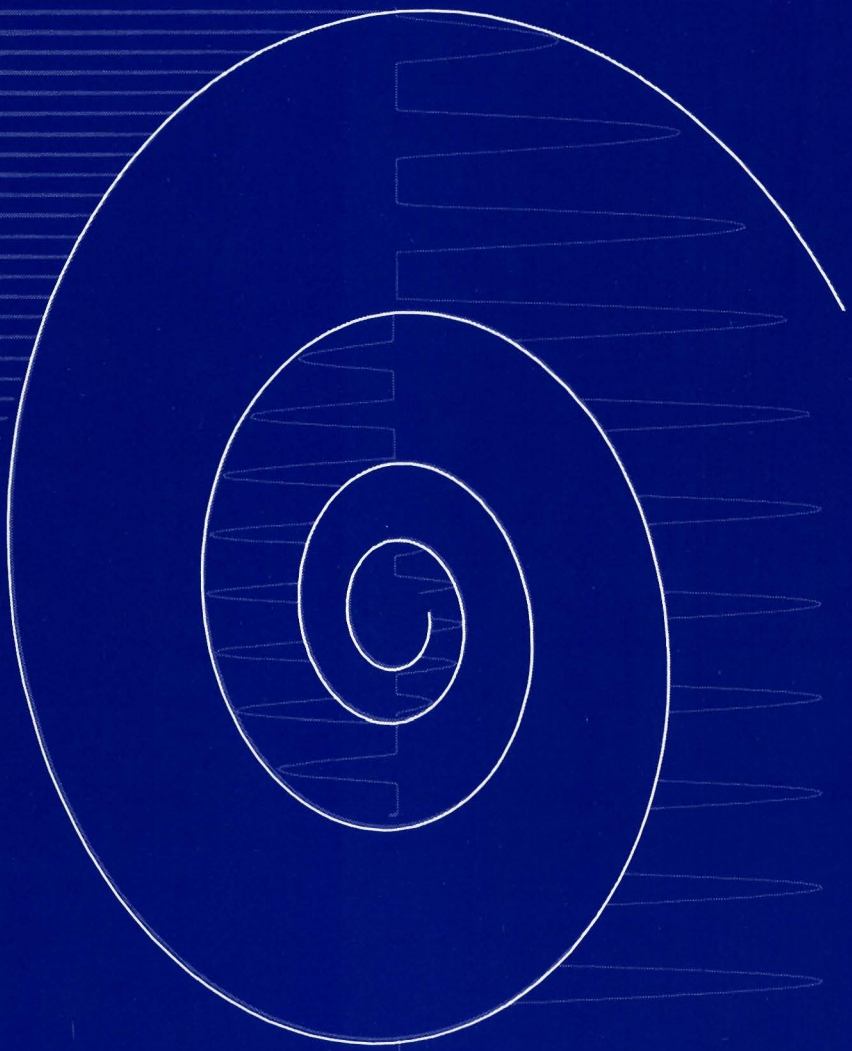
Physiological and psychoacoustical correlates of perceiving natural and modified speech

Physiological and psychoacoustical correlates of perceiving natural and modified speech



Reinier Kortekaas

Physiological and psychoacoustical correlates of perceiving natural and modified speech

Cover design: Katrien Muilwijk

Printing: PrintPartners Ipskamp, Enschede

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Kortekaas, Reinier Willem Leopold

Physiological and psychoacoustical correlates of perceiving natural and modified speech / by Reinier Willem Leopold Kortekaas. - Eindhoven: Technische Universiteit Eindhoven, 1997. - Proefschrift.
ISBN 90-3860-549-8
NUGI 743, 832
Subject Headings: Auditory models / Vowel-onset detection / Speech synthesis / PSOLA / Psychoacoustics

Physiological and psychoacoustical correlates of perceiving natural and modified speech

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. M. Rem, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 3 november 1997 om 16.00 uur

door

Reinier Willem Leopold Kortekaas

geboren te Velsen

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. A.J.M. Houtsma
en
prof.dr. R. Collier

Copromotor:
dr. A. Kohlrausch

"Waartoe men geen toegang heeft vanuit de achtergrond van zijn ervaring, daarvoor heeft men geen oor. Denken we ons nu een extreem geval in: dat een boek uitsluitend gaat over ervaringen, die geheel en al vallen buiten de mogelijkheid van een frequente of zelfs maar uitzonderingsgewijze beleving - dat het de eerste aanzet is tot een taal voor een nieuwe reeks van gewaarwordingen. In zo'n geval hoort men eenvoudigweg niets, wat de akoestische vergissing met zich meegbrengt dat waar niets te horen is, *ook niets bestaat...*"

Friedrich Nietzsche (1908), Ecce Homo.
Vertaling Pé Hawinkels, Uitgeverij De Arbeiderspers, p.57.

Contents

Chapter 1

General Introduction

Imagine two beings A and B living in a small world. A and B both feel a bit uneasy by the thought that they do not know what the other one knows, thinks, and feels. Fortunately, they have a means to send and receive messages. A likes to send messages of the form " $a \quad a \quad a \quad a$ " to show that A is there, and of the form " $ba \quad ab$ " to express their intimacy. A and B have grown accustomed to their conversations and are convinced that their messages are clear and unambiguous. The truth is that, no matter how much effort A puts into it, the "I'm here" messages are at best realized with variability such as " $a \quad a \quad a \quad a$ ". Even more confusing, when sending the "intimacy" message, A actually produces " $\check{b}a \quad a\hat{b}$ " even though A intends to refer to B and not to someone else. In addition, does A know that such a message is corrupted by messages from the world they live in, and is actually received by B as " $X\check{b}a \quad x \quad aX\hat{b}$ " ? Surprisingly, receiver B does not seem to be bothered too much by such variations in shape or integrity of the messages.

These small-world conversations illustrate a number of properties of human speech communication: speech "units" are never completely identical, may be produced at a high or a low rate, vary under influence of co-articulation and assimilation, vary as a result of prosody, context and emotional state of the speaker, and get fused with other sounds in the acoustical path from speaker to listener. This enumeration of factors that increase uncertainty about the intended message is not exhaustive. Considering the number of "noise sources", it is thus remarkable that the process of human speech perception is able to operate reliably and robustly. One usually realizes its versatility and strength only in the case of malfunctioning of one of its stages. A conventional way of labeling the successive stages of human speech processing in response to stimulation from the acoustical environment is: *hearing*, *understanding*, and *comprehension*. This bottom-up classification roughly demarcates (1) the (presumably) language-independent processing of acoustical information, (2) the transformation of processed acoustical information into streams of labeled speech sounds, i.e., phonemes, and (3) association of meaning to these

streams. This dissertation is concerned with studying aspects of speech perception at the first stage, *hearing*, in the context of perception of rhythm and intonation (Part I), and of speech-waveform manipulations (Part II).

Both studies presented in this dissertation investigate the perception of features present in the complex acoustic information conveyed by speech. In particular, Part I of this dissertation deals with automatic detection of vowel onsets. Vowel onsets are conceived of as important features for perception, and production, of intonation and rhythm in speech. The study presented in Part I addresses the question whether automatic detection can be performed on the basis of model simulations of the first stages of the auditory system, and which physical characteristics of the speech signal should thereby be taken into account. In Part II of this dissertation a psychophysical evaluation of the Pitch-Synchronous OverLap-and-Add (PSOLA) technique, developed for speech-modification and synthesis purposes, is presented. In this study, the perception, or rather audibility, of distortions of the speech signal introduced by PSOLA is investigated. In this case, the speech-signal features, i.e. distortions, may be audible and may even become annoying, or may be masked or difficult to attend to due to the inherent complexity of speech signals. This study aims at psychoacoustically determining this audibility by a systematic study of the sensitivity of the human auditory system to these distortions. The studies on vowel-onset detection and PSOLA evaluation will be elaborated separately in the remainder of this Introduction.

1.1 Vowel-onset detection

Hermes (1990) defined vowel onsets in perceptual terms as the moment in the syllable at which a listener starts to perceive the vowel. Vowel onsets are conceived of as "land marks" in the speech signal playing an important role in the perception (and production) of intonation and rhythm in speech (Hermes, 1990; Eriksson, 1991). For instance, pitch movements are known to sometimes lend prominence to syllables in natural speech, provided that the movement is properly aligned to the vowel onset ('t Hart and Cohen, 1973; 't Hart and Collier, 1975; Hermes et al., 1997). Given this evidence for the prominent role of vowel onsets in speech production and perception, one may speculate that a robust mechanism of detecting vowel onsets contributes significantly to the robust speech-perception capacities of the human auditory system.

Because vowel onsets generally are short speech-signal intervals with substantial spectral change, they can be conceived of as a subclass of the so-called fast transitions in speech. For instance, the vowel onset in /pa/ is characterized by a transition from the noise burst /p/, having a more or less flat power spectral density, to the voiced vowel /a/, showing pronounced but rapidly changing maxima in its spectrum in the initial 20-40

msec following its onset. Several studies have provided evidence for the importance of fast transitions for human speech recognition by showing that such transitions convey information on the place of articulation of the consonant (e.g., Kewley-Port et al., 1983), or the identity of the vowel (e.g., Tekieli and Cullinan, 1979). In the vowel-onset detection algorithm described in Hermes (1990), the notion of fast transition plays a central role because detection is based on tracking simultaneous intensity increments in separate regions of the (preprocessed) short-term spectrum.

Against this background, the algorithm described in Hermes (1990) was developed for application in a system for teaching intonation to profoundly hearing-impaired children. Given the importance of vowel onsets for accentuation in speech, the training of hearing-impaired children in correctly aligning pitch movements to vowel-onsets was believed to contribute significantly to their speech in terms of intelligibility. Hermes (1990) first evaluated the algorithm for speech uttered by normal-hearing speakers and found that, for fluent speech, detection performance of approximately 90 % could be achieved, with a false-alarm rate of about 3 %. If tested on isolated words, however, detection performance improved to about 99 %, yet at the same time false-alarm rates increased dramatically to approximately 15 %. By resetting some of the parameters of the algorithm, the false-alarm rate could be reduced to about 4 %, while leaving the proportion of correctly detecting vowel-onset almost unaffected. In application to the speech of profoundly hearing-impaired children, however, the algorithm proved to be unsatisfactory in terms of false-alarm rates and detection scores, which urged for further investigations. Such investigations were conducted by modifying the algorithm: the original preprocessing stages of the spectrum were replaced by psychophysically-motivated stages (te Rietmole, 1991), or by cepstral analysis (Kaufholz, 1992). Neither of these adjustments could satisfactorily improve its performance. An underlying cause for this failure might have been that the influence of physical signal dimensions, such as intensity and spectral content, was not explicitly studied so that their specific role in human vowel-onset detection remained unclarified.

1.1.1 Problem statement

Given both the apparent importance of vowel onsets for perception of fluent speech and the fast-transition nature of vowel onsets, it is plausible to hypothesize that vowel onsets are somehow distinctively coded in the auditory pathway. Onset responses, i.e., neural firing activity associated with the onset of a particular acoustical stimulation, have been identified in various centers in the auditory pathway (Pickles, 1982). Such responses can be conceived of as signaling acoustic events, such as vowel onsets as suggested in Hermes (1990), to higher stages of processing in the brain. Model simulations of such responses may provide a good basis for automatic vowel-onset detection. In fact, the

algorithm presented by Hermes basically mimicked some processes in the first stages of auditory processing, although a in functional, rather than a physiological way. Knowledge of neural processing of auditory information, in as far as it is "materialized" in computational models, has been restricted to the VIII-th (auditory) nerve for a long time (for an overview, see for example Hewitt and Meddis, 1991). Over recent years, an increasing effort has been put into modeling cell responses in higher stages such as the cochlear nucleus, the first stage in the central nervous system, onto which the auditory nerve directly and exclusively projects (Rhode and Greenberg, 1992).

Also at the level of the auditory nerve, stimulus onsets are coded differently than steady-state parts of stimuli, due to short-term adaptation (e.g., Smith and Zwislocki, 1975). Model simulations of such adaptation processes are therefore potentially useful for vowel-onset detection. Moreover, physiological measurements have been reported in the literature on responses of VIII-th nerve fibers in response to natural speech syllables (e.g., Sinex and Geisler, 1983; Carney and Geisler, 1986), so that model simulations may be compared to physiological data. Models of cochlear-nucleus responses have also been under development (e.g., Arle and Kim, 1991; Meyer, 1993b), but physiological data for cell responses to speech-like stimuli are sparse. Responses to "conventional" acoustic stimuli, such as pure tones and noise bands, are richly reported in the literature. In summary, the problem dealt with in the study presented in Part I of the thesis can be stated as follows: "Does the physiological literature on cell responses in the auditory nerve and the cochlear nucleus support the hypothesis that vowel onsets are coded distinctively in these two stages of the auditory system? Is it possible to automatically detect vowel onsets in such cell responses simulated in computational models?"

1.1.2 Outline of Part I

Chapter 2 of this dissertation will present a vowel-onset detection study in which a model of cell responses in the cochlear nucleus (CN) developed by Meyer (1993b) was employed. As will be argued in that Chapter, the use of simulated *auditory nerve responses* for vowel-onset detection is less plausible if detection is based on firing-rate information. A detection scheme was developed which operated on so-called Transient Chopper responses which are, of all response types observed physiologically in the CN, presumably most suited for vowel-onset detection in terms of firing-rate information. The validity of the detection scheme was examined by comparing its vowel-onset predictions to aurally determined onsets for a large database of natural read speech. By way of comparison, the same database was used to re-evaluate the algorithm used by Hermes (1990), and to evaluate a conventional pattern-recognition method, i.e., multilayer perceptrons.

The contribution of this study to the understanding of vowel-onset detection is twofold: first, literature data on neural representations of vowels, and vowel onsets, in the first

stages of auditory processing are explored. In this way, the assumed potential of onset responses is addressed. Second, the relative contributions to vowel-onset detection of the physical speech characteristics intensity, harmonicity, and spectral content are determined.

1.2 Psychophysical evaluation of PSOLA

Part II of this dissertation is devoted to a perceptual evaluation of the Pitch Synchronous OverLap-and-Add (PSOLA) method (Hamon et al., 1989; Charpentier and Moulines, 1989; Moulines and Charpentier, 1990; Moulines and Laroche, 1995). This technique was developed for (simultaneous) modification of pitch and duration of natural speech, achieving a remarkable combination of high synthesis quality and low computational complexity. This low complexity can be accomplished by simple time-domain operations of the speech waveform, which had been regarded by many researchers in speech synthesis, prior to the introduction of PSOLA, as inappropriate for yielding high synthesis quality. The (almost) independent control over pitch and duration has made it a suitable tool for both speech-perception research and speech-synthesis systems. For instance in diphone synthesis, PSOLA can be applied to impose a desired intonation contour and segmental durations onto the signal composed of concatenated diphones (e.g., Moulines and Charpentier, 1990).

Throughout this dissertation, the evaluations are restricted to the time-domain implementation (TD-PSOLA) which has been favored in practical systems over the frequency-domain variant (FD-PSOLA). TD-PSOLA will be referred to as PSOLA in the following. In contrast to Linear Predictive Coding (LPC), for instance, the PSOLA technique is non-parametric (Moulines and Laroche, 1995). PSOLA is computationally attractive, it requires only a (local) pitch estimation of the speech signal for alignment of the waveform operations, but may be less efficient in storage requirements because waveforms are stored uncoded.

The intelligibility of the modified, or synthesized, speech is generally very high which indicates that the technique poses no problems at the *understanding* and *comprehension stages*. As a rule of thumb, this is true for modifications up to a factor of two, which in the case of pitch modification corresponds to an octave. At the level of the *hearing stage*, however, problems have been reported: roughness, hoarseness, and "tube-effects" (as if speaking through a tube) are known to to occur as annoying artefacts in an unpredictable fashion. If such artefacts occur, the speech sample is often discarded or, if artefacts occur frequently, it is often speculated that PSOLA may be applicable to *some* voices, but not to others. In addition, factors such as recording conditions and speaking style are thought to be influential, although the extent to which these factors

would determine perceived quality is uncertain.

Because the modified speech generally has a natural-sounding character as a result of using the unparametrized waveform, one may speculate that unexpected artefacts (roughness) or unnatural artefacts (tube effect) have a high probability of being "heard out". Conversely, if a synthesis technique yields "machine-like" speech, which is the case for some LPC synthesizers, then listeners might predominantly attend to the message *per se*, rather than to naturalness of the acoustic speech information. McGee (1964) studied the effect of signal distortions, i.e. low-pass, high-pass, or band-pass filtering, on perceived quality. Quality was measured using preference judgements and semantic-differential scales. McGee found that perceived quality could be represented in a two-dimensional semantic space, with dimensions "Intelligibility Factor" and "Naturalness Factor". Such a space suggests that these two factors are independent implying that, if listeners judge highly intelligible synthesized speech, they may perform a ranking along the "Naturalness" dimension. In doing so, unexpected or unnatural artefacts may be weighted heavily. Low-level auditory processes in the "hearing stage" are often thought of as being independent of higher-level (cognitive) processes. These higher-level processes selectively attend to the information provided by the low-level processes, which in the case of high-quality synthesis may even be the detection of small signal degradations. A study of low-level auditory processes to determine the sensitivity of the human auditory system to, for the present study, PSOLA-induced distortions seems therefore justified.

To some extent, the perceptual importance of "Naturalness" may have its counterpart in the methodological difficulties encountered in evaluating text-to-speech (TTS) systems (van Bezooijen and Pols, 1990). For many years of evaluating TTS systems, segmental or overall intelligibility tests have been used exclusively, probably because of their ease in conducting and interpreting. However, such tests may no longer differentiate between high-quality systems yielding highly intelligible speech, even though listeners may be able to consistently rank these systems in terms of preference. As van Bezooijen and Pols (1990) argue, diagnostic tests which include judgements on the perceived prosodic structure of messages and on the "voice quality" may better serve as differentiators. The measurement of "voice quality", or perhaps more general sound quality, in itself is not free of methodological pitfalls. Nevertheless, recent computational models for objective quality assessment seem to be able to replicate experimental data, in particular in the context of speech coding and transmission (Beerends and Stemerdink, 1994; Hansen and Kollmeier, 1996). Interestingly, such models are based on psychoacoustically motivated preprocessing of the speech signal (*"hearing stage"*).

A number of studies have been reported in which an evaluation of PSOLA through listening experiments was part of the investigations. First, listening experiments have

been carried out as tests for modifications of PSOLA (e.g., Boëffard and Violaro, 1994), often measuring the subjects' preference for the modified over the original scheme. By measuring preference only, a diagnosis of the relation between algorithmic modifications on the one hand, and perception and quality judgement on the other hand, is rather limited. Second, evaluations of PSOLA as part of speech-synthesis systems have been reported, often measuring several aspects such as acceptance and intelligibility (e.g., Klaus et al., 1997), or intelligibility only (Sydeserff et al., 1992). Sound-quality aspects strictly pertaining to PSOLA modification might, however, be confounded by implementation-specific aspects of, for instance, diphone concatenation. In summary, none of these studies was aimed at understanding in perceptual terms why PSOLA, in its original implementation, often yields high synthesis quality but sometimes fails.

1.2.1 Problem statement

The motivation for the study presented in part II of this dissertation was the appraisal that the perceptual consequences of speech-waveform modification were unsufficiently understood. In signal-processing terms the PSOLA waveform operations are expected to introduce distortions to the speech signal under modification. Apparently such distortions, if not resulting in the artefacts mentioned above, either are phonetically less relevant (Klatt, 1982) and can be easily ignored, or are not perceivable due to limitations of the auditory system. The latter aspect may be dealt with in a systematic psychoacoustical study. By determining the auditory sensitivity to the typical PSOLA distortions, a better insight into the perceptual consequences of speech-waveform manipulations may be achieved, for instance in terms of tolerance to inaccuracy of the operations. As a more general goal, this better insight may contribute to a further understanding of the robust speech-perception mechanisms described above.

PSOLA is generally assumed to preserve the spectral envelope of a speech signal under modification, but this is only roughly correct. The spectro-temporal changes induced by PSOLA may therefore be described psychophysically in terms of the potential spectral or temporal cues a listener may use to discriminate between modified and unmodified signals. The time-domain PSOLA operations are centered around so-called "pitch markers" which indicate pitch-period boundaries. Pitch-marker positioning is generally considered to be a crucial factor in determining synthesis quality, but to the best of our knowledge no perceptual data have been published on the auditory sensitivity to pitch-marker positioning. In summary, the problem addressed in Part II may be stated as follows: "Is it possible to psychophysically identify the perceptual effects of PSOLA-induced distortions, and to relate these effects to properties of natural speech?"

1.2.2 Outline of Part II

The study on evaluation of PSOLA is presented in Chapters 3 to 5. In **Chapter 3** the auditory sensitivity for discriminating PSOLA-modified synthetic single-formant stimuli from unmodified stimuli will be discussed. The perceptual mechanisms involved in discrimination will be characterized as the detection of differences in either excitation patterns or modulation depth. In **Chapter 4** these findings will be extended by studying discrimination sensitivity in the case of double-formant stimuli and in the case of vocal perturbation, simulated by introducing jitter and shimmer. In order to verify whether these results obtained for synthetic signals can be generalized to natural speech, **Chapter 5** will report on a number of experiments of Chapter 3 and 4 replicated for sustained natural vowels.

The contribution of the study presented here to a better understanding of speech-waveform manipulation are (1) determination of auditory discrimination cues that subjects are able to use; (2) verification of these findings by psychoacoustical modeling, and (3) determination of the auditory sensitivity to changes in pitch-marker positioning.

Chapter 2

Vowel-onset detection[1]

2.1 Abstract

An algorithm for detection of vowel onsets in fluent speech was presented by Hermes [J. Acoust. Soc. Am. **87**, 866-873 (1990)]. Performance tests showed that detection was good for fluent speech, although the parameter settings had to be modified for application to well-articulated speech. One of the purposes of the algorithm was application to speech by deaf persons, for which it failed completely. In order to improve the algorithm and to make it more generally applicable, two alternative detection strategies have been explored in the present study. These strategies were (a) simulation of Transient-Chopper responses in the cochlear nucleus and (b) training of multi-layer perceptrons. Two large databases of read speech have been used for performance comparison of the original algorithm and the new strategies. The strategy based on simulating cochlear-nucleus responses is found both to result in a higher false-alarm rate than the original algorithm and to be rather level-dependent. On the other hand, the performance of a multi-layer-perceptron network, trained on mel-scaled spectra, is comparable to the performance of the Hermes algorithm. In more general terms, the results suggest that temporal information on intensity and (rough) spectral envelope are important for human vowel-onset detection behavior. Information on harmonicity can be used as a secondary source of information to avoid detection of mainly unvoiced, non-vowel onsets.

[1]This chapter is a slightly modified version of Kortekaas, R.W.L., Hermes, D.H. and Meyer, G.F. (1996) "Vowel-onset detection by vowel-strength measurement, cochlear-nucleus simulation, and multilayer perceptrons," J. Acoust. Soc. Am. **99**, 1185-1199.

2.2 Introduction

A prominent characteristic of speech signals is the presence of simultaneous frequency modulation (FM) and amplitude modulation (AM). On a suprasegmental level, FM can be found in the pitch contour, whereas AM is present in the syllabic structure. The corresponding modulation frequencies, disregarding phenomena like micro intonation, are typically low: in general, the upper bounds are 8 Hz for FM and 5 Hz for AM (e.g., Plomp, 1984). On the segmental level, FM and AM are most prominent in the so-called 'fast transitions' that occur in the succession of two phonemes or in diphthongs. The rate of those modulations often exceeds by far the rate of suprasegmental FM and AM. There is growing evidence that fast transitions are important for phoneme recognition. Especially in the case of a plosive-vowel combination, a short portion of the speech signal (typically 20 - 40 ms) appears to contain sufficient information for determination of the place of articulation of the consonant (e.g., Kewley-Port et al., 1983), or the identity of the vowel (e.g., Tekieli and Cullinan, 1979). In general, much perceptually relevant information is present in speech portions which show substantial spectral change within a short time interval (Strange et al., 1983; Furui, 1986; Nossair and Zahorian, 1991).

Hermes (1990) defines vowel onsets perceptually as the moment in the syllable at which a listener starts to perceive the vowel. He determined vowel onsets by using a so-called gating paradigm ('t Hart and Cohen, 1964) in which short, windowed portions of the speech signal are isolated and listened to. Using this paradigm, a phonetician can determine vowel onsets with an average accuracy of at least 20 ms. Agreement among different phoneticians will generally be high, but may vary for different phonemic contexts. In the Hermes (1990) study, vowel onsets are assumed to coincide with speech portions which show large increments in intensity in separated frequency channels of the auditory system. As such, vowel onsets can be conceived of as a subset of the class of fast transitions.

House (1990) showed that speech segments with considerable spectral change play an important role in tonal perception in speech. The perceptual identity of a pitch movement, i.e. which syllable it accentuates and how much it contributes to the prominence of the syllable, depends on the temporal position of the pitch movement with respect to those segments with considerable spectral change. Of these segments, the vowel onset appears to play the most important role ('t Hart and Cohen, 1973; 't Hart and Collier, 1975; House, 1990). Furthermore, vowel onsets appear to play a primary role in perceiving rhythm in speech (Rapp, 1971; Allen, 1972; Cole and Scott, 1973; Eriksson, 1991). In phonology, the vowel onset corresponds to the transition from syllable onset to syllable rhyme. At the syllable level, this division is generally believed to be basic to the structure of the syllable (Pike, 1947; Selkirk, 1982). The theoretical issue of the importance of the vowel onset for speech perception will be readdressed in the Discussion.

In spite of these arguments concerning the relevance of vowel onsets to speech perception, little attention has been paid to the automatic detection of vowel onsets. Hermes (1990) introduces the concept of 'vowel strength', which is a measure for the presence of both a formant structure, i.e. with pronounced maxima, as well as a harmonic structure, i.e. a line structure, in the amplitude spectrum. Strong increments in vowel strength are supposed to signal the presence of a vowel onset. Based on these ideas, Hermes (1990) presents an algorithm for the automatic detection of vowel onsets in natural speech, which we will refer to here as Vowel-Strength Measurement (VSM). The algorithm was found to perform quite satisfactorily for fluent speech: approximately 10 % of vowel onsets were missed, most of which before unaccented schwas. For well articulated isolated words, some parameter settings of the algorithm had to be changed for a satisfactory performance; otherwise, the number of false alarms increased unacceptably. The algorithm was applied in a teaching system of intonation to profoundly deaf persons (Kaufholz, 1992). In this case as well, the performance of the algorithm was judged to be unacceptable. As the aim of this system was to improve the intonation of deaf persons, the system could not be optimized for deaf speech. Two independent strategies for improving VSM were attempted: first, psychophysically-inspired processing stages were introduced (te Rietmole, 1991), and second, cepstral analysis was included (Kaufholz, 1992). Neither strategy could substantially reduce the number of missed onsets without simultaneously increasing the number of false alarms. More generally, the relative contributions of physical speech-signal characteristics like intensity, harmonicity, and spectral content, remained unclear.

This chapter will describe further investigations of automatic vowel-onset detection by comparing the performance of VSM to two alternative automatic detection strategies. The aim of the investigations is both to obtain more insight into the underlying signal dimensions which play a role in human vowel-onset detection, and to develop a better detection scheme. The rationale for comparing different schemes is that such a method may help to derive what information is required to model human vowel-onset detection. The chapter will focus on the description and evaluation of the alternative strategies. In the evaluation, both missed-onset and false-alarm rates as well as phonemic context of occurrence will be analyzed.

As alternative vowel-onset detection strategies we have chosen (a) to apply a model of the Cochlear Nucleus (CN) developed by Meyer (1993b,a) and (b) to train Multi-Layer Perceptron (MLP) networks. In the case of the CN model, a detection scheme will be presented that is based on simulated Transient-Chopper responses. This detection scheme will be referred to as Cochlear Nucleus Simulation (CNS). The MLP networks have been trained both with the simulated Transient-Chopper responses and with conventional mel-scaled amplitude spectra. We have compared VSM, CNS, and the MLP

networks by means of performance tests using two speech databases. Both databases consisted of Dutch read speech uttered by nonprofessional speakers: 18 and 24 speakers, respectively, with an equal number of male and female speakers. Recording conditions of the databases were good and excellent, respectively. The results give support to the hypothesis that the important signal dimensions for human vowel-onset detection are both the increment of intensity as well as the spectral envelope, and that harmonicity serves as an additional source of information. The CNS detection scheme yields missed-onset rates slightly higher than VSM, and false-alarm rates that are considerably higher. Moreover, the performance of the CNS scheme is seen to be substantially level dependent. On the other hand, the performance of a detection scheme based on an MLP network, with mel-scaled spectra as input, is competitive with the performance of VSM.

2.3 Models

In this section, the detection schemes will be presented: first, the VSM algorithm will be briefly reviewed. Second, both the auditory model and the corresponding vowel-onset detection scheme CNS will be discussed, and finally, the MLP-network architecture will be presented.

2.3.1 Vowel-strength measurement

The VSM algorithm is based on measurement of vowel strength. This measure expresses (a) the weighted contribution of the harmonics to the pitch of a speech segment[2], and (b) the degree to which a formant pattern is present in the preprocessed amplitude spectrum of a pitch period within that segment. Vowel strength is measured every 10 ms. Both (a) and (b) show a high correlation to intensity so that vowel-onset detection in VSM is based on spectral envelope, harmonicity, and intensity.

In VSM, vowel onsets are associated with rapid increments in vowel strength. Such increments are detected by finding the maxima of the smoothed derivative of the sequence of vowel-strength measurements. The impulse response of the smoothed-derivative filter has a bipolar character; it is the sum of two Gaussians shifted in time and of opposite sign, starting with a positive excursion. The effective duration of the filter is approximately 100 ms. A similar filter is applied in the CNS detection scheme and the MLP networks. An example of the time course of vowel strength in given in Figure

[2]In Hermes (1990, p. 868) it is stated that "it is necessary to suppress the contribution of the unvoiced parts. This is achieved by weighing the result of the measurement of the combined strength of the spectral peaks with the maximum value of the subharmonic sum spectrum". This phrasing gives the impression that this feature was only applied to avoid some false detections in noisy speech segments. It appears now, however, that this feature was essential for the good performance of the algorithm.

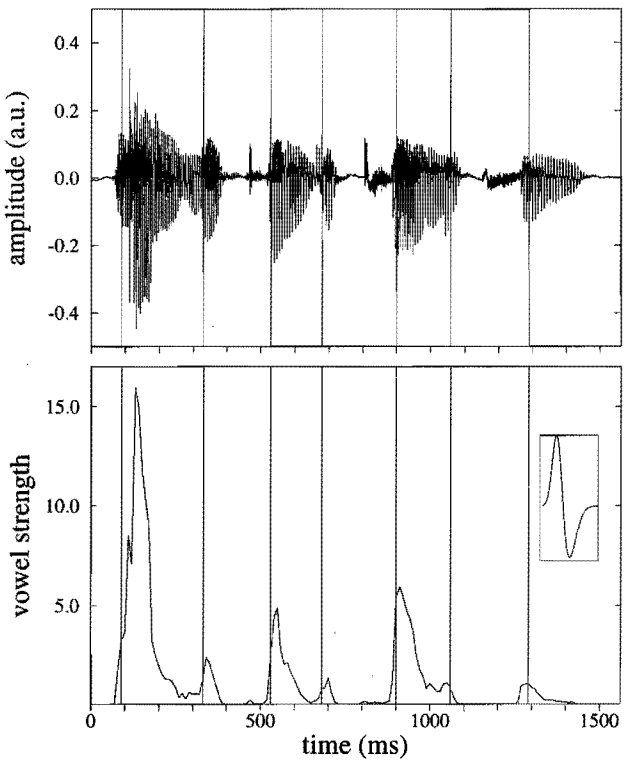


Figure 2.1: *Waveform of PM-sentence 13 (top) and corresponding vowel strength, as measured by VSM, as a function of time (bottom). Aurally detected vowel onsets are marked by vertical lines. The impulse response of the smoothed-derivative onset filter is shown in the inset of the bottom panel. Note that the time axes of vowel strength and the onset filter do not match; the effective duration of the onset filter is approximately 100 ms.*

2.1 (PM-sentence 13 "Eindelijk kwam de trein op gang", see section 'Materials'). The smoothed-derivative filter is shown in the inset of Figure 2.1. For further details of the VSM algorithm, the reader is referred to Hermes (1990).

2.3.2 Detection based on simulated Transient-Chopper responses

If vowel onsets play a role in human speech perception, it is reasonable to expect that the information required for this process is encoded in the auditory pathway. The auditory model (Meyer, 1993b,a; Ainsworth and Meyer, 1994) applied in this study comprises simulations of responses in the first two stages in the neural auditory pathway: the cochlear or Auditory Nerve (AN) and the Antero Ventral Cochlear Nucleus (AVCN).

For these stages, however, no physiological observations have been reported in the literature that demonstrate any enhancement of vowel-onsets in neural responses (Kortekaas and Meyer, 1994). Nevertheless, simulation of neural encoding of (speech) signals may be appropriate for automatic vowel-onset detection; the spectro-temporal resolution of such simulations is approximately the same as that of the auditory system. This means that such simulations may contain information which is perceptually more important. Moreover, several authors have demonstrated the perceptual importance of onsets and offsets, which are enhanced in simulated neural encoding and therefore more easily detectable (e.g., Darwin, 1984; Summerfield and Culling, 1992). Furthermore, spectral contrast is enhanced, too, which will be important for tracing resolved harmonics and formant frequencies.

