finding that detailed cues give a better account of the listeners' classification behavior than the gross cues.

An important difference between Krull's approach and our approach is that Krull aims to account for perceptual confusions only, while we have simulated the complete classification process. We have set up a formal classification model which generates a response to every stimulus. Moreover, our simulation not only enables us to evaluate the perceptual relevance of acoustic cues, but, in addition, allows for interpretation in terms of how the cues are actually integrated in the classification process.

6.5.4 Coonclusions

In this chapter, we have presented a simulation of the classification behavior of the listeners in the perception experiment of the chapter 5. A number of detailed and gross cues have been measured on the stimuli, and these cues have bee mapped onto the observed perceptual data using a formal model of human classification behavior. The results have shown that the detailed cues, such as formant transitions, give a better account of the perceptual data than the gross cues. The best-performing models have been interpreted in terms of the acoustic boundaries which are associated with the perceived linguistic contrasts,

Appendix 6.A
Locus equations

Table 6.XI: Locus equations calculated separately for the two speakers and separately for voiced and unvoiced stops. "V/UV" indicates voiced stops or unvoiced stops, and "POA" stands for place of articulation. For the velar place of articulation separate locus equations were calculated for the back velars (vowel contexts /a, u/), and the fronted velars (vowel contexts /i, y/).

V/UV	speaker	POA	locus equation
UV	1	labial	$F2_o = 0.755F2_{st} + 0.236$
UV	1	dental	$F2_o = 0.328F2_{st} + 1.167$
UV	1	fronted velar	$F2_o = 0.496F2_{st} + 1.001$
UV	1	back velar	$F2_o = 1.731F2_{st} - 0.492$
UV	2	labial	$F2_o = 0.827F2_{st} + 0.122$
UV	2	dental	$F2_o = 0.150F2_{st} + 1.423$
UV	2	fronted velar	$F2_o = 0.594F2_{st} + 0.709$
UV	2	back velar	$F2_o = 1.191F2_{st} + 0.080$
V	1	labial	$F2_o = 0.747F2_{st} + 0.293$
V	1	dental	$F2_o = 0.416F2_{st} + 1.073$
V	2	labial	$F2_o = 0.817F2_{st} + 0.183$
V	2	dental	$F2_o = 0.462F2_{st} + 0.937$

Appendix 6.B
Acoustic cues measured on stimuli with unvoiced stops

Table 6.XII: Values of detailed cues, measured on the original utterances containing the unvoiced stops of speaker 1. The values of the detailed cues for the no-burst stimuli are identical to the values for the original utterances, except that there are no burst cues. The values of the detailed cues for the burst-only stimuli are identical to the values for the original utterances, except that there are only burst cues. The values of the detailed cues for the mixed-burst stimuli are identical to the values for the original utterances, taking the appropriate formant or locus-equation values from one utterance and the burst values from another. In this table and the following tables, the first 2 characters in the stimulus name indicate the stop-vowel utterance in phonetic notation, the number 1 or 2 indicates the token number, "org", "nob", and "bur" indicate original, no-burst, and burst-only stimulus, and "mxbp", "mxbt", and "mxbk" indicate a mixed-burst stimulus, with the burst of /p/, /t/, or /k/, respectively. All formant frequencies are listed in Hz (left number) as well as ERB (right number).

stim	L_b (dB)	F_{bp} (ERB)	L_{bp} (dB)	l_b (ms)	$F2_o$ (Hz,	ERB)	$F2_{st}$ (Hz,	ERB)	$F3_o$ (Hz,	ERB)	$F3_{st}$ (Hz,	ERB)	D_l (ERB)	D_d (ERB)	D_v (ERB)
pa1org	83.2	19.3	58.9	10.5	1120	16.5	1330	17.8	2400	22.7	2550	23.2	0.622	2.757	2.273
pa2org	86.7	13.0	68.2	11.9	1160	16.8	1358	18.0	2380	22.6	2596	23.4	0.513	2.535	2.253
ta1org	87.6	19.1	67.8	9.2	1360	18.0	1394	18.2	2400	22.7	2523	23.1	0.333	1.382	1.700
ta2org	87.9	20.9	66.3	8.0	1480	18.7	1349	17.9	2500	23.0	2514	23.1	1.023	0.659	1.088
ka1org	91.1	20.5	72.7	16.2	1820	20.4	1367	18.0	2200	21.9	2550	23.2	2.259	0.941	0.150
ka2org	89.2	21.0	73.0	14.4	1900	20.7	1349	17.9	2240	22.1	2468	22.9	2.600	1.315	0.157
pi1org	84.9	25.6	63.2	9.7	1960	21.0	2165	21.8	2180	21.9	3009	24.6	0.296	0.334	0.425
pi2org	89.7	20.1	64.7	8.6	1800	20.3	2128	21.7	2060	21.4	2963	24.5	0.148	0.279	0.990
ti1org	94.6	23.0	76.5	10.0	1980	21.1	2156	21.8	2500	23.0	2945	24.4	0.383	0.425	0.334
ti2org	92.1	14.2	71.5	10.8	1980	21.1	2128	21.7	2480	23.0	2872	24.2	0.455	0.454	0.284
ki1org	91.9	23.9	73.6	17.6	2040	21.3	2119	21.6	2940	24.4	2963	24.5	0.668	0.710	0.045
ki2org	92.5	24.7	74.6	15.5	2080	21.5	2138	21.7	2920	24.4	2936	24.4	0.742	0.836	0.064
py1org	88.2	19.8	73.3	21.6	1500	18.8	1697	19.8	2100	21.6	2101	21.6	0.071	1.080	1.535
py2org	88.6	19.1	69.8	17.2	1540	19.0	1615	19.4	2040	21.3	2092	21.5	0.352	0.756	1.178
ty1org	97.5	27.6	82.3	20.0	1740	20.0	1651	19.6	2240	22.1	2110	21.6	1.002	0.140	0.339
ty2org	97.1	27.1	82.8	25.3	1620	19.4	1679	19.7	2200	21.9	2156	21.8	0.465	0.459	0.929
ky1org	94.4	20.1	81.4	12.6	1700	19.8	1688	19.7	2020	21.2	2101	21.6	0.740	0.097	0.588
ky2org	97.9	20.4	84.1	12.5	1960	21.0	1661	19.6	2140	21.7	2156	21.8	1.723	1.062	0.537
pu1org	81.5	15.3	63.2	13.2	920	15.0	835	14.3	2520	23.1	2312	22.4	0.356	3.508	0.143
pu2org	82.9	20.0	59.6	25.7	900	14.8	807	14.0	2240	22.1	2321	22.4	0.372	3.635	0.021
tu1org	92.9	27.9	70.0	6.5	1440	18.5	771	13.7	2300	22.3	2358	22.5	3.499	0.113	2.310
tu2org	94.4	27.9	71.6	5.0	1640	19.5	761	13.6	2500	23.0	2339	22.5	4.407	1.217	3.031
ku1org	94.1	14.4	82.9	15.3	800	14.0	780	13.8	1940	20.9	2358	22.5	0.182	4.468	0.268
ku2org	90.2	13.3	74.7	14.3	880	14.7	761	13.6	1720	19.9	2339	22.5	0.486	3.744	0.249

Table 6.XIII: Values of gross cues, measured on the original and no-burst stimuli containing the unvoiced stops of speaker 1.

stim	T_o (dB/ERB)	ΔT_o (dB/ERB/ms)	T_{st} (dB/ERB)	F_o^{mfp} (ERB)	F_{st}^{mfp} (ERB)	L_o (dB)	ΔL_o (dB)	L_{st} (dB)
pa1org	-0.104	-0.559	-1.435	19.9	17.4	1.60	-3.851	3.97
pa2org	-0.631	-0.327	-1.488	15.6	13.7	3.94	-0.226	4.55
ta1org	-0.092	-0.435	-1.713	20.4	13.7	4.47	-2.672	5.13
ta2org	0.282	-0.521	-1.505	20.9	13.5	5.56	-3.723	3.72
ka1org	0.286	-0.435	-1.489	21.5	17.3	6.17	-2.850	4.07
ka2org	0.191	-0.293	-1.740	21.3	13.7	6.24	-1.950	4.96
pi1org	0.447	0.118	1.054	20.7	25.4	3.56	0.516	6.88
pi2org	0.065	0.227	1.315	19.1	25.7	3.13	-0.454	7.49
ti1org	0.587	0.087	0.483	23.4	24.5	3.91	-0.036	4.70
ti2org	0.207	0.214	0.896	24.0	24.8	3.46	0.046	5.19
ki1org	0.764	0.088	1.029	23.5	25.2	5.59	0.021	6.12
ki2org	1.255	-0.041	0.926	24.6	24.9	6.28	0.127	5.32
py1org	-0.097	-0.025	0.194	19.0	20.7	4.26	0.437	6.43
py2org	-0.004	-0.053	0.051	23.2	20.7	1.94	-1.310	5.15
ty1org	0.987	-0.121	0.403	23.5	20.6	2.96	-0.774	4.66
ty2org	0.873	-0.060	0.364	24.1	20.0	4.55	-0.906	5.44
ky1org	-0.224	0.058	0.016	20.2	20.0	6.70	0.036	5.97
ky2org	-0.049	-0.041	0.069	20.7	19.8	6.38	0.203	5.99
pu1org	-0.047	-0.570	-2.346	14.9	13.7	2.67	1.176	7.08
pu2org	0.054	-0.498	-2.381	14.5	12.6	2.79	0.817	8.37
tu1org	0.686	-0.558	-2.365	20.6	12.7	3.80	-3.320	7.72
tu2org	1.287	-0.690	-2.499	25.3	13.1	6.57	-3.521	7.70
ku1org	-1.596	-0.279	-3.120	14.1	13.1	7.30	-0.057	8.59
ku2org	-0.896	-0.437	-3.012	12.6	13.3	8.46	-0.449	8.30
pa1nob	-1.558	-0.160	-1.418	16.0	17.4	4.76	-0.516	4.18
pa2nob	-1.788	-0.013	-1.524	14.8	13.6	4.14	-0.200	4.60
ta1nob	-1.039	-0.195	-1.721	18.5	13.7	0.98	0.137	5.24
ta2nob	-0.950	-0.198	-1.505	13.2	13.5	5.02	1.228	3.90
ka1nob	-0.555	-0.301	-1.558	18.6	17.3	3.45	-0.335	4.29
ka2nob	0.229	-0.409	-1.798	22.8	13.7	2.60	-2.340	4.92
pi1nob	0.557	0.117	1.027	21.2	25.2	6.13	-0.304	6.97
pi2nob	-0.123	0.304	1.132	20.6	25.4	7.06	-0.602	6.51
ti1nob	0.838	0.032	0.625	23.5	25.2	5.22	-0.443	5.18
ti2nob	0.815	0.044	0.957	21.8	25.3	4.68	0.192	4.40
ki1nob	0.857	0.064	0.892	24.5	25.1	7.26	-0.471	5.77
ki2nob	1.115	0.008	0.751	14.8	24.8	-1.62	-3.116	4.81
py1nob	0.439	-0.206	-0.103	21.5	20.1	4.12	-0.152	6.57
py2nob	0.590	-0.224	0.160	19.6	20.0	4.62	0.507	5.08
ty1nob	1.524	-0.356	0.300	20.6	20.3	1.94	1.061	4.86
ty2nob	1.497	-0.352	0.097	20.4	19.8	1.42	1.161	6.52
ky1nob	0.615	-0.195	0.080	18.5	20.2	2.42	1.070	5.68
ky2nob	0.121	-0.134	0.210	21.3	20.4	6.41	-0.025	5.77
pu1nob	-1.395	-0.203	-2.402	14.3	13.3	5.95	0.275	7.54
pu2nob	-1.618	-0.177	-2.715	14.4	13.1	6.50	0.312	8.19
tu1nob	-0.408	-0.275	-2.298	13.2	12.6	5.16	-0.300	8.05
tu2nob	0.296	-0.443	-2.456	19.1	13.2	3.87	-3.126	7.63
ku1nob	-1.205	-0.430	-3.394	15.9	13.1	6.04	-0.078	8.89
ku2nob	-1.801	-0.224	-3.136	15.0	13.2	7.90	-0.363	8.52

Table 6.XIV: Values of gross cues, measured on the mixed-burst stimuli containing the unvoiced stops of speaker 1.

stim	T_o (dB/ERB)	ΔT_o (dB/ERB/ms)	T_{st} (dB/ERB)	F_o^{mfp} (ERB)	F_{st}^{mfp} (ERB)	L_o (dB)	ΔL_o (dB)	L_{st} (dB)
pa1mxbt	-0.092	-0.563	-1.410	20.4	17.4	4.47	-5.534	4.03
pa2mxbt	0.285	-0.592	-1.437	20.9	13.7	5.56	-5.151	4.82
pa1mxbk	0.269	-0.624	-1.525	21.5	13.9	6.26	-5.444	3.93
pa2mxbk	0.191	-0.534	-1.517	21.3	13.7	6.24	-5.012	4.52
ta1mxbp	-0.104	-0.427	-1.710	19.9	13.7	1.60	-0.873	5.13
ta2mxbp	-0.631	-0.245	-1.483	15.6	13.3	3.94	-0.940	4.18
ta1mxbk	0.286	-0.478	-1.688	21.5	13.7	6.17	-3.867	5.02
ta2mxbk	0.191	-0.447	-1.487	21.3	13.4	6.24	-3.974	4.03
ka1mxbp	-0.104	-0.388	-1.552	19.9	17.1	1.60	-0.522	4.24
ka2mxbp	-0.631	-0.103	-1.760	15.6	13.9	3.94	-1.124	4.93
ka1mxbt	-0.092	-0.397	-1.555	20.4	17.1	4.47	-1.986	4.21
ka2mxbt	0.287	-0.384	-1.759	20.9	13.7	5.57	-2.074	5.03
pi1mxbt	0.587	0.077	1.055	23.4	25.4	3.91	-0.015	6.87
pi2mxbt	0.207	0.176	1.326	24.0	25.5	3.46	-0.307	7.06
pi1mxbk	0.764	0.004	1.159	23.5	25.4	5.59	-0.495	7.10
pi2mxbk	1.275	-0.153	1.259	24.6	25.4	6.45	-1.398	6.59
ti1mxbp	0.447	0.128	0.489	20.7	24.5	3.56	0.336	4.73
ti2mxbp	0.061	0.253	0.925	19.1	24.8	3.14	-1.380	5.54
ti1mxbk	0.764	0.014	0.447	23.5	25.0	5.59	-0.465	4.35
ti2mxbk	1.255	-0.089	0.887	24.6	21.7	6.28	-0.771	4.89
ki1mxbp	0.461	0.172	0.972	20.7	25.1	3.48	-0.251	6.04
ki2mxbp	0.062	0.297	0.830	19.1	24.9	3.14	-1.780	5.08
ki1mxbt	0.582	0.136	0.970	23.4	25.1	3.92	0.343	5.99
ki2mxbt	0.205	0.260	0.828	23.8	25.0	3.46	0.787	4.69
py1mxbt	0.987	-0.273	0.135	23.5	20.7	2.96	-2.114	6.36
py2mxbt	0.873	-0.206	0.102	24.1	20.8	4.55	-2.359	5.22
py1mxbk	-0.224	-0.004	-0.076	20.2	20.1	6.70	-0.228	6.22
py2mxbk	-0.049	-0.040	0.144	20.7	20.9	6.38	-0.193	4.98
ty1mxbp	-0.097	0.130	0.396	19.0	20.8	4.26	0.345	4.68
ty2mxbp	-0.007	0.103	0.261	23.2	19.9	1.92	-0.235	5.92
ty1mxbk	-0.224	0.157	0.378	20.2	20.3	6.70	-0.313	5.18
ty2mxbk	-0.049	0.104	0.167	20.7	19.8	6.38	-0.288	6.37
ky1mxbp	-0.097	0.032	-0.029	19.1	20.0	4.26	0.719	6.29
ky2mxbp	-0.004	-0.029	-0.061	23.2	19.5	1.94	-0.556	6.46
ky1mxbt	0.987	-0.213	-0.032	23.5	20.0	2.96	-1.681	6.31
ky2mxbt	0.873	-0.161	-0.077	24.1	19.5	4.55	-1.683	6.82
pu1mxbt	0.701	-0.776	-2.375	20.6	13.6	3.81	-5.311	7.34
pu2mxbt	1.299	-0.980	-2.556	25.3	13.1	6.59	-5.904	8.28
pu1mxbk	-1.596	-0.173	-2.311	14.1	13.6	7.30	-0.058	7.09
pu2mxbk	-0.896	-0.361	-2.476	12.6	12.7	8.46	-0.505	8.48
tu1mxbp	-0.047	-0.309	-2.294	14.9	12.7	2.67	0.486	7.71
tu2mxbp	0.056	-0.160	-2.395	14.5	12.7	2.79	-0.151	8.23
tu1mxbk	-1.596	0.109	-2.139	14.1	12.6	7.30	-0.745	8.13
tu2mxbk	-0.896	-0.019	-2.574	12.6	12.9	8.46	-1.273	8.00
ku1mxbp	-0.051	-0.695	-3.157	14.9	13.1	2.72	1.175	8.62
ku2mxbp	0.054	-0.547	-3.187	14.5	13.2	2.79	0.898	8.73
ku1mxbt	0.700	-0.942	-3.355	20.6	13.1	3.81	-5.579	8.81
ku2mxbt	1.281	-1.074	-3.102	25.3	13.1	6.57	-6.241	8.55

Table 6.XV: Values of gross cues, measured on the burst-only stimuli containing the unvoiced stops of speaker 1.

stim	T_o (dB/ERB)	F_o^{mfp} (ERB)	L_o (dB)
pa1bur	-0.100	19.9	1.58
pa2bur	-0.632	15.6	3.94
ta1bur	-0.092	20.4	4.46
ta2bur	0.281	20.9	5.55
ka1bur	0.287	21.5	6.17
ka2bur	0.190	21.3	6.24
pi1bur	0.449	20.7	3.55
pi2bur	0.063	19.1	3.15
ti1bur	0.586	23.4	3.91
ti2bur	0.207	24.0	3.46
ki1bur	0.763	23.5	5.59
ki2bur	1.255	24.6	6.28
py1bur	-0.099	19.1	4.25
py2bur	-0.006	23.2	1.93
ty1bur	0.987	23.5	2.96
ty2bur	0.873	24.1	4.55
ky1bur	-0.223	20.2	6.70
ky2bur	0.072	20.8	6.77
pu1bur	-0.045	14.9	2.68
pu2bur	0.058	14.5	2.61
tu1bur	0.700	20.6	3.81
tu2bur	1.306	25.3	6.59
ku1bur	-1.598	14.1	7.30
ku2bur	-0.897	12.6	8.46

Table 6.XVI: Values of detailed cues, measured on the original utterances containing the unvoiced stops of speaker 2.

stim	L_b (dB)	F_{bp} (ERB)	L_{bp} (dB)	l_b (ms)	$F2_o$ (Hz,	ERB)	$F2_{st}$ (Hz,	ERB)	$F3_o$ (Hz,	ERB)	$F3_{st}$ (Hz,	ERB)	D_l (ERB)	D_d (ERB)	D_v (ERB)
pa1org	87.7	17.4	62.3	4.6	1120	16.5	1358	18.0	2120	21.6	2339	22.5	0.618	2.943	2.377
pa2org	81.1	13.3	58.3	15.0	1150	16.8	1404	18.3	2160	21.8	2413	22.7	0.591	2.706	2.368
ta1org	88.0	13.0	66.4	9.5	1440	18.5	1459	18.6	2240	22.1	2358	22.5	0.480	1.053	1.365
ta2org	91.5	18.6	67.4	7.6	1420	18.3	1450	18.5	2320	22.4	2358	22.5	0.429	1.158	1.410
ka1org	88.6	21.2	72.0	14.7	1760	20.1	1404	18.3	1760	20.1	2294	22.3	1.895	0.604	0.029
ka2org	93.8	21.0	80.6	14.9	1780	20.2	1431	18.4	1980	21.1	2257	22.2	1.862	0.676	0.013
pi1org	89.2	17.9	63.6	8.1	1880	20.6	2110	21.6	2180	21.9	3394	25.7	0.043	0.629	0.302
pi2org	83.2	16.2	58.8	6.5	1880	20.6	2156	21.8	2160	21.8	3431	25.8	0.080	0.597	0.399
ti1org	91.4	20.7	76.9	16.2	1860	20.5	2174	21.8	2480	23.0	3486	25.9	0.194	0.497	0.512
ti2org	93.1	23.4	77.4	17.1	1880	20.6	2183	21.9	2660	23.6	3450	25.8	0.152	0.578	0.455
ki1org	101.5	25.6	86.8	46.8	2020	21.2	2183	21.9	2840	24.1	3349	25.5	0.289	1.166	0.049
ki2org	99.6	25.5	86.8	47.4	2000	21.1	2193	21.9	2920	24.4	3505	25.9	0.202	1.077	0.041
py1org	85.8	21.4	68.7	18.3	1660	19.6	1743	20.0	2100	21.6	2037	21.3	0.362	0.120	0.350
py2org	88.2	21.4	75.4	29.9	1660	19.6	1761	20.1	2020	21.2	2037	21.3	0.305	0.133	0.393
ty1org	97.3	21.7	82.7	24.5	1700	19.8	1780	20.2	2140	21.7	2119	21.6	0.390	0.046	0.271
ty2org	97.3	22.1	84.4	26.8	1660	19.6	1817	20.4	2080	21.5	2119	21.6	0.131	0.173	0.525
ky1org	99.2	21.0	85.4	27.0	1800	20.3	1872	20.6	2060	21.4	2064	21.4	0.456	0.443	0.082
ky2org	95.9	21.0	81.8	25.5	1780	20.2	1771	20.1	2020	21.2	2055	21.4	0.697	0.425	0.075
pu1org	93.8	15.7	77.3	21.6	700	13.0	734	13.3	1980	21.1	2018	21.2	0.218	5.987	1.630
pu2org	89.3	17.2	70.5	14.3	800	14.0	688	12.9	1880	20.6	2009	21.2	0.803	5.011	0.624
tu1org	96.8	25.8	78.1	10.2	1660	19.6	670	12.7	2120	21.6	2046	21.3	5.275	0.712	3.747
tu2org	99.4	24.4	82.8	8.9	1600	19.3	688	12.9	1960	21.0	2073	21.4	4.916	0.390	3.375
ku1org	97.2	14.1	85.5	27.8	860	14.5	706	13.1	1240	17.3	2046	21.3	1.092	4.492	0.367
ku2org	97.4	14.9	88.0	30.9	900	14.8	642	12.4	1840	20.5	2037	21.3	1.774	4.121	0.347

Table 6.XVII: Values of gross cues, measured on the original and no-burst stimuli containing the unvoiced stops of speaker 2.

stim	T_o (dB/ERB)	ΔT_o (dB/ERB/ms)	T_{st} (dB/ERB)	F_o^{mfp} (ERB)	F_{st}^{mfp} (ERB)	L_o (dB)	ΔL_o (dB)	L_{st} (dB)
pa1org	-0.206	-0.453	-1.698	17.8	13.5	4.05	-0.417	4.41
pa2org	-0.389	-0.340	-1.710	20.9	13.6	1.52	-2.917	5.50
ta1org	-0.572	-0.226	-1.450	18.9	13.5	2.99	-0.071	4.84
ta2org	-0.325	-0.293	-1.509	19.9	13.4	2.16	0.035	5.27
ka1org	-0.319	-0.050	-1.681	20.0	13.3	4.87	0.023	6.12
ka2org	0.171	-0.178	-1.348	21.0	13.5	6.50	-0.550	4.20
pi1org	-0.156	0.223	1.340	18.1	21.8	5.10	-4.585	3.89
pi2org	-0.262	0.253	1.266	16.5	21.8	2.77	-4.592	4.19
ti1org	0.228	0.121	1.064	21.0	21.9	4.81	0.043	6.10
ti2org	0.301	0.079	1.010	24.6	25.8	3.07	0.114	6.11
ki1org	1.209	0.014	0.715	25.3	21.7	8.59	-0.158	7.11
ki2org	1.223	-0.061	0.701	25.4	21.8	8.78	-0.292	7.03
py1org	-0.691	0.236	0.116	12.6	20.8	3.60	-3.612	7.91
py2org	-0.590	0.183	-0.127	12.6	20.6	3.47	-3.461	7.68
ty1org	0.468	-0.032	0.240	23.4	21.2	4.30	0.026	7.73
ty2org	0.917	-0.180	0.056	23.4	21.0	6.68	-0.767	7.78
ky1org	-0.052	0.016	0.202	20.2	21.0	7.30	0.074	8.06
ky2org	-0.302	0.101	0.120	21.1	20.8	4.04	0.937	7.75
pu1org	-1.047	-0.466	-3.466	16.6	13.1	5.52	-2.207	9.38
pu2org	-0.188	-0.571	-4.397	21.2	12.9	1.97	-3.797	9.76
tu1org	0.984	-0.417	-2.565	25.0	13.5	4.94	-4.226	8.67
tu2org	1.493	-0.557	-1.340	24.5	21.2	7.81	-4.355	-7.75
ku1org	-2.053	-0.066	-2.866	14.2	13.2	7.85	-0.041	9.32
ku2org	-1.530	-0.125	-3.233	14.0	13.1	6.30	0.319	9.60
pa1nob	-1.486	-0.090	-1.689	13.2	13.5	7.15	-0.428	4.74
pa2nob	-1.354	-0.084	-1.713	16.8	13.5	5.54	-0.596	5.59
ta1nob	-1.136	-0.091	-1.443	15.9	13.5	1.79	-0.605	5.09
ta2nob	-1.128	-0.077	-1.688	16.0	13.4	1.74	-0.524	5.57
ka1nob	-0.434	-0.102	-1.613	20.8	13.2	4.49	-0.252	6.25
ka2nob	-0.282	-0.117	-1.274	21.5	13.5	3.57	-0.212	4.23
pi1nob	-0.352	0.313	1.396	20.7	21.8	7.32	-0.383	3.68
pi2nob	0.157	0.152	1.366	21.0	21.8	6.61	-0.021	3.78
ti1nob	1.144	-0.127	1.027	24.9	21.8	6.52	-0.621	5.82
ti2nob	0.946	-0.037	0.705	24.0	21.8	6.79	-0.821	7.41
ki1nob	1.726	-0.295	0.682	25.9	21.8	7.09	-0.532	6.85
ki2nob	0.836	-0.074	0.827	25.7	21.8	7.05	-0.956	6.90
py1nob	0.348	-0.079	0.174	21.9	20.9	7.64	-0.274	6.74
py2nob	-0.010	-0.017	-0.076	21.2	20.8	8.04	-0.284	7.68
ty1nob	0.501	-0.078	0.292	21.9	21.2	7.84	-0.302	7.62
ty2nob	0.169	-0.013	0.282	21.8	21.2	7.39	-0.167	8.00
ky1nob	0.013	-0.005	0.104	21.1	20.9	8.00	-0.031	8.11
ky2nob	-0.412	0.121	0.154	21.5	20.9	7.67	0.019	7.58
pu1nob	-1.139	-0.446	-3.718	16.6	12.8	5.54	-2.647	9.98
pu2nob	-0.998	-0.368	-3.765	13.6	12.7	4.58	0.848	10.02
tu1nob	0.610	-0.364	-2.340	21.6	13.3	5.03	-0.926	8.78
tu2nob	0.233	-0.247	-1.373	22.3	21.3	4.01	-1.219	-6.58
ku1nob	-1.712	-0.264	-3.470	14.5	12.8	6.08	0.685	10.09
ku2nob	-1.845	-0.114	-2.229	15.1	21.2	7.29	0.064	-13.49

Table 6.XVIII: Values of gross cues, measured on the mixed-burst stimuli containing the unvoiced stops of speaker 2.