The initial motivation for investigating simulations of peripheral auditory processes was based on the observation of overshoot phenomena in the AN, like short-term adaptation (e.g., Smith and Zwislocki, 1975; Eggermont, 1985). The hypothesis was that, due to short-term adaptation, rapid intensity increments and decrements in auditory channels are strongly enhanced in the firing profiles of cochlear nerve fibres. In the present context, rapid increments correspond to vowel onsets. The idea was that vowel onsets can be detected by measuring such strong simultaneous increments in channels corresponding to formant regions. It should be noted that an alternative approach would be to concentrate on phase locking in the auditory nerve. Several authors have demonstrated that the spectral content of speech sounds is preserved in the phase-locked activity of cochlear nerve fibres (e.g., Delgutte and Kiang, 1984a,b,c; Carney and Geisler, 1986).

Nerve fibres generally show a limited dynamic range, normally extending 20-30 dB of differential sensitivity (e.g., Evans and Palmer, 1980). Steady-state stimulation at levels that exceed this range results in saturation of the fibre and as a consequence, loss of differential coding. The dynamic ranges of single nerve fibres generally are too small to provide differential coding over the whole dynamic range of speech. On the other hand, the auditory system seems to be provided with a continuum of nerve fibres with different thresholds and dynamic ranges (Liberman, 1978). Often a categorization of nerve fibres into a low- and a high-threshold population is made. A combination of the two populations seems necessary for coding spectral information in terms of discharge rate over the whole dynamic range of speech.

Such a combination is found for the so-called stellate cells in the AVCN (e.g., Rhode and Greenberg, 1992). Stellate cells receive excitatory input from both low- and high-threshold AN-fibres whose characteristic frequencies are within 1 Bark of the stellate cell (Rhode and Smith, 1986; Blackburn and Sachs, 1990; Rhode and Greenberg, 1992). Moreover, the cell receives inhibitory input from a relatively wide receptive field. The

response patterns that were physiologically recorded from stellate cells are often characterized as 'Transient Chopper' or 'Chop-T'. Such a response pattern is characterized by initial regularity of discharge ('chopping') followed by a rapid transition to irregularity. Under stimulation with pure tones, the dynamic range of Chop-T cells is comparable to the limited dynamic range of AN-fibres. Nevertheless, physiological recordings have shown that discharge profiles of Chop-T cells represent the spectrum of (speech) sounds over a wide dynamic range. Therefore, simulation of Chop-T responses seems to be appropriate for vowel-onset detection (Kortekaas and Meyer, 1994). Blackburn and Sachs (1990) showed, for instance, that the formants of the synthetic vowel /ɛ/ are represented with high contrast in Chop-T discharge rates over a dynamic range of 35 to 75 dB sound-pressure level. Such contrasts were not observed in the rate profiles of AN-fibres. These differences in spectral contrast can be modeled by lateral inhibition which is present in the case of Chop-T neurons, but not in the case of AN-fibres.

2.3.2.A Auditory model

This section briefly describes the auditory model which consists of a peripheral part and a Chop-T-response simulation part (Meyer, 1993b,a). The peripheral part consists of the following stages:

1. Gamma-tone filterbank with 32 4th-order IIR filters (de Boer, 1969; Darling, 1991). Center frequencies range from 0.1 to 4.6 kHz at 0.5 Bark spacing. Each channel has a bandwidth of one Equivalent Rectangular Bandwidth (ERB) (Compernolle, 1991).

2. Filter-output scaling for:

   - Human hearing-threshold adjustment (Fay, 1988).
   - Dynamic-range extension (see below).

3. Inner-hair-cell model (Meddis, 1986, 1988).

4. Spike generation on the basis of expected firing rates.

The hearing thresholds are calculated for each channel by a polynomial fit to the data reported in Fay (1988)[3]. The dynamic-range-extension stage is introduced for simulation of two populations of nerve fibres; instead of adjusting the parameters of the Meddis (1988) model, the signals that drive the inner-hair-cell model are scaled. The two populations are specified as follows, where relative firing threshold denotes the absolute firing threshold of the fibre relative to the absolute hearing threshold (i.e. sensation level):

[3]The equation reads : $TH(c_f) = 4.0758c_f^{-1} + 17.4741 - 45.2252c_f + 45.7596c_f^2 - 19.5892c_f^3 + 4.1071c_f^4 - 0.4133\,c_f^5 + 0.0159\,c_f^6$, where c_f denotes center frequency in kHz, and $TH(c_f)$ denotes absolute hearing threshold in dB SPL.

- low-threshold fibres: relative fibre threshold of 0 dB, dynamic range of 30 dB, and spontaneous activity of 50 spikes s^{-1}.

- high-threshold fibres: relative fibre threshold of 15 dB, dynamic range of 50 dB, and spontaneous activity of 15 spikes s^{-1}.

Simulation of CN responses is based on a point-neuron model. The membrane potential is controlled by the Goldman-Hodgkin-Katz equation (e.g., Brown, 1991) as a function of concentration gradients and membrane permeability. Each simulated neuron receives excitatory input from neurons having the same center frequency. In addition, the neuron receives inhibitory input from neurons which have center frequencies between one Bark below, and two Bark above the neuron's center frequency. Both for excitation and inhibition, afferent neurons are from both populations. Instead of generating action potentials for the Chop-T simulation, we concentrate on the extracellular potential of the neuron relative to its firing threshold. We will refer to this potential as 'activity'. The output of the Chop-T simulation is the activity, as a function of time, of an array of 23 neurons. Best frequencies of those neurons are in the range from 0.2 to 2.6 kHz, with 0.5 Bark spacing. For more details on the model, the reader is referred to Meyer (1993b,a) and Ainsworth and Meyer (1994).

2.3.2.B Vowel-onset-detection scheme

The vowel-onset-detection scheme was developed with the aim of applying phonetic knowledge to the process of detecting vowel onsets in simulated Chop-T responses. We will refer to this scheme as Cochlear Nucleus Simulation (CNS) in the following. In the scheme, the simulated neuron activity is averaged over two frequency bands which roughly correspond to the regions of the first and second formant (see Figure 2.2). The corresponding best-frequencies are 0.2-1.1 kHz for the first-formant, and 0.9-2.6 kHz for the second-formant band, respectively. These two bands can be conceived as a rough representation of the spectral envelope within the first and second formant region. The underlying idea of defining two formant areas is the assumption that vowel-onsets are characterizedby simultaneous and strong increase of activity in these two bands. The scheme traces such increases of activity ('vowel-onset candidates') and applies a number of criteria to discard onsets other than vowel onsets.

The averaged activity is low-pass filtered by means of leaky integration to obtain the signals $A_L(t)$ and $A_H(t)$ for the first and second-formant band, respectively. The -3 dB point of the LP filter is at $\sim$ 25 Hz. Subsequently, we take the smoothed derivative of $A_L(t)$ and $A_H(t)$, as described in Section 'Vowel-strength measurement', which results in the signals $O_L(t)$ and $O_H(t)$. Vowel-onset candidates are found at those instances at

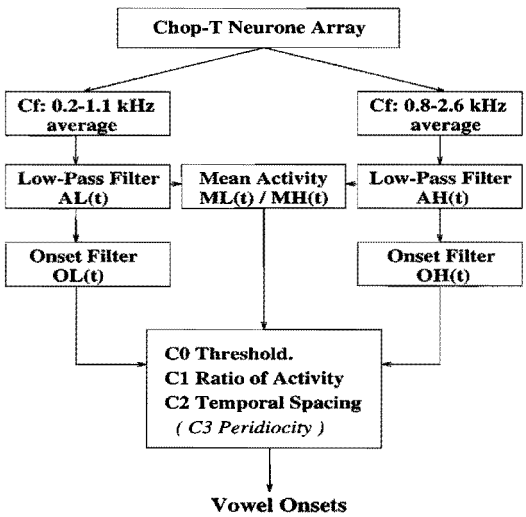


Figure 2.2: *Schematic representation of the CNS (and CNS-ACF) detection scheme.*

which (a) both $O_L(t)$ and $O_H(t)$ are greater than zero, (b) $O_L(t)$ has a local maximum. The condition that $O_H(t)$ should also have a local maximum was found to be too strict.

A vowel-onset candidate detected at time t_c should match the following criteria (parameter settings will be given below) :

C0 Threshold: To exclude irrelevant fluctuations of activity in $A_L(t)$, the criterion reads:

$$O_L(t_c) \geq O_{TH},$$

where O_{TH} is a parameter. A similar criterion is applied in VSM.

C1 Ratio of Activity: In order to discard candidates that signal phoneme classes other than vowels, the following criterion is introduced:

$$\alpha \leq \frac{M_L(t_c)}{M_H(t_c)},$$

where α is a parameter and $M_L(t_c)$ and $M_H(t_c)$ denote the mean activity in $A_L(t)$ and $A_H(t)$ over 45 ms following the onset at t_c. The lower bound α is introduced to discard onsets of fricatives which have concentration of spectral energy in the high region. An upper bound for the ratio of activity, which could exclude onsets for nasals, was found not to contribute significantly.

C2 Temporal Spacing: Some phoneme classes, e.g., liquids and semivowels, yield

multiple candidates during an interval of continuous increase of activity, i.e. $O_L(t) > 0$ for all t within the interval. This criterion consists of two parts:

- Within such an interval, a vowel onset detected at t_1 is discarded if a subsequent vowel onset is detected at t_2, with $t_2 > t_1$.
- If a candidate at t_2 is separated from a preceding vowel onset at t_1 by a period of decrease or absence of activity, then the candidate at t_2 is accepted if:

$$t_2 - t_1 > \Delta,$$

where t_1, t_2 like above and Δ is a parameter. The vowel onset at t_1 is not discarded. A similar criterion is applied in VSM.

Unlike VSM, the CNS scheme does not include information about the harmonicity of the signal. Modulation transfer functions measured for Chop-T responses show that phase locking to the AM envelope is preserved up to approximately 400 Hz (e.g., Rhode and Greenberg, 1992). This means that information about harmonicity of the speech signal can be derived from a broad frequency range. A rather ad-hoc solution to incorporate a harmonicity criterion into CNS is to calculate the short-term autocorrelation of the activity of each Chop-T neuron. The autocorrelation is calculated for a 10 ms signal window. The individual autocorrelations are combined to obtain the summary autocorrelogram of the whole neuron array (e.g., Meddis and Hewitt, 1991). The magnitude of the maximum of the summary autocorrelogram is taken to represent the 'strength of harmonicity'. This measure generally shows a high correlation to the instantaneous level of the input signal. In this extended scheme, referred to as CNS-Auto-Correlogram Function (CNS-ACF) in the following, a criterion is introduced for periodicity:

C3 Periodicity: For each vowel-onset candidate, the corresponding strength of periodicity should be above ACF_{TH} % of the maximum strength of periodicity observed in the utterance. Here, ACF_{TH} is a parameter.

An example of the signals $A_L(t)$, $A_H(t)$, and the autocorrelation peak is given in Figure 2.3 for PM-sentence 13 ("Eindelijk kwam de trein op gang"; see Section 'Materials'). Note that the C3 criterion requires that the whole utterance is analyzed before vowel onsets can be detected. Instead, the maximum strength of periodicity could also be determined for a constant interval of, e.g., 200 ms. Such an interval may require, however, that (a) pauses in utterances can be detected and (b) that the Signal-to-Noise Ratio (SNR) is always high.

In summary, vowel-onset detection in the CNS scheme is based on changes of intensity and (rough) spectral envelope. In addition to these characteristics, detection in the CNS-ACF scheme is based on estimation of harmonicity.

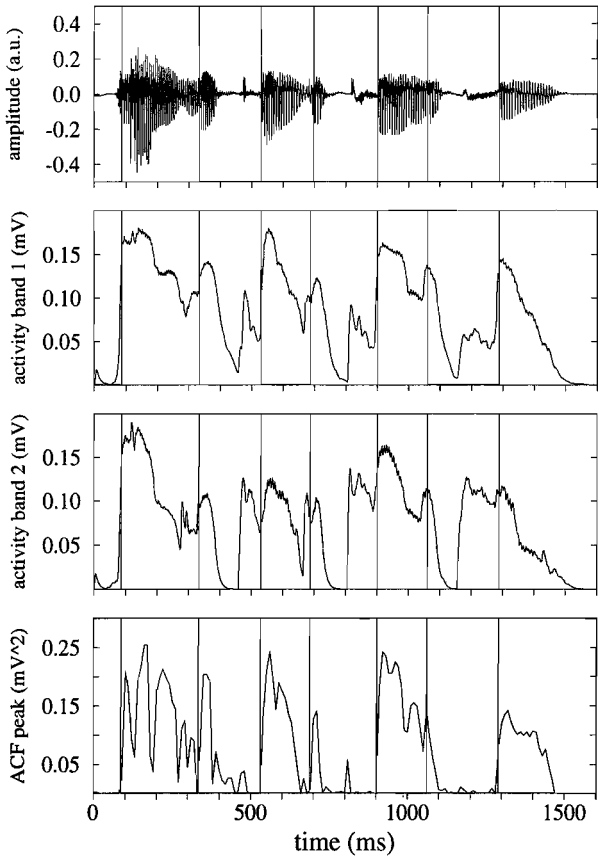


Figure 2.3: *Waveform of PM sentence 13 (top), activity in band 1 and band 2 at 55 dB SPL (middle panels), and magnitude of the peak of the summary autocorrelogram of the Chop-T neuron array at 55 dB SPL (bottom). Aurally detected vowel onsets are marked by vertical lines.*

2.3.2.C Scaling of input signals

The computational model of the auditory periphery and the cochlear-nucleus responses is non-linear, which makes input scaling necessary. Optimizing and testing the CNS and CNS-ACF schemes (and MLP networks) was done at 35, 55, and 75 dB SPL, where the root-mean-square of each of the sentences was normalized. Note that both schemes are partially based on intensity-level information, but that the rate-intensity functions of the simulated neurons behave non-linearly.

2.3.2.D Parameter setting

The parameters presented above were optimized for the T-sentence database containing 377 vowel onsets (see Section 'Materials'). Setting all parameters to zero, i.e. accepting

O_{TH}	0.00075	mV/ms
α	0.8	
Δ	45	ms
ACF_{TH}	10	%

Table 2.1: *The parameters of the CNS and CNS-ACF schemes.*

all vowel-onset candidates, resulted in 356 correct detections and 136 false alarms (at 55 dB SPL input level; see Section 'Scaling of input signals'). Using the criteria mentioned above, the vowel-onset-detection scheme has to perform a 'yes-no' task for each vowel-onset candidate. A method for analyzing the performance of the scheme as a function of parameter settings is the Receiver-Operating-Characteristic (ROC) curve (e.g., Green and Swets, 1966). By setting all other parameters to zero, individual parameters were optimized by finding the value that optimally reduces the number of false alarms, while keeping the number of correct detections almost unaffected. This means that the combination of the different optimized criteria may drastically reduce the false-alarm rate, provided the different criteria affect different vowel-onset candidates. The parameter settings listed in Table 2.1 were derived by finding optimal values compromizing for both 55 and 75 dB SPL input level.

Figure 2.4 shows ROC curves for CNS and CNS-ACF where each of the parameters is varied while all other parameters are set to the optimal values as listed in Table 2.1. These curves show the sensitivity of both schemes to variation of each of the parameters. This sensitivity can be derived to a first approximation from the area under the ROC curve (Green and Swets, 1966). It should be noted that the ROC curves are plotted with raw numbers as units, whereas ROC curves usually display probabilities. Moreover, for all parameters except O_{TH}, only part of the ROC curve is shown. In general, the individual optimal parameter settings determined by the ROC-curve method described above are seen to be also appropriate in case all criteria are applied. This indicates that the criteria can be conceived of as being more or less independent. From Figure 2.4b it can be derived that the contribution of the C3 criterion is important in case no spectral-envelope information, i.e. the C1 criterion, is used. If the C1 criterion is applied then the C3 criterion contributes by additionally discarding some false alarms while keeping the correct-detection rate almost unaffected (see Figure 2.4d).

2.3.3 Training Multi-Layer Perceptron (MLP) networks

MLP networks have proven to be robust techniques for pattern classification in speech-recognition tasks (e.g., McCulloch and Ainsworth, 1988; Markowitz, 1993). A weak point of the CNS (-ACF) detection scheme is that the decision boundary for the vowel versus non-vowel categories may not be optimal. In other words, the criteria defined on the basis

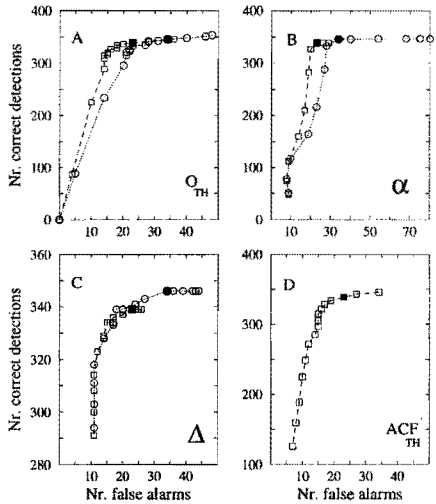


Figure 2.4: *ROC curves for each of the parameters of CNS (circles) and CNS-ACF (squares). Filled symbols represent the parameter settings used in the evaluations and listed in Table 2.1. In all cases, the smallest parameter values correspond to the rightmost data points on the ROC-curve. (A) ROC curve for parameter O_{TH} with parameter range of 0 to 0.003 mV/s in steps of 0.00025 mV/s. Additional values are 0.004, 0.008, 0.016, and 0.032 mV/s. (B) ROC curve for α where the range is 0 to 1.5 with a stepsize of 0.05. (C) ROC curve for Δ with values are set in the range 5 to 100 ms with increments of 5 ms. (D) ROC curve for ACF_{TH} plotted for the range 0 to 75% in steps of 5%. These data are obtained for the T-sentence database at 55 dB SPL (see section 'Materials').*

of general phonetic knowledge may not be optimal for separating the two categories. In this respect, MLP networks are used (a) to determine whether the information required for vowel-onset detection is present in the Chop-T representation, and (b) to investigate whether this information can be retrieved from conventional speech-analysis techniques. MLP networks are used in this study as vowel versus non-vowel classifiers, where two classification experiments have been performed:

- MLP networks have been trained with the Chop-T representation for the following input-pattern configurations:
  - (a) the full-resolution, 23-channel representation.
  - (b) the two-formant-band representation as applied in CNS(-ACF).
  - (c) a single unit being the sum of all 23 channels.

  In the case of (b), performance results inform about the amount of information

required for vowel-onset detection with respect to spectral envelope. In the case of (c), only intensity information was supplied to the network.

- MLP networks have been trained with conventional mel-scaled amplitude spectra. The choice of mel-scaled spectra was motivated by the similarity of these spectra to the simulated cochlear-nucleus representation except for the non-linearities of the auditory model. The mel-scaled spectra consisted of 23 points covering approximately the same frequency range as in VSM and CNS(-ACF), namely 200-2600 Hz. Three input-pattern configurations have been investigated:
  - (a) the RMS of each utterance was normalized so that individual spectra contained intensity-level information.
  - (b) each mel-scaled spectrum was normalized so that no intensity-level information was used.
  - (c) all components of individual mel-scaled spectra were summed to obtain a measure of the instantaneous level.

  For input patterns under condition (a), all spectra contained intensity-level information with respect to the whole utterance. In contrast, input patterns under (b) did not contain intensity-level information but only contained spectral envelope information. Finally under condition (c), where the RMS of the utterance was normalized, patterns only consisted of a measure of the instantaneous level.

2.3.3.A Network architecture

In all classification experiments, the output of the MLP networks consisted of a single unit representing the presence of a vowel. The unit's output range was between zero and one, where zero indicates that the presence of a vowel, given the input pattern, is highly unlikely. In the case of both the mel-scaled spectra and the Chop-T representation, input patterns consisted of 1, i.e. the sum of spectral components, or 23 units, i.e. the full resolution. Performance was evaluated for the number of hidden units ranging between 0 and 10. If no hidden units were used, the network did not learn, whereas 10 hidden units caused 'over-training'. Best results on the training database (see Section 'Materials') were obtained for 2 to 5 hidden units: except for the single input condition, where the network encorporated 2 hidden units, all results to be described below were obtained with networks having 5 hidden units.

2.3.3.B Network training

All input patterns were calculated over 25.6 ms long time slices by either calculating an FFT of the windowed speech signal, or by binning the activity within channels in the

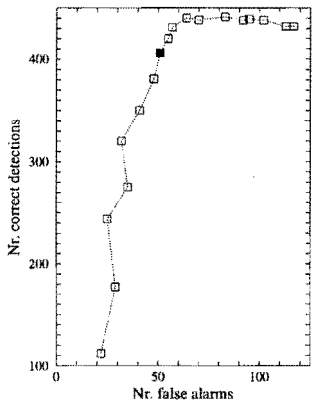


Figure 2.5: *ROC curve for the threshold parameter of the MLP schemes. The parameter-value range is 0.1 (rightmost data point on the ROC-curve) to 0.95 (leftmost point) with a stepsize of 0.05. The filled symbol represents the value of 0.6 as applied in the experiments. These data are obtained for the PM-sentence database (see section 'Materials').*

Chop-T representation. The input patterns were presented to the networks without any further preprocessing. The frequency range of the input patterns was 200 to 2600 Hz both for the mel-spectra as well as the Chop-T representation. Training patterns for the vowel category were taken at aurally detected vowel onsets (see Section 'Performance evaluation') and at 25.6 ms after those onsets. Each training sample was checked visually to exclude erroneous training data like very short vowels or early detections. In all, 1345 training patterns were used: 744 and 601 patterns for the vowel and non-vowel category, respectively. The training patterns for the non-vowel category were chosen in two passes: initially a small set of non-vowels and silences was chosen. After training, the network was tested on the T-sentence database (see section 'Materials') and patterns that the network erroneously classified as vowels were added to the non-vowel-pattern set. Then, the network was retrained with the vowel set and the extended non-vowel set.The MLP was trained using standard back-propagation with a learning rate of 0.0005 and a momentum term of 0.1. To prevent over-training we set an error threshold to 0.05. The networks were trained in steps of 100 cycles until the summed error in the vowel versus non-vowel classification no longer declined, usually for 400 to 600 cycles.

2.3.3.C Vowel-onset detection

Sentences from the training set were processed in steps of 1 ms. The activation of the output unit, representing the likelihood of the presence of a vowel, was first thresholded at 0.6. This threshold value was determined by trial-and-error, and applied to all training and testing conditions. A pseudo ROC curve is depicted in Figure 2.5 for a

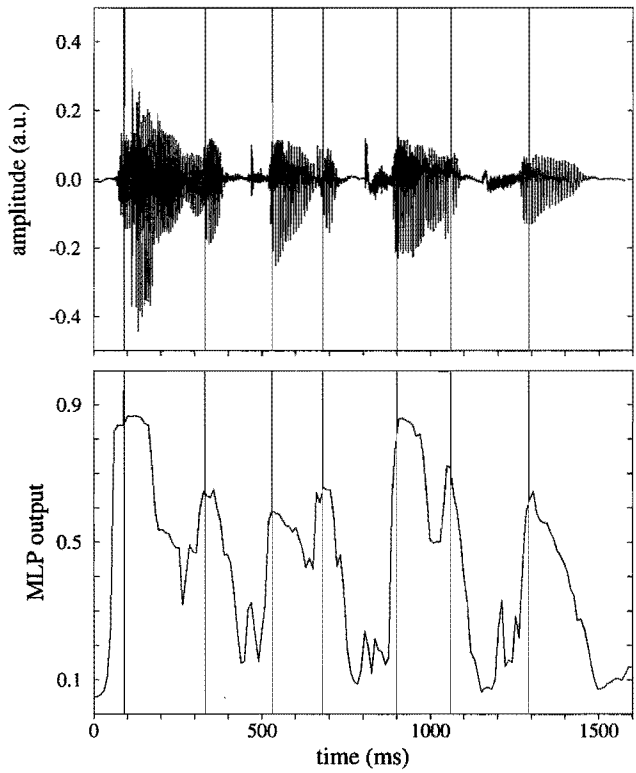


Figure 2.6: *Waveform of PM-sentence 13 (top) and corresponding MLP output as a function of time (bottom). The MLP is trained on mel-scaled spectra, with normalization of the RMS of each sentence. Aurally detected vowel onsets are marked by vertical lines.*

particular condition, namely, input patterns consisting of 23-channel mel-scaled spectra, with normalization of the RMS of the sentence. The ROC curve is determined by evaluation on the PM-sentence database (see Section 'Materials').

Like the detection of vowel onsets in the sequence of vowel strength in VSM, we then applied the smoothed derivative filter (see Section 'Vowel-strength measurement') to trace the local maxima. An example of the output of an MLP as a function of time, based on mel-scaled spectra, is given in Figure 2.6 (PM-sentence 13, see section 'Materials').

2.4 Model Evaluations

2.4.1 Materials

Hermes (1990) evaluated VSM for a database consisting of 28 Dutch sentences spoken by 9 male and 9 female speakers, all nonprofessional native speakers of Dutch. They were instructed to read the sentences without taking special care of articulation. This database will be referred to as the T-sentences. As described in the Introduction, a trained phonetician traced the vowel onsets by means of the gating technique. The gating technique is based on listening to a short portion of the speech signal by windowing the signal, typically of 20 to 40 ms duration, and shifting this window in time through the speech signal ('t Hart and Cohen, 1964). Vowel onsets obtained in this manner are called 'actual onsets', 377 of which were detected in the T-sentence database.

Because both the CNS (-ACF) and the MLP schemes have been optimized or trained on the T-sentences, an alternative speech database was required for comparison of performances. For this purpose, we made a random selection of 28 out of 560 sentences from the Plomp and Mimpen (1979) set which has its application in diagnostic audiology. We will refer to this selection as the PM-sentences. The 28 sentences were re-recorded with 14 male and 14 female nonprofessional speakers resulting in both a male- and a female-speaker version of each sentence. Instructions for reading were similar to those of the T-sentences. Actual onsets in these 56 utterances were determined by an experienced phonetician, who traced 466 vowel onsets.

2.4.1.A Performance evaluation

The number of missed onsets and false alarms are calculated for each of the detection schemes by determining the cross-coincidence between actual and algorithmically detected vowel onsets. We adopt the criterion proposed in Hermes (1990) that algorithmically detected onsets should be within ± 60 ms to the actual onsets. Missed-onset and false-alarm rates will be presented as a proportion of the total number of actual onsets. For indication of accuracy, we will also give figures that express the proportion of correct algorithmically detected onsets that are within ± 20 ms to the actual onsets. These results will be presented as 'Accuracy' in the next section. The results for the T-sentences have been evaluated for both rates and contexts of occurrences for missed onsets and false alarms. For reasons of clarity, we will only present missed-onset and false-alarm rates. On the basis of these rates, we will select a number of schemes that will be evaluated in more depth for the PM-sentences.

		Test level (dB SPL)	Missed onsets (%)			False alarms (%)			Accuracy (%)		
VSM			8			3			91		
CNS		35	16			6			86		
		55	8			9			87		
		75	10			15			84		
CNS-ACF		35	24			3			87		
		55	10			6			88		
		75	11			8			84		
MLP	Mel-spectra		10			10			80		
	Normalized		9			36			67		
	Summed		9			9			77		
			channels			channels			channels		
MLP	Chop-T		1	2	23	1	2	23	1	2	23
	Trained on	35	14	10	12	7	5	3	82	82	85
	test level	55	16	8	10	5	8	9	84	82	82
		75	13	9	8	6	18	14	85	79	77
	Trained on	35	11	14	17	11	15	12	77	74	78
	all levels	55	8	18	14	29	25	24	72	69	73
		75	10	20	8	30	26	24	70	65	73

Table 2.2: *Missed-onset and false-alarm rates (as percentages of the number of actual onsets) found for the detection schemes, for the T-sentence database. 'MLP Mel-spectra' refers to training the MLP with mel-scaled spectra, with normalization of the RMS of* the whole sentence. *In the case of 'MLP Normalized',* each individual *mel-scaled spectrum is normalized. 'MLP Summed' refers to the single input unit containing the summed spectral values. 'MLP Chop-T' refers to training with the cochlear nucleus simulation. See text for a description of the 'channels 1-2-23' conditions.*

2.4.2 Results

2.4.2.A The T-sentences

Performance scores for the detection schemes are listed in Table 2.2. The figures given for VSM performance are the same as in Hermes (1990).

For the CNS scheme, a difference between sound-pressure-level conditions can be observed: if all sentences are normalized to 35 dB SPL, the number of missed onsets is about twice the missed-onset rate for 55 or 75 dB SPL. The missed-onset rates for the

latter two levels are comparable to VSM performance. The false-alarm rate is substantially lower for the 35 dB SPL condition than for the two other levels. Many more false alarms are obtained with CNS, for the 55 and 75 dB SPL conditions, than with VSM. However, if harmonicity information is used in the detection process, as in CNS-ACF, then the false-alarm rate can be seriously reduced. This effect is most prominent for the 75 dB SPL condition, where analysis of false-alarm occurrences showed that the harmonicity criterion especially rejected onsets for unvoiced plosives. The reduction of false alarms does not affect the missed-onset rate significantly, except for the 35 dB SPL condition. Accuracy figures for the CNS and CNS-ACF schemes are generally lower than the accuracy figure for VSM.

The MLP scheme, with mel-scaled spectra training, yields a missed-onset rate comparable to VSM performance (10%) and a higher false-alarm rate (10%) for the condition where each sentence is normalized. If instead each spectrum is normalized, a comparable missed-onset rate (9%) is obtained while at the same time the false-alarm rate increases dramatically (36%). These results indicate that level differences within a sentence contribute to vowel-onset detection. Accuracy figures for these MLP schemes are also slightly worse than the VSM accuracy figure. An interesting result is obtained for the 'MLP-Summed' condition, which shows comparable missed-onset and false-alarm rates as the mel-spectra case (sentence RMS normalized). This finding indicates that instantaneous level or, more specifically, change of instantaneous level contains much information for vowel-onset detection. Analysis of phonemic context of missed onsets, however, reveals that the MLP trained on the full-resolution mel-spectra has a relatively higher missed-onset rate of unaccented /ə/-vowels, and a lower missed-onset rate of accented vowels.

Finally, we will discuss the performance results obtained with MLP networks trained with the simulated Chop-T responses, in either the 1, 2, or 23-channel representation. For all representations we applied two training conditions: (a) the network was trained and tested on a single sound-pressure level, and (b) the network was trained on all levels but tested on a single level. In the case of training condition (a) we see that the missed-onset rates are comparable for the 2 and 23-channel representation, and that the false-alarm rate is slightly higher for the 2-channel representation. These results indicate that the 2-channel representation, despite its reduced spectral resolution, contains almost as much information for vowel-onset detection as the 23-channel representation. Comparing the results of this training condition to the CNS model, we find performances which are generally comparable, especially at 55 and 75 dB SPL. The accuracy of the CNS model is overall higher by approximately 5 %. If instead only a measure of the instantaneous intensity is used, like in the single-channel condition, the evaluations generally show higher missed-onset and lower false-alarm rates.

For the 2 and 23-channel representations, application of training condition (b) yields an overall trend similar to training condition (a) discussed above. However, missed-onset and false-alarm rates are generally worse than for condition (a): they are higher by a factor of two in the case of the 2-channel representation, and moderately higher at low levels in the case of the 23-channel representation. The false-alarm rates are higher by a factor of two to three for both representations at nearly all levels. These results indicate that the spectral representation in (simulated) Chop-T responses is fairly stable over a wide dynamic range as missed-onset rates for the 23-channel representation are comparable for conditions (a) and (b). The increase of missed-onset rate for the 2-channel representation in condition (b) is possibly due to the absence of absolute-level cues within the training set. The same factor probably underlies the increase in false-alarm rate for both representations under condition (b). The results for the single-input condition are in general agreement with this observation; in contrast to the reduction in missed onsets relative to the condition of training and testing on a single level, the false-alarm rate increases dramatically. This finding indicates that the absolute level of (summed) activation is not a very reliable cue.

On the basis of these results, we made a selection of schemes that were tested on the PM-sentence database: the VSM, CNS-ACF and MLP (mel-scaled spectra training and sentence-level normalization) schemes will be compared. Their performances will be analyzed in more detail in terms of phonemic context of occurrence of missed onsets and false alarms.

2.4.2.B The PM-sentences

Vowel-onset-detection performances for the PM-sentence database are listed in Table 2.3. Missed-onset rates are found to be almost identical to the rates reported above for the T-sentence database. Not only do rates show a high similarity, but phonemic categories of missed onsets correspond as well (see below for a listing of categories for the PM-sentence database). An exception to the latter finding applies to the CNS-ACF scheme: a higher occurrence of missed onsets for /i/-sounds is measured for the PM-sentence database. On the other hand, comparing false-alarm rates for both databases shows that rates for the PM database are generally higher. We suspect that this finding results from differences between both database in the speakers' care of articulation. Hermes (1990) reported a similar increase of false-alarm rate when testing VSM, optimized for fluent speech, on a database containing carefully articulated, isolated nonsense words. In addition, a contribution to the effect may have come from differences in recording quality, which is higher for the PM database. As opposed to the performance of VSM and CNS-ACF, the MLP scheme shows a rather stable performance.

An analysis of occurrences of missed onsets is given in Table 2.4. For all detection

	Test level (dB SPL)	Missed onsets (%)	False alarms (%)	Accuracy (%)
VSM		7	9	90
CNS-ACF	35	20	9	92
	55	10	10	91
	75	10	14	90
MLP Mel-spectra		9	14	88

Table 2.3: *Missed-onset and false-alarm rates (as percentage of the number of actual onsets) found for the detection schemes, for the PM-sentence database. MLP Mel-spectra refers to network training with mel-scaled spectra, with normalization of the RMS of* the whole sentence.

schemes main categories of missed onsets are /ə/, /i/, /ɪ/. The VSM-scheme does show a moderate number of missed onsets for /e/, but not for the categories /ɑ/, /y/, and /u/. For CNS-ACF and MLP, these latter categories are seen to give rise to a number of missed onsets. Missed onsets for /ə/-sounds generally occurred in unaccented syllables. It is surprising to find that the high vowels /i/, /ɪ/, and to a lesser extent /ɛ/ and /y/, cause difficulties in detection for all schemes. For all schemes, this may be explained by the observation of a delayed detection coinciding with the high second formant reaching its full resonance. Such a detection was typically delayed by more than 60 ms and was thus regarded as a false alarm. We will refer to such phonemic contexts as long vowels.

Table 2.5 lists the phonemic categories of occurrences of false alarms. The main general categories are long-vowel, /r/, unvoiced-fricative, and nasals contexts. To a slightly lesser extent, false alarms are given for the categories voiced fricatives and /ə/-like. In the case of the latter category, the phonetician did not mark these contexts as vowel onsets because they were poorly articulated. Despite the introduction of a harmonicity criterion in CNS-ACF, we also find a large number of false alarms for unvoiced plosives, even though the false-alarm rate for unvoiced plosives is already largely reduced by this criterion. Also in the case of the MLP scheme, unvoiced plosives and unvoiced fricatives are found to evoke a number of false alarms. With an exception of the unvoiced contexts, the overall distribution of false-alarm categories is in good agreement with the data presented in (table I in Hermes, 1990).

2.5 Discussion

2.5.1 Detection schemes

The occurrences of both missed onsets and false alarms in the cases of all three schemes are found to be distributed over more or less the same phonemic categories. This is

	VSM	CNS-ACF(35)	CNS-ACF(55)	CNS-ACF(75)	MLP
/ə/	17	37	18	19	15
/i/	6	12	9	10	11
/ɪ/	1	13	5	3	5
/e/	4				1
/ɛ/		3	2	2	
/o/	1	2	1	1	1
/ɔ/	1	2	1	2	
/a/	1	4	1	2	3
/ɑ/	1	12	2	1	2
/y/	1	3	2	2	3
/u/	1	1	2	3	3
/ɛi/		1	1	1	
/ø/		1	1	1	
Σ	34	91	45	47	44

Table 2.4: *Occurrences in raw numbers of phonemic contexts of missed onsets, for the PM-sentence database. 'CNS-ACF(35)' indicates results obtained for 35 dB SPL input level.*

an interesting finding given the fact that the schemes differ considerably in their pre-processing of the input signals. The main category of missed onsets is 'ə-like', occurring in unaccented syllables, which is in agreement with data presented in Hermes (1990). Furthermore, all three schemes have difficulties in detecting onsets of high vowels. In the case of false-alarms, a similar distribution is found as presented in Table 2.1 in Hermes (1990), although the present evaluations show higher false-alarm rates for /l/-contexts and nasals. As is mentioned in Hermes (1990), some categories of false-alarms can function as vowels in other phonetic contexts, e.g., in the context of syllabic consonants or in other languages. The fact that all three schemes show similar trends may therefore support the conclusion that the presence of these false alarms has a phonological background.

Within the scope of improving the automated vowel-onset detection, it is found that an MLP network trained with mel-scaled spectra, with normalization of the RMS of the whole sentence, is competitive with the VSM scheme. In terms of improvement of the automatic vowel-onset detection, the CNS-ACF scheme does not seem to be an obvious candidate; missed-onset and false-alarm rates are generally higher than for VSM, and the performance of the scheme is found to depend on level. However, this may have a perceptual counterpart. For speech, a comfortable loudness level appears to be well specified. The determination of actual onsets with the gating technique was normally

	VSM	CNS-ACF(35)	CNS-ACF(55)	CNS-ACF(75)	MLP
Long vowel	18	21	16	16	15
/ə/-like	3	3	3	3	4
/r/	4	7	8	8	8
/k/,/p/,/t/	4	1	5	16	15
/l/	2	3	6	5	4
/n/,/m/	5	3	4	5	4
/j/,/w/		1	2	3	1
/b/,/d/				2	1
/χ/,/s/	4		1	3	10
/z/,/v/	2	1	2	4	1
others				2	
Σ	42	40	47	67	63

Table 2.5: *Occurrences in raw numbers of phonemic contexts of false alarms, for the PM-sentence database. 'CNS-ACF(35)' indicates results obtained for 35 dB SPL input level.*

done at a comfortable loudness level, which presumably was substantially higher than 35, and moderately lower than 75 dB sound-pressure level. If the determination had been done at other than comfortable levels, the correspondence between the detection behavior of CNS(-ACF) and the human listener may have been substantially better. In this respect, the performance of the CNS(-ACF) scheme presumably is more comparable to human vowel-onset detection.

It should be noted that the determination of vowel onsets by the human listener was done under conditions with high signal-to-noise-ratio (SNR) levels, where 'noise' refers to background noises. Likewise, the models were trained and tested with speech signals having high and very high SNR levels for the T and PM-sentence databases, respectively. This means that model performances may deteriorate if trained and/or tested under worse SNR levels. The experiments reported in this Chapter have not been repeated for such levels, however. Ainsworth and Meyer (1994) investigated speech *recognition* performances using Hidden Markov Models (HMMs) trained and tested on a number of simulated-response representations of various stages of auditory processing, namely the auditory nerve and cochlear nucleus. Training and testing was done for SNR levels up to 0 dB. Recognition performances were compared to human recognition scores. Of all simulated representations, it was found that the results obtained by using Chop-T representations were most like human performance. This finding may implicate that the simulated representation of Chop-T responses also provides a reasonably robust representation for vowel-onset detection under low SNR conditions. It should be stressed,

though, that this is merely speculative at present and requires experimental evidence.

2.5.2 Signal characteristics

The comparative tests give more insight into the relative importance for vowel-onset detection of the signal-characteristics intensity, spectral envelope, and harmonicity.

2.5.2.A Intensity

The importance of intensity may be derived from comparing vowel-onset-detection models and training conditions. First, for MLP networks trained on mel-scaled spectra, comparable performance is obtained for the conditions of (a) a single input, i.e. the sum of all 23 spectral points, and (b) the full-resolution 23-channel input. In both training conditions, the RMS of each sentence processed are normalized. This finding indicates that increments of intensity are important for vowel-onset detection. Second, for the condition that each individual spectrum is normalized before being processed by an MLP network, detection performance is moderate compared to the condition that the RMS of each sentence is normalized. Third, if the MLP networks are trained with Chop-T representations at all levels and tested on a single sound-pressure level, then the false-alarm rate is substantially higher than with training on a single level. Finally, the difference between the MLP scheme trained on the 2-channel Chop-T representation and the CNS scheme is that in the latter the derivative of activation plays a key role: vowel-onset candidates are selected at local maxima of the derivative and its time course is explicitly used as a criterion. Comparing the results for the MLP and CNS, it is found that performances are comparable for the condition that the MLP is trained and tested on a single level. With training and testing on all levels, however, the MLP performance is seen to be substantially worse. Assuming that training of the MLP constructed an optimal classification boundary based on the rough 2-channel spectral envelope, this finding indicates that the derivative of activation can be conceived of as a more robust source of information than spectral envelope *per se*. In sum, the comparison of different models and training conditions indicates that increments of intensity relative to sentence level are of prime importance for modeling human vowel-onset detection. This observation is in agreement with the fact that also in VSM, intensity is an important source of information.

2.5.2.B Spectral envelope

An MLP network trained with the 2-channel Chop-T representation is seen to exhibit reasonable performance, provided training and testing is done on a single sound-pressure level only. A similar, rather crude spectral weighing is the basis of the CNS(-ACF) schemes, which also show reasonable performance. These findings give support to the

hypothesis that vowel-onset detection predominantly relies on the rough spectral envelope. On the other hand, false-alarm rates for the CNS-ACF scheme prove to be substantially lower than for CNS, especially at higher sound-pressure levels. This finding indicates that rough spectral envelope is sufficient for detecting vowel-onset candidates, but that more information is required for correct rejection of false alarms. It should be noted, though, that hardly anything is known about integration of activity over such large bandwidths as in the CNS(-ACF) scheme.

2.5.2.C Harmonicity

The false-alarm rate obtained with the CNS-ACF scheme is substantially lower than the rate obtained with the CNS scheme, while the missed-onset rate is practically identical. This finding may give support to the hypothesis that harmonicity information plays an important role in modeling vowel-onset detection. On the other hand, mel-scaled spectra do not show a harmonic structure, yet the corresponding MLP performance is found to be satisfactory. This may mean that (a) harmonicity information is present in the mel-scaled spectra in some other way, or that (b) in human vowel-onset detection, harmonicity information is only used to reject some onsets other than vowel onsets. Possibility (a) can be verified by adding harmonicity information to the MLP input, and comparing performance with and without such an extra input unit. Possibility (b) would be in line with the decrease of false-alarm rate observed for CNS-ACF relative to CNS.

2.5.3 Onsets in (simulated) Chop-T responses

To our knowledge, there are no physiological data reported in the literature pointing at the specific coding of transitions such as vowel onsets. This means that the perceptual information used for the specific detection of vowel onsets cannot directly be derived from measured or simulated response patterns.

Initially, we measured the presence of a formant structure ("vowel strength") in the simulated Chop-T representation. In this array of simulated neuron responses, the formant structure is enhanced by lateral inhibition resulting in a stronger spectral contrast. Blackburn and Sachs (1990) observed that spectral contrast is preserved over a wide dynamic range. However, these measurements were obtained by taking the long-term average of discharge activity during steady state, which may not adequately describe discharge activity at the onset. If spectral contrast is preserved over a broad dynamic range, then we can also expect vowel strength to be stable over a broad range. The measurements did not provide support for this expectation, as especially at high signal levels (75 dB SPL), contrast in vowel strength diminished between vowel and non-vowel contexts. As a result, the false-alarm rate increased considerably, which indicated that the relative contribution of onsets other than vowel onsets started to increase.

	Test level (dB SPL)	Missed syllables (%)	Falsely detected syllables (%)
VSM		5	8
CNS-ACF	35	16	5
	55	7	8
	75	7	12
MLP Mel-spectra		7	11

Table 2.6: *Rates of* missed syllables *and* falsely detected syllables *(as percentage of the number of actual vowel onsets) found for the detection schemes, for the PM-sentence database. MLP Mel-spectra refers to network training with mel-scaled spectra, with normalization of the RMS of* the whole sentence.

2.5.4 Syllabification

In this study, a vowel onset was considered to be correctly detected if the detection algorithm signaled a vowel onset within 60 ms of a vowel onset indicated by a phonetician. This rather strong demand can be loosened by considering vowel-onset-detection algorithms as a means for syllabification. For syllabification it is necessary to have one onset per syllable, but this onset does not necessarily have to coincide with the vowel onset. Hunt (1993) presents a study of syllabification by detecting onsets and offsets of vowels by means of recurrent neural networks. The following results were reported: 6% missed syllables and 5% falsely detected syllables, with an accuracy figure for vowel-onset and offset detection of 87%. For the purpose of comparison, Table 2.6 presents missed-onset and false-alarm rates if the three schemes of the present study are applied to detect syllables, instead of vowel onsets. These rates were found by taking one algorithmically detected vowel-onset per syllable, while allowing that the onset does not coincide with an actual vowel onset. If there are more than one detections per syllable, these are regarded as falsely detected syllables.

In comparison to vowel-onset detection (Table 2.3), the CNS-ACF scheme benefits most from this redefinition of the task: in general, the missed-onset and false-alarm rates decrease with approximately 4%. Analysis of the phonetic contexts showed that this improvement especially results from correct syllabification in long vowel contexts, where this context is often seen to cause false alarms in vowel-onset detection. The proportion of falsely detected syllables still is slightly higher than the data presented by Hunt (1993). This may, however, also have resulted from the use of another database (i.e., TIMIT), analogously to the differences found in the false-alarm rate for the T and PM databases in the present study.

2.5.5 Perceptual mechanism and phonological structure

It is well known that speech can induce a rhythmic beat which runs synchronously with the syllables which make up the speech signal. Various early studies have shown that the moment of occurrence of this rhythmic beat is close to the vowel onset of the syllable (Rapp, 1971; Allen, 1972; Cole and Scott, 1973). Allen (1972, p. 72) mentions that "the rhythmic beats were closely associated with the onsets of the nuclear vowel of the stressed syllables, but precede those vowel onsets by an amount positively correlated with the length of the initial consonant(s) of the syllable". Cole and Scott (1973) reported the importance of "vowel transitions" for the perception of the temporal order of syllables. Studies like these into the rhythmic structure of speech has led to the concept of the perceptual moment of occurrence of the syllable, or its P-center (Morton et al., 1976). Marcus (1981) showed that the length of the onset and the rhyme of the syllable affects its P-center, so that the P-center cannot exactly be identified with the vowel onset. Pompino-Marshall (1989, 1990) developed a detailed model for algorithmic P-center determination, in which he estimates the P-center from the acoustic waveform on the basis of a weighted average of onsets and offsets in the course of the specific-loudness of the speech sounds. Since the strongest onsets in a syllable occur in the neighborhood of the vowel onset, where the oral cavity opens and the formants start to rise, this model predicts that the P-center will be close to the vowel onset. In the Pompino-Marshall model, secondary onsets and offsets in the course of the syllable will move the estimated P-center forward or backward. For a number of syllables, this model is able to predict the shift of the P-center quite accurately. The shifts are never larger than at most a few dozens of ms, however. Therefore, the vowel onset appears to be a good first approximation for the location of the P-center. Also for the rhythmic production of syllables, it is concluded from various studies that, among other phonetic events, the vowel onset best corresponds with the moment speakers use in timing their syllables (Eriksson, 1991).

Phonologically, the syllable has been divided into onset and rhyme. Although this binary division has been supplemented by a division of the rhyme into nucleus (or peak) and coda (Pike, 1947; Selkirk, 1982), it is generally assumed that the boundary between onset and rhyme is more important (Treiman, 1986). In metrical phonology, the syllable is divided into a weak onset and a strong rhyme. The latter in its turn is then divided into a strong nucleus and a weak coda. This implicates that the strongest transition is from the weak onset to the strong rhyme. The transition from nucleus to coda is of a lower level; the transition between two syllables, though of a higher level, takes place from the weak coda of the first syllable to the weak onset of the second. The latter transition is subject to many coarticulatory phenomena, even if the two syllables belong to different words. Therefore, the syllable onset, the transition from one syllable

to the next, is much less well defined than the transition from onset to rhyme. In this respect it is interesting to note that the three algorithms tested, though very different in nature, show errors in the same phonetic contexts, which, as argued above, might have a phonological background.

From both a phonetic and a phonological point of view, it is therefore concluded that the best candidate for timing the syllable is the transition from onset to rhyme. In production and perception, the actually realized and perceived temporal moment of occurrence is closely linked with the vowel onset, though higher order processes can apparently shift the actual moment of occurrence for a few tens of ms. Large shifts occur only when the syllable onset contains (sonorant) segments with strong onsets. For isolated syllables, it may eventually not be too difficult to estimate these shifts from the acoustic speech signal preceding and following the vowel onset. But also in this case, one has to start from the vowel onset. In running speech, the situation is much worse. The syllable onset is often badly defined, in any case much worse than the vowel onset. It is as yet impossible to estimate how much onsets and rhymes will contribute to the rhythmic beat perceived. It is a well known fact that perceived rhythmicity, even in poetry and music, does often not correspond with regular intervals between any well known acoustic event of the sound signal. Nevertheless, we are very sensitive for small changes in the temporal structure of speech signals. In general, it can be concluded that higher order, "top-down" processes push the speech events into a rhythmic frame in which the perception of rhythmicity occurs before it is clearly present in the temporal series of acoustic events (For a review, see Eriksson, 1991). In this situation, it is likely that vowel onsets do not precisely correspond with the perceptual moment of occurrence of the syllable. Taking all arguments together, however, they come out as the first phonetic events to be considered in studying perceived rhythmicity in speech.

2.6 Conclusions

In this Chapter, three methods for automatic detection of vowel onsets in speech have been presented and evaluated. One of these methods, namely vowel-strength measurement, was presented in an earlier publication. For evaluation, two databases of read Dutch speech uttered by non-professional male and female speakers have been used. Both databases can be characterized as having a good recording quality with high signal-to-noise ratios. For all methods and corresponding parameter settings, missed-onset rates are found to be better than 25%, and false-alarm rates are not seen to exceed 30%. For the best performing schemes, missed-onset and false-alarm rates are found to be in the order of 10%. In spite of the substantial differences between the three methods, an analysis of the phonetic contexts of missed onsets and false alarms has shown that these contexts generally match.

The method of vowel-strength measurement has been found to be overall best performing. This method is not only based on spectral-envelope information, but also depends to a great extent on both intensity as well as harmonicity information.

The second method presented is based on simulation of Chop-T responses found in the anteroventral cochlear nucleus. Reasonably good results were obtained by using a detection scheme that post-processes the simulated responses, where rough spectral-envelope, intensity, and harmonicity information are used. However, this method was seen to be rather input-level dependent which may have, as has been argued in the Discussion, a perceptual counterpart.

The third detection method is based on training multi-layer perceptrons having a single hidden layer. Best results for these MLPs are obtained if the input to the network consists of mel-scaled spectra. Intensity information, i.e. the distribution of intensities over the whole sentence, has been found to be an important source of information in this method. Training the MLPs with the simulated Chop-T responses only resulted in fair performance if training and testing was done at the same input signal sound-pressure level.

In summary, the performance results for the different methods and parameter settings support the hypothesis that the main information required for automatic vowel-onset detection are (a) rough spectral envelope and (b) intensity. Harmonicity information can be conceived of as an additional source of information to reduce the false-alarm rate.

Chapter 3

Psychoacoustical evaluation of PSOLA. Single-formant synthetic stimuli[1]

3.1 Abstract

This Chapter presents two experiments dealing with a psychoacoustical evaluation of the Pitch Synchronous OverLap and Add (PSOLA) technique. This technique has been developed for modification of duration and fundamental frequency of speech and is based on simple waveform manipulations. Both experiments were aimed at deriving the sensitivity of the auditory system to the basic distortions introduced by PSOLA. In experiment I, manipulation of fundamental frequency was applied to synthetic single-formant stimuli under minimal stimulus-uncertainty, level roving, and formant-frequency roving. In experiment II, the influence of the positioning of the so-called "pitch-markers" was studied. Depending on the formant and fundamental frequency, experimental data could be described reasonably well by either a spectral intensity-discrimination model or a temporal model based on detecting changes in modulation of the output of a single auditory filter. Generally, the results were in line with psychoacoustical theory on the auditory processing of resolved and unresolved harmonics.

[1]This chapter is a slightly modified version of Kortekaas, R.W.L. and Kohlrausch, A. (1997) "Psychoacoustical evaluation of the pitch-sychronous overlap-and-add speech-waveform technique using single-formant stimuli," J. Acoust. Soc. Am. **101**, 2202-2213.

3.2 Introduction

Over the past few decades, considerable research activities have concentrated on instrumental modification of the fundamental frequency and duration of natural speech. These types of modification enable the manipulation of speech prosody: modification of duration typically alters speech rhythm and tempo, whereas modification of fundamental frequency changes intonation. Characteristics not strictly pertaining to prosody, such as phonemic content and voice quality, ideally remain unaffected by these modifications. Numerous techniques have been proposed with the general aim of both maximizing intelligibility and perceived synthesis quality and minimizing computational complexity. A class of digital-signal-processing techniques with generally low complexity is the so-called OverLap and Add (OLA) framework (Rabiner and Schafer, 1978). For instance, time-domain OLA, where all operations are performed on the waveform itself, has been successfully applied not only in speech manipulation (e.g., Roucos and Wilgus, 1985) but also in other fields such as music synthesis (e.g., Roads, 1988).

This chapter will focus on Pitch Synchronous OverLap and Add (PSOLA; Moulines and Charpentier, 1990; Moulines and Laroche, 1995), which is a variant of time-domain OLA[2]. The main feature of PSOLA is that the OLA operations are aligned to the (quasi-)periodicity of the input speech signal. PSOLA has found widespread application, e.g., in modules for text-to-speech synthesis and as a tool for fundamental speech-perception research (Moulines and Laroche, 1995). PSOLA-manipulated natural speech is generally characterized not only by high intelligibility but also by high synthesis quality. This finding is remarkable given the fact that, as will be described in the following section, the technique is based on rather rough signal operations.

Despite the generally satisfactory synthesis quality of PSOLA, annoying artefacts are sometimes introduced. Although a strict categorization is difficult, these artefacts can often be described as hoarseness and roughness of the synthesized signal. In addition, artefacts similar to comb-filtering are observed in practice. As far as we know, the occurrence of these artefacts cannot be predicted beforehand. This unpredictability is, in our opinion, caused to a great extent by the lack of knowledge of the perceptual effects of the (PS)OLA operations, a view that was also expressed by Moulines and Laroche (1995).

Even if PSOLA manipulation of a speech signal does not lead to the perception of either of the artefacts mentioned above, the manipulation does affect its spectral content. To explain the success of PSOLA manipulation, one may hypothesize that these spectral changes are either perceptually subliminal or, within the context of speech perception,

[2]Although a frequency-domain (FD-PSOLA) variant has also been proposed, the time-domain version (TD-PSOLA) has been commonly preferred due to its computational efficiency.

phonetically less relevant (c.f., Klatt, 1982). This Chapter addresses the first hypothesis by determining the detectability of the spectral changes and by deriving the auditory cues involved in this detection process. Such a psychoacoustical basis is probably important for the long-term aim of increasing the predictability of audible (and annoying) artefacts. In addition, psychophysical evaluation may also increase knowledge about the auditory processing of speech.

In the experiments, synthetic single-formant signals (Klatt, 1980) were PSOLA manipulated. Apart from their application in speech synthesizers, single or multiple formant signals have been used to determine, e.g., JNDs in formant frequency (for a recent overview, see Lyzenga and Horst, 1995) and in fundamental frequency (Flanagan and Saslow, 1958; Klatt, 1973). Single-formant signals are used here to derive the sensitivity of the auditory system to the "basic distortions" introduced by the PSOLA operations. To establish the link to "classical" psychoacoustics, Appendix A presents experimental results concerning the auditory sensitivity to basic distortions when manipulating pure tones.

In experiment I, the perceptual effects of fundamental frequency ($F0$) manipulation for three levels of stimulus uncertainty, mimicking particular aspects of natural speech, are investigated. In experiment II, the perceptual effects of the "pitch-marker" location are studied (see General Methods). Both experiments focus on $F0$ modification only because, in practice, this type of manipulation is more likely to result in annoying artefacts than the manipulation of duration. In addition, the experimental data are compared with predictions of two model simulations: a model based on detecting intensity differences between excitation patterns (Durlach et al., 1986; Gagné and Zurek, 1988) and a model based on the discrimination of modulation depth within a single auditory filter.

3.3 General methods

3.3.1 The PSOLA technique

The PSOLA technique is a time-domain variant of the so-called OverLap-Add (OLA) technique for analysis-synthesis (Rabiner and Schafer, 1978; Allen and Rabiner, 1977). OLA generally consists of three steps: (1) decomposition of a signal into separate, but often overlapping, segments, (2) optional modification of these segments, and (3) recombination of the segments by means of overlap-adding. PSOLA consists only of steps (1) and (3). A short introduction to PSOLA will be presented here; for further details the reader is referred to Moulines and Charpentier (1990) and Moulines and Laroche (1995).

Figure 3.1(a) shows the waveform of a synthetic single-formant signal as used in both experiments. This signal is decomposed into separate segments in analysis step (1)

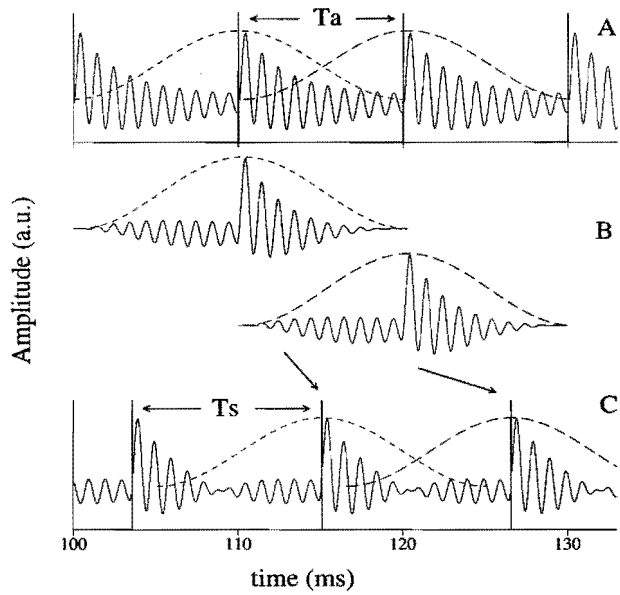


Figure 3.1: *Illustration of the PSOLA technique: Panel (a) shows the waveform of a synthetic 1000-Hz single formant signal with a fundamental of 100 Hz. At the pitch-marker locations, indicated by thick vertical lines, the signal is decomposed by means of Hanning windowing. The interval between two pitch markers is indicated by T_a. Two segments are shown in panel (b). These segments are recombined by means of overlap-adding at the new pitch-marker positions indicated by thick vertical lines in panel (c). These pitch markers are regularly spaced at 11.5 ms, indicated by T_s, which gives a fundamental frequency of 87 Hz.*

by windowing it at particular time instances. These instances, represented by vertical lines in Fig. 3.1(a), are positioned *pitch synchronously* and are called "pitch markers". Pitch markers are determined either manually by inspection of the speech waveform or automatically by means of some local $F0$ estimation (e.g., Ma et al., 1994; Smits and Yegnanarayana, 1995). Figure 3.1(b) shows two segments extracted from the input signal. The maxima of the Hanning (raised-cosine) windows coincide with the pitch markers. The window duration depends on the temporal spacing between pitch markers; consecutive windows have 50% overlap. Because adjacent windows sample-wise add up to one, the input signal can be restored perfectly. Note that in natural speech windows will typically be asymmetrical due to variation in $F0$.

Segment recombination in synthesis step (3) is performed after *defining* a new pitch-marker sequence. In Fig. 3.1(c), the new sequence is represented by vertical lines. An

output signal is synthesized by first assigning a decomposed segment to each of the new pitch markers and then performing the sample-wise overlap-add operation. Manipulation of fundamental frequency is achieved by changing the time intervals between pitch markers. In Fig. 3.1(c), for instance, these intervals are increased, leading to the percept of a lower pitch. Modification of duration, on the other hand, is achieved by either repeating or omitting segments. Note that, in principle, modification of fundamental frequency also implies a modification of duration.

3.3.2 Terminology

In the experiments manipulation was investigated for signals having a constant $F0$. This means that the pitch markers in the decomposition and synthesis phase are positioned at regular intervals. These intervals will be denoted by T_a and T_s, respectively. Analogous with the fundamental frequency, we introduce the "window rates" $F_{wa} = 1/T_a$ and $F_{ws} = 1/T_s$. In the experiments the analysis window rate F_{wa} was fixed and F_{ws} was the experimental parameter. In what follows, experimental results will be presented as a function of ΔF given by:

$$\Delta F = \frac{F_{ws} - F_{wa}}{F_{wa}} \times 100\ \%.$$

For positive and negative values of ΔF, the symbols ΔF^+ and ΔF^- will be used.

Under some experimental conditions the perceptual effects of pitch-marker location were investigated. The pitch-marker location will be denoted by the parameter ΔP. As will be described in section 3.4.1, the single-formant signals are generated by exciting a formant filter with a regular pulse train. The parameter ΔP indicates the shift of the pitch markers relative to the excitatory pulses. This shift will be given as a percentage of T_a. In Fig. 3.1(a), for instance, the pitch markers coincide with the formant-filter excitations so that $\Delta P = 0$ %. Because the formant filter is minimum phase, the amplitude maxima of the signal in Fig. 3.1(a) are only slightly delayed relative to the maxima of the Hanning window. On the other hand, if $\Delta P = 50$ % the pitch markers are located *between* excitations of the formant filter. In that case maxima of the input signal and the Hanning windows are maximally misaligned.

3.3.3 Distortions in pure tones

First, we consider PSOLA manipulation of a single pure tone which is thought of as a component of a harmonic spectrum. Figure 2(a) depicts a pure tone of carrier frequency $f_c = 1000$ Hz, assumed to be the 10th harmonic of a 100-Hz fundamental. After decomposition at intervals of $T_a = 10$ ms and overlap-adding, the signal shown in Fig. 3.2(b) is synthesized where ΔF = - 2.44 % ($F_{ws} = 97.56$ Hz, $T_s = 10.25$ ms). In contrast to

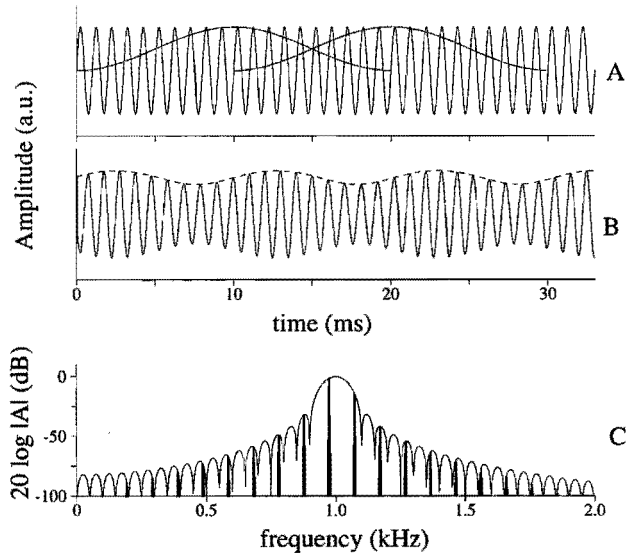


Figure 3.2: *A pure tone signal of f_c = 1000 Hz, shown in panel (a), is decomposed into segments by windowing the signal at a rate of F_{wa} = 100 Hz. These segments are recombined at a rate of F_{ws} = 97.56 Hz to synthesize the signal in panel (b). The dashed line represents the Hilbert envelope of the synthesized signal. Panel (c) shows the log-amplitude spectra of a single segment (thin solid line) and the synthesized signal (thick vertical lines).*

the original pure tone, this signal shows amplitude modulation (AM) in its envelope and frequency modulation (FM) in its fine structure. For a sinusoidal input signal these two changes are the basic distortions introduced by PSOLA. Experimental results relating to the auditory sensitivity to these distortions will be presented in Appendix A. The AM of the envelope is partly caused by the fact that adjacent Hanning windows do not sum up to one if $T_a \neq T_s$. This can be compensated for by using a "synthesis window". The perceptual relevance of using such a window will be discussed in the Discussion.

Alternatively, we can describe the distortions in the spectral domain. Time-domain multiplication (windowing) results in frequency-domain convolution of the spectra of the Hanning window and the pure tone (e.g., Rabiner and Schafer, 1978). The thin solid line in Fig. 3.2(c) depicts the log-amplitude spectrum of a single segment decomposed from the original pure tone of Fig. 3.2(a). The overlap-adding operation in synthesis, which extends the signal periodically in the time domain, is equivalent to *resampling* the (complex) spectrum of a single segment (Moulines and Laroche, 1995). The log-amplitude spectrum of the synthesized signal in Fig. 3.2(b) is shown by the line spectrum in Fig. 3.2(c). The spectral lines are harmonics of F_{ws} = 97.56 Hz. For example,

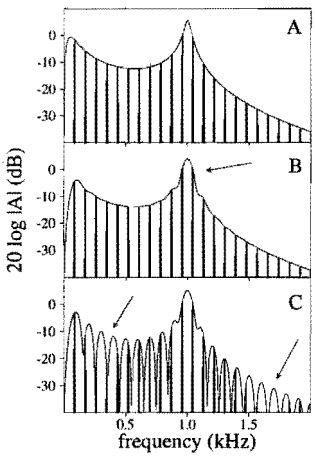


Figure 3.3: *Panel (a) shows the log-amplitude spectrum of a 1000-Hz single-formant signal with a fundamental of 87 Hz. The thin line represents the amplitude transfer function of the formant filter including pre-emphasis. Panel (b) shows the log-amplitude spectrum for a PSOLA-manipulated single-formant signal, shifted in fundamental frequency from 100 to 87 Hz. Parameter ΔP was set to 0 %. The thin solid line now represents the log-amplitude spectrum of a single segment decomposed from the input signal. Arrows indicate frequency regions of maximal difference between the spectra in panels (a) and (b). Panel (c) shows the spectrum of a PSOLA-manipulated signal but now with ΔP set to 50 %. Arrows indicate notches introduced in the log-amplitude spectrum.*

the strongest harmonic has a frequency of $10 \times F_{ws} = 975.6$ Hz. In other words, the introduction of AM and FM has a spectral counterpart in terms of the interaction of introduced components (Goldman, 1948)[3].

3.3.4 Distortions in single-formant signals

The experiments deal with the discrimination of PSOLA-manipulated and unmanipulated single-formant signals. Such signals intrinsically have a harmonic structure so that the introduction of side components *per se* cannot be a cue for discrimination. In fact, cues may be changes in both spectral envelope and phase relations between harmonics. These changes will be illustrated below.

Figure 3.3(a) shows the log-amplitude spectrum of an unmanipulated single-formant signal with an $F0$ of 87 Hz, a formant frequency of 1000 Hz, and a formant bandwidth of 50 Hz. This spectrum also shows the effects of pre-emphasis applied to the form-

[3]Strictly speaking, this only applies if the pure tone is a harmonic of F_{wa}.

ant signal (see section II.A.1). The corresponding phase spectrum (not shown here) is approximately linear except for a phase jump of π rad around the formant frequency.