stim	T_o (dB/ERB)	ΔT_o (dB/ERB/ms)	T_{st} (dB/ERB)	F_o^{mfp} (ERB)	F_{st}^{mfp} (ERB)	L_o (dB)	ΔL_o (dB)	L_{st} (dB)
pa1mxbt	-0.572	-0.350	-1.878	18.9	13.5	2.99	-0.880	5.84
pa2mxbt	-0.322	-0.373	-1.724	19.9	13.6	2.26	-1.938	5.55
pa1mxbk	-0.314	-0.402	-1.877	20.0	13.4	4.85	-3.370	5.34
pa2mxbk	0.172	-0.478	-1.710	21.0	13.6	6.47	-4.041	5.50
ta1mxbp	-0.204	-0.350	-1.423	17.8	13.5	4.08	-0.459	4.78
ta2mxbp	-0.389	-0.254	-1.537	20.9	13.4	1.52	-0.990	5.92
ta1mxbk	-0.314	-0.267	-1.481	20.0	13.6	4.85	-0.647	5.41
ta2mxbk	0.171	-0.395	-1.536	21.0	13.4	6.50	-2.425	5.92
ka1mxbp	-0.194	-0.159	-1.614	17.9	13.2	4.08	-0.258	6.17
ka2mxbp	-0.390	-0.032	-1.348	20.9	13.5	1.51	0.771	4.20
ka1mxbt	-0.571	-0.017	-1.645	18.7	13.3	2.91	0.432	5.87
ka2mxbt	-0.322	-0.085	-1.301	19.9	18.3	2.16	0.591	3.82
pi1mxbt	0.228	0.071	1.392	21.0	21.7	4.81	0.469	3.23
pi2mxbt	0.301	0.047	1.182	24.6	21.7	3.07	-0.702	5.30
pi1mxbk	1.209	-0.077	1.358	25.3	25.9	8.59	-0.864	8.48
pi2mxbk	1.223	-0.068	1.257	25.4	25.9	8.78	-0.470	8.51
ti1mxbp	-0.153	0.236	0.886	18.1	21.8	5.11	-3.897	7.14
ti2mxbp	-0.213	0.273	0.836	16.5	21.7	2.66	-3.634	7.11
ti1mxbk	1.209	-0.001	0.808	25.3	21.7	8.59	-0.247	7.18
ti2mxbk	1.223	-0.055	1.109	25.4	21.8	8.78	-0.371	5.96
ki1mxbp	-0.154	0.245	0.699	18.1	21.8	5.10	-4.025	6.79
ki2mxbp	-0.203	0.220	0.851	16.5	21.8	2.63	-3.667	6.64
ki1mxbt	0.228	0.147	0.771	21.0	21.8	4.80	-0.059	6.73
ki2mxbt	0.301	0.066	0.892	24.6	25.9	3.07	0.152	6.91
py1mxbt	0.468	-0.061	0.472	23.4	20.9	4.28	-0.314	7.82
py2mxbt	0.913	-0.203	0.006	23.4	20.6	6.68	-1.509	7.62
py1mxbk	-0.051	0.064	0.384	20.2	20.8	7.31	-0.202	7.61
py2mxbk	-0.302	0.085	-0.014	21.1	20.7	4.04	0.729	7.52
ty1mxbp	-0.688	0.270	0.157	12.6	21.1	3.60	-3.491	7.43
ty2mxbp	-0.591	0.219	0.000	12.6	21.0	3.47	-3.979	7.75
ty1mxbk	-0.052	0.102	0.198	20.2	21.2	7.30	-0.207	7.93
ty2mxbk	-0.302	0.119	0.005	21.1	21.0	4.02	0.681	7.84
ky1mxbp	-0.691	0.196	0.141	12.6	21.0	3.60	-4.822	8.10
ky2mxbp	-0.590	0.200	0.044	12.6	20.7	3.47	-4.306	7.87
ky1mxbt	0.490	-0.111	0.157	23.4	21.0	4.21	-0.072	8.24
ky2mxbt	0.910	-0.198	0.143	23.4	20.8	6.67	-1.089	7.81
pu1mxbt	1.000	-1.037	-3.679	25.0	12.9	4.98	-8.253	9.65
pu2mxbt	1.493	-1.032	-4.288	24.5	12.9	7.81	-7.617	9.81
pu1mxbk	-2.053	-0.191	-3.361	14.2	13.1	7.85	0.008	9.28
pu2mxbk	-1.534	-0.181	-4.327	14.0	13.1	6.31	0.196	9.59
tu1mxbp	-1.047	0.172	-2.370	16.6	13.6	5.52	-0.864	8.62
tu2mxbp	-0.188	-0.076	-1.777	21.2	12.5	1.97	0.056	10.29
tu1mxbk	-2.053	0.337	-2.071	14.2	13.6	7.85	-1.807	8.31
tu2mxbk	-1.533	0.189	-1.742	14.0	12.6	6.31	-1.291	9.59
ku1mxbp	-1.047	-0.309	-2.947	16.6	13.1	5.52	-1.074	9.46
ku2mxbp	-0.188	-0.503	-2.501	21.2	21.2	1.97	-3.003	-16.24
ku1mxbt	0.982	-0.946	-3.189	25.0	12.9	4.94	-7.590	9.83
ku2mxbt	1.493	-0.998	-2.403	24.5	12.6	7.81	-8.149	9.89

Table 6.XIX: Values of gross cues, measured on the burst-only stimuli containing the unvoiced stops of speaker 2.

stim	T_o (dB/ERB)	F_o^{mfp} (ERB)	L_o (dB)
pa1bur	-0.194	17.8	4.06
pa2bur	-0.302	20.9	1.53
ta1bur	-0.573	18.9	2.99
ta2bur	-0.326	19.9	2.16
ka1bur	-0.318	20.0	4.87
ka2bur	0.172	21.0	6.49
pi1bur	-0.155	18.1	5.10
pi2bur	-0.263	16.5	2.76
ti1bur	0.228	21.0	4.81
ti2bur	0.300	24.6	3.07
ki1bur	1.209	25.3	8.59
ki2bur	1.224	25.4	8.78
py1bur	-0.692	12.6	3.60
py2bur	-0.587	12.6	3.47
ty1bur	0.493	23.4	4.25
ty2bur	0.918	23.4	6.67
ky1bur	-0.052	20.2	7.30
ky2bur	-0.302	21.1	4.04
pu1bur	-1.045	16.6	5.52
pu2bur	-0.188	21.2	1.98
tu1bur	0.984	25.0	4.94
tu2bur	1.493	24.5	7.81
ku1bur	-2.059	14.2	7.85
ku2bur	-1.532	14.0	6.30

Appendix 6.C
Acoustic cues measured on stimuli with voiced stops

Table 6.XX: Values of detailed cues, measured on the original utterances containing the unvoiced stops of speaker 1. The values of the detailed cues for the no-burst stimuli are identical to the values for the original utterances, except that there are no burst cues. The values of the detailed cues for the mixed-burst stimuli are identical to the values for the original utterances, taking the appropriate formant or locus-equation values from one utterance and the burst values from another. In this table and the following tables, the first 2 characters in the stimulus name indicate the stop-vowel utterance in phonetic notation, the number 1 or 2 indicates the token number, "org" and "nob" indicate original and no-burst stimulus, and "mxbb", and "mxbd" indicate a mixed-burst stimulus, with the burst of /b/, or /d/, respectively. All formant frequencies are listed in Hz (left number) as well as ERB (right number).

stim	L_b (dB)	F_{bp} (ERB)	L_{bp} (dB)	l_b (ms)	$F2_o$ (Hz, ERB)		$F2_{st}$ (Hz, ERB)		$F3_o$ (Hz, ERB)		$F3_{st}$ (Hz, ERB)		D_l (ERB)	D_d (ERB)
ba1org	79.6	17.0	59.4	5.8	1120	16.5	1101	16.3	2480	23.0	2606	23.4	0.023	2.401
ba2org	89.2	20.0	70.3	9.0	1230	17.2	1404	18.3	2520	23.1	2688	23.6	0.545	2.258
da1org	92.0	14.0	73.1	8.1	1530	18.9	1165	16.8	2520	23.1	2606	23.4	1.731	0.139
da2org	91.6	17.3	72.6	11.9	1680	19.7	1394	18.2	2580	23.3	2624	23.4	1.459	0.127
bi1org	77.2	20.5	61.1	10.0	1920	20.8	2229	22.1	2360	22.5	2963	24.5	0.128	0.309
bi2org	85.2	19.5	67.1	10.3	1940	20.9	2147	21.7	2260	22.2	2862	24.2	0.142	0.100
di1org	87.6	13.0	71.0	12.6	2050	21.4	2174	21.8	2570	23.3	3000	24.6	0.428	0.276
di2org	94.8	20.8	82.0	13.7	2000	21.1	2174	21.8	2510	23.1	2771	23.9	0.270	0.088
by1org	88.4	23.3	68.3	11.7	1580	19.2	1688	19.7	2080	21.5	2101	21.6	0.103	0.885
by2org	91.7	19.2	75.7	14.2	1590	19.3	1661	19.6	2090	21.5	2092	21.5	0.226	0.790
dy1org	93.7	19.8	74.7	7.4	1710	19.9	1697	19.8	2260	22.2	2119	21.6	0.579	0.300
dy2org	94.9	20.0	79.5	10.4	1690	19.8	1688	19.7	2300	22.3	2156	21.8	0.531	0.374
bu1org	80.1	15.1	68.0	14.9	950	15.2	752	13.5	2320	22.4	2404	22.7	0.649	3.002
bu2org	75.9	13.4	59.9	9.0	820	14.1	798	13.9	2350	22.5	2394	22.7	0.493	4.178
du1org	92.6	19.0	74.2	6.8	1420	18.3	780	13.8	2350	22.5	2431	22.8	3.075	0.132
du2org	95.7	25.7	76.0	10.8	1450	18.5	817	14.1	2240	22.1	2440	22.8	3.009	0.213

Table 6.XXI: Values of gross cues, measured on the original, no-burst, and mixed-burst stimuli containing the voiced stops of speaker 1.

stim	T_o (dB/ERB)	ΔT_o (dB/ERB/ms)	T_{st} (dB/ERB)	F_o^{mfp} (ERB)	F_{st}^{mfp} (ERB)	L_o (dB)	ΔL_o (dB)	L_{st} (dB)
ba1org	-0.559	-0.316	-2.089	16.7	12.5	2.57	0.220	7.80
ba2org	-0.491	-0.274	-1.692	19.8	13.4	4.11	-3.298	5.45
da1org	-0.181	-0.218	-2.117	20.7	13.3	4.00	-2.602	5.79
da2org	-0.104	-0.252	-1.587	20.7	13.3	2.13	-1.576	5.22
bi1org	0.395	0.095	1.453	21.1	25.9	4.19	0.430	8.03
bi2org	-0.025	0.253	0.820	19.6	25.4	3.62	-0.058	6.22
di1org	-0.390	0.353	1.475	12.9	25.7	5.18	-3.878	7.87
di2org	0.277	0.181	1.017	20.9	24.8	5.04	0.006	4.56
by1org	0.459	-0.138	0.155	23.4	19.9	5.12	-2.290	6.79
by2org	0.042	-0.028	0.170	25.9	20.2	1.91	-0.998	6.80
dy1org	0.014	0.065	0.313	14.1	20.7	2.10	-3.636	6.47
dy2org	0.292	-0.032	-0.081	20.2	20.0	4.16	0.614	6.94
bu1org	-0.901	-0.373	-2.677	15.3	13.9	3.85	1.000	7.41
bu2org	-0.363	-0.493	-2.651	13.3	13.3	2.73	1.098	8.26
du1org	0.273	-0.380	-2.585	19.9	13.9	3.79	-2.327	7.75
du2org	0.569	-0.456	-2.295	21.5	12.7	1.22	-2.364	6.39
ba1nob	-1.343	-0.098	-2.036	12.5	14.8	5.05	-0.144	7.95
ba2nob	-1.659	0.026	-1.649	13.4	23.2	-11.15	-0.441	5.36
da1nob	-0.678	-0.112	-2.101	13.1	19.5	2.65	-0.619	5.58
da2nob	-0.954	-0.074	-1.708	13.4	19.4	1.15	-0.647	5.42
bi1nob	0.049	0.223	1.324	25.8	20.8	5.54	0.005	6.96
bi2nob	0.037	0.278	0.757	25.4	21.2	5.59	-0.159	6.08
di1nob	0.896	-0.003	1.447	25.9	22.9	5.78	-0.024	8.25
di2nob	0.283	0.230	1.008	25.3	21.1	6.63	-0.611	4.63
by1nob	-0.312	0.115	0.052	19.9	19.4	4.82	0.372	7.17
by2nob	-0.287	0.084	0.090	19.9	19.3	5.33	0.317	7.23
dy1nob	0.751	-0.148	0.344	20.3	22.6	3.64	-0.307	6.58
dy2nob	0.390	-0.056	-0.145	19.9	19.9	4.72	0.447	7.05
bu1nob	-1.740	-0.196	-2.666	13.9	15.3	6.03	0.411	7.40
bu2nob	-0.544	-0.462	-2.702	13.3	16.4	3.01	0.430	8.34
du1nob	0.055	-0.359	-2.528	13.7	19.5	4.61	-2.515	7.79
du2nob	-0.379	-0.257	-2.344	12.6	18.5	3.70	-1.104	6.48
ba1mxbd	-0.179	-0.414	-2.077	12.5	20.7	4.06	-6.680	7.37
ba2mxbd	-0.095	-0.365	-1.690	13.4	20.7	2.19	-4.172	5.35
da1mxbb	-0.538	-0.132	-2.105	12.9	16.8	2.70	0.266	5.55
da2mxbb	-0.479	-0.169	-1.606	13.3	19.8	4.16	-1.458	5.21
bi1mxbd	-0.372	0.306	1.424	25.8	13.1	5.03	-2.900	7.96
bi2mxbd	0.278	0.147	0.861	25.5	20.9	5.04	0.199	6.33
di1mxbb	0.299	0.161	1.452	25.7	21.1	3.80	0.447	7.88
di2mxbb	-0.016	0.284	1.036	24.5	19.6	3.81	-0.273	5.13
by1mxbd	0.027	0.004	0.097	19.9	14.3	2.11	-2.268	7.04
by2mxbd	0.291	-0.086	0.193	20.2	20.2	4.16	0.655	6.75
dy1mxbb	0.501	-0.064	0.338	20.7	23.4	5.18	-1.498	6.47
dy2mxbb	-0.076	0.065	0.001	20.1	15.1	0.86	-2.389	6.72
bu1mxbd	0.279	-0.740	-2.694	13.9	19.9	3.83	-6.497	7.62
bu2mxbd	0.577	-0.739	-2.578	13.3	21.5	1.26	-4.192	8.28
du1mxbb	-0.835	0.000	-2.625	13.9	15.6	3.48	0.034	7.67

Table 6.XXII: Values of detailed cues, measured on the original utterances containing the voiced stops of speaker 2.