Figure 3.3(b) shows the log-amplitude spectrum of a PSOLA-manipulated signal obtained by generating a 1000-Hz formant signal with an $F0$ of 100 Hz, decomposing it at $F_{wa} = 100$ Hz, and resynthesizing it at $F_{ws} = 87$ Hz. ΔP here is set to 0 %. The Hanning-windowing operation has "smeared out" the spectral envelope: the bandwidth of the pronounced formant is increased to approximately F_{wa} Hz. The spectral slope, however, remains almost unaffected. Changes to the phase spectrum (not shown) are a phase shift of approximately $\pi/4$ rad for the two harmonics around the formant frequency. Figure 3.3(c) shows the spectrum of a signal synthesized with ΔP set to 50 %. Its spectral envelope is clearly discontinuous which introduces pronounced notches in spectral envelope after resampling. The notch depth depends monotonically on ΔP. The corresponding phase spectrum (not shown) is discontinuous as well.

3.4 Experiments

The main questions in these experiments are: (1) what are the thresholds for the discrimination of PSOLA-manipulated and unmanipulated single-formant signals, (2) what is the influence of pitch-marker location on discrimination performance, and (3) how do the discrimination results relate to psychoacoustical models.

3.4.1 Method

3.4.1.A Stimuli

For generation of the single-formant signals, a second-order digital resonator was implemented as proposed by Klatt (1980)[4]. This filter was excited by a pulse train with an $F0$ of 100 or 250 Hz. The window rate F_{wa} was accordingly set to these values. Low-pass characteristics of natural voicing and high-pass radiation at the mouth opening, as described in Klatt (1980), were included as pre-emphasis. Formant frequencies f_r were 500, 1000, and 2000 Hz with -3 dB bandwidths of 50, 50, and 100 Hz, respectively.

As a baseline experiment, a minimal stimulus-uncertainty condition was investigated in which f_r was fixed and the overall level L was set to 70 dB SPL. To increase stimulus

[4]The implementation described in Klatt (1980) is based on a sample rate of 10 kHz. Cox et al. (1989) address the issue of using other sample rates, such as 32 kHz used here, which leads to differences in spectral envelope. They propose to introduce additional poles in the resonance-filter transfer function to compensate for the high-frequency attenuation. In the present study no compensation is taken into account because just a single formant is simulated (the high-frequency mismatch is more severe for multiple formants). Nevertheless, the spectral difference at 5 kHz between a signal generated at 10 kHz and at 32 kHz amounts to 10 dB. Around the formant frequency, differences are within 0.5 dB.

uncertainty, level roving between intervals was applied in the second condition. The level rove was uniformly distributed in the range ± 5 dB. As a third condition, the overall level L was fixed but the formant frequency f_r was roved uniformly over a range of ± 2 Δf_r. Δf_r here denotes one JND in formant frequency for which Gagné and Zurek (1988) reported the following relation: $\Delta f_r = 0.079 f_r/\sqrt{Q}$, where Q is the Q-factor of the formant filter. The range of f_r roving was within the range measured in a study by Pisoni (1980) in which subjects were instructed to "reproduce" steady-state synthetic vowels.

To investigate the perceptual effects of pitch-marker positioning, ΔP was set to 0 and 50 % in experiment I. In experiment II, psychometric functions for ΔP were determined for two particular values of ΔF.

Stimuli were software generated on a Silicon Graphics Indigo workstation. The sampling frequency was 32 kHz. Apart from the built-in filters of the workstation, no additional anti-aliasing filtering was applied. Because the amplitude spectrum of the single-formant signals monotonically falls off to approximately -80 dB (relative to the formant peak) at the Nyquist frequency, no aliasing is to be expected. After DA conversion signal levels were adjusted by means of analog attenuation. Stimuli were presented to the subject, seated in a soundproof booth, over Beyer DT 990 headphones. Subjects responded via a keyboard and received immediate feedback. Stimulus duration was 300 ms, the first and last 25 ms were ramped using a Hanning window. The interval separation was 200 ms.

3.4.1.B Procedure

Psychometric functions were measured using a 3I3AFC odd-ball procedure with fixed levels of ΔF in each run. The odd-ball interval contained the PSOLA-manipulated single-formant signal. This signal was obtained by (1) generating a formant signal with an $F0$ of F_{wa} Hz, (2) decomposing this signal at a window rate of F_{wa} Hz, and (3) resynthesizing it at a rate of F_{ws} Hz. The reference intervals contained a single-formant signal generated directly with an $F0$ of F_{ws} Hz. For determination of the psychometric function, F_{ws} was varied according to:

$$F_{ws} = \frac{1}{T_a + \frac{n}{4}} [Hz], \tag{3.1}$$

where $n = \pm$ 1 ms, ± 2 ms, ± 3 ms, ...

Each run consisted of 15 trials. For each condition, i.e., a combination of ΔF, F_{wa}, and f_r, a total of 5 runs were performed of which the first run was omitted from the analysis. Each data point thus represents 60 trials. All conditions were measured once before the next set of runs was initiated. Mean values and standard deviations of the 4 runs are shown in the figures below. Instead of plotting percentage correct as a function of ΔF,

the Pc values were converted to d' using a conversion table (Macmillan and Creelman, 1991).

3.4.1.C Subjects

Three subjects (aged 25, 27, and 35) participated in the experiments. All subjects had normal pure-tone thresholds in quiet for the frequencies 500, 1000, and 2000 Hz. Unlike subject RK (the first author), subjects MB and KM had no or little experience in psychoacoustic listening experiments. All subjects performed experiments I and II for f_r = 1000 Hz. Subjects KM and RK performed experiment I for f_r = 2000 Hz. Results for f_r = 500 Hz (experiment I) were obtained only for subject RK.

3.4.2 Experiment I: influence of ΔF

3.4.2.A Minimal stimulus uncertainty

Psychometric functions for minimal stimulus uncertainty are shown in the lefthand panels of Figs. 3.4, 3.5, and 3.6. Figure 3.4 presents the data for F_{wa} = 100 Hz and f_r = 1000 Hz. Data points for ΔP = 50 %, indicated by filled squares, are generally far above the threshold d' = 1 for all subjects. For ΔP = 0%, however, the psychometric functions show a non-monotonic behavior. For all three subjects, subthreshold discrimination performance is found for ΔF = -16.66, -9.09, and +11.11 %. These values correspond to values for T_s of 12, 11, and 9 ms, respectively, while T_a equals 10 ms. Because the 50-Hz formant bandwidth is rather small it can be stated that each decomposed segment contains 20 periods of a 1000-Hz carrier (cf. Fig. 3.1). Setting T_s to an integer multiple of 1 ms, which is the period of the 1000-Hz carrier, thus results in an *in-phase* addition of the fine structure of adjacent windows. This results in minimal distortion of the temporal envelope of the signal. In spectral terms, setting T_s to an integer multiple of the carrier period results in a harmonic coinciding with the formant frequency, due to the resampling property of PSOLA.

Using the same line of reasoning for the case of f_r = 1000 Hz and F_{wa} = 250 Hz, subthreshold discrimination performance is predicted for ΔF = -20 and +33.33 %. This is confirmed by the psychometric functions shown in the lefthand panels in Fig. 3.5. Note also that the data for other ΔF values show ceiling effects to even greater extent than the data in Fig. 3.4.

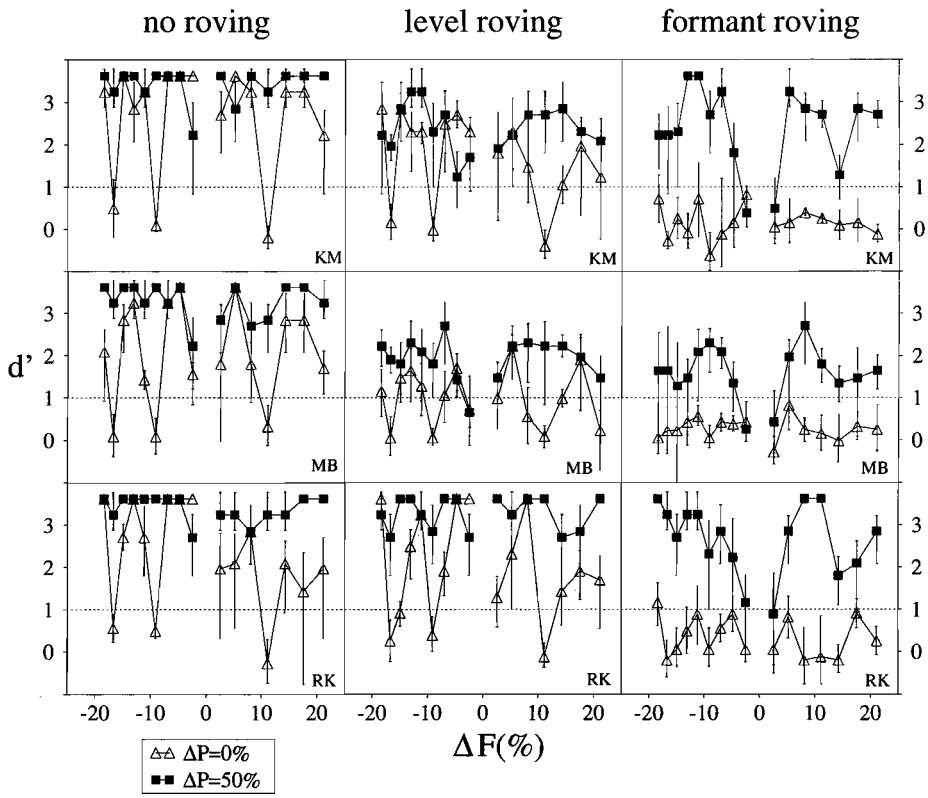


Figure 3.4: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated single-formant signals with formant frequency f_r = 1000 Hz and F_{wa} = 100 Hz. Mean data for ΔP = 0 % are shown by triangles, those for ΔP = 50 % by filled squares. Standard deviations are indicated by vertical lines. Lefthand panels show results for the minimal-stimulus-uncertainty condition. Center panels show psychometric functions for the level-roving condition, and the righthand panels data for roving of f_r.*

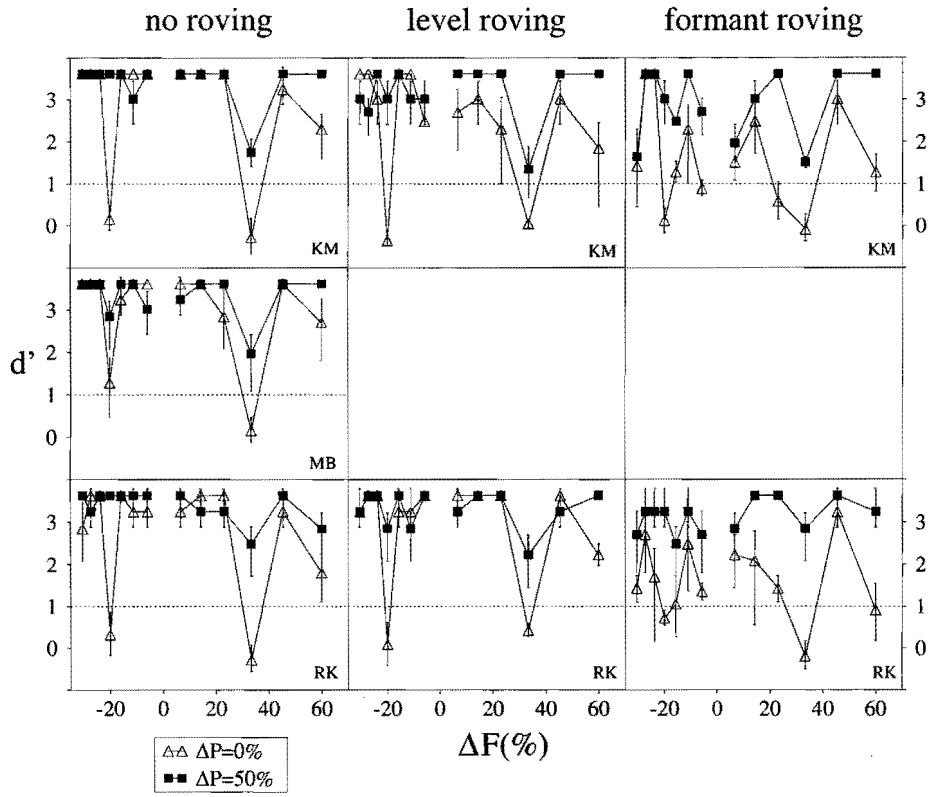


Figure 3.5: *Psychometric functions as in Fig. 3.4 but now for f_r = 1000 Hz and F_{wa} = 250 Hz.*

Data for the minimal stimulus-uncertainty conditions for f_r = 500 and 2000 Hz are shown in the lefthand panels of Fig. 3.6 (subjects KM and RK only). The argumentation presented above also holds for these two formant frequencies. For f_r = 2000 Hz, the general shapes of the psychometric functions are similar for both subjects, although the subjects are seen to have unequal difficulty in discrimination for ΔF^-.

3.4.2.B Level roving

The center panels of Figs. 3.4, 3.5, and 3.6 show psychometric functions for level roving. Starting with Fig. 3.4, i.e. f_r = 1000 Hz and F_{wa} = 100 Hz, it can be observed that discrimination performance is generally deteriorated relative to the minimal stimulus-uncertainty condition. Nevertheless, the general pattern of discrimination behavior is almost unaffected: the psychometric function for $\Delta P = 50$ % (squares) is almost always considerably above threshold. Moreover, the subthreshold data points for $\Delta P = 0$ % occur for the same ΔF values. Of all subjects, the performance of subject MB is seen to be affected most.

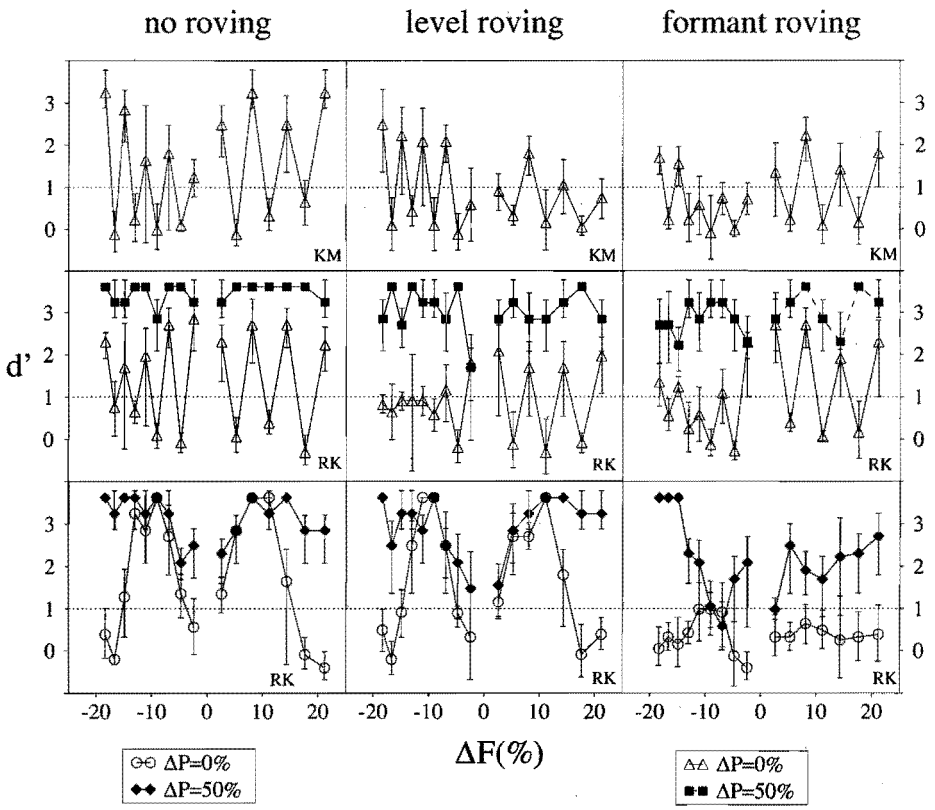


Figure 3.6: *Psychometric functions as in Figs. 3.4 and 3.5. Top and middle panels: psychometric functions for f_r = 2000 Hz and F_{wa} = 100 Hz. Bottom panels: psychometric function for f_r = 500 Hz. Circles here are data for ΔP = 0 %, filled diamonds data for ΔP = 50 %.*

Level roving does not have a great influence on discrimination performance for F_{wa} = 250 Hz and f_r = 1000 Hz, as shown in Fig. 3.5. The middle panels of Fig. 3.6 show the psychometric functions for f_r = 2000 and 500 Hz (F_{wa} = 100 Hz). For f_r = 2000 Hz, discrimination performance is reasonably affected, especially for ΔF^- for subject RK. On the other hand, performance for f_r = 500 Hz is as good with level roving applied as without level roving.

3.4.2.C Formant-frequency roving

The righthand panels in Fig. 3.4 show that, for f_r = 1000 Hz with F_{wa} = 100 Hz and ΔP = 0 %, discrimination performance under f_r roving drops below d' = 1 for *all* ΔF shifts. For ΔP = 50 %, performance is generally only moderately affected. A similar trend is observed for f_r = 500 Hz (Fig. 3.6) but, here, performance is also deteriorated

for ΔP = 50 % for some values of ΔF. For f_r = 2000 Hz, the alternating pattern is still observable for ΔF^+ if ΔP = 0 %. For ΔF^-, however, discrimination performance drops below threshold for almost all values. As is shown in the right panels of Fig. 3.5, roving of f_r has a moderate influence on performance for F_{wa} = 250 Hz

3.4.2.D Discussion

The psychometric functions show a clear interaction between f_r, F_{wa}, and F_{ws}. The pattern of these functions is not greatly influenced by roving of overall level which also occurs in natural speech. Roving of formant frequency, on the other hand, can drastically affect performance in the sense that the distortions introduced by PSOLA apparently are no longer usable cues for discrimination. This suggests that the non-steady-state nature of natural speech may explain part of the success of PSOLA. Because setting ΔP to 50 % provides strong and stable discrimination cues, the next experiment aims at determining discriminability as a function of ΔP.

3.4.3 Experiment II: variation of the pitch-marker position

Psychometric functions as a function of ΔP were measured for f_r = 1000 Hz and F_{wa} = 100 Hz. Two values of the $F0$ shift were selected for which results for ΔP = 0 % were below threshold: ΔF = -9.09 and +11.11 % (cf. Fig. 3.4). The parameter ΔP was varied in equal steps between -50 and 50 % (note that these two values are identical for strictly periodic signals). Psychometric functions were obtained for the minimal stimulus-uncertainty condition (all subjects) and for both roving conditions (ΔF = -9.09 %, and subjects KM and RK only).

Figure 3.7 shows the data for ΔF = -9.09 % (squares) and +11.11 % (circles) for the three stimulus uncertainty conditions. The data in Fig. 3.7 do not show systematic differences for the two ΔF values. Also, the psychometric functions are seen to be symmetric around ΔP = 0 %. Thresholds are approximately reached at $|\Delta P|$ = 25 %, which means that pitch markers not necessarily have to coincide exactly with either the filter excitation or the signal energy maximum. Moreover, these thresholds are reasonably stable under level and formant-frequency roving. The psychometric functions for subject KM become shallower with increasing stimulus uncertainty.

3.4.4 Model predictions

3.4.4.A Intensity discrimination

Gagné and Zurek (1988) used an intensity-discrimination model (Florentine and Buus, 1981) to account for JNDs in the resonance frequency of a single resonator. A model

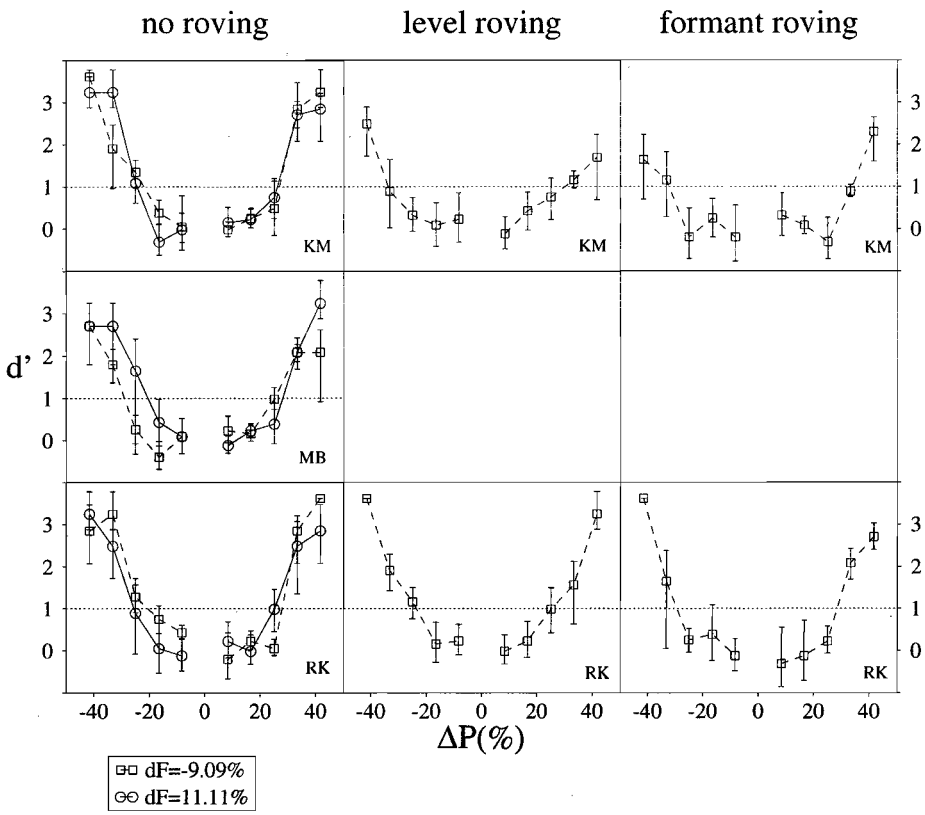


Figure 3.7: *Psychometric functions for F_{wa} = 100 Hz and f_r = 1000 Hz with ΔP as experimental parameter. Squares are data for ΔF = -9.09 %, circles data for ΔF = +11.11 %. Data for ΔP = 0 % and ± 50 % have already been shown in Fig. 3.4.*

of this kind takes only spectral cues into account. The model is based on a channel-wise determination of level differences between the excitation patterns of a reference and a signal. Channels here refer to critical bands. The model assumes that the partial sensitivity d'_i in channel i is proportional to the level difference $\Delta L_{E,i}$ between the two excitation patterns: $d'_i = k \cdot \Delta L_{E,i}$, where k is a constant which is the same for all channels. The overall sensitivity d' is derived from the partial sensitivities. In the single-band version of the model overall sensitivity d' is equal to the maximum of the partial sensitivities:

$$d' = max_{i=1,N}(d'_i) = k \cdot D_{max}, \tag{3.2}$$

where N is the number of channels. In the multiband version partial sensitivities are

	single band	multiband
KM	0.77	0.84
MB	0.78	0.82
RK	0.74	0.81

Table 3.1: *For both intensity-discrimination models, the square of the correlation coefficient of the linear regression, r^2, is tabulated for each of the subjects. Here, the ΔP experimental data for the minimal stimulus-uncertainty condition are used.*

optimally combined (Durlach et al., 1986):

$$d' = (\sum_{i=1}^{N} d_i'^2)^{1/2} = k \cdot D_{sum}. \tag{3.3}$$

Gagné and Zurek found that resonance-frequency JNDs could be best described by the single-band version of the model. In the present study both the single and the multiband model were implemented as a Gammatone filterbank (Patterson et al., 1987). Simulation of absolute hearing threshold was included by adding an "internal noise" value to the power estimate at the output of each filter (c.f., Moore and Glasberg, 1983).

According to formulas 3.2 and 3.3, d' is linearly related to D_{max} or D_{sum}. The predictive power of the models can thus be investigated by performing a linear regression on the experimental d' data in dependence on either D_{max} or D_{sum}. The regression equations were forced to intersect with the origin. As a measure of goodness of fit, the amount of explained variance, as expressed by the square of the correlation coefficient r, will be used.

First, the data of ΔP experiment II were used for finding the slopes k of the linear regression equations, for each subject individually. The minimal stimulus-uncertainty data from Fig. 3.7 for both values of ΔF were used for this regression. Data points were discarded if $d' \geq 3.62$, i.e., $Pc \geq 99$ %. The linear regression results, in terms of r^2, are presented in Table 3.1. For all subjects, the multiband model yields the highest value of r^2. Differences between the two models are, however, small. All regression slopes are significantly different from zero at the $p < 0.0001$ level. The slopes k for the multiband model are: 0.18, 0.15, and 0.17 for subjects KM, MB, and RK, respectively. The threshold value for D_{sum}, yielding $d' = 1$, is thus approximately equal to 6 dB. This value is a factor 2.5 higher than the value reported in Gagné and Zurek (1988). For the single-band model, the slopes k are 0.31, 0.24, and 0.28, respectively. At threshold $D_{max} \simeq 3.5$ dB, which is also a factor 2.5 higher than the value reported by Gagné and Zurek.

Second, the data of ΔF experiment I were used for linear regression for D_{sum}. Here

f_r (Hz)	500	1000	2000	1000 (F_{wa} = 250 Hz)
KM		0.48 *	0.45 **	0.54
MB		0.56 **		0.55
RK	0.61 **	0.54 **	0.40 **	0.79 **

Table 3.2: *Values of r^2 for linear regression on the d' data of the ΔF experiment using the multiband intensity-discrimination model. Significance levels p, indicating the probability that the slope of the linear regression equation is equal to zero, are indicated as follows: $* \; p < 0.05$, $** \; p < 0.01$, $*** \; p < 0.001$.*

again, only the data for the minimal stimulus-uncertainty conditions (lefthand panels in Figs. 3.4, 3.5, and 3.6) were used, with $\Delta P = 0$ %. The results are listed in Table 3.2. For $F_{wa} = 100$ Hz, r^2 is in the range 0.4 to 0.6. The slopes k are similar across subjects and are approximately 1, 0.8, and 1.4 for $f_r = 500$, 1000, and 2000 Hz, respectively. The threshold value for D_{sum} yielding $d' = 1$ is thus approximately 1 dB, which is clearly at variance with the results for the ΔP experiment. For $F_{wa} = 250$ Hz, however, k is approximately equal to 0.16, which is in good agreement with the slope found for the ΔP experiment. The high p levels, $p > 0.1$, for subjects KM and MB are probably due to the small number of data points resulting from ceiling effects in the ΔF experiment. As nearly all data points were far above threshold for $\Delta P = 50$ %, the corresponding r^2 values did not exceed 0.3 (not listed in Table 3.2).

3.4.4.B Modulation discrimination

This model is based on the discrimination of amplitude-modulation depth in the envelope of a single auditory-filter output. The auditory filter was simulated by a single Gammatone filter having a bandwidth of 1 ERB. The center frequency f_{cf} was varied over the range $[f_r - F_{WA}, f_r + F_{WA}]$, simulating off-frequency listening, in order to find the maximum difference between reference and signal. The maximum distances were mainly observed for filters centered at the boundaries of the f_{cf} range. The modulation indices M_{ref} and M_{sig} were calculated at the output of the filter [5]. The model is based on the assumption that sensitivity d' is governed by:

$$d' \; = \; k \cdot |M_{ref}^2 - M_{sig}^2| = k \cdot D_{mod} \;\; , \tag{3.4}$$

[5]Because the AM in the auditory filter output generally was not sinusoidal, the modulation index M was calculated:

$$M = \frac{\sqrt{2}\;\sigma_e}{m_e},$$

where σ_e and m_e are the standard deviation and average of the envelope of the output, respectively. The envelope is obtained by calculating a discrete Hilbert Transform. For an unmanipulated sinusoid of amplitude 1, $m_e = 1$ but $\sigma_e = 0$, so that $M = 0$. For a 100 % amplitude modulated sinusoid of amplitude 1, $\sigma_e = \frac{1}{\sqrt{2}}$ and $m_e = 1$, so that $M = 1$, as expected.

f_r (Hz)	500	1000	2000
KM		0.52 **	0.69 ***
MB		0.35 *	
RK	0.33 *	0.59 **	0.92 ***

Table 3.3: *Values of r^2 for linear regression on the d' data of the ΔF experiment using the modulation-discrimination model. Only data for F_{wa} = 100 Hz are shown. Significance levels p are indicated as in Table 3.2.*

where k is some constant. Moore and Sek (1992) found that, for low modulation rates (below 10 Hz), *detection* sensitivity was linearly related to the square of the modulation index. Wakefield and Viemeister (1990) used sinusoidally amplitude-modulated (SAM) noise and found almost linear relations between d' and $M_{ref}^2 - M_{sig}^2$. Because M_{ref}^2 may be smaller than M_{sig}^2 for the present signals, the absolute value of the difference was taken.

Table 3.3 shows the r^2 values for linear regression of D_{mod} on the experimental d' data of the ΔF experiment (F_{wa} = 100 Hz and ΔP = 0 % only). For f_r = 1000 Hz and ΔP = 0 %, the slope k is found to be approximately 7 for all subjects. This corresponds to a modulation-discrimination threshold of 0.14 for D_{mod}. The explained variance for subject MB, however, is rather low. For f_r = 2000 Hz, slope k is about 4 so that the threshold would be at $D_{mod} = 0.25$. These D_{mod} values are in reasonable agreement with the data reported in Wakefield and Viemeister (1990) for SAM noise, provided M_{ref} is large, i.e., 10 log $M_{ref}^2 \geq$ -5 dB. For the present signals, M_{ref} is indeed in this range.

3.5 Discussion

Although not explicitly verified experimentally, $|\Delta F|$ shifts as small as approximately 2 % may lead to detectable distortions, as can be inferred from the region around ΔF = 0 % in the psychometric functions in Figs. 3.4, 3.5, and 3.6. This finding agrees with the results presented in Appendix A for manipulation of pure tones. Remarkably, the psychometric functions for minimal stimulus-uncertainty and level-roving, provided ΔP = 0 %, were nonmonotonic, revealing a clear interaction between f_r, F_{wa}, and F_{ws}. This finding is not in agreement with the intuitive expectation that distortions are more easily detectable for larger shifts in $F0$. We will try to explain the discrimination results in terms of spectral and temporal cues by first comparing the present results with data from the literature and then discussing our modeling results.

3.5.1 Comparison with the literature

As a result of PSOLA manipulation with ΔP = 0 %, changes in the intensity of spectral

components, in combination with phase shifts, occur in the spectral region of f_r as was illustrated in Fig. 3.3. Changes in component intensities also occur due to changes in formant frequency. A number of studies (e.g., Gagné and Zurek, 1988; Kewley-Port and Watson, 1994; Lyzenga and Horst, 1995; Sommers and Kewley-Port, 1996) have explained formant-frequency JNDs in terms of profile analysis, i.e., in terms of discrimination of spectral shape (Richards et al., 1989; Zera et al., 1993). With minimal stimulus uncertainty and for $F_{wa} = 100$ Hz in the present study, (absolute) component-level differences between signal and reference maximally amount to 2.5, 2.5, and 1.5 dB for $f_r = 500$, 1000, and 2000 Hz, respectively. These values are valid for the range of ΔF investigated in the experiments. The lower value for $f_r = 2000$ Hz is a consequence of the larger formant bandwidth of 100 Hz.

Thresholds for the detection of level increments of single components of a complex tone of equal-amplitude harmonics were reported by Zera et al. (1993). For complex tones consisting of 60 harmonics of 100 Hz, level-increment thresholds for harmonics at 500, 1000, and 2000 Hz were found to be approximately 2, 2.5, and 4 dB, respectively. This means that for $f_r = 500$ and 1000 Hz, the level changes of individual harmonics due to PSOLA manipulation would be near detection threshold. For $f_r = 2000$ Hz, level changes would be below threshold. The psychometric functions in Figs. 3.4 and 3.6, however, showed that discrimination sensitivity generally was above $d' = 2$. For the fifth harmonic of 200 Hz, on the other hand, Henn and Turner (1990) and Zera et al. (1993) reported level-increment thresholds of 2 dB. For $F_{wa} = 250$ Hz and $f_r = 1000$ Hz, spectral-envelope level differences are maximally 10 dB so that level differences of single components were potential cues for discrimination.

Instead of just a single harmonic, however, the intensities of a number of harmonics are changed both as a result of PSOLA manipulation and by changing the formant frequency. Sommers and Kewley-Port (1996) found that salient cues for formant-frequency JNDs were mediated by the level changes of the three harmonics closest to the formant frequency. Using an excitation pattern model (Moore and Glasberg, 1987), they also found that formant-frequency JNDs under different conditions resulted in more or less constant level differences between excitation patterns. Sommers and Kewley-Port (1996) only investigated formants at 500 and 1350 Hz with an $F0$ of 200 Hz, so that harmonics around the formant frequencies were likely to be resolved. Particularly for $f_r = 2000$ Hz and $F_{wa} = 100$ Hz in the present study, harmonics are unresolved so that temporal cues may have been used for discrimination. Lyzenga and Horst (1995), who used formant-like signals around 2000 Hz with an $F0$ of 200 Hz, also proposed a temporal mechanism, at least for part of their JND data.

Roving of the *overall* stimulus level affected discrimination performance only to a small

degree. This finding is in agreement with results presented by Farrar et al. (1987) on formant-frequency discrimination using noise sources as input to the Klatt synthesizer. Because the overall level of the intervals was normalized in the minimal stimulus condition, differences in loudness between signal and reference may have been a cue (cf. Lyzenga and Horst, 1995). Taking into account the rather small spectral-envelope differences mentioned above, however, it is more likely that discrimination performance under level roving is affected by the increase in distracting stimulus uncertainty.

The present results show that f_r roving can affect discrimination performance considerably. Roving of f_r results in spectral-envelope level differences near f_r, not only between signal and references, but also between references. The distribution of level differences has a standard deviation of approximately 4 dB for all f_r-values and bandwidths under consideration. This value is of the same order of magnitude as the spectral-envelope differences introduced by PSOLA for $F_{wa} = 100$ Hz (see above). This means that if the harmonics around f_r are resolved, excitation-pattern differences between the two references are comparable to the differences between signal and references due to PSOLA. Discrimination performance can then be expected to drop below $d' = 1$, as observed for $f_r = 500$ and 1000 Hz. For $F_{wa} = 250$ Hz, level differences due to PSOLA can exceed those due to f_r roving (see above) so that discrimination is expected to be at most moderately influenced, which is in agreement with the experimental data. If, on the other hand, components are unresolved, as for $f_r = 2000$ Hz and $F_{wa} = 100$ Hz, the differences in the phase spectra between signal and references may become a cue. In other words, the effect of (in)coherent addition of subsequent segments is preserved in peripheral filtering. For ΔF^+, roving of f_r did indeed not deteriorate discrimination performance. It is not clear, however, why performance dropped below $d' = 1$ for ΔF^-.

In an additional, informal experiment, the phases of the components of the single-formant signal were randomized. For $f_r = 500$ Hz, phase randomization had only a small effect on discrimination performance. This was also observed for $f_r = 1000$ Hz with $F_{wa} = 250$ Hz. For $f_r = 1000$ and 2000 Hz ($F_{wa} = 100$ Hz), however, discrimination performance was below $d' = 1$. This provides additional evidence for the hypothesis that, for the latter two conditions, temporal cues played a dominant role.

In experiment II, the detection threshold was found to be $|\Delta P| \simeq 25$ %. The spectral "notch depth" at this value of $|\Delta P|$ is approximately 3 dB. Turner and Van Tasell (1984) found comparable thresholds for a notch with linear flanks on a dB scale, centered at 2120 Hz within the spectrum of a synthetic vowel with 120-Hz fundamental. If intensity discrimination determines detectability, however, then lower ΔP thresholds are to be expected if the components in the notch are resolved. The results of informal tests for $F_{wa} = 250$ Hz confirmed this expectation: with minimal stimulus uncertainty, ΔP

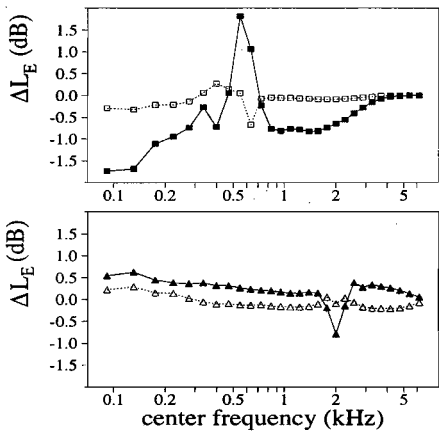


Figure 3.8: *Illustration of excitation-pattern level differences between unmanipulated and PSOLA-manipulated single-formant signals. The top panel shows the level differences $\Delta L_{E,i}$ for f_r = 500 Hz. Filled squares indicate level differences for ΔF = -11.1 %, open squares for ΔF = -2.4 %. The corresponding values of D_{sum} are 1.9 and 0.5 dB, respectively. The bottom panel shows level differences for f_r = 2000 Hz, where the filled and open triangles indicate differences for ΔF = 8.1 % and 11.1 %, respectively. For these ΔF values, D_{sum} is 0.8 and 0.5 dB, respectively.*

thresholds were considerably smaller than 25 %.

3.5.2 Models

In contrast to the findings of Gagné and Zurek (1988), the best results for the intensity-discrimination model were obtained here for the multiband version, although the differences between the single and multiband model were small. The multiband model could describe the data for the ΔP experiment reasonably well, which suggests that discrimination was based on profile analysis. In the case of the ΔF experiment, the descriptive power of this model depended on whether harmonics around f_r were resolved by peripheral filtering. For both conditions in which harmonics were resolved, i.e., f_r = 500 Hz (F_{wa} = 100 Hz) and f_r = 1000 Hz (F_{wa} = 250 Hz), reasonable r^2 values were obtained. For the latter condition, the regression slopes for the ΔP and ΔF data were almost identical. For f_r = 1000 Hz and F_{wa} = 100 Hz, where harmonics 9-11 around f_r are only just resolved, the r^2 values were reasonable but the regression slope was much smaller than for the ΔP data. For f_r = 2000 Hz and F_{wa} = 100 Hz, where harmonics 19-21 are unresolved, r^2 values were lowest.

Figure 3.8 illustrates excitation-pattern differences between a PSOLA-manipulated signal and the unmanipulated reference for f_r = 500 (top panel) and 2000 Hz (bottom

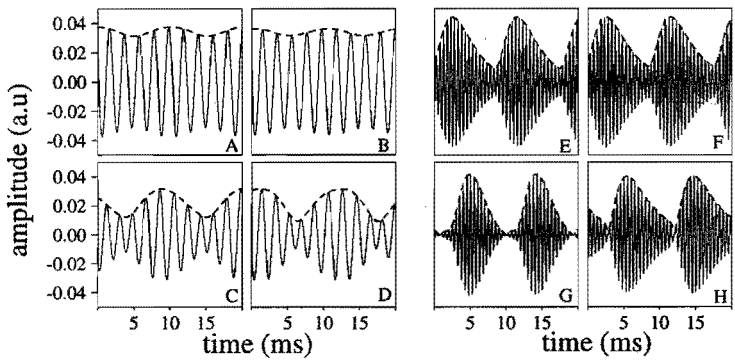


Figure 3.9: *Illustration of AM differences in the output of Gammatone filters. Panels A-D show portions of the output waveforms of a filter centered at 500 Hz (f_r = 500 Hz): panel A shows the waveform for ΔF = -2.4 % for the PSOLA-manipulated signal, panel B for the unmanipulated reference. The modulation-depth measure D_{mod} is equal to 0.003. Panels C and D show the corresponding output waveforms for ΔF = -11.1 % (D_{mod} = 0.04). Panels E-H show output waveforms of a filter centered at 2000 Hz (f_r = 2000 Hz): ΔF = 11.1 % in panels E and F (D_{mod} = 0.05) and ΔF = 8.1 % in panels G and H (D_{mod} = 0.7).*

panel). For both formant frequencies, ΔF conditions for which performance was above or below d' = 1 are indicated by the filled and open symbols, respectively. As the data in Fig. 3.8 suggest, excitation-pattern differences for above-threshold stimuli are larger if the harmonics around f_r are resolved (top) than for unresolved harmonics (bottom).

For the modulation-discrimination model r^2, values were highest for f_r = 2000 Hz and lowest for f_r = 500 Hz, as would be expected on the basis of unresolved and resolved harmonics, respectively. Figure 3.9 illustrates this expectation by showing, for the same ΔF values as in Fig. 3.9, the output of Gammatone filters centered at the formant frequencies 500 and 2000 Hz. In the experiments the ΔF value of the top panels resulted in performance below d' = 1. Accordingly, for both formants the difference in modulation depth is small. The ΔF value of the bottom panels resulted in above-threshold performance. Only for f_r = 2000 Hz, however, a substantial difference in modulation depth can be observed.

As the r^2 values were moderate for both models for f_r = 1000 Hz and F_{wa} = 100 Hz, a multicue model might be reasonable. This was not verified, however, because we judged the amount of experimental data insufficient for performing multiple regressions reliably.

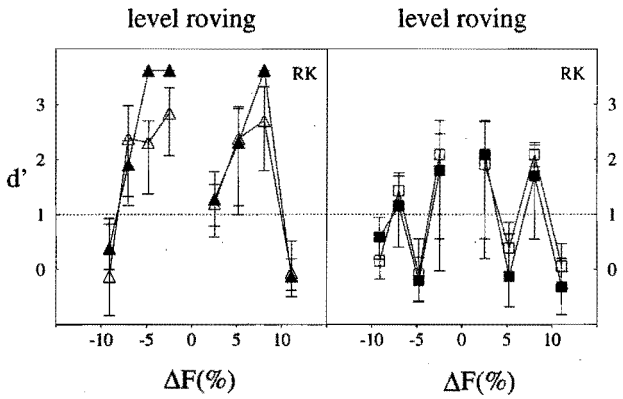


Figure 3.10: *Psychometric functions obtained by including a synthesis window (open symbols) and by the standard PSOLA operations (filled symbols, data already shown in Figs. 3.4 and 3.6). Data for f_r = 1000 Hz are shown in the lefthand panel, data for f_r = 2000 Hz in the righthand panel. In both cases, level roving was applied.*

3.5.3 Synthesis window

Some of the effects of envelope modulation introduced by PSOLA (cf. Fig. 3.2) can be canceled by applying a so-called "synthesis window". Such a window corrects for the fact that adjacent Hanning windows do not add up to one if $T_s \neq T_a$. A simple realization of such a window is to calculate the temporal envelope of the adjacent Hanning windows, spaced at intervals of T_s ms. By taking the reciprocal of this envelope and multiplying it with the PSOLA-manipulated (speech) signal, the degree of AM of the latter signal is reduced. Such an operation, however, does not correct for the AM introduced by out-of-phase addition of the fine structures of adjacent segments.

Experimental results for f_r = 1000 and 2000 Hz, obtained by including the synthesis window as described above, are shown in Fig. 3.10 by the open symbols (F_{wa} = 100 Hz with level roving, subject RK only). The filled symbols indicate corresponding data from experiment I. The synthesis window seems to cancel the ceiling effects for f_r = 1000 Hz, although the fact that performance is still above d' = 2 suggests that this effect is perceptually less relevant. Even in the case of f_r = 2000 Hz, a condition for which temporal cues presumably dominate detection performance, the two psychometric functions are basically identical.

3.5.4 Natural speech

In order to understand the perceptual effects of PSOLA manipulation of natural speech, the following aspects should, in our view, be additionally investigated. First of all,

natural speech is generally characterized by the presence of at least 2 to 3 formants, at least in vowels. The use of multiple-formant signals may inform about the way in which cues occurring in several frequency regions are combined. Secondly, it should be investigated to what extent the detectability of distortions is influenced by fluctuations in both spectral content and $F0$. These fluctuations can be either of a random nature (e.g., jitter) or more deterministic (e.g., formant and $F0$ trajectories). In addition, the perceptual consequences of errors in F0 estimation in natural speech, leading to incorrect pitch-marker positioning, should be investigated. Thirdly, the present experiments were performed under well-controlled acoustical conditions. The amplitude and phase transfer characteristics of (normal) playback rooms, however, will in their turn affect stimulus characteristics. It is conceivable that, under such listening conditions, the perceptual tolerance for the distortions introduced by PSOLA is actually increased.

3.6 Conclusions

(1) Discrimination thresholds as a function of ΔF (the shift in fundamental frequency) for PSOLA-manipulated and unmanipulated single-formant signals are found to be low: $|\Delta F| \leq 2$ %. Moreover, the psychometric functions reported here typically show an interaction between the formant and fundamental frequency.

(2) Roving of overall level does not seem to greatly affect discrimination performance as a function of ΔF. Roving of the formant frequency does impair performance: for formants at 500 and 1000 Hz (100 Hz fundamental), performance drops below $d' = 1$ for all ΔF values tested. For a fundamental of 250 Hz, performance seems to be only moderately influenced by formant roving.

(3) Discrimination thresholds as a function of ΔP (the pitch-marker location) reported here are approximately a quarter of the fundamental period. The discrimination cues that occur due to incorrect positioning of the pitch markers seem to be robust under level and formant frequency roving. These findings apply to signals with a fundamental of 100 Hz; ΔP thresholds for higher $F0$ values are expected to be lower.

(4) The discrimination data as a function of ΔP can be described well with an intensity-discrimination model. This model can also account moderately well for the experimental data as a function of ΔF, in the case of resolved harmonics around the formant frequency (e.g., a formant at 1000 Hz with a 250-Hz fundamental). The modeled discrimination sensitivities, however, generally differ across the ΔP and ΔF conditions. For unresolved harmonics, such as for a 2000-Hz formant and a 100-Hz fundamental, the modulation-discrimination model matches the experimental data reasonably well. These findings are in agreement with the psychoacoustical notion of different modes of processing for resolved and unresolved harmonics.

(5) As for natural speech, distortions introduced to signals with higher fundamental frequencies are expected to be more easily detectable (see conclusions **(2)** and **(3)**). In the case of low fundamental frequencies, the occurring phase cues are often subtle and may not be stable under different playback conditions.

3.7 Appendix: pure tones

The main question of this baseline experiment was to what extent the basic distortions described in section 3.3.3 are detectable by the human auditory system. Pure tones with carrier frequencies f_c = 500, 1000, and 2000 Hz were PSOLA-manipulated. The carriers were thought of as harmonics of a fundamental of 100 or 250 Hz. In order to determine the detectability of the introduced side components, the references were unmanipulated pure tones having the same frequency as the strongest component in the manipulated-tone spectrum (cf. Fig. 3.2(c)). Instead of taking all side component into account, the results presented below were obtained for signals consisting of the three strongest spectral components only. These results were compared with results for "real" PSOLA-manipulated tones and did not differ considerably. Three overall levels L were used: 45, 60, and 75 dB SPL in combination with a level rove uniformly distributed between -5 and +5 dB. Stimulus characteristics such as duration and stimulus generation were the same as described in section 3.4.1.A.

A two-down, one-up 3IFC adaptive procedure was used in which ΔF was varied adaptively for the determination of discrimination thresholds. After a learning phase, three measurements for each condition were collected whose mean and standard deviation will be presented below. Two subjects (MB and RK) participated in this experiment.

Figure 3.11 presents both the ΔF thresholds for subjects MB and RK and predictions from two models (the models are different from the models of experiments I and II, see below). Squares indicate data for L = 60 dB SPL (both subjects) and circles and triangles for L = 45 and 75 dB SPL, respectively (subject RK only). Generally, thresholds for both subjects are low: $|\Delta F| \leq 2$ %. Thresholds for ΔF^- tend to be lower than for ΔF^+. Carrier frequency does not greatly influence thresholds for F_{wa} = 100 Hz but does have an effect for F_{wa} = 250 Hz. Likewise, level does affect thresholds for F_{wa} = 100 Hz, except for f_c = 500 Hz, but influences thresholds for F_{wa} = 250 Hz.

In the spectral masking model ("SPEC"), the detection of the side components is determined by their masked threshold in the presence of the much stronger center component (cf. the "No Summation Model" of Hartmann and Hnath (1982)). The masked thresholds were estimated using the data in Schöne (1979)[6]. The modulation-detection model ("MOD") is based on determining[7] the AM and FM indices of the PSOLA-manipulated pure tone as a function of ΔF. As in Zwicker (1952), threshold predictions were only derived for F_{wa} = 100 Hz, with f_c = 1000 or 2000 Hz (where the "Phasengren-

[6]Predictions for spectral masking are based on Figure 3 in Schöne (1979)

[7]Visual inspection showed that, for the range of ΔF used here, both the envelope and the instantaneous frequency were modulated in a sinusoidal fashion. The AM index m and FM index $\Delta f/f_{mod}$ were therefore calculated using their definition in Zwicker (1952).

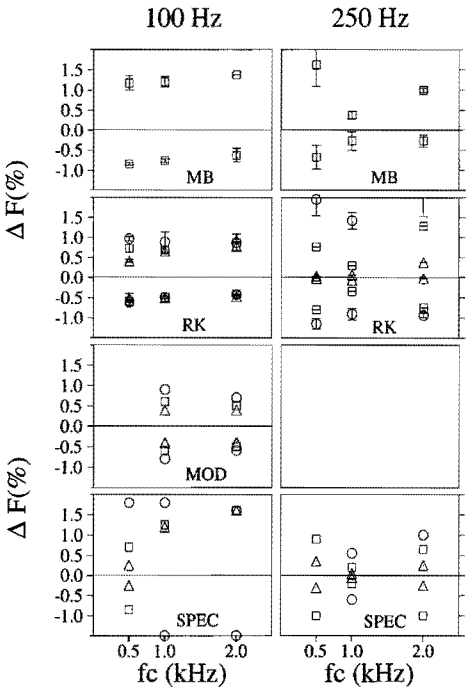


Figure 3.11: *Detection thresholds for subjects MB and RK for ΔF^+ and ΔF^- for the manipulation of pure tones. Lefthand panels give the data for F_{wa} = 100 Hz, righthand panels those for 250 Hz. Data for L = 45 dB SPL are shown by circles, those for 60 dB SPL by squares and those for 75 dB SPL by triangles. Standard deviations are indicated by vertical lines. Threshold predictions on the basis of a modulation-detection model are denoted by MOD. Predicted thresholds using a spectral masking model are marked SPEC.*

zfrequenz" is approximately 80 and 140 Hz, respectively). Although the signals used actually had mixed modulation (Hartmann and Hnath, 1982; Moore and Sek, 1992), the lowest prediction based on detection of either AM or FM was taken.

In the case of resolved harmonics, i.e., for f_c = 1000 and 2000 Hz with F_{wa} = 100 Hz (harmonics 9-11 and 19-21, respectively), the predictions based on modulation detection are in reasonable agreement with the experimental data. The predictions based on the spectral masking model are less accurate, especially for ΔF^-, with the possible exception of f_c = 500 Hz. For conditions with resolved harmonics (i.e., f_c = 500, 1000, and 2000 Hz with F_{wa} = 250 Hz, and f_c = 500 Hz with F_{wa} = 100 Hz), the spectral-masking-model predictions are qualitatively similar to the experimental data, also showing level dependency. The actual predicted thresholds, however, do not exactly match the experimental data, especially for L = 45 dB SPL.

Chapter 4

Psychoacoustical evaluation of PSOLA. Two-formant stimuli and vocal perturbation[1]

4.1 Abstract

This Chapter presents the results of listening experiments and psychoacoustical modeling aimed at evaluating the Pitch Synchronous OverLap-and-Add (PSOLA) technique. This technique can be used for simultaneous modification of pitch and duration of natural speech, using simple and efficient time-domain operations on the speech waveform. The first set of experiments tested the ability of subjects to discriminate double-formant stimuli, modified in fundamental frequency using PSOLA, from unmodified stimuli. Of the potential auditory discrimination cues, cues from the first formant were found to generally dominate discrimination performance. In the second set of experiments the influence of vocal perturbation, i.e. jitter and shimmer, on discriminability of PSOLA-modified single-formant stimuli was determined. The data show that discriminability deteriorates at most modestly in the presence of jitter and shimmer. With the exception of a few conditions, the trends in these data could be replicated by either using a modulation-discrimination or an intensity-discrimination model, dependent on the formant frequency. As a baseline experiment detection thresholds for jitter and shimmer were measured. Thresholds for jitter could be replicated by using either the modulation-discrimination or the intensity-discrimination model, dependent on the (mean) fundamental frequency of stimuli. The thresholds for shimmer could be accurately predicted for stimuli with a 250-Hz fundamental, but less accurately in the case of a 100-Hz fundamental.

[1]This chapter is based on Kortekaas, R. W. L. and Kohlrausch, A. (1997) "Psychoacoustical evaluation of PSOLA. II. Double-formant stimuli and the role of vocal perturbation," submitted for publication to J. Acoust. Soc. Am.

4.2 Introduction

For speech modification and speech synthesis purposes, high computational efficiency can be achieved using techniques that are based on speech-waveform operations in the time domain. Roucos and Wilgus (1985), for instance, presented speech-waveform manipulation schemes within the OverLap-and-Add (OLA) framework. Moulines and Charpentier (1990) proposed a Pitch-Synchronous OLA method (PSOLA) in which all overlap-and-add operations were aligned to the "local" pitch period. This time-domain technique[2] was developed for the (simultaneous) modification of both duration and pitch of natural speech. PSOLA has proven to be a practical method for speech modification and synthesis (e.g., Moulines and Laroche, 1995; Klaus et al., 1997).

Kortekaas and Kohlrausch (1997) argued that, despite the widespread application of PSOLA, the perceptual consequences of speech-waveform manipulation were not well understood. This lack of knowledge was reflected in the unpredictable occurrence of annoying artefacts such as roughness when using PSOLA. They presented results of a psychoacoustical evaluation of PSOLA that was aimed at determining the auditory sensitivity to the "basic distortions" introduced by PSOLA. Basic distortions referred to spectro-temporal changes of the speech signal that occur even if the OLA operations are *exactly* aligned to the local pitch periods. As such, basic distortions do not correspond to artefacts that arise due to, for instance, errors in estimating the local pitch periods. By measuring the sensitivity of the human auditory system to these basic distortions, the study was aimed at understanding the perceptual consequences of speech-waveform modification, rather than at explaining the occurrence of the artefacts mentioned above.

In Kortekaas and Kohlrausch (1997) the auditory sensitivity to these basic distortions was investigated through listening experiments in which single-formant stimuli (Klatt, 1980) were modified in fundamental frequency ($F0$). The $F0$ was kept constant over the duration of the stimulus. Spectral effects of this $F0$ modification are most prominent in the formant region, where changes in both harmonic levels and (relative) phases occur. The subjects' task was to discriminate PSOLA-manipulated from unmanipulated signals as a function of ΔF, the percentage shift in $F0$. Possible loudness cues were compensated for. The psychometric functions were typically found to be non-monotonic, with an almost sinusoidal pattern of discrimination sensitivity, which indicated a strong interaction between ΔF and the formant frequency. In the case of resolved harmonics in the formant region the discrimination performance could be modeled reasonably well by detection of changes in the spectral excitation pattern ("profile analysis"). In the case of unresolved harmonics, the data could be best explained on the basis of detection of

[2]The present study is restricted to the time-domain implementation (TD-PSOLA), although a frequency-domain variant was also proposed (FD-PSOLA; Charpentier and Moulines, 1989). PSOLA/TD is a registered trademark of France Telecom.

changes in modulation depth at the output of a simulated auditory filter. The present study was aimed at extending these results to more natural and, consequently, more complex stimuli.

Two separate aspects of natural speech were dealt with. First, discrimination sensitivity was measured for double-formant stimuli because natural vowels are characterized by at least two formants, If the available information from both formant regions is processed optimally, we expect discrimination performance to be higher than was measured for single-formant stimuli. This would imply that the auditory system is able to combine information from different spectral regions, even if these cues have a different nature, such as spectral cues for low formants and temporal cues for high formants. It should be stressed that the present study was not aimed at *exactly* determining the process of combining auditory information in the framework of signal-detection theory (Green and Swets, 1966). Instead, a general characterization of the discriminability of PSOLA-manipulated double-formant signals will be presented. This characterization will be compared to findings reported in the literature concerning differences in formant-frequency JNDs in the case of single or double-formant stimuli (Lyzenga, 1997).

Second, the extent to which discrimination performance is influenced by perturbations of either amplitude or $F0$ was investigated. These two perturbation types are usually referred to as shimmer and jitter, respectively (Pinto and Titze, 1990). Shimmer and jitter are often conceived of as characteristics of natural speech (Lieberman, 1961) and have been proposed as criteria to classify certain types of pathological voices (e.g., Askenfelt and Hammarberg, 1986). In the present study shimmer and jitter were simulated by introducing random fluctuations in the pulses exciting the formant filter (cf. Hillenbrand, 1987, 1988; Rozsypal and Millar, 1979). Even though this way of simulating shimmer and jitter may be consistent with quantitative measures based on the analysis of the speech waveform (Horii, 1980), the relevance of such measures is still under debate because of, e.g., interactions between vocal tract and glottis (Klatt and Klatt, 1990; Cranen, 1997; Schoentgen and De Guchteneere, 1997). Therefore, these shimmer and jitter simulations should be conceived of as a means to investigate in a controlled manner the role of perturbations in perceiving PSOLA manipulated stimuli, rather than as a true-to-nature simulation of speech production.

Intuitively, the presence of shimmer and jitter in the stimuli may be expected to affect the use of temporal cues identified in our previous study considerably. As mentioned above, these cues could be modeled as detection of changes of modulation depth at the output of auditory filters. Jitter and shimmer introduce additional modulations and may therefore hamper modulation-depth discrimination. The use of spectral cues, which were modeled by the detection of excitation-pattern differences, may intuitively

be expected to be hampered because shimmer and jitter introduce "noise" components between the harmonics in the spectrum (Klingholz and Martin, 1985). This possibly obscures PSOLA-induced excitation-pattern differences.

As a baseline experiment detection thresholds were measured for jitter and shimmer in single-formant stimuli, without any PSOLA modification. In the second part of the experiments, discriminability of PSOLA-modified single-formant signals under influence of perturbation was investigated. The two psychoacoustic models employed in our previous study were used again in an attempt to describe the auditory processes involved both in detection of perturbation and in discrimination of jittered or shimmered, PSOLA-modified stimuli.

4.3 General methods

The terminology used in the present Chapter is consistent with that in Kortekaas and Kohlrausch (1997). PSOLA-manipulated stimuli were obtained by (1) decomposing a single or double-formant signal into signal segments by pitch-synchronous windowing, and (2) recombining the decomposed signal segments by means of overlap-and-add. The decomposition was performed with Hanning (raised-cosine) windows which had a length of two pitch periods. The window maxima coincided with the "glottal pulses" that excited the formant filters (see section 4.4.1). In this way, the signal was regularly decomposed at an "analysis window rate" F_{wa} Hz. By means of overlap-add, the decomposed signal segments were recombined at a "synthesis window rate" F_{ws} Hz. The resulting $F0$ of the PSOLA-manipulated signal was thus equal to F_{ws} Hz. Discrimination performance was measured as a function of the shift in $F0$ expressed as a Weber fraction: $\Delta F = [(F_{ws} - F_{wa})/F_{wa}] \times 100$ %. For a detailed description of the PSOLA technique and its signal-processing aspects the reader is referred to Moulines and Laroche (1995).

4.4 Experiment 1: double-formant stimuli

The purpose of this experiment was to investigate the extent to which subjects' sensitivity in discriminating PSOLA-manipulated and unmanipulated formant signals is influenced by the presence of two formant regions. PSOLA-induced spectral changes in both formant regions yield cues for discrimination which, if optimally combined, are expected to increase discrimination sensitivity relative to the single-formant condition. If, on the other hand, the increase in stimulus complexity hampers the process of attending to particular cues, then discrimination sensitivity may actually decrease in the presence of two formants.

4.4.1 Method

4.4.1.A Stimuli

The test signal, which was subjected to PSOLA modification, and the reference signals were generated by filtering a "glottal-pulse train" by two second-order formant filters connected in series (Klatt, 1980). The glottal-pulse train was a pre-emphasized Dirac-impulse train having low-pass characteristics of natural voicing and high-pass characteristics of radiation at the mouth opening (Klatt, 1980). The test signal was first generated at $F0 = F_{wa} = 100$ Hz and subsequently PSOLA-modified to accomplish a shift in $F0$ to F_{ws} Hz. The reference signals were directly generated with an $F0$ of F_{ws} Hz. All signals were generated at a sample rate of 32 kHz[3].

Stimuli will be labeled as [F1,F2] in the case of fixed formant frequencies (expressed in Hz): the stimuli under investigation were [500,1000], [500,2000], and [1000,2000]. The -3 dB bandwidths were 50, 50, and 100 Hz for the 500, 1000, and 2000-Hz formant filters, respectively. Conditions in which F1 was roved while F2 was fixed will be labeled as [F1R,F2]. The F1 rove was uniformly distributed over a range of $\pm$ 2 $\Delta F1$, where $\Delta F1$ denotes one JND in $F1$ and amounts to $0.079\, F1/\sqrt{Q}$ (Gagné and Zurek, 1988). In this expression, Q denotes the Q-factor of the formant filter. The mean overall level L of the stimuli in all conditions was set to 70 dB SPL, with a level rove uniformly distributed in the range $\pm$ 5 dB. The "minimal stimulus uncertainty" condition presented in Kortekaas and Kohlrausch (1997), in which level and formant roving was absent, was not investigated here.

The spectral level difference between F1 and F2 can become rather large: the difference is approximately 12 dB for [500,1000], 30 dB for [500,2000], and 16 dB for [1000,2000]. Cues associated with the F1 region may therefore dominate the discrimination performance. By applying an additional formant filter, having the same resonance frequency and bandwidth as F2, the level difference is reduced to within a few dB, at the cost of somewhat reducing the bandwidth of F2. In a limited set of conditions, which will be labeled as [F1,F2D], the importance of equalizing the formant levels was investigated. In a pilot experiment it was verified that the general shape of the psychometric function for discriminating a PSOLA-manipulated single-formant was not affected by using the

[3]Klatt (1980) proposed an implementation of the formant filters for 10 kHz sample rate. Using higher sample rates results in a stronger roll-off of the spectrum to higher frequencies, due to the periodic nature of the discrete spectrum. This roll-off was compensated for by applying a "correction formant filter" twice (center frequency 5 kHz, bandwidth 2.2 kHz). In this way, the magnitude of the transfer functions for the 10 kHz and the 32 (48) kHz implementations differed by no more than 1 dB. Because the phase-transfer function of a Klatt formant filter is approximately linear for frequencies remote from the resonance frequency, the phase characteristic in the region of the "true" formants is preserved.

formant filter twice.

4.4.1.B Procedure

The experimental procedure was similar to the procedure described in Kortekaas and Kohlrausch (1997). Psychometric functions were measured using a 3I-3AFC procedure with a fixed value of ΔF in each run. A run consisted of 15 trials and the subject's task was to indicate, for each of these trials, which interval contained the odd-ball, i.e., the PSOLA-modified stimulus. The psychometric functions were only measured for positive values of ΔF by varying F_{ws} in the following fashion: $F_{ws} = 1/(T_a + \frac{n}{4})$ $[Hz]$, where $T_a = 1/F_{wa}$ and $n = 1, 2, 3, ..$ ms.

All signals were software generated on a Silicon Graphics Indigo workstation, D/A converted using the built-in board, and attenuated using TDT PA4 modules to have the appropriate sound pressure level. The signals were presented diotically to the subject over Beyer 990 headphones while they were seated in a soundproof booth. The subjects responded via a keyboard and received immediate feedback after each trial. Stimulus duration was 300 ms, with raised-cosine ramps of 25 ms at the beginning and end of the signal. Stimulus intervals were separated by 200 ms.

Because two out of three subjects had not participated in the previous experiments, psychometric functions were remeasured for the single-formant condition. For all combinations of formant frequency (fixed and roved) and ΔF, each subject performed 5 runs, each consisting of 15 trials. The first run of each condition was regarded as learning phase and thus not included in the analysis. In the double-formant experiment, 6 runs were performed, the first of which was also discarded. The means and standard errors of the means shown in the figures below thus represent 60 and 75 trials, respectively. These figures will display discrimination sensitivity d' as a function of ΔF which was obtained by means of a percentage correct to d' conversion table (Macmillan and Creelman, 1991).

4.4.1.C Subjects

Three subjects, who were aged between 23 and 28, participated in the experiments. All subjects had normal pure-tone thresholds in quiet for the frequencies 500, 1000, and 2000 Hz. All subjects had experience in other psychoacoustic listening experiments. Subject RK (the first author) additionally had ample experience with the present stimuli. Subjects JS and LB were paid an hourly wage for their participation.

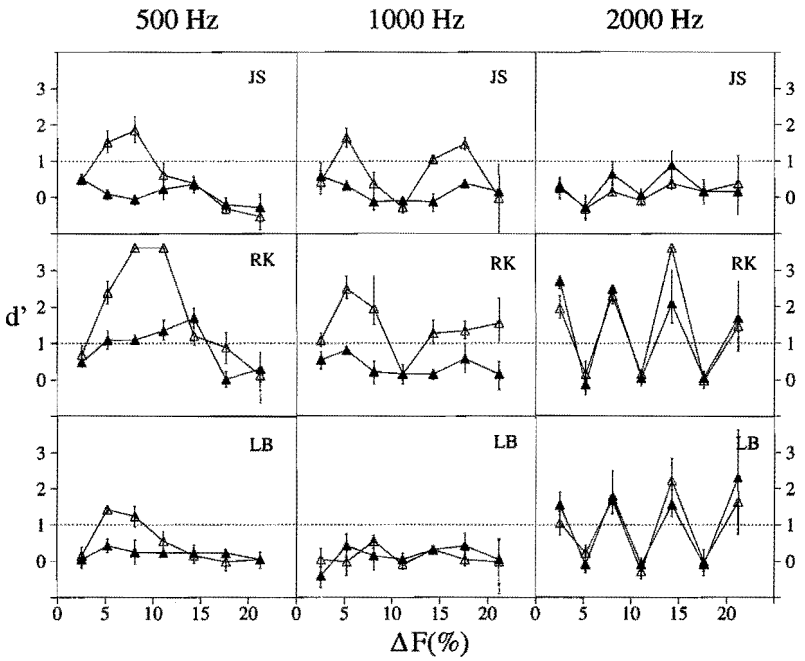


Figure 4.1: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated single-formant signals, with F_{wa} = 100 Hz. Formant frequencies are indicated above each column. Open symbols indicate conditions with constant first formant frequencies, closed symbols conditions with F1 roving. Standard errors of the mean are shown by vertical lines.*

4.4.2 Results

4.4.2.A Single-formant stimuli

Psychometric functions obtained for the single-formant stimuli are shown in Figure 4.1. The open symbols indicate conditions where the first formant was fixed, closed symbols indicate roved-F1 conditions. The alternating patterns of discrimination sensitivity as a function of ΔF, most easily observable for the 2000-Hz formant for subjects LB and RK, can be explained as follows. To a first approximation, each decomposed signal segment can be thought of as a modulated carrier having a frequency equal to the formant frequency. The maxima of the discrimination-sensitivity patterns shown in Figure 4.1 correspond to values of F_{ws} at which the carriers of adjacent decomposed segments are added incoherently in the synthesis phase. Conversely, the minima in the patterns correspond to F_{ws} values at which coherent addition of adjacent segments occurs. In our previous study, incoherent addition was shown to result in spectral auditory cues for a 500-Hz formant and in temporal cues for a 2000-Hz formant.