stim	L_b (dB)	F_{bp} (ERB)	L_{bp} (dB)	l_b (ms)	$F2_o$ (Hz, ERB)		$F2_{st}$ (Hz, ERB)		$F3_o$ (Hz, ERB)		$F3_{st}$ (Hz, ERB)		D_l (ERB)	D_d (ERB)
ba1org	74.4	15.1	54.9	5.0	1200	17.0	1275	17.5	2300	22.3	2532	23.1	0.122	1.792
ba2org	89.4	17.0	68.4	6.3	1210	17.1	1358	18.0	2380	22.6	2532	23.1	0.396	1.907
da1org	85.7	19.2	69.8	8.5	1570	19.2	1495	18.8	2470	22.9	2596	23.4	0.678	0.271
da2org	88.2	19.2	73.9	10.1	1560	19.1	1477	18.7	2530	23.1	2587	23.3	0.703	0.281
bi1org	94.7	20.4	81.0	12.4	1950	20.9	2183	21.9	2630	23.5	3358	25.6	0.053	0.015
bi2org	90.6	20.8	73.8	19.1	1990	21.1	2147	21.7	2700	23.7	3459	25.8	0.167	0.230
di1org	91.6	23.4	71.5	8.3	1980	21.1	2138	21.7	2680	23.6	3376	25.6	0.159	0.209
di2org	95.8	20.7	80.7	10.0	2000	21.1	2147	21.7	2690	23.7	3450	25.8	0.198	0.267
by1org	84.1	18.1	64.6	8.9	1520	18.9	1651	19.6	1980	21.1	1972	21.0	0.048	0.832
by2org	81.3	21.0	62.9	9.9	1600	19.3	1697	19.8	1920	20.8	2009	21.2	0.117	0.544
dy1org	102.0	24.3	89.1	15.6	1690	19.8	1807	20.3	2280	22.2	2009	21.2	0.112	0.353
dy2org	95.3	23.2	76.1	8.7	1740	20.0	1752	20.1	2310	22.4	2046	21.3	0.460	0.029
bu1org	80.4	15.1	60.7	7.2	770	13.7	725	13.3	2040	21.3	1963	21.0	0.040	3.833
bu2org	86.9	14.0	72.3	11.5	830	14.2	716	13.2	1890	20.7	1982	21.1	0.443	3.269
du1org	100.8	21.7	84.7	12.1	1370	18.1	798	13.9	2080	21.5	2000	21.1	2.965	0.381
du2org	98.4	25.7	80.4	6.8	1290	17.6	734	13.3	2090	21.5	1991	21.1	2.978	0.086

Table 6.XXIII: Values of gross cues, measured on the original, no-burst, and mixed-burst stimuli containing the voiced stops of speaker 2.

stim	T_o (dB/ERB)	ΔT_o (dB/ERB/ms)	T_{st} (dB/ERB)	F_o^{mfp} (ERB)	F_{st}^{mfp} (ERB)	L_o (dB)	ΔL_o (dB)	L_{st} (dB)
ba1org	-1.036	-0.167	-1.668	15.8	13.7	3.14	-0.007	5.41
ba2org	-0.590	-0.341	-1.634	19.3	13.9	2.33	-0.635	5.10
da1org	-0.438	-0.188	-1.432	19.1	13.6	3.39	-0.631	5.51
da2org	-0.828	-0.055	-1.361	14.6	13.7	2.49	0.457	5.49
bi1org	-0.176	0.274	1.368	20.2	25.8	6.97	-0.586	9.37
bi2org	0.121	0.129	1.170	23.7	21.9	2.95	0.342	4.16
di1org	0.328	0.217	1.449	23.1	22.0	2.63	0.011	3.42
di2org	0.548	0.134	1.756	21.0	25.9	6.03	-0.457	9.91
by1org	0.026	-0.013	-0.279	19.1	20.4	3.33	0.817	8.00
by2org	0.169	-0.063	0.209	21.7	20.4	1.83	1.250	7.65
dy1org	1.436	-0.282	0.289	23.5	20.8	6.70	-0.653	8.08
dy2org	0.631	-0.099	0.344	23.6	21.1	6.03	-1.019	8.24
bu1org	0.026	-0.673	-1.856	24.1	13.5	2.16	-7.243	8.06
bu2org	-0.149	-0.542	-1.940	13.3	12.8	2.51	1.434	8.25
du1org	0.648	-0.367	-2.168	21.6	13.9	3.11	-1.644	8.17
du2org	1.163	-0.542	-2.339	25.4	13.7	7.99	-5.017	8.34
ba1nob	-1.263	-0.109	-1.703	13.7	16.7	4.44	-0.238	5.80
ba2nob	-1.998	0.059	-1.733	13.7	16.7	1.00	0.752	5.79
da1nob	-0.542	-0.196	-1.409	13.6	19.3	2.36	-0.337	5.90
da2nob	-0.606	-0.147	-1.403	13.7	19.3	2.59	-0.161	5.37
bi1nob	0.479	0.129	1.423	25.9	21.9	4.81	0.020	9.00
bi2nob	0.242	0.129	1.265	21.7	23.5	4.61	-0.231	1.85
di1nob	0.723	0.157	1.516	21.9	23.7	5.66	-0.194	3.40
di2nob	0.763	0.100	2.131	21.8	23.6	6.36	-0.720	0.77
by1nob	-0.161	0.046	-0.135	20.4	20.0	5.05	0.605	7.95
by2nob	-0.372	0.115	0.238	20.9	19.6	5.63	0.306	8.06
dy1nob	0.764	-0.128	0.181	20.8	25.3	4.23	-2.843	8.03
dy2nob	0.942	-0.205	0.211	21.0	23.3	5.90	-0.670	7.88
bu1nob	-1.837	-0.166	-1.824	13.5	13.6	7.55	0.147	8.22
bu2nob	-1.712	-0.083	-2.630	13.1	14.1	6.26	0.378	8.81
du1nob	-0.033	-0.258	-2.375	13.7	21.5	4.88	-2.626	8.35
du2nob	0.121	-0.267	-2.365	13.9	21.3	4.21	-2.310	8.40
ba1mxbd	-0.440	-0.324	-1.643	13.7	19.1	3.41	-0.425	5.47
ba2mxbd	-0.820	-0.266	-1.503	14.0	14.6	2.53	0.787	4.95
da1mxbb	-0.559	-0.178	-1.405	13.6	15.4	2.51	-0.444	5.69
da2mxbb	-0.549	-0.144	-1.438	13.7	19.3	2.71	-0.139	6.16
bi1mxbd	0.334	0.146	1.417	25.9	23.1	2.66	0.504	9.51
bi2mxbd	0.551	0.032	1.463	25.9	21.0	6.04	-0.408	9.69
di1mxbb	-0.178	0.321	1.517	25.9	20.2	6.97	-0.771	9.91
di2mxbb	0.138	0.203	1.849	21.8	23.7	3.03	0.370	2.43
by1mxbd	1.429	-0.389	-0.243	20.4	23.5	6.70	-2.267	8.07
by2mxbd	0.635	-0.185	0.177	20.6	23.6	6.03	-2.675	7.61
dy1mxbb	0.044	0.094	0.050	20.8	19.1	3.40	0.083	8.16
dy2mxbb	0.183	0.036	0.322	21.1	21.7	1.92	1.345	8.23
bu1mxbd	0.647	-0.786	-1.857	13.1	21.6	3.12	-3.597	8.80
bu2mxbd	1.179	-0.881	-2.408	12.8	25.4	8.02	-7.120	9.06
du1mxbb	0.024	-0.240	-2.487	13.7	24.1	2.08	-3.873	8.40
du2mxbb	-0.144	-0.170	-2.095	13.9	13.5	2.31	-0.134	7.91

Chapter 7

General discussion and outlook

7.1 Introduction

The purpose of the research presented in this thesis is to study whether detailed or gross spectro-temporal properties are the primary cues in the perception of place of articulation of prevocalic stop consonants. In our approach we have emphasized two aspects. First of all, we wanted to use (manipulated) natural utterances. The primary reason for this choice was that past investigations employing synthetic stimuli have yielded unequivocal results, which may heave been partly caused by a certain unnaturalness of the stimuli. Secondly, we have attempted to carry out a complete simulation of the listeners' categorization behavior. The simulation was intended to establish which type of cues were used by the listeners, and how they were used.

A summary of the research is presented in the next chapter. In the current chapter we first of all elaborate on some of the methodological novelties introduced in this thesis. Next, in section 7.2.2, we will shortly discuss the phonetic insight provided by our study. In section 7.3 we will discuss the basic limitations of our research, and finally, in section 7.4, an experiment will be proposed which overcomes some of these limitations.

7.2 Value of our study

The present research may be valuable on two accounts, viz. the developed methodology and the phonetic insight that is gained.

7.2.1 Methodological value

Traditionally, knowledge on phoneme perception stems from three basic methodologies.

1. Acoustic analysis of natural speech.
   The general purpose of this type of study is to describe acoustic regularities and variability associated with phonemes (or other linguistic categories). Often, acoustic analysis is accompanied by automatic classification experiments in which the suitability of certain acoustic structures for classification is tested. Logically, from classification experiments it can be concluded that, if accurate automatic classification is possible from a certain set of acoustic properties, human listeners may actually make use of these cues while perceiving speech.

On the other hand, if the automatic classification based on a certain set of acoustic properties is unreliable, it can be concluded that listeners cannot base their phonetic perception on these cues using the specific classification method which is tested. It is important to realize that from a negative result it cannot be generally concluded that listeners do not base their phonetic perception on the tested set of cues, because an alternative classification strategy, using the same cues, may provide very different results (Smits and Ten Bosch, 1994c).

2. Perception of manipulated natural speech.
   Although it is generally impossible to manipulate the values of certain acoustic cues at will in natural speech, it is often possible to *delete* cues from utterances, or *exchange* them between utterances. As discussed in section 5.2, deleted-cue stimuli can be used to measure the necessity of the deleted cues and the sufficiency of the remaining cues. Conflicting-cue stimuli, on the other hand, can be used to measure the relative importance of the conflicting cues. Both types of experiments have the advantage that the natural variability of the speech signal is more or less preserved. The experiments have the major disadvantage, however, that the conclusions which can be drawn from the resulting perception data are generally very limited, pertaining only to the perceptual relevance of a cue (see chapter 5). How the *value* of a cue, like the onset frequency of F2, influences perception can generally not be derived from this type of experiment. Another disadvantage of this method is that the recorded speech material will naturally contain certain speaker characteristics, and the measured perceptual effects may differ substantially for speech material of different speakers, even though the manipulations are identical (e.g. Dorman *et al.*, 1977).

3. Perception of synthetic speech.
   Compared to the previous class of experiments, perception experiments with synthetic stimuli naturally have the advantage that individual cue values can, in general, be manipulated at will. Thus, synthetic stimulus continua can be constructed by systematically varying the values of one or more acoustic parameters. After presenting the stimuli to listeners, the resulting identification functions provide detailed information concerning the use of cues in the classification process, for instance with respect to the position of class boundaries in the cue space. On the other hand, this type of experiment has the drawback that, due to stylization, the stimuli may deviate strongly from natural utterances, and important acoustic structures may be absent or distorted. Furthermore, the stimuli may sound unnatural, which may affect the listeners classification behavior.

In our study, we have introduced a novel type of phonetic experiment which combines elements of all three classes of experiments discussed above. As discussed in earlier chapters, our approach essentially consists of 3 steps:

1. Perception experiment with manipulated natural utterances;
2. Measurement of acoustic cues;
3. Mapping of acoustic cues onto observed perceptual responses.

Clearly, step 1 and 2 correspond to experimental classes 2 and 1, respectively. Step 3 is intended to provide information on the classification process which is of the type of experimental class 3, that is, a quantitative description of the influence of cue values on phoneme perception.

It is our opinion that the experimental methodology used in the present research provides a new means to investigate various aspects of speech perception. Moreover, the general approach can be used to study any kind of human classification behavior, like the recognition of written language, visual recognition of objects, etc. The method can be generalized to the following 3 steps:

1. Perceptual classification experiment with manipulated natural stimuli;
2. Measurement of physical quantities on the stimuli;
3. Mapping of physical data onto categorical perceptual data using a formal model of human classification behavior.

7.2.2 Gained phonetic insight

With respect to our primary research question, we have found that detailed cues give a better account of the perceptual data than the the gross cues in all cases except for the burst-only stimuli. Furthermore, we have derived a number of linear functions of acoustic cues which correspond with the linguistic distinctions, as perceived by the listeners. We venture to speculate as follows. First of all, it seems to be the case that the formants, especially F2, have been explicitly used by the listeners. All of the best model fits contain formant information. Secondly, concerning the locus equations we are somewhat hesitant to draw conclusions on their explicit role in speech perception. For the unvoiced stops /p, t, k/, the locus-equation distances often outperformed the raw formant data. We speculate that it may be the case that the raw formant data are explicitly used by the listeners, but in a more complex way than our model is able to simulate. For instance, the perceptual system may be able to cope with the high acoustic variability of F2 and F3 for the velar category by employing non-linear class boundaries, or even disjunct response regions for allophonic variations of /k/. Our model cannot reproduce such strategies, hence the somewhat lower GOF-levels. However, the conversion of raw formant data into distances to locus-equations may provide a useful "preprocessing" of the formant data, in the sense that the distances are more or less linearly separable, and thus more suitable for our model. This speculation is supported by the observation that for the voiced stops, where the highly context-dependent velar class is absent, the locus equation preprocessing does not provide this advantage. Here, the labial and

dental class are more or less linearly separable using raw formant data, and in the simulations the raw formant data systematically outperform the locus equation data.

7.3 Limitations of our study

In this section, some of the limitations of our study are briefly discussed. First of all, as we have noted in the previous chapter, the results of our simulations are based upon a number of assumptions, which may be less than optimal. The two most important assumptions are:

1. The specific a-priori choice of acoustic cues, which may be particularly relevant for the gross cues. Although we have attempted to capture the most important gross spectro-temporal properties discussed in the literature, it may be the case that certain perceptually relevant structures have not been incorporated in our cues.

2. In our SLP-based model, certain assumptions are made. Class boundaries are assumed to be linear, and the shape of the probability "landscape" is dictated by the sigmoid function. As mentioned in the previous section, it may well be the case that these assumptions do not hold, especially when the velar class is involved. All SLP-model estimations presented in this thesis have also been carried out with the TLP, that is, using a hidden layer. However, this more powerful model only rarely produced higher GOF-levels, than for the SLP. This may, however, been caused by the fact that the number of data was rather limited, which prevents a generalizable estimation of a model with a large number of parameters.

Beside the assumptions made in the simulations, there is the issue of generalizability across speakers. The stop-vowel utterances used in this study were produced by two male speakers. We have simulated the classification behavior of the listeners as closely as possible, that is, fitting a model separately per experimental session. Preliminary experiments have shown that using the model for one speaker to predict the perceptual data for the other speaker results in rather low scores. It is, however, important to realize that it would indeed be surprising if the scores would be high. As discussed in the introductory chapter, listeners adapt to one particular speaker. Mullennix and Pisoni (1990) have shown that using several speakers within one experimental session affects the listeners' classification behavior. Fitting one model to both speakers' data would in fact be a simulation of a perception experiment in which stimuli of the two speakers would be randomly mixed in one session. In that case, however, the perceptual data would have been significantly different.