The psychometric functions for subject RK correspond reasonably well with those reported previously (Figures 4-6 in Kortekaas and Kohlrausch, 1997) but appear to be more pronounced in their variations than the functions for subjects JS and LB. This is probably due to the fact that subject RK was more familiar with the stimuli and that subjects JS and LB were only presented with stimuli that were roved in level. In our previous experiments, level roving did preserve the general pattern of the psychometric function but decreased discrimination sensitivity compared to the "minimal stimulus-uncertainty" condition.

The results for all three subjects for the $F1 = 500$ Hz (fixed and roving) conditions show similar patterns. In the case of fixed formant frequency, subject JS seems to be able to use discrimination cues for $F1 = 1000$ Hz. In the case of a 2000-Hz formant, however, this subject's data tend to show an alternating pattern although all values are below or close to threshold. Interestingly, subject LB shows the opposite discrimination performance. Differences between subjects in the ability to attend to cues were also reported in Kortekaas and Kohlrausch (1997).

4.4.2.B Double-formant stimuli

In Figure 4.2 the psychometric functions for the double-formant stimuli are shown. Open and filled symbols indicate experimental data for the [F1,F2] and [F1R,F2] conditions, respectively. The dashed curves indicate the discrimination performance predicted for [F1,F2] on the basis of the sensitivities determined for the single-formant stimuli with constant F1. These predictions are based on the mean d' values shown in Figure 4.1, under the assumption that the sensitivities for F1 and F2 are independent and that they are optimally combined so that $d' = \sqrt{(d'_{F1})^2 + (d'_{F2})^2}$. They should be thought of as first-order estimates due to the rather large variances. Provided these assumptions are valid, the difference in expected (dashed lines) and measured d' values (triangles) in the [F1,F2] condition indicates dominance of the F1 region. Because discrimination sensitivity in the case of roving the 500 or 1000-Hz formant was below $d' = 1$ with just a few exceptions, the expected sensitivities under the [F1R,F2] condition are expected to be practically identical to the performance for F2 alone (not shown in Figure 4.2).

As a general trend, discrimination performance seems to be predominantly determined by the first formant. For instance, the psychometric function for subject JS for the [500,1000] condition seems to show just a single maximum, even though this subject's discrimination performance showed two maxima for the 1000 Hz single-formant condition. The contribution of the second formant of 2000 Hz (as in [1000,2000]) is expected to be minimal on the basis of this subject's psychometric functions shown in Figure 4.1. Another example for the minor role of the higher formant are the data for the [1000,2000] condition for subject LB where the predicted sensitivities show a clear alternating pat-

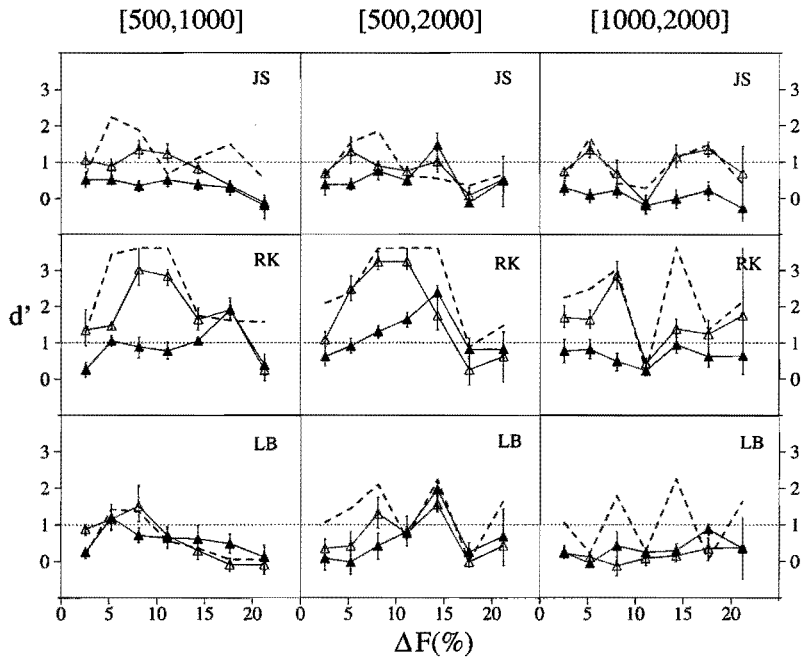


Figure 4.2: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated double-formant signals, with F_{wa} = 100 Hz. Formant-frequencies pairs are indicated above each column. Open symbols indicate conditions with constant F1, closed symbols conditions with roved F1. The dashed lines indicate the d' values predicted on the basis of an optimal combination of the d' values of the non-roved, single-formant condition. Standard errors of the mean are shown by vertical lines.*

tern while the observed d' values are essentially close to zero. Further evidence for F1 dominance can be inferred from the d' values for the [F1R,F2] condition; for all subjects, these are generally comparable to the roved single-F1 condition.

4.4.2.C Influence of level of F2

The dominance of the F1 region observed in Figure 4.2 may have its origin in the difference in levels of F1 and F2. Psychometric functions obtained for filtering the stimuli with "enhanced F2" (as described above) are shown in Figure 4.3. As in Figure 4.2 the dashed lines indicate expected thresholds. Because subject JS apparently could not use the auditory cues arising from changes in the 2000 Hz region (see Figure 4.1), this subject participated for the [500,1000D] condition. Equalization of the formant levels tends to introduce two maxima in this subject's psychometric function, indicating a contribution of F2 of 1000 Hz, although the effect is rather weak. Subjects RK and

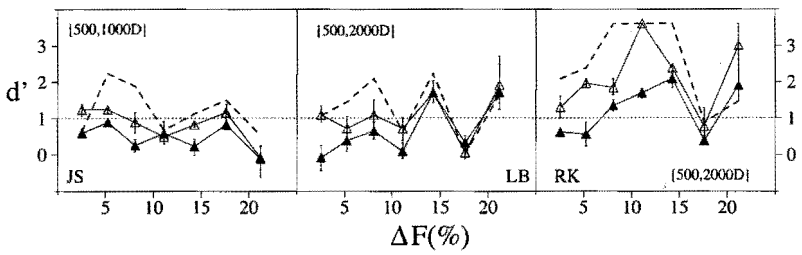


Figure 4.3: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated double-formant signals, with F_{wa} = 100 Hz. The F2 filter is applied twice. Formant-frequencies pairs are indicated in each panel separately. Symbols are used as in Figure 4.2.*

LB performed the experiment for the [500,2000D] condition. They seem to benefit from the higher level of F2 for the higher ΔF shifts only.

4.4.3 Discussion

In order to study the effectiveness of cues from the F2 region, it is necessary that the roving of F1 does not distort the spectrum around F2. Roving of F1 was performed by using a randomly chosen, but constant F1 for each of the three intervals within a trial. Because the phase transfer function of the Klatt formant filter is essentially linear for frequencies far from its resonance frequency, roving of F1 is not expected to alter the relative phases of neighboring harmonics around F2. Moreover, roving of F1 can be thought of as shifting the spectral envelope of the single formant along the frequency axis. This means that level changes of neighboring harmonics around F2 should be strongly coherent under F1 roving, which was confirmed by inspection of amplitude line spectra. Because such coherent level changes are overall-level changes rather than changes of spectral shape, roving of F1 was not expected to affect discrimination performance. This expectation is supported by the results of a study of Sommers and Kewley-Port (1996) who measured formant-frequency JNDs of single formants in the presence of three other formants, as a function of the number of harmonics varied in level (in accordance with the Klatt transfer function). Their results showed that level differences of only 2-3 harmonics in the formant region play a role in formant-frequency discrimination, even though many more harmonics undergo level changes.

As a general trend, auditory cues from the F1 region seem to dominate the discrimination process if both formant frequencies are fixed. In addition, if the first formant of 500 or 1000 Hz is roved, the general finding is that subjects can no longer discriminate between manipulated and unmanipulated signals despite the fact that, in principle, subjects could use cues from the F2 region. If the levels of both formants are more or less balanced,

two out of three subjects seem to be able to focus on information from the F2 region for the high but not the low ΔF shifts. It is unclear at present from where this dependency on ΔF may stem.

Lyzenga and Horst (1997) studied formant-frequency JNDs of single-formant synthetic stimuli having, among other conditions, a spectral envelope that was identical to the amplitude transfer-function of Klatt formant filters. Most of their data could be explained by using a model based on comparison of excitation patterns ("profile analysis"). At least part of the remaining data could be explained using a model based on detection of changes in the temporal envelope. Lyzenga extended these results for *double-formant* signals by measuring JNDs for both single and combined formant-frequency changes (chapter 5 in Lyzenga, 1997). In the case of combined changes Lyzenga proposed a model in which JNDs of both formant regions were assumed to be independent and to be combined in an optimal fashion. This model could accurately predict thresholds for combined changes on the basis of JNDs of single formants.

Those conditions for which the single-formant JNDs were in good agreement with predictions of a temporal model could be described better by Lyzenga's "modified place model" after introduction of an additional low formant. Under such conditions, JNDs of the second formant were generally higher than the corresponding single-formant JNDs. These findings are in line with the trend in the discrimination data observed here for the [F1,F2] conditions, which demonstrated an F1 dominance. Lyzenga hypothesized that, due to the increase in stimulus complexity, listeners are no longer able to attend to the temporal cues if spectral cues are available. It is uncertain whether this means that temporal cues are always inferior to spectral cues or whether the human auditory system is not able to combine both types of cues. The discrimination data for the high ΔF values in the [500,2000D] condition suggest that the temporal cues are not always inferior. Lyzenga (1997) did not report JNDs for synthetic "vowels" having equal levels for F1 and F2 which leaves room for hypothesizing that a role for temporal mechanisms should not be completely ruled out.

4.5 Experiment 2: jitter and shimmer

These experiments were aimed at determining the extent to which random perturbations in either $F0$ or amplitude affect the discriminability of PSOLA-manipulated single-formant signals. First, *detection thresholds* for jitter and shimmer in single-formant signals were measured. Second, psychometric functions were measured for the discriminability of PSOLA-manipulated single-formant signals as a function of the perturbation level. By simulating the detection and discrimination experiments using psychoacoustical models, an attempt was made to clarify the nature of the auditory processes.

4.5.1 Method

4.5.1.A Stimuli

Jitter and shimmer were introduced by randomly varying either the temporal positions or the amplitudes of the pulses in the glottal pulse train (see section 4.4.1.A). The temporal position of the i^{th} pulse was given by $P_i = i \cdot \overline{T} + J_{P,i}$, where $\overline{T}$ is the mean period duration. The jitter process J_P had a Gaussian distribution with zero mean and standard deviation σ_P which will be expressed as a percentage of $\overline{T}$. This way of generating jitter is consistent with a number of other studies (e.g., Cardozo and Ritsma, 1968; Rozsypal and Millar, 1979). In a similar fashion the (linear) amplitude of the i^{th} pulse was determined according to $A_i = \overline{A} + J_{A,i}$, where $\overline{A}$ is the mean pulse amplitude. J_A denotes a zero-mean, Gaussian process with standard deviation σ_A expressed as a percentage of $\overline{A}$.

In the jitter studies mentioned above, perturbation values $J_{P,i}$ within a realization of the jitter process were uncorrelated. In other words, such a sequence of perturbation values is expected to have an autocorrelation function which is only non-zero for zero lag. Hillenbrand (1988) tested the validity of this assumption of uncorrelatedness in a study on the influence of jitter and shimmer, simulated in synthetic vowels, on perceived roughness. To this end, Hillenbrand first measured autocorrelation functions for period duration and voice amplitude in natural sustained vowels. Instead of a single peak at zero lag, the measured autocorrelation functions had an exponential decay which indicates that the perturbation values in the sequences were correlated (for a recent review see Schoentgen and De Guchteneere, 1997). The difference in autocorrelation functions between correlated and uncorrelated sequences is reflected in their power spectra, because of the Fourier-pair relation between autocorrelation function and power spectrum. In this sense, the uncorrelated sequences have a white (flat) power spectrum, whereas the correlated sequences described in Hillenbrand (1988) have a low-pass (LP) characteristic. Hillenbrand argued that the power spectra "were similar, but not identical to $1/f$ spectra". Because the exact nature of $1/f$ processes is still unknown, Hillenbrand proposed an algorithm that yielded autocorrelation functions *approximating* the functions measured for sustained vowels.

In the present study, spectrally-white and LP perturbation sequences were generated in a slightly different way. Gaussian noise buffers were generated whose samples indicated the instantaneous jitter values $J_{P,i}$ or shimmer values $J_{A,i}$ in the case of *uncorrelated* sequences. By filtering these sequences using a LP filter, correlated sequences were obtained which again indicated instantaneous perturbation values. The cut-off frequencies of the filters were derived from the autocorrelation functions shown in Figure 2 in Hillenbrand (1988). More details of this procedure are given in the Appendix.

For shimmer, the method of introducing perturbation by modulating pulse amplitudes is consistent with the procedure described in Hillenbrand (1988). For jitter, however, both methods differ in the sense that Hillenbrand introduced perturbation to *period durations* rather than to pulse positions. Hillenbrand's procedure likely results in large deviations with respect to regular pulse positioning (at positions given by $i \cdot \overline{T}$); because the process is a random walk with respect to pulse positioning, deviations are bound to accumulate. This means that the shift of the PSOLA windows, regularly spaced at a rate of F_{wa} Hz, relative to the excitation pulses has a certain probability to steadily increase or decrease. Because our main interest was the influence of small irregularities in $F0$, *pulse positions* were jittered which has a limited effect on this "mismatch". The power spectrum of the corresponding perturbations in terms of *period duration*, however, is rather flat so that strictly speaking our simulations are not compatible with those of Hillenbrand.

First, for the baseline detection-threshold measurements, (single-) formant frequencies of 500, 1000, and 2000 Hz were used, with bandwidths of 50, 50, and 100 Hz, respectively. The $F0$ of all signals (i.e. without jitter) was 100 Hz, yielding a mean period duration $\overline{T}$ of 10 ms. For the 1000-Hz formant an additional condition was measured in which $F0$ was 250 Hz ($\overline{T} = 4$ ms). Thresholds were obtained for both the LP and the spectrally white perturbation conditions. The presentation sound-pressure level, range of level roving, and stimulus duration were the same as in Experiment I. Signals were generated at a sample rate of 48 kHz which was the highest rate obtainable given our hardware specifications.

Second, for the experiments dealing with discrimination of perturbed PSOLA-modified stimuli, formant frequencies of 500 and 2000 Hz were used with bandwidths as given above. These two formant frequencies were chosen here because in our previous study discrimination was found to be based on either spectral cues (500 Hz) or temporal cues (2000 Hz). In this way, the effect of perturbation on these two types of cues could be investigated separately. Psychometric functions were measured, with perturbation level as experimental parameter, for two ΔF values for both formant frequencies. For the 500-Hz formant, ΔF was either 8.11 %, yielding above-threshold to almost perfect discrimination in the absence of perturbation, or 17.65 %, yielding performance at chance level or close to threshold (see Figure 4.1). For the 2000-Hz formant, the respective ΔF values were 14.29 and 11.11 %. The test signal was first generated with an $F0$ of F_{wa} Hz including perturbation, and subsequently shifted in $F0$ using PSOLA. Reference signals were directly synthesized with an $F0$ of F_{ws} Hz, including (independent realizations of) perturbation. Note that in the case of jitter, the analysis window positions corresponded to the unjittered instants of excitation meaning that window maxima and filter excitations were no longer exactly aligned. This can be thought of as a simulation of small estimation errors in the local $F0$ in modification of natural speech.

4.5.1.B Procedure

In the (baseline) detection experiment, thresholds of jitter and shimmer were measured using an adaptive 3I-3AFC paradigm. Two of the intervals contained strictly periodic single-formant signals. The odd-ball interval contained a signal whose perturbation level, in terms of σ_P or σ_A, was adaptively varied. In each trial, a new realization of the perturbation process was applied to the odd-ball interval ("running jitter or shimmer"). The initial perturbation level of each adaptive measurement was sufficiently high to facilitate easy detection. After two consecutive correct responses, the perturbation level was decreased. After each incorrect response, the level was increased again so that the 70.7 % correct level was tracked ("two-down, one-up" Levitt, 1971). At each reversal from increasing to decreasing perturbation level, the step size was halved. For jitter, the initial step size was 0.8 % and 0.4 % for LP and white jitter, respectively. The minimum step sizes were 0.1 and 0.05 %, respectively. For shimmer, the initial step sizes were 4.5 and 2.0 % for LP and white shimmer, with minimal step sizes of 0.5625 and 0.25 %, respectively[4]. After reaching the minimal step size, the perturbation levels at 8 subsequent reversals were recorded. The median of these levels was taken as a threshold data point. For each condition, each subject performed five threshold measurements of which the first measurement was regarded as a practice run. The data shown below are the means and standard deviations of the four remaining measurements.

In the discrimination experiment, psychometric functions were measured for discriminability of perturbed, PSOLA-modified stimuli as a function of the level of perturbation. These data were only obtained for LP perturbation. Perturbation levels for jitter, in terms of σ_P, were 1.2, 2.4, 4.8, and 9.6 %, and for shimmer, in terms of σ_A, 6.75, 13.5, 27, and 54 %. The unrealistically high σ_A values could lead to changes of sign of the pulse amplitude and the subjects described these stimuli correspondingly as very rough and irregular (cf. Hillenbrand, 1988). These stimuli reminded one of the subjects of recordings of the first contact with Neil Armstrong on the moon. The same 3I-3AFC experimental procedure as described for Experiment I was used. Each subject performed 5 runs, each consisting of 15 trials. The first run of each condition was taken as learning phase and left out of the analysis. The means and standard errors of the mean shown in the figures below thus represent 60 trials, and will be given as d' values.

4.5.1.C Subjects

Four subjects, including the first author, participated in the experiments. They were aged between 23 and 29. Two subjects (MV and RK) participated in both the detection and discrimination experiments. All subjects had at least some experience in

[4]By using a sampling rate of 48 kHz, the sampling interval is approximately 21 μs which corresponds to approximately 0.21 % of the period for an $F0$ of 100 Hz.

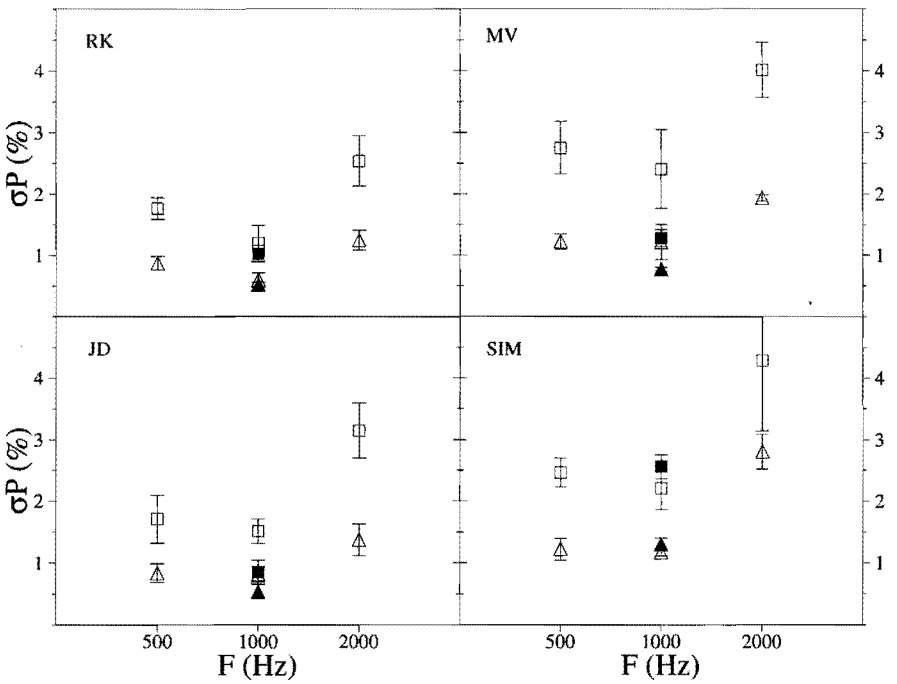


Figure 4.4: *Jitter detection thresholds for σ_P as a function of formant frequency. Open symbols indicate the $F0$ = 100 Hz condition, closed symbols the 250 Hz condition. Squares indicate data for LP jitter, triangles for white jitter. Standard deviations are shown by vertical lines. Model predicted thresholds, using the modulation-discrimination model, are shown in the lower right panel.*

psychoacoustic listening experiments. Except for subject RK the subjects were paid an hourly wage for their services.

4.5.2 Results

4.5.2.A Detection thresholds

Detection thresholds for jitter perturbation are shown in Figure 4.4 for the LP jitter (squares) and the white jitter condition (triangles). Open symbols are data obtained for $F0 = 100$ Hz, closed symbols for $F0 = 250$ Hz (1000-Hz formant only). Model predictions ("SIM"), which are displayed in the lower right panel, will be discussed later. For the LP and white jitter conditions separately, a low-to-high ranking of the thresholds would be: 1000 ($F0$ = 250 Hz), 1000 (100 Hz), 500, and 2000 Hz. Thresholds for the spectrally-white jitter condition are about a factor of two lower than the thresholds for LP jitter. Moreover, the standard deviations for the LP-jitter threshold data tend to be larger

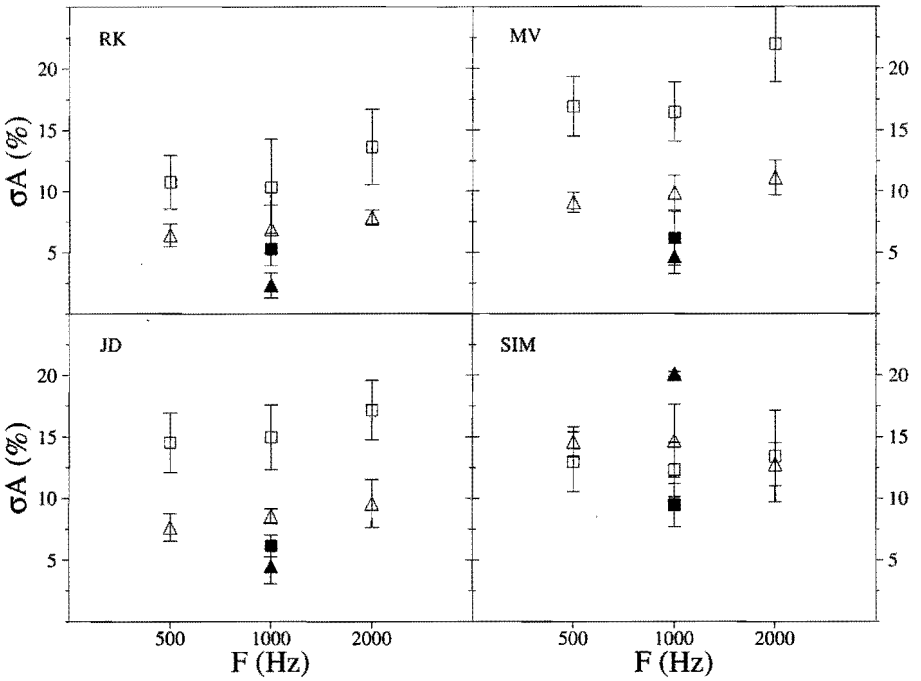


Figure 4.5: *Shimmer detection thresholds for σ_A as a function of formant frequency. Symbols are used as in Figure 4.4. Model predicted thresholds, using the modulation-discrimination model, are shown in the lower right panel.*

which is presumably a consequence of the larger variability across realizations of the LP-jitter process.

Figure 4.5 shows detection thresholds for shimmer perturbation using the same symbols as in Figure 4.4. Thresholds for the 2000-Hz-formant again tend to be highest, although the effect is less pronounced than for jitter. Comparable to jitter perturbation, thresholds for the white perturbation condition are lower than for the LP condition, and standard deviations tend to be larger for the latter condition. The high 250-Hz $F0$ condition yields the lowest thresholds which is a clear trend in the data of all subjects.

4.5.2.B Jittered PSOLA-manipulated signals

Figure 4.6 shows psychometric functions for subjects MV, AF, and RK for discrimination of a PSOLA-manipulated single-formant signal from an unmanipulated comparison as a function of perturbation level σ_P. Model predictions, shown by the filled symbols, will be treated in detail later.

Starting with the data of the 500-Hz formant (lower panels) and the baseline condition of $\sigma_P = 0$ %, it can be observed that all three subjects have no difficulty in discriminating

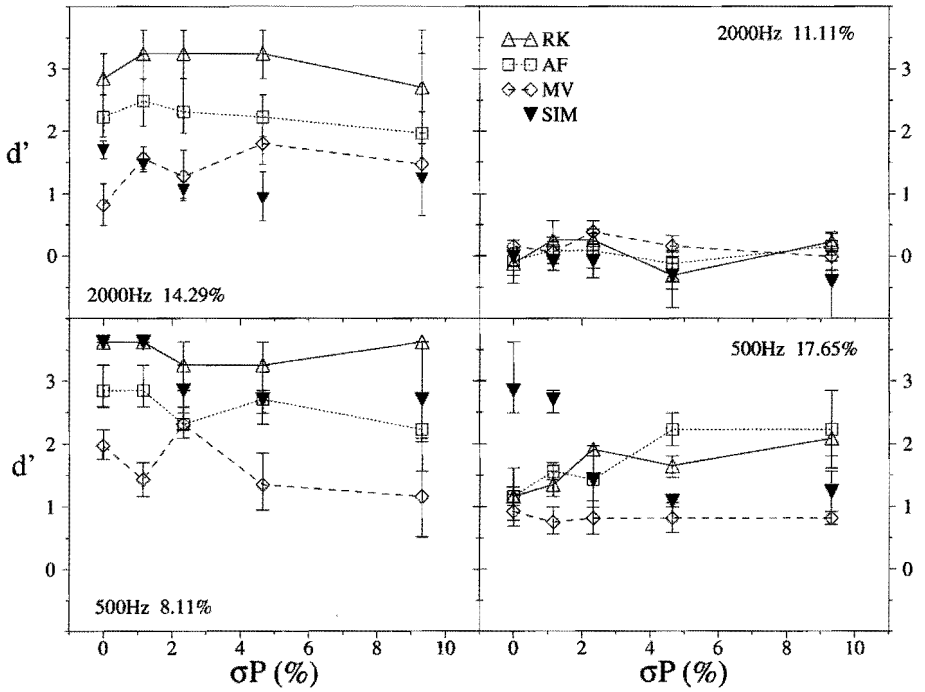


Figure 4.6: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated single-formant signals, with F_{wa} = 100 Hz, as a function of the jitter perturbation level σ_P. Closed symbols indicate model predicted sensitivities (see text for details). Standard errors of the mean are shown by vertical lines.*

the PSOLA-manipulated signal if ΔF = 8.11 % (lower left panel), and show near-threshold performance if ΔF = 17.65 % (lower right panel). For the corresponding stimulus conditions in experiment I, subjects JS and LB performed reasonably above threshold and at chance level, respectively, for these two ΔF values. The data of subject RK shown here are comparable to those displayed in Figure 4.1. Linear regression analysis for ΔF = 8.11 % yielded estimated slopes of the psychometric functions which were not significantly different from zero ($p > 0.15$) level for all subjects. This means either that jitter does not seriously hamper the use of the spectral discrimination cues, or that new cues become available at higher perturbation levels. An example of the latter explanation appears to occur for ΔF = 17.65 % where the psychometric functions for subjects AF and RK increase from $d' \simeq 1$ for no jitter to $d' \simeq 2$ for maximal jitter. The slopes of the psychometric function of these two subjects resulting from linear regression were about 0.1 (d' units per percent of σ_P) and were found to be significantly different from zero at the $p < 0.05$ and $p < 0.005$ levels, respectively. For subject MV the function

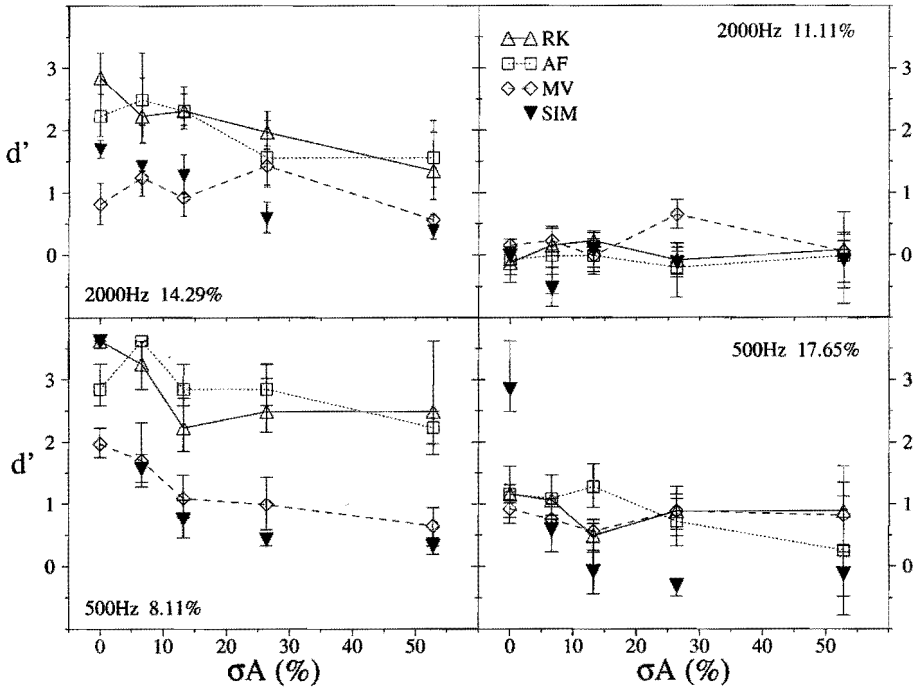


Figure 4.7: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated single-formant signals, with F_{wa} = 100 Hz, as a function of the shimmer perturbation level σ_A. Closed symbols indicate model predicted sensitivities (see text for details). Standard errors of the mean are shown by vertical lines.*

remains essentially flat around $d' = 1$ ($p > 0.8$).

The discrimination performance for the 2000 Hz-formant condition (upper panels) seems to be even more independent of jitter perturbation. For ΔF = 14.29 % (left panel) the psychometric function has a slight negative slope for subjects AF and RK. For ΔF = 11.11 % (right panel) the functions of all three subjects vary around $d' = 0$ for all perturbation levels. None of the psychometric functions have linear regression slopes significantly different from zero at the $p < 0.1$ level. These two findings suggest that, under these conditions, temporal discrimination cues remain in tact under jitter perturbation, and that an increase of perturbation level does not result in new PSOLA-induced discrimination cues. The variability in sensitivity among the three subjects is large, both for the 500 and 2000-Hz formant.

4.5.2.C Shimmered PSOLA-manipulated signals

Psychometric functions of discrimination performance under influence of shimmer per-

turbation are shown in Figure 4.7. The data for $\sigma_A = 0$ % are the same as those shown in Figure 4.6 for $\sigma_P = 0$ %. As a general trend, performance for $\Delta F = 8.11$ % under the 500 Hz-formant condition (lower left panel) decreases by about one to two d' units as a result of increasing σ_A from 0 to 54 %. Linear regression slopes for subjects AF and MV were about -0.02 (d' units per percent of σ_A) and were significantly different from zero at the $p < 0.01$ level. The slope for subject RK was not found to be significantly different from zero, although there may be a difference between the data for $\sigma_A = 0$ and 6.5 %, and the higher shimmer levels (not tested).

Performance for $\Delta F = 17.65$ % (lower right panel) remains either more or less constant near threshold (for subjects MV and RK; $p > 0.5$), or tends to decrease (subject AF; slope of -0.02, $p < 0.05$). For the 2000-Hz formant and $\Delta F = 14.29$ % (upper left panel) slopes of -0.02 and -0.03 were found for subjects AF and RK ($p < 0.05$ and $p < 0.005$, respectively), and an insignificant slope ($p > 0.25$) for subject MV. No significant effect of σ_P was found for $\Delta F = 11.11$ % shown in the upper right panel ($p > 0.75$). In general, the discrimination sensitivities for subjects AF and RK agree better than under the jitter condition. For both perturbation conditions, subject MV showed a discrimination performance about 1-2 d' units lower in sensitivity compared to the other subjects.

4.5.3 Modulation-discrimination model

Kortekaas and Kohlrausch (1997) described a modulation-discrimination model that estimated perceptual distances between PSOLA-manipulated and unmanipulated single-formant signals in terms of differences in their "effective modulation depth". The signals were filtered by means of a single Gammatone filter centered at or near the formant frequency. The temporal envelope of the filtered signal was calculated by performing a discrete Hilbert Transform. The effective modulation depth M was defined as $M = \sqrt{2} \cdot \sigma_e / m_e$, where σ_e is the standard deviation of the envelope, and m_e is its mean. As a distance measure between a reference and a PSOLA-manipulated test signal, the absolute value of the difference of their modulation powers was taken, i.e., $D_{mod} = |M_{ref}^2 - M_{sig}^2|$. For a formant of 2 kHz, satisfactory linear-regression fits were obtained between D_{mod} and discrimination sensitivity as a function of ΔF, using the equation $d' = k \cdot D_{mod}$. In this equation, k is some constant.

4.5.3.A Detection thresholds

The model was used here both to predict detection thresholds for σ_P and σ_A and to predict discriminability of PSOLA-manipulated stimuli under influence of perturbation. In the case of jitter detection, the modulation power M_{sig}^2 of a perturbed test signal may be larger than M_{ref}^2 of a strictly periodic reference due to the interaction of the (varying) instantaneous $F0$ and the formant structure in the spectral envelope (cf. Neelen, 1969).

In some sense, this resembles an FM-to-AM conversion (Zwicker, 1952). As for shimmer thresholds, M^2_{sig} is expected to be larger because shimmer introduces amplitude modulation to the pulse train.