How then do our results generalize to the perception of stop-vowel utterances of new speakers? At present, this issue remains unclear. Beside the often-discussed vocal-tract normalization, very little is actually known about the process of the listeners' adaptation to individual speakers. When discussing our general model in the introductory chapter, we have suggested that speaker adaptation may take

place at the level of cue measurement - cues may be measured differently for different speakers - as well as in the classifier, for example by shifting class boundaries (altering the SLP's biases), or by rotating the class boundaries (weighing various cues differently for different speakers). Clearly, there is ample room for additional research on the problem of speaker adaptation in consonant perception.

7.4 Proposal for an additional experiment

In this section, we will propose an experiment which does not suffer from two of the basic limitations mentioned in the previous section, namely making a-priori assumptions on the details of the cue-measurement procedures, as well as on the nature of the classification process. Rather than measuring the perceptual relevance of a set of predefined detailed and gross cues, the proposed experiment aims to measure the perceptual relevance of various *levels of detail* in the spectrogram. The approach is as follows. Like in our experiments, natural utterances are used, and from these utterances conflicting-cue stimuli are constructed. However, instead of splicing time segments of conflicting utterances together, we will construct signals by combining detailed time-frequency from one utterance with gross time-frequency information from another utterance.

Technically, this can be implemented as follows. First of all, two conflicting utterances, e.g. /pa/ and /ta/, are recorded and the instant of burst onset is determined for both signals. Next, a short-time Fourier transform (STFT, Rabiner and Shafer, 1978) is performed, with the first window centered at the instant of burst onset. The STFT is carried out with a fixed window length of, say, 6 ms (roughly corresponding to the wideband spectrogram) and fixed - preferably small - window shift. This transformation yields a spectro-temporal representation for each utterance, which consists of a short-time Fourier amplitude (STFA) and phase (STFP). We concentrate on the STFA, as this corresponds to the wideband spectrogram which is generally thought to contain nearly all of the perceptually relevant information. The STFA of each utterance is first converted into an "auditory spectrogram" by warping the frequency axis as described in section 6.2.2, and transforming the linear amplitudes to a dB-scale using Eq. –2. The resulting representation, indicated by STFA', is subdivided into a detailed and a gross part by using a two-dimensional filtering technique. To this end, we perform a low-pass filtering through a convolution of the spectral levels along the time axis (temporal smoothing) as well as along the frequency axis (spectral smoothing). With this filtering operation, a time constant and a frequency constant is associated which represents the effective window length of the convolution window along the time axis and the frequency (ERB) axis. The resulting smoothed version of STFA' is called the *gross* part of STFA', and will be indicated by STFA'_g. Subtracting STFA'_g from STFA' yields the *detailed* part of STFA', indicated by STFA'_d. After the STFA's of both utterances have been subdivided into gross and detailed parts using identical smoothing constants, the gross part of utterance 1 and the detailed part of utterance 2 are summed, and vice versa. Finally, these representations are to be transformed into waveforms. First, the inverse operations of the frequency and amplitude warping

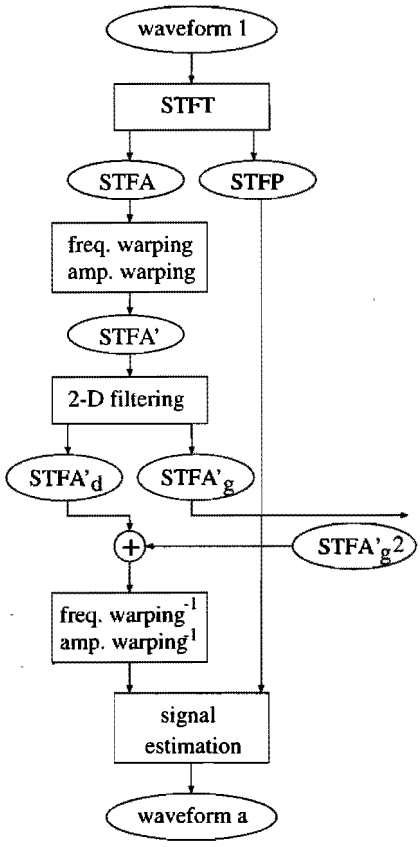

Figure 7.1: Flow diagram for the signal processing procedures for the proposed experiment. Ovals represent signal representations, squares represent processing steps.

are performed. Next, each of the STFAs is coupled with one of the original, unmanipulated STFPs in order to perform the inverse STFT. Fundamentally, it seems preferable to use the STFP which corresponds to the detailed STFA, because the short-time Fourier phase spectrum has been shown to contain information on details along the frequency axis (Yegnanarayana, 1978), as well as along the time axis (Smits and Yegnanarayana, 1995). Next, the amplitude-phase pairs are entered into the iterative signal estimation method from the STFA, using the iterative procedure described in Griffin and Lim (1984) and Veldhuis and He (1994). The comprehensive operation is summarized in Figure 7.1. The signal-processing procedure has been implemented and extensively described by Jonkers (1993) in a Master's thesis project.

The resulting signals contain detailed spectro-temporal information of one utterance and gross spectro-temporal information of another utterance. By creating

several such signals from two original utterances by varying the time-smoothing constant and the frequency-smoothing constant, a two-dimensional stimulus continuum can be created. Presenting such a continuum to listeners will result in a two-dimensional identification function. The boundary in the continuum at which the percept changes from one stop consonant to the other, or where the transition in the identification function is steepest, indicates the level of detail in the spectrogram which contains the perceptually most relevant information.

The procedure described above provides the means to create *conflicting-cue* stimuli in which the gross spectro-temporal information cues one place of articulation, while the detailed spectro-temporal information cues another. The *deleted-cue* version of this approach has been carried out by Ter Keurs *et al.* (1992, 1993a, 1993b), Drullman (1995), and Drullman *et al.* (1994a, 1994b). Ter Keurs has studied the effect of spectral smearing on speech reception. Her results show that intelligibility of sentences is not impaired for spectral smearing up to 1/3 octave. With smearing over 1/3 octave, intelligibility is increasingly reduced. Furthermore, for consonants spectral smearing most strongly affects perception of place of articulation. Drullman studied the effect of reducing slow or fast modulations in the speech signal. His results show that intelligibility of sentences is not impaired when temporal fluctuations higher than 16 Hz are canceled. When slower fluctuations are deleted, intelligibility decreases. Decreasing slow fluctuations, while preserving rapid fluctuations does not affect sentence intelligibility for cut-off frequencies up to 4 Hz. Stop consonants appear to be mostly confused with either fricative consonants or glides. Place-of-articulation confusions also occur often.

Unfortunately, information concerning the perceptually most relevant level of detail in the spectro-temporal representation of stop consonants can hardly be deduced from the data of these studies. The reason for this is that the stimulus sets as well as response sets contained nearly all Dutch consonants, and apart from the major consonantal confusions mentioned above, the confusion matrices did not show very clear patterns.

References

Aarts, E.H.L., and Korst, J. (**1989**) *Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing.* (Wiley, Chichester).

Agresti, A. (**1990**) *Categorical data analysis.* (Wiley, New York).

Ainsworth, W.A. (**1968**) "Perception of stop consonants in synthetic CV syllables," Language and Speech **11**, 139–155.

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (**1989**) *Statistical modeling in GLIM* (Clarendon Press, Oxford).

Ashby, F.G. (**1992**) "Multidimensional models of categorization," In: F.G. Ashby (Ed.), *Multidimensional models of perception and cognition.* (Lawrence Erlbaum, Hillsdale, New Jersey).

Ashby, F.G., and Lee, W.W. (**1991**) "Predicting similarity and categorization from identification," Journal of Experimental Psychology: General **120**, 150–172.

Ashby, F.G., and Maddox, W.T. (**1993**) "Relations between prototype, exemplar and decision bound models of categorization," Journal of Mathematical Psychology **37**, 372–400.

Ashby, F.G., and Perrin, N.A. (**1988**) "Toward a unified theory of similarity and recognition," Psychological Review **95**, 124–150.

Atlas, L.E., Kooiman, W., and Loughlin, P.J. (**1990**) "New nonstationary techniques for the analysis and display of speech transients," Proc. Int. Conf. on Acoustics, Speech and Signal Processing **1**, 385–388.

Atlas, L.E., Loughlin, P.J., and Pitton, J.W. (**1991**) "Truly nonstationary techniques for the analysis and display of voiced speech," Proc. Int. Conf. on Acoustics, Speech and Signal Processing **1**, 433–436.

Atlas, L.E., Loughlin, P.J., and Pitton, J.W. (**1992**) "Signal analysis with cone kernel time-frequency distributions and their applications to speech," In: Boashash (Ed.), *Time-frequency signal analysis, methods and applications.* (Longman Cheshire, Melbourne).

Barry, W.J. (**1984**) "Place-of-articulation information in the closure voicing of plosives," J. Acoust. Soc. Am. **76**, 1245–1247.

Blumstein, S.E., Isaacs, E., and Mertus, J. (**1982**) "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **72**, 43–50.

Blumstein, S.E., and Stevens, K.N. (**1979**) "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.

Blumstein, S.E., and Stevens, K.N. (**1980**) "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Blumstein, S.E., and Stevens, K.N. (**1981**) "Phonetic features and acoustic invariance in speech," Cognition **10**, 25–32.

Boashash, B. (Ed.) (**1992**) *Time-frequency signal analysis, methods and applications* (Longman Cheshire, Melbourne).

ten Bosch, L., and Smits, R. (**1994**) "On error criteria in perception modeling," In preparation.

Casacuberta, F., and Vidal, E. (**1987**) "A nonstationary model for the analysis of transient speech signals," IEEE Trans. on Acoustics, Speech and Signal Processing **35**, 226–228.

Choi, H.-I., and Williams, W.J. (**1989**) "Improved time-frequency representation of multicomponent signals using exponential kernels," IEEE Trans. on Acoustics, Speech and Signal Processing **37**, 862–871.

Chomsky, N., and Halle, M. (**1968**) *The sound pattern of English.* (Harper and Row, New York).

Cohen, L., and Pickover, C.A. (**1986**) "A comparison of joint time-frequency distributions for speech signals," IEEE Proc. Int. Symp. on Circuits and Systems **1**, 42–45.

Cohen, L. (**1989**) "Time-frequency distributions - A review," Proc. of the IEEE **77**, 941–981.

Cole, R.A., and Scott, B. (**1974**) "Toward a theory of speech perception," Psychological Review **81**, 348–374.

Cole, R.A., and Scott, B. (**1974**) "The phantom in the phoneme: Invariant cues for stop consonants," Perception & Psychophysics **15**, 101–107.

Cooper, F.S. (1950) "Research on reading machines for the blind," In: P. Zahl (Ed.), *Blindness: Modern approaches to the unseen environment.* (Princeton University Press, Princeton, N.J.), 512–543.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J. (**1952**) "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.

Crystal, T.H., and House, A.S. (**1988**) "The duration of American-English stop consonants: an overview," J. of Phonetics **16**, 285–294.

Darwin, C.J., and Gardner, R.B. (**1987**) "Perceptual separation of speech from concurrent sounds," In: M. Schouten (Ed.) *The psychophysics of speech perception.* (Martinus Nijhoff, Dordrecht).

Davis, K., and Kuhl, P.K. (**1992**) "Best exemplars of English velar stops: a first report," Proc. Int. Conf. on Spoken Language Processing, 495–498.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (**1955**) "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Delgutte, B. (**1986**) "Analysis of French stop consonants using a model of the peripheral auditory system," In: J. Perkell and D. Klatt (Eds.), *Invariance and variability in speech processes*, (Lawrence Erlbaum, Hillsdale, NJ), 163–177.

Dogil, G., and Wokurek, W. (**1989**) "Wigner distribution analysis of stop conso-

nant release transients labials, velars and labio-velars," Proc. of the Speech Research '89 International Conference in Budapest, 1–4.

Dogil, G., and Wokurek, W. (**1991**) "Wigner time-frequency representation for major places of articulation in stop consonants," Proc. 12th Int. Congress of Phonetic Sciences **3**, 390–393.

Dorman, M.F., Studdert-Kennedy, M., and Raphael, L.J. (**1977**) "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Perception & Psychophysics **22**, 109–122.

Drullman, R. (**1995a**) "Temporal envelope and finestructure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

Drullman, R. (**1995b**) "Intelligibility of temporally degraded speech: A study on the significance of narrowband temporal envelopes," Doctoral dissertation, Free University of Amsterdam.

Drullman, R., Festen, J.M., and Plomp, R. (**1994a**) "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Drullman, R., Festen, J.M., and Plomp, R. (**1994b**) "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

Duquesnoy, A.J., and Plomp, R. (**1980**) "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," J. Acoust. Soc. Am. **68**, 537–544.

Edwards, T.J. (**1981**) "Multiple features analysis of intervocalic English plosives," J. Acoust. Soc. Am. **69**, 535–547.

Efron, B. (**1982**) *The Jackknife, the Bootstrap and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Elman, J.L., and McClelland, J.L. (**1986**) "Exploiting lawful variability in the speech wave," In: J. Perkell and D. Klatt (Eds.), *Invariance and variability in speech processes*, (Lawrence Erlbaum, Hillsdale, NJ), 360–380.

Fant, G. (**1960**) *Acoustic theory of speech production.* (Mouton, The Hague).

Fant, G. (**1973**) *Speech sounds and features.* (MIT Press, Cambridge, MA).

Fischer-Jørgensen, E. (**1954**) "Acoustic analysis of stop consonants," Miscellenea Phonetica **2**,42–59.

Fischer-Jørgensen, E. (**1972**) "Tape-cutting experiments with Danish stop consonants in initial position," Annu. Rep. Inst. Phon., Univ. Copenhagen **6**, 104–168.

Flanagan, J. (**1972**) *Speech analysis synthesis and perception* (Springer, New York, 2nd edition).

Forrest, K., Weismer, G., Milenkovic, P., Dougall, R.N. (**1988**) "Statistical analysis of word-initial voiceless obstruents: Preliminary data," J. Acoust. Soc. Am. **84**, 115–123.

Fowler, C.A. (**1986**) "An event approach to the study of speech perception from a direct-realist approach," J. of Phonetics **14**, 3–28.

Fox, R.A., and Feth, L.L. (**1992**) "Identification of initial stop consonants processed by the Patterson-Holdsworth ASP model," In: Y. Cazals and K. Horner (Eds.), *Auditory physiology and perception.* (Pergamon Press, Oxford).

Fukunaga, K. (**1972**) *Introduction to statistical pattern recognition.* (Academic

Press, New York).

Fukunaga, K., and Kessell, D.L. (**1971**) "Estimation of classification error," IEEE Transactions on Computers **20**, 1521–1527.

Garudadri, H., Gilbert, J.H.V., Benguerel, A.P., and Beddoes, M.P. (**1987**) "Invariant acoustic cues in stop consonants: A cross-language study using the Wigner distribution," J. Acoust. Soc. Am. **82**, S55.

Ghitza, O. (**1993**) "Processing of spoken CVCs in the auditory periphery. I. Psychophysics," J. Acoust. Soc. Am. **94**, 2507–2516.

Gill, P.E., and Murray, W. (**1976**) *Minimization subject to bounds on the variables.* National Physical Laboratory report NAC 72.

Glasberg, B.R., and Moore, B.C.J. (**1990**) "Derivation of auditory filter shapes from notched-noise data," Hearing Research **47**, 103–138.

Grenier, Y. (**1983**) "Time-dependent ARMA modeling of nonstationary signals," IEEE Trans. on Acoustics, Speech and Signal Processing **31**, 899–911.

Griffin, D.W., and Lim, J.S. (**1984**) "Signal estimation from modified short-time Fourier transform," IEEE Trans. on Acoustics, Speech and Signal Processing **32**, 236–243.

Halle, M., Hughes, G.W., and Radley, J.-P.A. (**1957**) "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107–116.

Harris, C.M. (**1953**) "A study of the building blocks in speech," J. Acoust. Soc. Am. **25**, 962–969.

Harris, F.J. (**1978**) "On the use of windows for harmonic analysis with the discrete Fourier transform," Proc. of the IEEE **66**, 51–83.

Haykin, S. (**1994**) *Neural networks – A comprehensive Foundation.* (Macmillan College Publishing Company, New York)

Hays, W.L. (**1994**) *Statistics* (Holt, Rinehart and Winston, Plymouth UK).

Henton, C., Ladefoged, P., and Maddieson, I. (**1992**) "Stops in the world's languages," Phonetica **49**, 65–101.

Hertz, J., Krogh, A., and Palmer, R.G. (**1991**) *Introduction to the theory of neural computation.* (Addison-Wesley, Redwood City).

Hlawatsch, F., and Boudreaux-Bartels, G.F. (**1992**) "Linear and quadratic time-frequency signal representations," IEEE Signal Proc. Magazine, 21–67.

Hoffman, H.S. (**1958**) "Study of some cues in the perception of the voiced stop consonants," J. Acoust. Soc. Am. **30**, 1035–1041.

Hornik, K., Stinchcombe, H., and White, H. (**1989**) "Multilayer feedforward networks are universal approximators," Neural Networks **2**, 359–366.

Howitt, A.W. (**1991**) "Application of the Wigner Distribution to speech analysis," *Speech communication group working papers VII, MIT, Res. Lab. of Electr.*, 23–46.

Jakobson, R., Fant, G., and Halle, M. (**1952**) *Preliminaries to speech analysis.* (MIT Press, Cambridge, MA).

Jongman, A., Blumstein, S.E., and Lahiri, A. (**1985**) "Acoustic properties for dental and alveolar stop consonants: a cross language study," J. of Phonetics **13**, 235–251.

Jongman, A., and Miller, J.D. (**1991**) "Method for the location of burst-onset

spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants," J. Acoust. Soc. Am. **89**, 867–873.

Jonkers, H.G. (**1994**) *Designing a software tool for research on the perception of speech*, unpublished Master's Thesis, IPO-Report 965.

Keating, P., and Lahiri, A. (**1993**) "Fronted velars, palatalized velars and palatals," Phonetica **50**, 73–101.

Keating, P. A., Byrd, D., Flemming, E., and Todaka, Y. (**1994**) "Phonetic analyses of word and segment variation using the TIMIT corpus of American English," Speech Communication **14**, 131–142.

ter Keurs, M. (**1992**) "Intelligibility of spectrally smeared speech," Doctoral dissertation, Free University of Amsterdam.

ter Keurs, M., Festen, J.M., and Plomp, R. (**1992**) "Effects of spectral smearing on speech reception I," J. Acoust. Soc. Am. **91**, 2872–2880.

ter Keurs, M., Festen, J.M., and Plomp, R. (**1993a**) "Effects of spectral smearing on speech reception II," J. Acoust. Soc. Am. **93**, 1547–1552.

ter Keurs, M., Festen, J.M., and Plomp, R. (**1993b**) "Limited resolution of spectral contrasts and hearing loss for speech in noise" J. Acoust. Soc. Am. **94**, 1307–1314.

Kewley-Port, D. (**1982**) "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. **72**, 379–389.

Kewley-Port, D. (**1983**) "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.

Kewley-Port, D., and Luce, P.A. (**1984**) "Time-varying features of initial stop consonants in auditory running spectra: A first report," Perception & Psychophysics **35**, 353–360.

Kewley-Port, D., Pisoni, D.B., and Studdert-Kennedy, M. (**1983**) "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **73**, 1779–1793.

Klatt, D.H. (**1989**) "Review of selected models of speech perception," In: W. Marslen-Wilson (Ed.), *Lexical representation and process*, (MIT Press, Cambridge, MA), 169–226.

Kodera, K., Gendrin, R., and De Villedary, C. (**1978**) "Analysis of time-varying signals with small BT values," IEEE Trans. on Acoustics, Speech and Signal Processing **26**, 64–76.

Krull, D. (**1990**) "Relating acoustic properties to perceptual responses: A study of Swedish voiced stops," J. Acoust. Soc. Am. **88**, 2557–2570.

Kruskal, J.B. (**1964**) "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**, 1–27.

Kuhl, P.K. (**1992**) "Speech prototypes: Studies on the nature, function, ontogeny and phylogeny of the "centers" of speech categories," In: Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure*, (IOS Press, Amsterdam), 239–264.

Lahiri, A., Gewirth, L., and Blumstein, S.E. (**1984**) "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391–404.

Lahiri, A., and Marslen-Wilson, W. (**1991**) "The mental representation of lexical form: A phonological approach to the recognition lexicon," Cognition **38**, 245–294.

LaRiviere, C., Winitz, H., and Herriman, E. (**1975**) "Vocalic transitions in the perception of voiceless initial stops," J. Acoust. Soc. Am. **57**, 470–475.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (**1967**) "Perception of the speech code," Psychological Review **74**, 431–461.

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (**1954**) "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychological Monographs **68**, 1–13.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F.S. (**1959**) "Minimal rules for synthesizing speech," J. Acoust. Soc. Am. **31**, 1490–1499.

Liberman, A.M., and Mattingly, I.G. (**1985**) "The motor theory of speech perception revised," Cognition **21**, 1–36.

Lindblom, B. (**1961**) "Accuracy and limitations of sona-graph measurements," Proc. of the 4th Int. Congress of Phonetic Sciences, 188–202.

Lindblom, B. (**1986**) "On the origin and purpose of discreteness and invariance in sound patterns," In: J. Perkell and D. Klatt (Eds.), *Invariance and variability in speech processes*, (Lawrence Erlbaum, Hillsdale, NJ), 493–510.

Lindholm, J.M., Dorman, M., Taylor, B.E., Hannley, M.T. (**1988**) "Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **83**, 1608–1614.

Lippman, R.P. (**1987**) "An introduction to computing with neural nets," IEEE ASSP Magazine **4**, 4–22.

Liu, S. (**1993**) "Locating landmarks in utterances for speech recognition," J. Acoust. Soc. Am. **93**, 2320.

Loughlin, P.J., Atlas, L.E., and Pitton, J.W. (**1992**) "Advanced time-frequency representations for speech processing," In: M. Cooke, S. Beet, and M. Crawford (Eds.) *Visual representations of speech signals*, (Wiley, Chichester), 27–53.

Luce, R.D. (**1963**) "Detection and recognition," In: R.D. Luce, R.R. Bush, and S.E. Galanter (Eds.), *Handbook of mathematical psychology, vol. 1, ch. 3*, (Wiley, New York).

Mack, M., and Blumstein, S.E. (**1983**) "Further evidence of acoustic invariance in speech production: The stop-glide contrast," J. Acoust. Soc. Am. **73**, 1739–1750.

Massaro, D.W. (**1988**) "Some criticisms of connectionist models of human performance," Journal of Memory and Language **27**, 213–234.

Massaro, D.W., and Friedman, D. (**1990**) "Models of integration given multiple sources of information," Psychological Review **97**, 225–252.

Massaro, D.W., and Oden, G.C. (**1980**) "Evaluation and integration of acoustic features in speech perception," J. Acoust. Soc. Am. **67**, 996–1013.

McClelland, J.L., Rumelhart, D.E., and the PDP research group (**1986**) *Parallel distributed processing: Explorations in the microstructure of cognition.* (MIT press, Cambridge, MA).

Medin, D.L., and Schaffer, M.M. (**1978**) "Context theory of classification learning," Psychological Review **85**, 207–238.

Miller, G.A., and Nicely, P.E. (**1955**) "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Miller, J.L. (**1981**) "Some effects of speaking rate on phonetic perception," Phonetica **38**, 159–180.

Minsky, M., and Papert, S. (**1969**) *Perceptrons: An introduction to computational geometry.* (MIT Press, Cambridge, MA).

Monsen, R.B., and Engebretson, A.M. (**1983**) "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction," J. Speech and Hearing Res. **26**, 89–97.

Mullennix, J.W., and Pisoni, D.B. (**1990**) "Stimulus variability and processing dependencies in speech perception," Perception & Psychophysics **47**, 379–390.

Nathan, K.S., Lee, Y.-T., and Silverman, H.F. (**1991**) "A time-varying analysis method for rapid transitions in speech," IEEE Trans. Signal Processing **39**, 815–824.

Nathan, K.S., and Silverman, H.F. (**1991**) "Classification of stops based on formant transitions prior to release," Proc. Int. Conf. on Acoustics, Speech and Signal Processing **1**, 445–448.

Nosofsky, R.M. (**1986**) "Attention, similarity, and the identification-categorization relationship," Journal of Experimental Psychology: General **115**, 39–57.

Nossair, Z.B., and Zahorian, S.A. (**1991**) "Dynamical spectral features as acoustic correlates for initial stop consonants," J. Acoust. Soc. Am. **89**, 2978–2991.

Oden, G.C. (**1979**) "A fuzzy logical model of letter identification," J. of Exp. Psychology: Human Perception and Performance **5**, 336–352.

Oden, G.C., and Massaro, D.W. (**1978**) "Integration of featural information in speech perception," Psychological Review **85**, 172–191.

Ohde, R.N. (**1984**) "Fundamental frequency as an acoustic correlate of stop consonant voicing", J. Acoust. Soc. Am. **75**, 224–230.

Ohde, R.N., and Sharf, D.J. (**1977**) "Order effect of acoustic segments of VC and CV syllables on stop and vowel identification," J. of Speech and Hearing Research **20**, 543–554.

Ohde, R.N., and Sharf, D.J. (**1981**) "Stop identification from vocalic transition plus vowel segments of CV and VC syllables: A follow-up study," J. Acoust. Soc. Am. **69**, 297–300.

Ohde, R.N., and Stevens, K.N. (**1983**) "Effect of burst amplitude on the perception of stop consonant place of articulation," J. Acoust. Soc. Am. **74**, 706–714.

Öhman, S.E.G. (**1966**) "Coarticulation in VCV utterances: Spectrographic measurements," J. Acoust. Soc. Am. **39**, 151–168.

Perkell, J.S., and Klatt, D.H. (Eds.) (**1986**) *Invariance and variability in speech processes* (Lawrence Erlbaum, Hillsdale).

Pisoni, D.B. (**1992**) "Some comments on invariance, variability and perceptual normalization in speech perception," Proc. Int. Conf. on Spoken Language Processing, 587–590.

Pitton, J.W., Atlas, L.E., and Loughlin, P.J. (**1994**) "Applications

of positive time-frequency distributions to speech processing," IEEE Trans. on Speech and Audio Processing **2**, 554–566.

Pols, L.C.W. (**1979**) "Coarticulation and the identification of initial and final plosives," In: J. Wolff and D. Klatt (Eds.), *ASA 50 Speech Communication Papers*, (Acoust. Soc. Am., New York), 459–562.

Pols, L.C.W., and Schouten, M.E.H. (**1978**) "Identification of deleted consonants," J. Acoust. Soc. Am. **64**, 1333–1337.

Pols, L.C.W., and Schouten, M.E.H. (**1981**) "Identification of deleted plosives: The effect of adding noise or applying a time window (A reply to Ohde and Sharf)," J. Acoust. Soc. Am. **69**, 301–303.

Quinlan, P. (**1991**) *Connectionism and Psychology.* (Harvester Wheatsheaf, Hemel Hempstead).

Rabiner, L.R. and Schafer, R.W. (**1978**) *Digital processing of speech signals* (Prentice Hall, Englewood Cliffs).

Rand, T.C. (**1971**) "Vocal tract size normalization in the perception of stop consonants," Haskins Laboratories Status Report on Speech Research **SR 25,26**, 141–146.

Repp, B.H., and Lin, H.B. (**1988**) "Acoustic properties and perception of stop consonant release transients," J. Acoust. Soc. Am. **85**, 379–396.

Riley, M.D. (**1989**) *Speech time-frequency representations* (Kluwer Academic Publishers, Norwell).

Saleh, B.E.A., and Subotic, N.S. (**1985**) "Time-variant filtering of signals in the mixed time-frequency domain," IEEE Trans. on Acoustics, Speech and Signal Processing **33**, 1479–1485.

Schatz, C.D. (**1954**) "The role of context in the perception of stops," Language **30**, 47–56.

Schouten, M.E.H., and Pols, L.C.W. (**1981**) "Consonant loci: A spectral study of coarticulation. Part III," J. of Phonetics **9**, 225–231.

Schouten, M.E.H., and Pols, L.C.W. (**1983**) "Perception of plosive consonants - The relative contributions of bursts and vocalic transitions," In: M. van den Broecke, V. van Heuven, and W. Zonneveld (Eds.), *Sound structures: Studies for Antonie Cohen*, (Foris, Dordrecht), 227–243.

Searle, C.L., Jacobson, J.Z., and Rayment, S.G. (**1979**) "Stop consonant discrimination based on human audition," J. Acoust. Soc. Am. **65**, 799–809.