The lower right panels of Figures 4.4 and 4.5 show predicted thresholds for σ_P and σ_A obtained by letting the model take the role of "listener". At each trial, the model calculated M^2 for each of the three intervals. In order to limit detection sensitivity, an independent "internal noise" value was added to each of the M^2 values. To this end a zero-mean Gaussian noise was used with standard deviation σ_{int}. The interval that deviated most in terms of M^2 from both others was regarded as the odd-ball. In the framework of signal-detection theory (Green and Swets, 1966), (detection) sensitivity can be taken to obey $d' = D_{mod}/\sigma_{int}$. Kortekaas and Kohlrausch (1997) reported detection thresholds ($d' = 1$) for D_{mod} of 0.14 for 1000 Hz, and of 0.25 for 2000 Hz, so that σ_{int} should be equal to these values. In the simulated detection experiments, however, these internal noise levels were found to be much too high. The model simulations of Figures 4.4 and 4.5 were run instead with $\sigma_{int} = 0.03$ for the 2000-Hz formant, and 0.02 in the other conditions. This conspicuous difference in sensitivity will be discussed in section 4.5.5.C. Further details of the adaptive procedure were the same as for the "real" experiment.

The detection thresholds shown in Figure 4.4 replicate the general trends observed in the data of the subjects, at least for a fundamental of 100 Hz. Simulation thresholds for $F0 = 250$ Hz are too high, although the ranking of the thresholds is in line with the experimental data. Quantitatively the simulation thresholds are closest to those of subject MV. Note that by choosing a lower level of σ_{int} of the internal noise, the simulated thresholds may correspond better to the data of RK and MV. The simulated thresholds for shimmer shown in Figure 4.5, however, match the experimental data to a lesser extent, although the simulated thresholds do show an independence of formant frequency, probably even to greater extent than the experimental data. Thresholds for the LP and white-shimmer condition lie much closer to each other than in the experimental data. The thresholds for the high 250 Hz $F0$ condition are much too high and in addition show the wrong ranking.

4.5.3.B Discrimination of PSOLA-modified stimuli

As for the influence of perturbation on the *discrimination* of a PSOLA-manipulated 2-kHz formant signal, it is expected that high values of σ_P and σ_A decrease the difference in modulation power and thus deteriorate discrimination performance. Because this task was essentially the same as the earlier task of discriminating unperturbed stimuli, σ_{int} was set to 0.25 as deduced in our previous study. The model-predicted psychometric functions are shown as the closed symbols in the top panels of Figures 4.6 and 4.7. The

simulated functions for jitter (upper panels in Figure 4.6) capture the general trend in the experimental data, in that discrimination varies within about one d' unit as a function of σ_P. In absolute terms, discrimination sensitivity for $\Delta F = 14.29$ % (upper left panel in Figure 4.6) is rather low compared to the data of the subjects, although they also show large variation in sensitivity. The predicted psychometric functions for the shimmered 2000-Hz formant, displayed in the upper panels of Figure 4.7, are in line with the observations for jitter just described.

4.5.4 Intensity-discrimination model

The predicted thresholds and psychometric functions shown in the lower panels of Figures 4.6 and 4.7 for the 500-Hz formant were obtained by employing an intensity-discrimination model (Gagné and Zurek, 1988; Kortekaas and Kohlrausch, 1997) as "listener". In this model, stimuli are submitted to a critical-band analysis which is performed by means of a Gammatone filter bank (Patterson et al., 1987). Subsequently the power in each band is determined, yielding an "excitation pattern". Discrimination of two stimuli is based on detecting level differences in their excitation patterns. In the present study, the multiband model was applied whose distance measure D_{sum} is equal to $(\sum_{i=1}^{N} \Delta L_{E,i}^2)^{1/2}$, where $\Delta L_{E,i}$ is the level difference in channel i. Because the integration time equals the stimulus duration of 300 ms in the present study, this model only takes spectral cues into account. In the simulated 3IFC experiments D_{sum} was corrupted by a Gaussian internal noise with a standard deviation of 1 dB, a value which was found to correspond to threshold in our previous study.

4.5.4.A Discrimination of PSOLA-modified stimuli

For jitter (lower panels in Figure 4.6), simulated discrimination sensitivities display only a moderate influence of increasing σ_P for $\Delta F = 8.11$ %. Remarkably, the sensitivity in the case of $\Delta F = 17.65$ % is seen to drop from $d' > 2$ for the two lowest σ_P values to just above threshold, which is not in agreement with the experimental data. More or less in contrast, the model psychometric functions for shimmer (lower panels in Figure 4.7) do show the "correct" trend, at least for $\sigma_A > 0$ %, but suggest that the model sensitivity is too low.

4.5.5 Discussion

4.5.5.A Detection thresholds

Cardozo and Ritsma (1968) reported jitter detection thresholds for band-pass filtered pulse trains using a filter with a Q-factor of 3 and slopes of 34 dB per octave. The sequences of pulse-position perturbations had a white spectrum. Thresholds were found

to decrease with increasing stimulus duration up to approximately 100 ms, and to remain constant for longer durations. In the latter case, Cardozo and Ritsma reported thresholds for pulse sequences with a (mean) $F0$ of 100 Hz and filtered at 1000 Hz center frequency of $\sigma_P = 0.5$ %. For a higher center frequency of 1400 Hz the threshold was about 0.7 %. These two results are in line with the thresholds reported here: for a 1000-Hz and 2000-Hz formant thresholds are in the range 0.5-1.0 % and 1.2-2.0 %, respectively. For high $F0$ levels of 333 Hz Cardozo and Ritsma (1968) reported thresholds for a "formant" at 1400 Hz of about 0.2 %, which is a factor of two to three lower than the present thresholds of 0.4-0.6 % for a 1000-Hz formant with a 250 Hz fundamental. Neelen (1969) measured thresholds for a center frequency of 500 Hz (of a band-pass filter with 50 Hz bandwidth) of 0.5 %. This value is also a factor of two lower than the present range of 0.8-1.1 %. Pollack (1971a) used rectangular pulses randomly varied in duration by $\pm\Delta T$, without any additional "formant" filtering. The value of ΔT at threshold was in the order of 1-2 %. Pollack (1971a) also measured shimmer thresholds by randomly varying the amplitudes of the rectangular pulses by $\pm\Delta A$. For stimulus $F0$ and durations comparable to the present study, detection thresholds were 10 - 20 % which is a factor two higher than the range of 5 - 10 % reported here.

The fact that the jitter thresholds presented here generally are a factor two higher may be due to differences in paradigms: for instance, Cardozo and Ritsma (1968) used a adaptive 2IFC paradigm tracking the 72 % correct level, which corresponds to $d' \simeq 1$. In the present study, the 70.7 % correct level in a 3IFC paradigm corresponds to $d' \simeq 1.28$ (Macmillan and Creelman, 1991). Additionally, the difference in thresholds may have been caused by the limited temporal resolution of the digital signal representation, relative to the analogue signal generation of the older studies.

4.5.5.B Discrimination

The initial hypothesis in the discrimination experiments was that the introduction of jitter and shimmer in the single-formant signals would severely affect the subjects' discrimination performance. The psychometric functions shown in Figures 4.6 and 4.7 do not support this hypothesis: even though performance tends to deteriorate for larger values of σ_P and σ_A for $\Delta F = 14.29$ % (2000 Hz) and $\Delta F = 8.11$ % (500 Hz), the size of this effect is limited to approximately one d' unit. In other words, the PSOLA-manipulated signals remain discriminable.

For the $\Delta F = 17.65$ % (500 Hz) condition, an unexpected increase in discrimination sensitivity for subjects AF and RK could be observed for larger values of σ_P. An explanation for this increase might be that jitter perturbation actually had a different effect on the the test and reference signals, due to the interaction between harmonics and the (formant) spectral envelope (Neelen, 1969). This can be reasoned as follows:

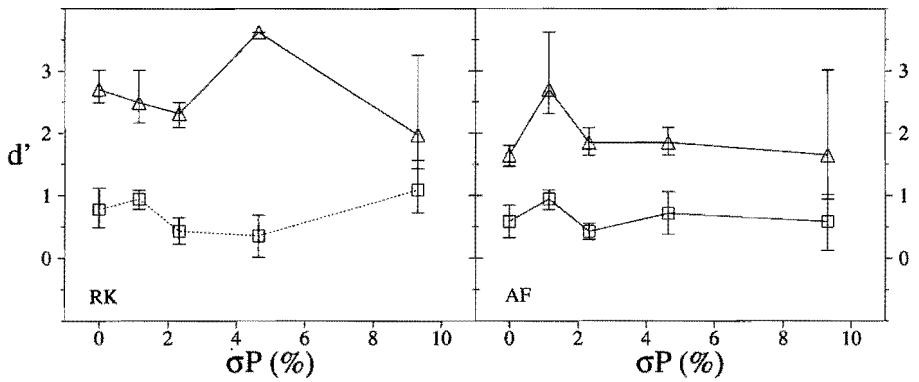


Figure 4.8: *Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated single-formant signals, with F_{wa} = 100 Hz, as a function of the shimmer perturbation level σ_A and under formant roving. Triangles indicate data for the 2000 Hz formant (ΔF = 14.29 %), squares for the 500 Hz formant (ΔF = 17.85 %). Standard errors of the mean are shown by vertical lines.*

jittering of the pulses can be thought of as randomly varying the instantaneous $F0$. This means that the "harmonics" of the signal will be shifted randomly "under" the spectral envelope. In the case of the test signal prior to PSOLA-modification, the strong fifth harmonic of the 100 Hz $F0$ coincides with the formant frequency of 500 Hz. Roving of $F0$ thus results only in moderate AM at the output of the auditory filter, because side harmonics remain relatively weak. A necessary condition is that the fluctuations in $F0$ are small compared to the bandwidth of the formant. In the case of ΔF = 17.65 % (500 Hz) in the absence of jitter, there is a strong harmonic at about 470 Hz and a somewhat weaker harmonic at 590 Hz. In this case, random fluctuations of $F0$ influence the relative levels of these harmonics significantly so that, given an ERB of about 70 Hz at 500 Hz center frequency, considerable AM in the output of an auditory filter can be expected. The difference in modulation depths between the test and reference intervals may therefore have been an additional cue for larger values of σ_P.

This explanation can be tested by *roving* the formant frequency; this alters the relative levels of harmonics, also for the test-signal interval. Because roving of the 500-Hz formant for ΔF = 17.65 % caused discrimination performance to drop to close to, or below threshold (see Figure 4.1), the expected performance is close to threshold and independent of σ_P. If this explanation is satisfactory for the 500-Hz formant case, it may be questioned whether discrimination for large values of σ_P was also based on differences in modulation depth due to perturbation in the case of the 2000-Hz formant for ΔF = 14.29 %. Here too, roving of the formant frequency should in principle make

such cues ineffective. The performance in this case is expected to be independent of σ_P and generally far above threshold (see Figure 4.1, at least for subjects LB and RK). Figure 4.8 shows the psychometric functions obtained for roving the formant frequency by $\pm 3\,\Delta F1$ (Gagné and Zurek, 1988). As predicted, the psychometric function for the 500-Hz formant, indicated by squares, remains flat around $d' \simeq 0.5$ ($p > 0.75$ for AF, and $p > 0.5$ for RK). For the 2000-Hz formant, indicated by triangles, the functions have more or less the same shape as those shown in Figure 4.6, and the corresponding linear regression slope for subject AF is not significantly different from zero ($p > 0.75$). Although a slight negative slope may be present in the data for subject RK, the slope does not significantly differ from zero ($p > 0.25$), which is probably due to the perfect discrimination for $\sigma_P = 4.8$ %. In general, discrimination performance is well above threshold which suggests that discrimination was not based on additional cues in the fixed-formant condition.

4.5.5.C Modeling

With a few exceptions for shimmer, the general trends in the experimental psychometric functions, showing discriminability of PSOLA-modified stimuli as a function of perturbation level, could be replicated using the same models as in our previous study. In other words, this finding supports the hypothesis that subjects are able to persist in using spectral cues for the low formants, and temporal cues for the high formants in the presence of perturbation. In addition, the assumption that detection of perturbation can be modeled by detecting changes in modulation depth in a single auditory filter was found to be fruitful for $F0 = 100$ Hz in the case of jitter and also, but to a lesser extent, of shimmer.

The model thresholds for jitter and shimmer for the 250 Hz $F0$ condition, however, were found to be much too high. Pollack (1971b) argued that jitter detection for stimuli with such short pulse-to-pulse intervals relies on spectral cues, rather than on temporal cues. The intensity-discrimination model yielded the following thresholds (with σ_{int} set to 1 dB): thresholds for σ_P for white jitter of 0.8 % (standard deviation 0.07 %) and for LP jitter of 1.6 % (1.17 %), and for σ_A thresholds for white shimmer of 5.4 % (0.1 %) and for LP shimmer of 4.4 % (1.32 %). In all four cases predicted thresholds are in better agreement with the experimental data, although the prediction for LP jitter is somewhat higher than expected. The order of the thresholds for LP and white shimmer seems to be wrong, albeit that the thresholds appear to lie close to each other in the experimental data as well (especially for subjects MV and JD).

The modulation-discrimination model was based on two assumptions: (1) jitter and shimmer detection essentially is a one-channel process, and (2) the decision variable solely depends on first and second-order statistics of the temporal envelope within that

channel. The first assumption might explain the observed difference in internal noise required to yield satisfactory threshold predictions in the present study, and good linear fits to discrimination performance in our previous study. In the absence of perturbation, the spectral changes due to PSOLA are essentially restricted to the formant area making it likely that a limited number of auditory filters will be attended to. Jitter and shimmer perturbation, however, are wide-band processes which (coherently) affect all harmonics of the glottal source. Under the assumption that the internal noises in different channels are uncorrelated, the differences in effective modulation depths may be optimally combined across channels. This means that, in order to achieve the same discrimination sensitivity, the internal noises per individual channel may be larger. Combination of modulation information across auditory filters seems to be present in detecting sinusoidal modulation of broadband noises (Eddins, 1993; Dau et al., 1997).

As for the second assumption, the use of merely first and second-order envelope statistics, there is growing evidence that the human auditory system is able to perform an analysis of the distribution of modulation frequencies within auditory channels (e.g., Houtgast, 1989). In order to simulate such processing, Dau et al. (1997) proposed a modulation filterbank acting on the output of simulated auditory filters. For the detection of jitter and shimmer, such a spectral decomposition might separate the intrinsic modulation around $F0$ of the (quasi-)harmonic formant signal, and the modulation probably lower in frequency due to the random perturbations. This may be a topic for future investigations.

4.6 General discussion

In this section some of the results of experiments I and II will be discussed in terms of their relevance for PSOLA modification of natural speech.

A prominent outcome of experiment I was the dominance of the low F1 region in discriminating PSOLA-modified double-formant stimuli. It seems plausible to assume that such a dominance also occurs when modifying *natural* vowels in $F0$. The cues associated with PSOLA-induced changes in the low F1 region will have a spectral nature (Kortekaas and Kohlrausch, 1997), at least for $F0$ in the normal range of natural speech (about 80 to 400 Hz). These cues vanish, however, by roving the low formant in formant frequency. In natural (steady) vowels, the formant frequency may fluctuate in time implying that differences in instantaneous harmonic levels may be of the same order as the instantaneous spectral changes introduced by PSOLA. If the PSOLA-induced differences are greater, then discrimination may still be hampered by the rate of the formant fluctuations; if the rate is too high, the auditory system may not be able to reliably integrate PSOLA-induced spectral information. In contrast, temporal cues associated

with changes in the high formant region were found to be much less affected by formant roving. These cues, however, are also less likely attended to due to the F1 dominance. In sum, the F1 dominance may explain part of the success of PSOLA when applied to natural speech.

Pinto and Titze (1990) advocated the use of standardized perturbation measures. For jitter, one can adopt the Mean Rectified jitter measure[5] (Horii, 1980) according to which the jitter levels applied in the discrimination experiments were approximately 0, 0.75, 1.5, 3.0 and 6.0 %. Likewise, according to the Mean Rectified shimmer measure the shimmer perturbation levels were 0, 0.3, 0.6, 1.2, and 2.4 dB. Both the jitter and shimmer ranges are comparable to the range of investigation in Hillenbrand (1988). The *detection thresholds* for the 500 and 1000-Hz formants shown in Figures 4.4 and 4.5 would correspond to $MR(J)$ values of 0.5-0.8 % for LP and white jitter, and $MR(S)$ values of 0.75-1 dB for white shimmer and 0.2-0.3 dB for LP shimmer. The difference between white and LP shimmer thresholds is reflected in the disparity between the roughness rating presented in Figure 5 in Hillenbrand (1988): for both LP and white shimmer, the corresponding $MR(S)$ values resulted in roughness ratings of about 20 on a scale from 10-90. Horii (1980) measured jitter and shimmer in three natural sustained vowels uttered by thirty-one speakers and found $MR(J)$ and $MR(S)$ values of 0.6-0.7 % and 0.3-0.4 dB, respectively. According to our threshold measurements, these perturbation levels would be at detection threshold. The perturbation levels used in determining the influence of jitter and shimmer of discriminating PSOLA-induced changes would range from close to, to far above the levels measured in healthy voices. This suggests that vocal perturbation, as a source of variation present in natural speech, presumably does not contribute to the success of PSOLA by hampering the use of auditory cues. As was pointed out in the Introduction, however, the plausibility of making a distinction between shimmer and jitter, and of associating these perturbations to the voice source exclusively, is still an issue under debate.

[5]Horii (1980) defined Mean Rectified jitter as

$$MR(J) = \frac{1}{N-1} \sum_{i=1}^{N-1} | T_i - T_{i-1} | ,$$

where T_i is the i^{th} period duration. Mean Rectified shimmer was defined as

$$MR(S) = \frac{20}{N-1} \sum_{i=1}^{N-1} \left| log_{10} \frac{A_i}{A_{i-1}} \right| ,$$

where A_i denotes the waveform peak in the i^{th} period.

4.7 Conclusions

(1) Discrimination of PSOLA-modified double-formant stimuli is predominately determined by the availability of cues in the region of F1. If the level differences in F1 and F2 due to the spectral tilt are canceled, cues from the F2 region can be used in a restricted number of conditions.

(2) The detection thresholds for LP jitter or shimmer are generally higher than the thresholds measured for spectrally-white perturbation. Thresholds are found to depend on formant frequency to a somewhat greater extent for jitter than for shimmer. In agreement with data reported in the literature, thresholds for the high $F0$ conditions are generally lowest.

(3) The detection thresholds for jitter perturbation of single-formant signals can be modeled reasonably well by analyzing the effective modulation depth at the output of simulated auditory filters. For shimmer, thresholds are moderately well predicted in quantitative terms, although differences between the LP and white shimmer conditions are not replicated.

(4) The jitter and shimmer detection thresholds for the high $F0$ condition can be accurately predicted by applying an intensity-discrimination ("excitation-pattern") model.

(5) Although jitter and shimmer perturbations do influence discrimination sensitivity, the general pattern of discrimination performance of PSOLA-modified stimuli is maintained under influence of jitter and shimmer.

(6) In discriminating PSOLA-modified from unmodified stimuli, additional modulation cues can be introduced as a result of the interaction between the formant frequency and the harmonics of the signal subjected to PSOLA modification.

4.8 Appendix

Sequences of instantaneous perturbation values $J_{P,i}$ or $J_{A,i}$ were obtained by (1) generating Gaussian noise buffers, and (2) LP filtering these buffers in the case of correlated sequences. To determine the parameters of LP filtering, the autocorrelation functions in figure 2 of Hillenbrand (1988) were parametrized by the function $R_{nn}[l] = e^{-\beta|l|}$, where l is the lag (in periods) and $|.|$ denotes absolute value. R_{nn} denotes the *normalized* autocorrelation. A close fit to the measured functions was obtained by setting parameter β to 0.23 for the synthesis of jitter, and to 0.046 for the synthesis of shimmer. The corresponding power spectra have a LP shape similar to the power spectrum of a noise filtered by a second-order LP filter with a cut-off frequency of 4 and 1 Hz, respectively[6]. Note that for white jitter and shimmer perturbation parameter β has to be set to ∞ resulting in a single peak at $R_{nn}[0]$.

The relation between *pulse-position* and *period-duration* perturbation can be described as follows. The i^{th} period duration is by definition equal to

$$\begin{aligned} D_i &= P_i - P_{i-1} \\ &= \overline{T} + J_{P,i} - J_{P,i-1} \\ &= \overline{T} + J_{D,i}, \end{aligned}$$

where P_i is the i^{th} pulse position and $J_{P,i}$ is the i^{th} perturbation value. The mean of J_D is zero and its standard deviation can be related to the standard deviation of J_P as follows: $\sigma_{J_D} = \sqrt{2(1-e^{-\beta})}\ \sigma_{J_P}$, which reduces to $\sqrt{2}\ \sigma_{J_P}$ in the case of white jitter. Because J_D is based on a difference of adjacent samples in a J_P sequence, its power spectrum is expected to be a high-pass filtered version of the power spectrum of J_P. Therefore, if J_P has a low-pass spectrum, the power spectrum of J_D is essentially flat.

[6]The sample rate of the perturbation sequence is assumed to correspond to $F0$.

Chapter 5

PSOLA pitch-marker positioning in synthetic versus natural speech signals[1]

5.1 Abstract

This Chapter presents the results of psychophysical experiments dealing with pitch-marker positioning within the Pitch Synchronous OverLap and Add (PSOLA) framework. Sustained natural vowels were PSOLA-modified in fundamental frequency. The experiments were aimed at determining the auditory sensitivity to (1) deterministic shifts of either all or single pitch markers within a sequence, and (2) random shifts of all pitch markers ("jitter"). As for deterministic shifts of all pitch markers, the results were in reasonable agreement with results obtained previously for synthetic formant signals. For deterministic shifts of single pitch markers, thresholds depended on position in the sequence. Detection thresholds for jittered shifts were comparable to thresholds for detecting jitter in pulse trains. The ranking of the thresholds for these three conditions indicated that the auditory system is more sensitive to dynamic (modulation) cues rather than to static (timbral) cues arising from shifts in pitch-marker positioning.

[1]This chapter is a slightly modified version of Kortekaas, R.W.L. and Kohlrausch, A. (1997) "Psychophysical evaluation of PSOLA: Natural versus synthetic speech," that will appear in Proceedings EUROSPEECH'97, Rhodos Greece, September 22-25.

5.2 Introduction

The Pitch-Synchronous OverLap and Add (PSOLA) technique is a well-known method for time-scale and pitch-scale modification of natural speech (Moulines and Laroche, 1995). The present Chapter focuses on the time-domain implemantation (TD-PSOLA) which will be referred to as PSOLA in the following. Even though PSOLA has found widespread application, little is known about the perceptual consequences of the PSOLA operations such as pitch-marker positioning.

PSOLA modification of an input speech signal is based on determining pitch markers, which indicate boundaries of local pitch periods in the case of voiced speech. Pitch-marker positioning is commonly assumed to be an important factor for synthesis quality. The present study was aimed at psychophysically determining the sensitivity of the human auditory system to shifts in pitch-marker position. The findings of these experiments provide information about the required accuracy of pitch-marker determination.

The results reported on here are an extension to those presented in Kortekaas and Kohlrausch (1997) where, among other things, the role of pitch-marker shifts in synthetic single-formant signals (Klatt, 1980) was studied . Pitch-marker shifts were defined as shifts relative to the pulses that excited the formant filter. All pitch markers were shifted equally. Detection thresholds for pitch-marker shifts were about 25 % of the fundamental period, for a fundamental frequency ($F0$) of 100 Hz. For an $F0$ of 250 Hz thresholds were (informally) measured to be approximately 10 %. Discrimination performance was also seen to increase monotonically with increasing pitch-marker shift. This performance could be described well by a psychoacoustic model based on excitation pattern differences (Gagné and Zurek, 1988).

In experiment 1 of the present Chapter, similar experiments were performed using natural sustained vowels (Houben, 1996). In contrast to the synthetic signals, such signals fluctuate (slightly) in $F0$, formant frequencies and level over time. The aim was to determine whether this non-stationarity also resulted in monotonicity and in comparable thresholds of discrimination. In experiment 2 discrimination thresholds for shifting single pitch makers were measured. In contrast to shifting all pitch markers, such a shift yields auditory cues that dynamically vary over the duration of the signal. In experiment 3 thresholds were measured for detecting random shifts ("jitter") imposed on the pitch-marker sequence. Such shifts can be conceived of as small errors in the (local) $F0$ estimate and also introduce dynamic cues.

5.3 General methods

5.3.1 Pitch markers

The pitch markers of the natural sustained vowels were determined by (1) estimating the local $F0$ and (2) determining the local energy maxima (Ma et al., 1994). For each signal this resulted in a sequence of pitch markers P_i^a $(i = 1, N)$, where superscript "a" indicates ""analysis" and N is the number of pitch markers in the sequence. In experiment 1 each pitch marker P_i^a was shifted over a relative amount ΔP given by:

$$\overline{P_i^a} = \begin{cases} P_i^a + \Delta P\,(P_{i+1}^a - P_i^a) & \text{if } \Delta P \geq 0 \\ P_i^a - \Delta P\,(P_i^a - P_{i-1}^a) & \text{otherwise} \end{cases}$$

ΔP thus expresses a fraction of the local fundamental period and will be presented as percentage in the following. In experiment 2 just a single pitch marker was shifted according to the formula given above.

In experiment 3 pitch markers were randomly shifted according to:

$$\overline{P_i^a} = P_i^a + \mathbf{J} \cdot T0$$

where $T0$ is the mean fundamental period. $\mathbf{J}$ is a spectrally white, Gaussian random variable with zero mean and standard deviation σ. This standard deviation will be presented as a percentage of $T0$.

5.3.2 Original signals

A (non-professional) male speaker uttered a sequence of isolated vowels /a/ and /i/ at semitone intervals over one octave. Recordings were made in a low-reverberant, quiet room using B&K microphones and a DAT recorder. The signals were stored on computer hard disk after low-pass filtering at 8 kHz (sample rate 48 kHz). One realization of both the vowel /a/ and /i/ from the sequence was chosen that fell within the normal register of the speaker. Taking the pitch-marker intervals as measures of the (instantaneous) $F0$, the vowel /a/ had an average of 161 Hz with a range 158 to 163 Hz. The vowel /i/ was slightly more stable: average $F0$ 166 Hz and a range of 165 to 167 Hz.

5.3.3 Experimental stimuli

Using PSOLA, two modified vowels /a/ were synthesized having an $F0$ of 127 and 195 Hz. These stimuli will be referred to as /a/-low and /a/-high, respectively. Likewise, the $F0$ of the vowel /i/ was modified to 129 and 196 Hz (/i/-low and /i/-high). These $F0$ shifts amount to approximately 3 semitones. The vowels were synthesized with a constant $F0$, i.e., the synthesis pitch-marker sequence P_i^s had constant intervals. Stimuli were

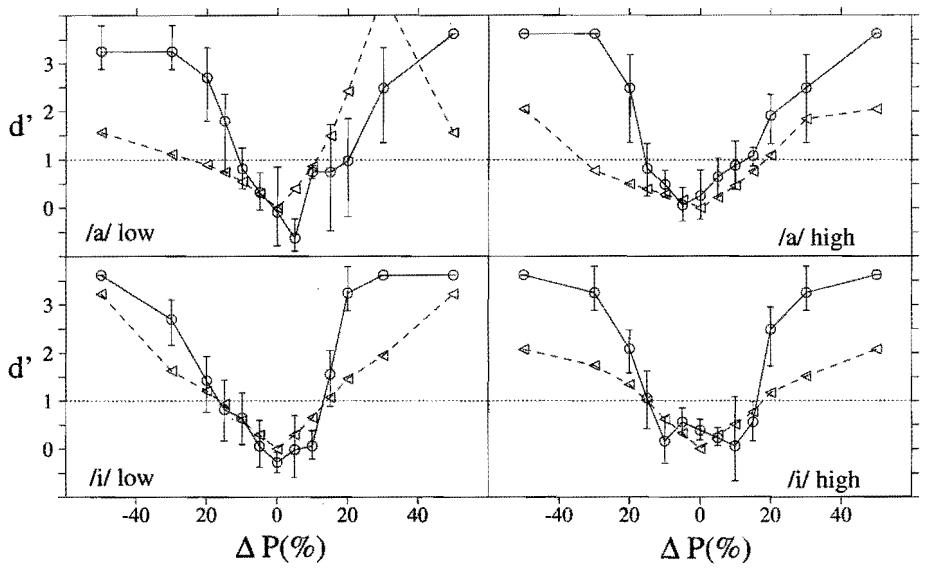


Figure 5.1: *Psychometric functions for subject MH (circles) showing discrimination sensitivity as a function of ΔP. Vertical bars indicate standard deviations. Model predictions are shown by the triangles.*

400 ms in duration where the first and last 25 ms were ramped using a raised-cosine window. The ramp duration in experiment 2 was 15 ms. The stimuli were presented to the subjects over Beyer DT990 headphones with a mean overall presentation level of 70 dB SPL. On each presentation the level was roved within ± 5 dB in order to rule out the use of possible loudness cues. Subjects were seated in a sound-proof booth and received immediate feedback after each trial.

5.3.4 Measurement procedures

In all experiments PSOLA-modified vowels using a shifted or jittered pitch-marker sequence ("test") had to be discriminated from vowels modified using the original sequence ("reference"). A 3I-3AFC paradigm was used in which one test and two reference stimuli were presented to the subject in random order. The subject's task was to indicate which interval contained the deviant "test" stimulus.

In experiment 1 psychometric functions were measured in which discriminability was determined as a function of ΔP. For each condition, all subjects performed at least one set of measurements, containing 15 trials, as a practice. The data presented below are the means (and standard deviations) over the final four sets of measurements. Instead of presenting percentage correct Pc, data will be presented in terms of the discrimination index d'. The discrimination threshold corresponds to $d' = 1$.

In experiment 2 and 3 an adaptive 3I-3AFC paradigm was used to measure thresholds for ΔP (single shifts) and σ, respectively. The level of ΔP or σ was decreased after two correct responses, and raised after one incorrect response ("two-up, one-down"). The minimal step size for ΔP and σ was 1.25 % and 0.1 %, respectively. Note that with a sampling rate of 48 kHz and a (mean) fundamental period of about 6 ms, shifting a pitch marker over one sample corresponds to $\Delta P \simeq 0.3$ %. All subjects performed at least one threshold measurement for each condition as a practice. The data presented below are the means and standard deviations of the last three measurements.

5.4 Uniform shifts

5.4.1 Results

Three subjects participated in this experiment. Subjects mostly reported timbral cues ("nasality") as their discrimination criterion. Because the data were sufficiently similar among subjects, only the psychometric functions for subject MH are shown in Figure 5.1. As for the synthetic single-formant signals (Kortekaas and Kohlrausch, 1997), the psychometric functions show monotonicity as a function of ΔP. The shape of the functions does not reveal a systematic difference between raising or lowering $F0$, or between /a/ en /i/.The detection thresholds (i.e., the value of ΔP for which $d' = 1$) amount to about 15 %. In Kortekaas and Kohlrausch (1997) detection thresholds were reported of approximately $\pm$ 25 % for an $F0$ of 100 Hz. Thresholds for higher $F0$ values were expected to be lower. Because the $F0$ of the natural vowels is about 160 Hz, this finding is thus in agreement with the previous results.

5.4.2 Modeling

The psychometric functions for reported in Kortekaas and Kohlrausch (1997) could be described well by using an intensity-discrimination model Gagné and Zurek (1988). Such a model calculates the difference between the excitation patterns of the "test" and "reference" stimulus. The excitation pattern is derived by analyzing the stimulus by means of an (auditory) filterbank and calculating the power within each channel. In this way the model outcome only depends on the power spectrum of the stimulus.

Figure 5.1 also shows the psychometric function predicted by a multiband model in which all channels of the filterbank are taken into account. This model was gauged to the previously reported psychometric functions for synthetic stimuli. Except for negative shifts in the case of raising the vowel /a/, all thresholds are predicted rather accurately. With just a few exceptions, however, the predicted psychometric functions for values above the threshold lie substantially below the measured d' values.

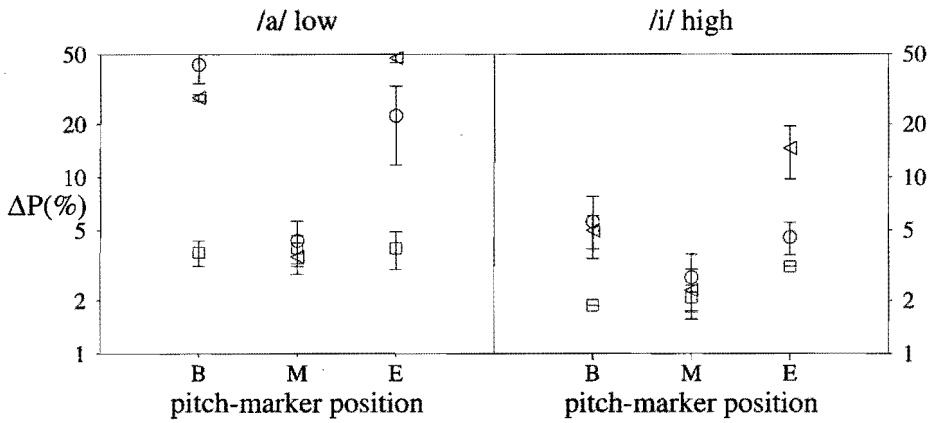


Figure 5.2: ΔP *thresholds for shifting single pitch markers as a function of pitch-marker position (see text for details). Data for subject HH are shown by circles, for PK by triangles, and RK by squares. Vertical bars indicate standard deviations.*

5.5 Single shifts

Threshold measurements were performed for single shifts of the fourth, $(N/2)^{th}$, and $(N-4)^{th}$ pitch marker (denoted by "B", "M", and "E", respectively). These conditions were investigated for the /a/-low and /i/-high condition.