Seneff, S. (**1988**) "A joint synchrony/mean-rate model of auditory speech processing," J. of Phonetics **16**, 55–76.

Shepard, R.N. (**1958**) "Stimulus and response generalization: tests of a model relating generalization to distance in psychological space," Journal of Experimental Psychology **55**, 509–523.

Silverman, H.F., and Lee, Y.-T. (**1987**) "On the spectrographic representation of rapidly time-varying speech," Computer Speech and Language **2**, 63–86.

Slis, I.H., and Cohen, A. (**1969**) "On the complex regulating the voiced-voiceless distinction I," Language and Speech **12**, 80–102.

Smits, R. (**1994**) "Accuracy of quasistationary analysis of highly dynamic speech

signals," J. Acoust. Soc. Am. **96**, 3401–3415.

Smits, R. and ten Bosch, L. (**1994a**) "The multi-layer perceptron as a model of human categorization behavior. I. Theory," Submitted to J. Math. Psychology.

Smits, R. and ten Bosch, L. (**1994b**) "The multi-layer perceptron as a model of human categorization behavior. II. Practical aspects," Submitted to J. Math. Psychology.

Smits, R. and ten Bosch, L. (**1994c**) "A note on classification experiments in acoustic phonetics," Submitted to J. of Phonetics.

Smits, R., ten Bosch, L., and Collier, R. (**1995a**) "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants: I. Perception experiment," Submitted to J. Acoust. Soc. Am.

Smits, R., ten Bosch, L., and Collier, R. (**1995b**) "Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants: II. Modeling and evaluation," Submitted to J. Acoust. Soc. Am.

Smits, R., and Yegnanarayana, B. (**1995**) "Determination of instants of significant excitation in speech using group delay function," Accepted for IEEE Trans. on Speech and Audio Processing.

Sommers, M.S., Nygaard, L.C., and Pisoni, D.B. (**1992**) "The effects of speaking rate and amplitude variability on perceptual identification," J. Acoust. Soc. Am. **91**, 2SP11, 2340.

Stevens, K.N. (**1980**) "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Am. **68**, 836–842.

Stevens, K.N. (**1994**) "Phonetic evidence for hierarchies of features," In: P.A. Keating (Ed.), *Phonological structure and phonetic form. Papers in laboratory phonology III*, (Cambridge University Press, Cambridge, UK).

Stevens, K.N., and Blumstein, S.E. (**1978**) "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

Stevens, K.N., and Blumstein, S.E. (**1981**) "The search for invariant acoustic correlates of phonetic features," In: P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the study of speech*, (Lawrence Erlbaum, Hillsdale, NJ), 1–39.

Stevens, K.N., Manuel, S.Y., Shattuck-Hufnagel, S., and Liu, S. (**1992**) "Implementation of a model for lexical access based on features," Proc. Int. Conf. on Spoken Language Processing, 499–502.

Suomi, K. (**1985**) "The vowel dependence of gross spectral cues to place of articulation of stop consonants in CV syllables," J. of Phonetics **13**, 267–285.

Suomi, K. (**1987**) "On spectral coarticulation in stop-vowel-stop syllables: implications for automatic speech recognition," J. of Phonetics **15**, 85–100.

Sussman, H.M. (**1991**) "The representation of stop consonants in three-dimensional acoustic space," Phonetica **48**, 18–31.

Sussman, H.M., McCaffrey, H.A. and Matthews, S.A. (**1991**) "An investigation of locus equations as a source of relational invariance for stop place of articulation," J. Acoust. Soc. Am. **90**, 1309–1325.

Tekieli, M.E., and Cullinan, W.L. (**1979**) "The perception of temporally segmented vowels and consonant-vowel syllables," J. of Speech and Hearing Re-

search **22**, 103–121.

Veldhuis, R.N.J., and He, H. (**1994**) "Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform," submitted to Speech Communication.

Velez, E.F., and Absher, R.G. (**1989**) "Transient analysis of speech signals using the Wigner time-frequency representation," Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 2242–2245.

Velez, E.F., and Garudadri, H. (**1992**) "Speech analysis based on smoothed Wigner-Ville distribution," In: Boashash (Ed.), *Time-frequency signal analysis, methods and applications.* (Longman Cheshire, Melbourne).

Walley, A.C., and Carrell, T.D. (**1983**) "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.

Weenink, D. (**1991**) "Aspects of neural nets," Proc. Inst. Phon. Sci. Amsterdam **15**, 1–25.

Wickelgren, W.A. (**1976**) "Phonetic coding and serial order," In: E. Carterette and M. Friedman (Eds.) *Handbook of perception, vol. 7: Language and Speech.* (Academic Press, New York).

van Wieringen, A. (**1995**) *Perceiveing dynamic speechlike sounds: psycho-acoustics and speech perception.* Doctoral dissertation, University of Amsterdam.

Wilks, S.S. (**1935**) "The likelihood test of independence in contingency tables," Annals of Mathematical Statistics **6**, 190–196.

Winitz, H., Scheib, M.E., and Reeds, J.A. (**1972**) "Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech," J. Acoust. Soc. Am. **51**, 1309–1317.

Wokurek, W. (**1991**) "Comments on 'On the spectrographic representation of rapidly time varying speech'," Computer Speech and Language **5**, 1–10.

Wokurek, W., Hlawatsch, F., and Kubin, W. (**1987**) "Wigner distribution analysis of speech signals," Proc. Int. Conf. on Digital Signal Processing, Florence, Italy, 294–298.

Yegnanarayana, B. (**1978**) "Formant extraction from linear prediction phase spectra," J. Acoust. Soc. Am. **63**, 1638–1640.

Zhao, Y., Atlas, L.E., and Marks, R.J. II (**1990**) "The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals," IEEE Trans. on Acoustics, Speech and Signal Processing **38**, 1084–1091.

Zue, V.W. (**1976**) *Acoustic characteristics of stop consonants: a controlled study.* Tech. Rep. 523, Lincoln Lab, MIT, Cambridge, MA.

Summary

The perception of place of articulation of stop consonants constitutes a long-standing problem in phonetic research. During the last 50-odd years many acoustic structures have been described which correlate with phonetic distinctions and/or which demonstrably influence the perceived place of articulation when they are artificially manipulated. Examples of such phonetically relevant acoustic structures, or *cues*, are the frequency of the second formant at voicing onset and the global spectral tilt at consonantal release. Despite the considerable research effort devoted to the problem of stop-consonant perception many issues remain unsolved, such as the question which cues are perceptually most important and how the various cues are actually combined by the perceptual system in order to reach a linguistic classification. Particularly for stop consonants the individual cues show a large variability depending on the phonetic context, and no single acoustic property has been shown to entertain a one-to-one relationship with a perceived linguistic class.

In the research on acoustic cues for place of articulation of stop consonants, the issue whether *detailed* or *gross* spectro-temporal structures are perceptually most important has received increasing attention in recent years. For the purpose of this study, detailed cues are defined as acoustic structures which result from measurements with a relatively high resolution in time or frequency, such as the length of the release burst or formant frequencies. Gross cues, on the other hand, result from acoustic measurements with a relatively low resolution, like the global spectral tilt, or the existence of a broad spectral prominence.

The research presented in this thesis investigates the relative importance of detailed and gross acoustic structures for the perception of place of articulation in prevocalic stop consonants. In addition, the research aims to model how the most relevant cues are combined and mapped onto the linguistic classes by the perceptual system. Two methodological aspects are emphasized in our study. First of all, we use (manipulated) natural utterances in our experiments in order to preserve the natural variability in the speech signal. Secondly, we make a complete simulation of the behavior of listeners during a stop-consonant classification task. The simulation is complete in the sense that it includes the basic processing steps of initial signal representation, extraction and combination of cues, and classification.

In our investigation a three-step paradigm is adopted. First, a perception experiment is carried out in which manipulated natural stop-vowel utterances are presented to listeners for classification. This first step results in a body of perceptual data, organized in a stimulus-response matrix. The second and third step constitute the simulation of the listeners' behavior in the experiment. The second

step is the measurement of a number of detailed and gross spectro-temporal cues on the stimuli, resulting in a body of acoustical data. The third step is the mapping of the acoustical data onto the perceptual data using a formal model of human classification behavior. A comparison of the resulting levels of goodness-of-fit for the detailed cues and for the gross cues indicates which of the two gives the better account of the observed behavior of the listeners. In addition, an interpretation of the best-fitting models leads to a mathematical approximation of the cue-combination and response-selection strategies used by the listeners.

Before the paradigm described above can be applied, a number of methodological hurdles are taken. Firstly, chapter 2 addresses the issue whether detailed spectro-temporal structures, such as formant frequencies, can be measured accurately enough when the signal spectrum changes very rapidly, as is the case for stop consonants. In particular, the accuracy is assessed of the two most widely used speech analysis tools - the spectrogram and Linear Prediction - for the measurement of formant frequencies in stop consonants. Analysis of various dynamic signals shows that when a long analysis window, like 25 ms, is used, the quality of the representation may be impoverished. A number of unwanted effects may occur, such as staircase-like formant tracks, flattening-off of formants close to voicing onset, and bending of the formant towards a strong energy concentration in the release burst. The parameters that have the largest influence on the quality of the representation are the length of the analysis window, the transition rate of the formant, the fundamental frequency, and the position and energy of the release burst. It is shown that the most accurate analysis using a quasi-stationary method is made when windows are positioned pitch-synchronously. A quantitative analysis of the influence of the afore-mentioned parameters provides evidence that no deviations due to the quasi-stationarity assumption occur when the effective length of the analysis window is not larger than the pitch period. Therefore, the wideband spectrogram is a reliable speech-analysis tool because it meets this condition for fundamental frequencies up to about 370 Hz.

The second methodological issue in this thesis is the formulation of an appropriate mathematical model which can be used for the mapping of the acoustic cues onto the observed perceptual responses. In chapter 3, a model of human categorization behavior is presented which is based on the multi-layer perceptron (MLP). First the modeling behavior of the single-layer perceptron (SLP) is studied through an analysis of the mathematical expressions and a discussion of a number of theoretical examples. A number of similarities and differences between the SLP and the well-known similarity-choice model (SCM) are discussed and it is shown that the SLP and SCM coincide in a certain limit case. Finally, the theory that we have developed for the SLP is extended to the two-layer perceptron (TLP). It is shown that, compared to the SLP, the TLP has substantially increased modeling power, but it can become hard to interpret. A linearization of the sigmoid functions in the hidden nodes is introduced, which facilitates interpretation.

In chapter 4, a number of practical methods are presented for estimating the parameters and the goodness-of-fit of the MLP-based model of human categorization behavior. A measure of goodness-of-fit is defined which is interpreted as a gener-

alized "percent correct score". The "leaving-one-out method", a cross-validation technique which is commonly used in the field of statistical pattern recognition, is adopted to estimate the generalizability of a model estimation. Finally, the methodology is illustrated by a practical example which deals with the perception of release bursts that are excised from prevocalic stop consonants. The danger of overfitting is demonstrated, and the best fitting model is interpreted using the theory presented in chapter 3.

The experimental treatment of the research questions formulated earlier is presented in chapters 5 and 6. First a perception experiment is carried out with "burst-spliced" stop-vowel utterances. This experiment is described in chapter 5. From a number of stop-vowel utterances, "burst-only", "burst-less" and "cross-spliced" stimuli are created and presented to listeners. The results of the experiment show that the relative perceptual importance of burst and transitions highly depends (1) on the stop consonants from which burst and transitions originate, (2) on the vowel context, and (3) on the voiced or unvoiced nature of the stop. Velar bursts are generally stronger in cueing place of articulation than are other bursts. The dental transitions appear to be weaker than labial or velar transitions. In front-vowel contexts the release burst dominates the perception of place of articulation, while in non-front vowel contexts the formant transitions are generally dominant. Finally, the bursts of unvoiced stops are perceptually more important than the bursts of voiced stops.

Chapter 6 describes the simulation of the listeners' behavior in the perception experiment presented in chapter 5. First, a number of detailed and gross spectro-temporal cues are measured on all stimuli. The set of detailed cues includes the frequencies of F2 and F3 at voicing onset and in the vowel and a number of release-burst parameters. The set of gross cues includes the global spectral tilt and the strength of the mid-frequency peak just after consonantal release, and the change of these measures into the stationary part of the utterance. Next, these cues are mapped onto the observed perceptual data using the formal model of human classification behavior introduced in chapters 3 and 4. The results of the model fits show that the detailed cues, such as formant transitions combined with certain burst parameters, give a better account of the perceptual data than the gross cues. The best-performing models are interpreted in terms of the acoustic boundaries which are associated with the perceived linguistic contrasts. These boundaries are linear functions of 5 or 6 acoustic cues, which are phonetically interpretable. The boundaries give a quantitative description of the often-discussed "trade-off" relation between the various cues for perception of place of articulation in stop consonants.

In the final chapter it is discussed how the research presented in this thesis may be valuable on two accounts. Firstly, with respect to our primary phonetic research question, we have found that detailed cues give a better account of the perceptual data than the the gross cues, and we have described the linguistic distinctions perceived by the listeners in terms of linear functions of acoustic cues. Secondly, we have developed an novel experimental paradigm for research on phoneme perception. The paradigm combines aspects of three traditional experimental approaches, viz. acoustic analysis, perception of manipulated natural utterances, and percep-

tion of synthetically prepared speech-like sounds. An important advantage of our paradigm over existing ones is the possibility to model the actual cue-measurement, cue-combination, and classification strategies of listeners via a perception experiment in which manipulated *natural utterances* are used.

Samenvatting

De perceptie van de plaats van articulatie van plofklanken vormt een oud probleem in het fonetisch onderzoek. Gedurende de laatste 50 jaar zijn een groot aantal akoestische structuren beschreven die correleren met fonetische distincties en/of aantoonbaar de waargenomen plaats van articulatie beïnvloeden als zij worden gemanipuleerd. Voorbeelden van dergelijke fonetisch relevante akoestische strukturen - ook wel *cues* genoemd - zijn de frekwentie van de 2e formant bij steminzet en de globale spectrale helling op het moment dat een afsluiting van het spraakkanaal wordt losgelaten. Ondanks dat er een grote hoeveelheid onderzoek is gewijd aan het probleem van de perceptie van plofklanken zijn er nog vele vragen onopgelost, zoals de vraag welke cues perceptief het belangrijkst zijn en hoe verschillende cues worden gecombineerd door het perceptieve systeem bij het maken van een linguïstische classificatie. De individuele cues vertonen, in het bijzonder voor plofklanken, een grote variabiliteit afhankelijk van de fonetische context, en er blijkt geen cue te bestaan die een één-op-één relatie vertoont met een waargenomen linguïstische klasse.

In het onderzoek aan akoestische cues voor plaats van articulatie van plofklanken is de vraag naar het relatieve belang van *fijnschalige* en *grofschalige* spectrotemporele structuren recentelijk steeds meer in de belangstelling komen te staan. Ten behoeve van ons onderzoek definiëren wij een fijnschalige cue als het resultaat van een akoestische meting met een relatief hoog oplossend vermogen in tijd of frekwentie, zoals de lengte van de ruisplof of formantfrekwenties. Een grofschalige cue wordt gedefinieerd als het resultaat van een akoestische meting met een relatief lage resolutie, zoals de globale spectrale helling, of de sterkte van een brede spectrale prominentie.

Het onderzoek beschreven in dit proefschrift heeft als doel om het relatieve belang vast te stellen van fijnschalige en grofschalige akoestische structuren voor de perceptie van plaats van articulatie in prevocalische plofklanken. Tevens wordt geprobeerd om te modelleren hoe de meest relevante cues door het perceptieve systeem worden gecombineerd en afgebeeld op de linguïstische klassen. In ons onderzoek worden twee methodologische aspecten benadrukt. Allereerst worden in de experimenten (gemanipuleerde) natuurlijke uitingen gebruikt om de natuurlijke variabiliteit van het spraaksignaal zo veel mogelijk te behouden. Ten tweede maken we een volledige simulatie van het gedrag van de luisteraars in een classificatie experiment van plofklanken. De simulatie is in die zin volledig, dat deze een aantal essentiële verwerkingsstappen bevat, te weten initiële signaalrepresentatie, extractie en combinatie van cues, en classificatie.

Het experimentele paradigma bestaat uit drie stappen. Eerst wordt een percep-

tie experiment uitgevoerd waarin gemanipuleerde natuurlijke uitingen voor classificatie worden aangeboden aan luisteraars. Deze eerste stap levert een stimulus-respons matrix op. De tweede en derde stap vormen de simulatie van het classificatiegedrag van de luisteraars in het perceptie experiment. De tweede stap is het meten van een aantal fijnschalige en grofschalige cues aan de stimuli, met als resultaat een hoeveelheid akoestische data. De derde stap is de afbeelding van de akoestische data op de perceptieve data met behulp van een formeel classificatiemodel. Een vergelijking van de resulterende niveaus van *goodness-of-fit* voor de fijnschalige en voor de grofschalige cues laat zien welke van de twee het beste het gedrag van de luisteraars verklaart. Bovendien leidt een interpretatie van de beste modellen tot een wiskundige benadering van de cue-combinatie strategieën en respons-selectie strategieën van de luisteraars.

Voordat de bovenstaande aanpak kan worden uitgevoerd moeten nog enkele methodologische hindernissen worden genomen. In hoofdstuk 2 wordt onderzocht of fijnschalige spectro-temporele structuren, zoals formantfrekwenties, nauwkeurig genoeg gemeten kunnen worden als het signaalspectrum zeer snel verandert, zoals bij plofklanken. In het bijzonder wordt de nauwkeurigheid bepaald voor de twee meest gebruikte spraakanalysegereedschappen - het spectrogram en lineaire predictie - voor het meten van formant frekwenties in plofklanken. Analyses van een aantal dynamische signalen laat zien dat de kwaliteit van de representatie kan verslechteren als een lang analysevenster (bijvoorbeeld 25 ms) wordt gebruikt. Een aantal ongewenste effekten kunnen dan optreden, zoals trapvormige formantsporen, afgeplatte formantsporen dicht bij het moment van steminzet, en het afbuigen van een formant naar een sterke energieconcentratie in de ruisplof. De parameters die de grootste invloed hebben op de kwaliteit van de representatie zijn de lengte van het analysevenster, de transitiesnelheid van de formant, de grondtoon van het spraaksignaal, en de energie en positie van de ruisplof. Ons onderzoek toont aan dat de meest nauwkeurige metingen kunnen worden gemaakt met quasi-stationaire methoden als de vensters "pitch-synchroon" worden gepositioneerd. Een kwantitatieve analyse van de invloed van eerder genoemde parameters laat zien dat er geen meetonnauwkeurigheden optreden t.g.v. de aanname van quasi-stationariteit als de effectieve lengte van het analysevenster niet groter is dan de pitch-periode. Het breedband-spectrogram is een betrouwbaar spraakanalysegereedschap omdat het aan deze voorwaarde voldoet voor grondfrekwenties tot 370 Hz.

Het tweede methodologische probleem dat wordt behandeld is het formuleren van een geschikt wiskundig model dat gebruikt kan worden voor het afbeelden van de akoestische cues op de perceptieve responsen. In hoofdstuk 3 wordt een model van menselijk classificatiegedrag gepresenteerd dat gebaseerd is op het *multi-layer perceptron* (MLP). Eerst wordt het modelgedrag van het *single-layer perceptron* (SLP) bestudeerd via de analyse van de relevante wiskundige expressies en het bespreken van een aantal theoretische voorbeelden. Een aantal overeenkomsten en verschillen tussen de SLP en het bekende *similarity-choice model* (SCM) worden besproken en het wordt aangetoond dat de SLP en de SCM in een bepaald limietgeval samenvallen. Tenslotte wordt de theorie die we hebben ontwikkeld voor de SLP uitgebreid naar het *two-layer perceptron* (TLP). We laten zien dat de TLP

een grotere modelleerkracht heeft dan de SLP, maar dat dit ten koste kan gaan van interpreteerbaarheid. We introduceren een linearisatie van de sigmoide functies in de *hidden nodes*, waardoor interpretatie wordt vergemakkelijkt.

In hoofdstuk 4 worden een aantal praktische methoden gepresenteerd waarmee de parameters en de *goodness-of-fit* van het classificatiemodel kunnen worden geschat. Er wordt een *goodness-of-fit*-maat gedefinieerd die geïnterpreteerd kan worden als een gegeneraliseerde percent-correct score. De *leaving-one-out* methode, een cross-validatie techniek die veel wordt gebruikt in het gebied van de statistische patroonherkenning, wordt aangepast voor het schatten van de generaliseerbaarheid van een model na training. Tenslotte wordt de methode geïllustreerd met een praktisch voorbeeld betreffende de perceptie van ruisplofjes die zijn geïsoleerd uit prevocalische plofklanken. Het gevaar voor *overfitting* wordt gedemonstreerd, en het best kloppende model wordt geïnterpreteerd met behulp van de theorie uit hoofdstuk 3.

De experimentele aanpak van de eerder geformuleerde onderzoekvraag wordt besproken in hoofdstukken 5 en 6. Allereerst wordt een perceptie-experiment uitgevoerd met zogenaamde *burst-spliced* plofklank-klinker uitingen. Dit experiment wordt beschreven in hoofdstuk 5. Uitgaande van een aantal plofklank-klinker uitingen worden stimuli gecreëerd die ofwel uitsluitend de ruisplof, ofwel de gehele uiting behalve de ruisplof, ofwel de ruisplof van de ene uiting en de rest van de andere uiting bevatten. Deze stimuli worden aangeboden aan luisteraars voor classificatie. De resultaten van het experiment laten zien dat het relatieve perceptieve belang van de ruisplof en de formanttransities sterk afhankelijk is (1) van de plofklank waarvan de plof en transities afkomstig zijn, (2) van de klinkercontext, en (3) van het stemhebbende of stemloze karakter van de plofklank. Velaire ruisplofjes bevatten over het algemeen sterkere cues voor plaats van articulatie dan de andere ruisplofjes. Dentale transities zijn zwakker dan labiale of velaire transities. De ruisplof domineert de waargenomen plaats van articulatie in de context van een voor-klinker, terwijl in andere klinkercontexten de formanttransities domineren. Tenslotte vinden we dat de ruisplofjes van stemloze plofklanken perceptief belangrijker zijn dan die van stemhebbende plofklanken.

Hoofdstuk 6 beschrijft de simulatie van het classificatiegedrag van de luisteraars in het perceptie-experiment van hoofdstuk 5. Eerst worden een aantal fijnschalige en grofschalige cues gemeten aan alle stimuli. De set van fijnschalige cues bevat onder andere de frekwenties van F2 en F3 bij steminzet en in de klinker, plus een aantal ruisplofparameters. De set van grofschalige cues bevat onder andere de globale spectrale helling en de sterkte van de mid-frekwentie piek direct na loslating van de afsluiting, en de verandering van deze maten over de tijd. Vervolgens worden deze cues afgebeeld op de perceptieve data met behulp van het eerder geïntroduceerde formele model van menselijk classificatiegedrag. De resultaten van de modelschattingen laten zien dat de fijnschalige cues, zoals de formantfrekwenties gecombineerd met enkele ruisplofparameters, een betere beschrijving geven van de perceptieve data dan de grofschalige cues. The beste modellen worden geïnterpreteerd in termen van de akoestische grenzen die horen bij de waargenomen linguïstische contrasten. De grenzen zijn lineare functies van 5 of 6 cues, met een duidelijke fonetisch interpre-

tatie. De grenzen geven een kwantitatieve beschrijving van de vaak bediscussieerde "ruilkoers" relatie tussen de verschillende cues voor plaats van articulatie in plofklanken.

In het laatste hoofdstuk wordt aangegeven hoe het onderzoek dat beschreven wordt in dit proefschrift waardevol kan zijn op twee punten. Allereerst hebben we met betrekking tot de primaire fonetische vraagstelling gevonden dat het classificatiegedrag van de luisteraars beter kan worden beschreven op basis van fijnschalige cues dan op basis van de grofschalige cues. Tevens hebben we de waargenomen linguïstische distincties beschreven in termen van lineaire functies van akoestische cues. Ten tweede hebben we een nieuw experimenteel paradigma ontwikkeld voor onderzoek aan fonetische perceptie. Het paradigma combineert aspecten van drie traditionele experimentele benaderingen, te weten akoestische analyse, perceptie van gemanipuleerde natuurlijke uitingen, en perceptie van synthetische spraakachtige klanken. Een belangrijk voordeel van ons paradigma ten opzichte van de bestaande methoden is de mogelijkheid om via een perceptie-experiment met gemanipuleerde *natuurlijke uitingen* een wiskundige modellering te maken van hoe luisteraars cues auditief verwerken om te komen tot een classificatie.

Curriculum Vitae

Roel Smits was born on February 10th, 1964 in Geldrop. In 1982 he got his "VWO"-diploma from the Eckart College in Eindhoven. Subsequently he started his study in Physics at the Eindhoven University of Technology. His graduation project on auditory masking of tone glides was carried out at the Institute for Perception Research. In 1988 he got his "ingenieur"-degree. In 1989 he worked as a research assistant with Prof. F. van Nes at the Institute of Perception Research and with Prof. C. Darwin at Sussex University in the United Kingdom. In 1990 he started working as "OIO" with Prof. R. Collier on a PhD-project entitled "Scale-space coding of spectro-temporal representations of speech", which was funded by NWO (the Dutch Science Foundation). This thesis was written as a part of the project.

At present Roel Smits is with the Dept. of Phonetics and Linguistics of University College London in the United Kingdom.

Stellingen

behorende bij het proefschrift
Detailed versus gross spectro-temporal cues
for the perception of stop consonants
van Roel Smits

1. De metingen van Kewley-Port (*J. Acoust. Soc. Am.* **72**, 1982) aan formantovergangen in initiële plosieven hebben een grotere onnauwkeurigheid dan de gepretendeerde 60 Hz, vanwege de grote vensterlengte en vensterverschuiving die gebruikt zijn bij de analyse van het spraakmateriaal.

2. De conclusies gerapporteerd door Suomi (*J. of Phonetics* **13**, 1985) betreffende het ontbreken van invariantie in globale spectrale structuren zijn niet geoorloofd op grond van zijn onderzoeksresultaten.

3. Met het verschuiven van de invariantie in het spraaksignaal, via invariantie in neuro-motor commando's voor articulatie, naar invariantie in articulatorische *intenties*, schuift de motor theorie van spraakperceptie het rijk der religie binnen, aangezien de theorie vrijwel onfalsifieerbaar is geworden.

4. Het "fuzzy-logical model of perception" van Oden en Massaro zou zowel een zinvoller onderzoekgereedschap worden, als minder vrije parameters bevatten, indien de afbeelding van fysische cues naar interne representaties expliciet zou worden gemodelleerd via een eenvoudige wiskundige transformatie, bijvoorbeeld met een sigmoide-achtige functie.

5. Het succes van bepaalde spraakrepresentaties, zoals de *korte tijd Fourier amplitude* en *linear predictive coding*, is zo groot dat zij heden ten dage de ontwikkeling van spraakonderzoek in zekere mate vertragen.

6. Het feit dat een "stimulus-cue ruimte" met een dimensie hoger dan 2 moeilijk interpreteerbaar is, heeft het onderzoek aan spraakperceptie parten gespeeld. Bij het ontbreken van akoestische invariantie in 1 of 2 dimensies wordt te gauw geconcludeerd dat invariantie volledig ontbreekt, en worden complexe "decoderings"-mechanismen gepostuleerd. Toch is er geen reden om aan te nemen dat het menselijke perceptieve systeem niet in staat is om geluiden te classificeren in hoogdimensionale ruimten.

7. Bij het ontwikkelen van een lesprogramma voor een onderzoekschool is de bijdrage van degenen die het programma gaan volgen - hoog-opgeleide mensen met een kritische houding en grote ervaring in het volgen van onderwijs - onmisbaar. Als deze bijdrage spontaan wordt gegeven dient deze positief te worden ontvangen.

8. Een goede wetenschapper onderscheidt zich van minder goede wetenschappers in het vermogen om vraagstellingen te formuleren eerder dan in het vermogen om meningen of "zekerheden" te poneren. Het zou daarom informatiever zijn over de kwaliteit van de promovendus om het proefschrift vergezeld te doen gaan van een aantal vraagstellingen dan van een aantal stellingen.

9. Het is voor de voortgang van, en creativiteit in wetenschappelijk onderzoek van belang dat er ruime gelegenheid blijft bestaan voor onderzoek waarvoor op geen enkele wijze het maatschappelijke of commerciële nut hoeft te worden aangetoond.

10. Het is ironisch dat medewerkers en studenten van de Technische Universiteit Eindhoven troost zoeken bij koffie-automaten met een klokje dat de tijd uitdrukt in de eenheid guldens.

11. Geen gedicht kan de poëzie van de maandagochtend zo treffend uitdrukken als een hondedrol naast je bureau.