Three subjects participated in the experiment and the results are shown in Figure 5.2. The cue they reported mostly was a (rough) discontinuity in the "test" stimulus. The thresholds for shifting the middle pitch marker ("M") are comparable for the three subjects. These thresholds are about 2 to 5 % which is a factor 3 lower than for uniform shifts. This suggests that the auditory system is more sensitive to the dynamic changes introduced by single shifts than to the (almost) static cues introduced by uniform shifts. The thresholds for the /a/-low condition seem to be higher than for the /i/-high condition.

Thresholds for the "B" and "E" conditions are generally higher than for "M". The variability among subjects is, however, rather high, especially for the /a/-low condition. This finding is in agreement with data presented in Cardozo and Ritsma (1968). In that study pulse trains were used as stimuli and thresholds were measured for (random) changes in the inter-pulse intervals. In the case of shifting single pulses, higher thresholds were reported when shifting leading and trailing pulses rather than central pulses (see figure 5 in Cardozo and Ritsma (1968)).

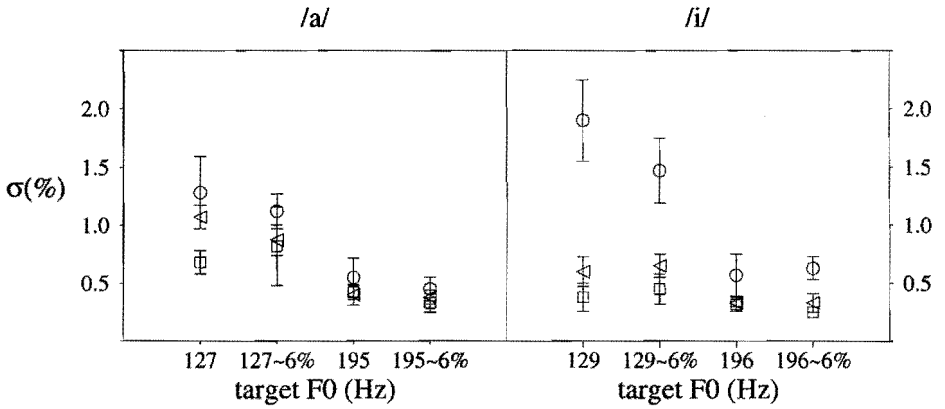


Figure 5.3: *Detection thresholds for σ as a function of target F0. Roving of target F0 is indicated as in "127~6 %" (see text for details). Symbols are used as in Figure 5.2.*

5.6 Jittered pm sequences

The participating subjects were the same as in experiment 2. They reported roughness (for moderate and large σ) and unsteadiness (near threshold) as discrimination cues. The σ thresholds shown in Figure 5.3 for jitter RMS are about 0.5 to 1 %. The lowest thresholds are observed for the /a/-high and /i/-high conditions. These two findings are in agreement with the data on jitter detection of (filtered) pulse trains presented in Cardozo and Ritsma (1968). The ΔP thresholds for single "M" shifts are about 4 times σ at threshold. Therefore, discrimination at near-threshold levels may have been based on detecting single shifts.

Figure 5.3 additionally shows thresholds measured with roving of the synthesized ("target") $F0$ over $\pm$ one semitone (6 %). The resulting thresholds do not differ markedly from the constant-$F0$ condition which indicates that subjects did not base their discrimination on differences in pitch.

5.7 Conclusions

The discrimination performance for uniform pitch-marker shifts in sustained natural vowels is qualitatively similar to the performance for synthetic formant signals reported previously. The performance could at least be partly explained by model predictions. The auditory sensitivity for single shifts is found to be position dependent. Central shifts are most easily detectable and thresholds are about three times lower than for uniform shifts. The thresholds for random (jittered) shifts are lowest. Jittered shifts can provoke clear sensations of roughness even at rather low levels of the jitter RMS.

Chapter 6

Some unresolved issues

In this section a general discussion of the two studies described in Part I and II of this dissertation will be presented. The limitations of the present studies will be described, and some directions for further research will be proposed. For a discussion of the main findings of this dissertation, the reader is refered to the Summary.

6.1 Vowel-onset detection

Part I of this dissertation was concerned with automatic vowel-onset detection in natural speech. The study presented in Chapter 2 will be discussed in terms of (1) the robustness of vowel-onset detection, (2) the focus on firing rate information in the simulated neural responses, and (3) the method of perceptually measuring actual vowel onsets.

6.1.1 Robustness of vowel-onset detection

As was mentioned in Chapter 2, the speech signals used to train and evaluate the automatic vowel-onset detection schemes were characterized by high Signal-to-Noise Ratios (SNRs). As was argued in the General Introduction, vowel-onset detection may be important for the human speech-perception mechanism to operate robustly. This implies that detection should not be affected too much by concurrent speech signals or by background noises. Therefore, future research may focus on vowel-onset detection performance in the presence of intervening background noise, for instance in terms of temporal accuracy of detection. Preferably, such noises should have spectral distributions similar to the (long-term) spectral distribution of natural speech. Because multiple sound sources are presented simultaneously in such an experiment, the listener has to segregate the vowel onset to be determined from the interfering noises. In a review of perceptual mechanisms thought to influence grouping of auditory information, Darwin and Carlyon (1995) report that, among other things, $F0$ differences and onset asynchronies can serve

to segregate concurrent vowels. In the case of $F0$ differences, Assmann and Summerfield (1990) showed that identification scores increase markedly if the $F0$ difference between concurrent vowels increases from 0 to 0.5 semitones. The models described in Chapter 2 were found to benefit from harmonicity information, i.e., a rough voiced/unvoiced decision, but it is expected that pitch information should be exploited to a greater extent for segregation of concurrent sound sources. In their present state, the models are likely to fail for vowel-onset detection in the presence of competing sounds. In the case of segrągation based on onset asynchronies, Darwin (1981) found that staggering the first three formants of synthetic vowels up to 100 ms does not affect identification of these vowels, but does influence the number of "voices" heard. Almost all models presented in Chapter 2 assume simultaneous onsets within separate frequency regions which, in order to be grouped into one voice, should thus occur within at most 100 ms. The temporal accuracy of human vowel-onset detection of 60 ms reported in Hermes (1990) does not contradict this conclusion. Moreover, the results reported by Darwin (1981) might imply that human vowel-onset detection does not rely on the speech utterances being recognized, provided it also plays a role in grouping processes; in the experiments of Darwin (1981) several "voices" are heard, so that segregation takes place, even though the vowels are identified correctly. In this sense, vowel-onset detection might be a *primitive grouping mechanism* (Bregman, 1990).

6.1.2 Firing rate information

The vowel-onset detection schemes based on simulation of auditory processing in the cochlear nucleus focussed on firing rate information. Using this information, it was argued that firing patterns observed in the auditory nerve presumably were less appropriate for vowel-onset detection; due to the limited dynamic range of about 20-30 dB for individual fibers, the spectral profiles in terms of rate representation are restricted to a certain dynamic range due to saturation. As was briefly mentioned in Chapter 2, phase-locked activity in fibers also may represent characteristic features of the speech spectrum. The so-called *average localized synchrony measures* (ALSM) computed on the basis of physiologically measured period histograms show speech-spectrum features, such as formants, for a dynamic range up to 60-90 dB (Delgutte and Kiang, 1984a,b,c; Carney and Geisler, 1986). Fiber responses have been observed to "tune in" to formant frequencies (e.g., Sinex and Geisler, 1983). Physiological and behavioral studies, however, have shown that phase locking accuracy deteriorates above 1 kHz, but this does not necessarily mean that formants below about 2.5 kHz cannot be detected in phase-locked responses. Future research may therefore investigate whether vowel onsets can be detected in (simulated) auditory nerve responses as sudden increments in synchronized activity, simultaneously occurring in different frequency regions in the F1-F3 range. The calculation of the ALSM measure, however, is expected to be technically

more difficult than the detection of changes in firing rate.

6.1.3 Determining actual onsets

Some issues concerning the method of perceptually determining actual onsets will be discussed here. The reference vowel onsets ("actual onsets" in Chapter 2) were measured by gating a speech signal. Performance consistency in detecting actual onsets was informally verified for the trained phonetician, but little is known about consistency across different listeners. Informally, we did notice detection differences in some phonetic contexts for listeners with different native tongue. The question is whether this finding contradicts the assumption that human vowel-onset detection is a primitive ("hard wired") mechanism. In addition to the possible differences between listeners, one may question whether the gating technique itself biased the results. First, this technique is expected to introduce spectral splatter by windowing the speech signal over a short duration. This splatter reduces the spectral resolution of the "internal representation" of the gated stimulus, which would be in agreement with our finding that only global spectral content is necessary for automatic vowel-onset detection. Because windows of 20-40 ms are used, the spectral smearing is expected to be limited to about 25-50 Hz which is only a small effect. Second, and presumably more important, the whole sentence could be listened to in determining the actual onsets and the speech signal could be visually inspected by the phonetician. One could argue that this additional information somewhat confounded the aural detection of vowel onsets; in perceiving spoken language, such information is usually not available. Actual-onset detection was also performed at a comfortable loudness level whereas the models based on CN responses showed a dependency on overall sound-pressure-level of presentation. It may therefore be useful to test for such dependencies in human performance as well.

We propose an experiment in which some of these issues might be investigated. Imagine a listening experiment in which a subject is presented with two intervals (2I-2AFC), and has to respond which interval contains a vowel. The intervals contain gated portions from syllables, where the window function used for gating is sufficiently smooth (for instance, a plateau of a certain duration and raised-cosine ramps). Reference intervals are taken from the consonant portion of the syllable. Using signal-detection theory (Green and Swets, 1966), the vowel onset can be derived from the psychometric functions that are obtained from the 2I-2AFC experiment. Such an experimental paradigm enables to investigate the required integration time (by varying the gating-window length), the role of phonetic context (by using different syllables, for instance in random order), and of presentation level (by varying the sound pressure level). If the prerequisite of naturalness of the stimuli is somewhat loosened, original syllables may also be resynthesized by using, for instance, a Harmonic/Stochastic model (McAulay and Quartieri, 1986; Dutoit and

Leich, 1993, see below). Using such a model enables to study in detail the perceptual role of physical characteristics such as intensity, spectral content, and harmonicity.

6.2 Psychoacoustical evaluation of PSOLA

The aim of the study presented in part II of this dissertation was to try to establish the psychoacoustical basis for the apparent success of certain speech-waveform manipulations. In the following, the study presented in Chapters 3 to 5 will be discussed in the context of (1) signal-processing aspects of the psychophysical experiments, (2) the development of PSOLA variants, (3) acoustical listening conditions, (4) influence of the experimental paradigms, and (5) important properties of natural speech.

6.2.1 Signal-processing aspects

The strategy to psychophysically determine the auditory sensitivity to PSOLA-induced spectro-temporal changes was to gradually increase the complexity of the stimuli under investigation, while trying to establish the relation between physical characteristics of the stimuli and the way they are perceived. Except for the perturbated stimuli in Chapter 4 and the natural vowels in Chapter 5, the signals under modification always had a constant fundamental frequency ($F0$) and an invariant spectral content. In addition, the synthesis $F0$ was constant so that the spectra of an original and modified signal had a clear-cut relation, as was described and illustrated in Chapter 3. From a signal-processing point of view, however, this relation may be regarded as one of the "trivial cases" of the PSOLA technique due to the fact that its probability of occurrence, when modifying natural speech, is very low; speech has a highly dynamic character.

Dynamic behavior manifests itself for instance in $F0$ contours. In terms of PSOLA operations, a variable $F0$ implies decomposition windows of variable length and asymmetrical shape. Suppose that a signal with a slowly decreasing $F0$, but constant spectral envelope, is "monotonized" by PSOLA modification. The spectrum of each PSOLA window will not be as smooth and symmetrical as shown in Chapter 3, but will nevertheless have a strong main lobe and weaker side lobes. Therefore, the effect of this $F0$ shift on either the excitation pattern or the temporal pattern in auditory filters is expected to be basically similar to the static $F0$ case. Note that the interaction between the decreasing $F0$ and the formant resonances may result in differences in "gain" between decomposed segments. This in its turn will be reflected as low-frequency modulations in the envelopes of auditory filters in response to the "monotonized" stimuli. In the case of fast variations of $F0$, the complexity of such signals probably neccesitates the use of psychoacoustical models for characterizing the perceptual consequences of PSOLA application (see section "Further aspects of natural speech" below).

6.2.2 PSOLA variants

The study presented here was based on the original time-domain implementation proposed in Charpentier and Moulines (1989), except for scaling factors for normalizing the (maximum) amplitude or energy within each decomposed segment. Dutoit (1997) provides a recent review of practically all new variants of PSOLA which have been proposed over the last few years. These variants include hybrid models combining Linear Predictive Coding (LPC) and Time Domain (TD) PSOLA. In the latter variant, TD-PSOLA can be applied on the speech waveform after resynthesis by means of the LPC filter, or on the LP residual obtained after inverse filtering.

Dutoit and Leich (1993) described a system for resynthesizing Text-to-Speech (TTS) database segments, such as diphones, prior to the actual speech synthesis using PSOLA. The resynthesis is based on decomposing speech frames into harmonic and stochastic components (a so-called H/S model). The advantage of this decomposition is that the frames can be resynthesized with *known* harmonic phases and amplitudes, and subsequently recombined by means of OLA to be stored as TTS segment. In this way, the positioning of the PSOLA windows in synthesis is facilitated because each TTS segment is resynthesized with known phase relations. For instance, the TTS segments can be resynthesized with cosine phases yielding a signal with high peak factor. In addition, the technique enables spectral interpolation, or normalization, between TTS segments.

For the role of (harmonic) component phases in resynthesis of the TTS segments, the results of Chapter 5 showed that low jitter levels of pitch-marker positioning can already lead to a roughness percept. This finding supports the re-synthesis strategy of Dutoit and Leich (1993). Nevertheless, the results of that chapter do not seem to support the necessity of positioning pitch markers *exactly* at, e.g., the instant of glottal closure as suggested by Charpentier and Moulines (1989). Instead, constancy in pitch-marker intervals seems to be more important in perceptual terms, provided pitch markers are positioned in the "vicinity" of the energy maximum within a period. This may partly explain the success of Pitch-Inflected OverLap-and-Add (PIOLA; Meyer et al., 1993), in which pitch markers are positioned automatically on the basis of a local $F0$ estimate. No manual correction to the pitch-marker positions is performed afterwards. The results of Chapter 3 also showed, however, that "nasality cues" arise if pitch markers and glottal excitations deviate by more than 10-20 % of the local pitch period, so that the use of more than just the local $F0$ information seems to be necessary.

The resynthesis system of Dutoit and Leich (1993) also offers the possibility to equalize initial phases of all PSOLA segments prior to resynthesis. As the results of the double-vowel experiment in Chapter 4 suggest, phase differences between PSOLA segments are perceptually less important when modifying natural speech; the corresponding temporal

cues for the high formants are generally dominated by the spectral cues of the stronger low formants. Note that the results of Chapter 4 deal with *static* phase differences: the auditory sensitivity to shifting single segments was found to be high in Chapter 5, but also to be dependent on position in the stimuli. Shifts in the middle of a stimulus were most easily detected. Therefore, phase mismatch at the boundary between TTS segments, and in the rhyme of a syllable, may have a greater impact on perceived continuity than static or slowly varying phase shifts.

6.2.3 Acoustical listening conditions

The experimental data presented in Chapters 3-5 were obtained in acoustically "optimal" conditions; subjects were seated in soundproof booths and received acoustical stimuli over high-quality headphone sets. In this way, background noises, other than noises produced by the subject, were attenuated by at least 20-30 dB. Because the amplitude and phase transfer function of the headphones can be considered as flat in the restricted spectral region of interest for the present study, about $\pm 2 \cdot F0$ around the formant frequency, the spectra of the synthetic signals will be presented undistorted at opening of the ear canal. In practical applications of PSOLA outside the laboratory, however, the signal spectra may be disturbed by bandlimited and non-linearly distorting transmission channels, for instance telephone lines or coding algorithms. In this case, the perceptual tolerance to PSOLA-induced distortions may increase (see also the discussion of Chapter 2). In addition, room acoustics may play a role: in a reverberant sound field, component phases are known to be randomized. Also, in such sound fields in large rooms, the sound-pressure levels are known to vary considerably due to vectorial addition of sound waves with random phases (for a discussion of room acoustics and speech perception see, e.g., Plomp, 1984). Schroeder (1954) found that the theoretical uncertainty of the sound-pressure level of a pure tone at a large distance from a loudspeaker amounts to about 6 dB, which is comparable to the PSOLA-induced component-level changes. These two aspects of room acoustics may very well increase the tolerance to PSOLA-induced spectro-temporal changes when applied in "normal" acoustical environments, and therefore contribute to explaining why PSOLA application often yields satisfactory quality.

6.2.4 Experimental paradigms

In Chapters 3-5, a 3I-3AFC paradigm was used throughout the experiments. Such a paradigm provides a listener with two references which facilitates discrimination. Additionally, such a paradigm does not force a listener to cognitively label the stimuli as "PSOLA-modified" or "unmodified". Performance was often observed to be close to perfect discrimination. If instead a one-interval "yes/no" experiment had been performed,

where the subject has to indicate whether he or she perceived a PSOLA-modified or unmodified stimulus, performance scores for differentiating between PSOLA-modified and unmodified stimuli might have been worse, even if these scores were corrected for differences in experimental paradigm; as was noted in Chapter 3, the temporal (phase) cues were often rather subtle. In modifying natural speech, where an $F0$-shifted reference signal does not exist, such cues can be expected to play a minor role. On the other hand, roughness cues reported in Chapter 5 for pitch-marker jitter, or nasality cues reported in Chapter 3 for constant pitch-marker shifts, might be adequately used in such a "yes/no" task; these cues potentially are annoying artefacts in modifying natural speech.

6.2.5 Further aspects of natural speech

As mentioned above, the present study has not taken into account a number of important dynamic features of speech such as $F0$ contours, formant trajectories, and fast transitions from, e.g., plosives to vowels. A systematic psychoacoustical study of the perceptual impact of such features under PSOLA modification might be limited to analyzing such signals using (advanced) psychoacoustical models; psychoacoustical literature data for signals with comparable complexity are sparse if not totally absent. Nevertheless, such signals may serve as a means to test the applicability of these models. For instance, the audibility of a spectral-envelope mismatch between TTS segments (see above) might be tested by examining the modulation patterns in auditory filters; if the spectral mismatch is large, large modulations are expected to show up in the filter responses. By relating the depths of these modulations to psychoacoustical data of modulation discrimination, as was done in Chapter 3, predictions of audibility of this spectral mismatch might be derived. Similarly, jitter perturbation thresholds, and to a less extent shimmer thresholds, could be accurately modeled by comparing effective modulation depths (see Chapter 4). By comparing the effective modulation depths of an original and a PSOLA-modified signal, a prediction of "perceived roughness" might be derived. To this end, the differences in effective modulation depth might have to be compared across channels (Terhardt, 1974). Psychoacoustical models may also be used to analyze PSOLA-modified signals with (slowly) varying $F0$ as described in section 6.2.1. In this way, "thresholds" for rates of $F0$ change might be derived below which PSOLA-modified stimuli with varying $F0$ are expected to yield the same auditory discrimination cues as stimuli with constant $F0$.

PSOLA modification of duration of unvoiced speech has been a problem often dealt with by ad-hoc methods. By repeating decomposed unvoiced segments a (locally) periodic signal is constructed which elicits a (faint) pitch percept. An ad-hoc method to reduce this effect is to time-invert every odd repeated PSOLA segment. Phase randomization of each PSOLA segments may be another, although more expensive method. Background

noises which are uncorrelated with the speech signal may in turn also give rise to such pitch percepts. By repeating a segment, the spectrum of the modified (background) noise locally becomes harmonic with harmonics coinciding with the harmonics of the modified target speech signal. On a larger time scale, however, the levels of the harmonics of the "repeated noise" fluctuate randomly. It may therefore be interesting to investigate what the "critical Signal-to-Noise ratio" is at which the speech signal and the background noises no longer are fused after (drastically) increasing the duration. Such an experiment may be a direction for further psychophysical research on PSOLA.

Bibliography

Ainsworth, W. A., and Meyer, G. F. (**1994**). "Recognition of plosive syllables in noise: Comparison of an auditory model with human performance," J. Acoust. Soc. Am. **96**, 687–694.

Allen, G. A. (**1972**). "The location of rhythmic stress beats in english: An experimental study 1," Language Speech **15**, 72–100.

Allen, J. B., and Rabiner, L. R. (**1977**). "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE **65**, 1558–1564.

Arle, J. E., and Kim, D. O. (**1991**). "Neural modeling of intrinsic spike–discharge properties of cochlear nucleus neurons," Biol. Cybernetics **64**, 273–283.

Askenfelt, A., and Hammarberg, B. (**1986**). "Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures," J. Speech Hear. Res. **29**, 50–64.

Assmann, P. F., and Summerfield, A. Q. (**1990**). "Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Beerends, J. G., and Stemerdink, J. A. (**1994**). "A perceptual speech quality measure based on a psychoacoustic sound perception," J. Aud. Eng. Soc. **42**, 115–123.

Blackburn, C. C., and Sachs, M. B. (**1990**). "The representation of the steady-state vowel sound /ɛ/ in the discharge patterns of cat anteroventral cochlear nucleus neurons," J. Neurophys. **63**, 1191–1212.

Boëffard, O., and Violaro, F. (**1994**). "Improving the robustness of text–to–speech synthesizers for large prosodic variations," in *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis (New–York, 12–15 sept.)*, 111–118.

Bregman, A. S. (**1990**). *Auditory scene analysis: The perceptual organisation of sound*, Cambridge, MA: Bradford Books, MIT Press.

Brown, A. G. (**1991**). *Nerve Cells and Nervous Systems*, Springer-Verlag, London.

Cardozo, B. L., and Ritsma, R. J. (**1968**). "On the perception of imperfect periodicity," IEEE Transactions on Audio and Electroacoustics **16**, 159–164.

Carney, L. H., and Geisler, C. D. (**1986**). "A temporal analysis of auditory nerve fiber responses to spoken stop consonant–vowel syllables," J. Acoust. Soc. Am. **79**, 1896–1914.

Charpentier, F., and Moulines, E. (**1989**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *EUROSPEECH'89, Paris*, volume 2, 13–19.

Cole, R. A., and Scott, B. (**1973**). "Perception of temporal order in speech. the role of vowel transitions," Can. J. Psychol. **27**, 441–449.

Compernolle, D. S. J. v. (**1991**). "Development of a computational auditory model," IPO report 784, Institute for Perception Research Eindhoven.

Cox, N. B., Ito, M. R., and Morrison, M. D. (**1989**). "Technical considerations in computation of spectral harmonics-to-noise ratios for sustained vowels," J. Speech and Hearing Res. **32**, 203–218.

Cranen, B. (**1997**). *personal communication*.

Darling, A. M. (**1991**). "Properties and implementation of the gammatone filter: a tutorial," in *Speech, Hearing and Language. Work in Progress*, volume 5, 43–61, University of London, Dept. Phonetics and Linguistics.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset time," Quart. J. Exp. Psych. **33A**, 185–208.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound: constraints on for mant perception," J. Acoust. Soc. Am. **76**, 1636–1647.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in: *Hearing*, edited by B.C.J. Moore, Academic Press, 387–424.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration," J. Acoust. Soc. Am. **xx**, in press.

de Boer, E. (**1969**). "Reverse correlation ii: initiation of nerve inpulses in the inner ear," Proc. Kon. Ned. Acad. Wet. **72**, 129–151.

Delgutte, B., and Kiang, N. Y. S. (**1984a**). "Speech coding in the auditory nerve. I: Vowel-like sounds," J. Acoust. Soc. Am. **75**, 866–878.

Delgutte, B., and Kiang, N. Y. S. (**1984**b). "Speech coding in the auditory nerve. II: Processing schemes for vowel-like sounds," J. Acoust. Soc. Am. **75**, 879–886.

Delgutte, B., and Kiang, N. Y. S. (**1984**c). "Speech coding in the auditory nerve. III: Voiceless fricative consonants," J. Acoust. Soc. Am. **75**, 887–896.

Durlach, N. I., Braida, L. D., and Ito, Y. (**1986**). "Towards a model for discrimination of broadband sounds," J. Acoust. Soc. Am. **80**, 63–72.

Dutoit, T. (**1997**). *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht.

Dutoit, T., and Leich, H. (**1993**). "MBR-PSOLA: Text-to-Speech synthesis based on an MBE re-synthesis of the segments database," Speech Comm. **13**, 435–440.

Eddins, D. A. (**1993**). "Amplitude modulation detection of narrow-band noise: Effects of absolute bandwidth and frequency region," J. Acoust. Soc. Am. **93**, 470–479.

Eggermont, J. J. (**1985**). "Peripheral auditory adaptation and fatigue: a model oriented review," Hear. Res. **18**, 57–71.

Eriksson, A. (**1991**). *Aspects of Swedish speech rhythm*, Ph.D. thesis, PhD. thesis, University of Göteborg, Sweden.

Evans, E. F., and Palmer, A. R. (**1980**). "Relationship between the dynamic range of cochlear nerve fibres and their spontaneous activity," Exp. Brain Res. **40**, 115–118.

Farrar, C. L., Reed, C. M., Ito, Y., Durlach, N. I., Delhorne, L. A., Zurek, P. M., and Braida, L. D. (**1987**). "Spectral-shape discrimination. I. Results from normal hearing listeners for stationary broadband noises," J. Acoust. Soc. Am. **81**, 1085–1092.

Fay, R. R. (**1988**). *Hearing in Vertebrates: a Psychophysics Data Book*, Hill-Fay Associates, Winnetka.

Flanagan, J. L., and Saslow, M. G. (**1958**). "Pitch-discrimination for synthetic vowels," J. Acoust. Soc. Am. **30**, 435–442.

Florentine, M., and Buus, S. (**1981**). "An excitation-pattern model for intensity discrimination," J. Acoust. Soc. Am. **70**, 1646–1654.

Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am. **80**, 1016–1025.

Gagné, J., and Zurek, P. M. (**1988**). "Resonance-frequency discrimination," J. Acoust. Soc. Am. **83**, 2293–2299.

Goldman, S. (**1948**). *Frequency analyis, modulation and noise*, McGraw-Hill Book Company, INC., New York, Toronto, London.

Green, D., and Swets, J. (**1966**). *Signal detection theory and psychophysics*, John Wiley & Sons, New York, London, Sydney.

Hamon, C., Moulines, E., and Charpentier, F. (**1989**). "A diphone synthesis system based on time–domain prosodic modifications of speech," in *Proc. IEEE Int. Conf. Acoust. Speech and Sig. Proc.*, volume 3, 238–241, Glasgow.

Hansen, M., and Kollmeier, B. (**1996**). "Implementation of a psychoacoustical preprocessing model for sound quality measurement," in *ESCA Tutorial and Workshop on The Auditory Basis of Speech Perception*, 79–82, Keele.

Hartmann, W. M., and Hnath, G. M. (**1982**). "Detection of mixed modulation," Acustica **50**, 297–312.

Henn, C. C., and Turner, C. W. (**1990**). "Pure-tone increment detection in harmonic and inharmonic backgrounds," J. Acoust. Soc. Am. **88**, 126–131.

Hermes, D. J. (**1990**). "Vowel-onset detection," J. Acoust. Soc. Am. **87**, 866–873.

Hermes, D. J., Beaugendre, F., and House, D. (**1997**). "Temporal alignment of accentuation boundaries in Dutch," in *Proc. ESCA Workshop on Intonation, 18–20 sept.*, accepted for publication, Rhodes, Greece.

Hewitt, M. J., and Meddis, R. (**1991**). "An evaluation of eight computer models of mammalian inner hair-cell function," J. Acoust. Soc. Am. **90**, 904–917.

Hillenbrand, J. (**1987**). "A methodological study of perturbation and additive noise in synthetically generated voice signals," J. Speech Hear. Res. **30**, 448–461.

Hillenbrand, J. (**1988**). "Perception of aperiodicities in synthetically generated voices," J. Acoust. Soc. Am. **83**, 2361–2371.

Horii, Y. (**1980**). "Vocal shimmer in sustained phonation," J. Speech Hear. Res. **23**, 202–209.

Houben, M. (**1996**). "Psycho-acoustical evaluation of pitch-marker positioning in natural speech," internal IPO report **1132**.

House, D. (**1990**). *Tonal Perception in Speech*, Lund University Press, Lund.

Houtgast, T. (**1989**). "Frequency selectivity in amplitude-modulation detection," J. Acoust. Soc. Am. **85**, 1676–1686.

Hunt, A. (**1993**). "Recurrent neural networks for syllabification," Speech Commun. **13**, 323–332.

Kaufholz, P. A. P. (**1992**). "Improvement of the vowel-onset-detection algorithm in the ipo intonation meter," IPO report 870, Institute for Perception Research Eindhoven.

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.

Kewley-Port, D., and Watson, C. S. (**1994**). "Formant-frequency discrimination for isolated english vowels," J. Acoust. Soc. Am. **95**, 485–496.

Klatt, D. H. (**1973**). "Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception," J. Acoust. Soc. Am. **53**, 8–16.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D. H. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: a first step," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc.*, volume 2, 1278–1281, Paris.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Klaus, H., Fellbaum, K., and Sotscheck, J. (**1997**). "Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache," Acustica united with Acta Acustica **83**, 124–136.

Klingholz, F., and Martin, F. (**1985**). "Quantitive spectral evaluation of shimmer and jitter," J. Speech Hear. Res. **28**, 169–174.

Kortekaas, R. W. L., and Kohlrausch, A. (**1997**). "Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli," J. Acoust. Soc. Am. **101**, 2202–2213.

Kortekaas, R. W. L., and Meyer, G. F. (**1994**). "Vowel-onset detection using models of the auditory periphery and the nucleus cochlearis: physiological background," IPO report 963, Institute for Perception Research Eindhoven.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Liberman, M. C. (**1978**). "Auditory nerve response from cats raised in a low noise chamber," J. Acoust. Soc. Am. **63**, 442–455.

Lieberman, P. (**1961**). "Perturbation in vocal pitch," J. Acoust. Soc. Am. **33 (5)**, 597–603.

Lyzenga, J. (**1997**). *Discrimination of simplified vowel spectra*, Ph.D. thesis, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Lyzenga, J., and Horst, J. W. (**1995**). "Frequency discrimination of bandlimited harmonic complexes related to vowel formants," J. Acoust. Soc. Am. **98**, 1943–1955.

Lyzenga, J., and Horst, J. W. (**1997**). "Frequency discrimination of stylized synthetic vowels with a single formant," accepted for J. Acoust. Soc. Am. .

Ma, C., Kamp, Y., and Willems, L. F. (**1994**). "A Frobenius norm approach to glottal closure detection from the speech signal," IEEE Trans. Speech Audio Proc. **2**, 258–265.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection theory: a users guide*, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney.

Marcus, S. (**1981**). "Acoustic determinants of perceptual centre (P-center) location," Percept. Psychophys. **30**, 247–256.

Markowitz, J. (**1993**). "Listening with intelligence," AI Expert **8**, 38–45.

McAulay, R. J., and Quartieri, T. F. (**1986**). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust. Speech Signal Process. **34**, 744–754.

McCulloch, N., and Ainsworth, W. A. (**1988**). "Speaker independent vowel recognition using multi-layer perceptions," in *Proceedings of the 7th FASE Symposium*, volume 8, 851–858, Edinburgh.

McGee, V. E. (**1964**). "Semantic components of the quality of processed speech," J. Speech Hear. Res. **7**, 310–323.

Meddis, R. (**1986**). "Simulation of mechanical to neural transduction in the auditory receptor," J. Acoust. Soc. Am. **79**, 702–711.

Meddis, R. (**1988**). "Simulation of auditory-neural transduction: further studies," J. Acoust. Soc. Am. **83**, 1056–1063.

Meddis, R., and Hewitt, M. J. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. **89**, 2866–2882.

Meyer, G. F. (**1993**a). "CNET – point neurone simulator," Technical report TR93-01, Dept. Computer Science, University of Keele .

Meyer, G. F. (**1993**b). *Models of neurones in the ventral cochlear nucleus: signal processing and speech recognition*, Ph.D. thesis, (unpublished), Dept. Communication and Neuroscience, University of Keele, Keele, Great–Britain.

Meyer, P., Rühl, H. W., Krüger, R., Kugler, M., Vogten, L. L. M., Dirksen, A., and Belhoula, K. (**1993**). "PHRITTS – a Text–to–Speech synthesizer for the german language," in *EUROSPEECH93*, 877–890.

Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

Moore, B. C. J., and Glasberg, B. R. (**1987**). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Hearing Research **28**, 209–225.

Moore, B. C. J., and Sek, A. (**1992**). "Detection of combined frequency and amplitude modulation," J. Acoust. Soc. Am. **92**, 3119–3131.

Morton, J., Marcus, S. M., and Frankish, C. R. (**1976**). "Perceptual centers (P-centers)," Psychol. Rev. **83**, 405–408.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.

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

Neelen, J. J. M. (**1969**). "Audibility of jitter in pulse trains as affected by filtering - II," in *IPO annual progress report*, volume 4, 23–29.

Nossair, Z. B., and Zahorian, S. A. (**1991**). "Dynamical spectral features as acoustic correlates for the initial stop consonant," J. Acoust. Soc. Am. **89**, 2978–2991.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). "An efficient auditory filterbank based on the gammatone function," in *Appendix B of SVOS Final Report: The auditory Filterbank*, volume APU report 2341.

Pickles, J. O. (**1982**). *An Introduction to the physiology of hearing*, Academic, London.

Pike, K. L. (**1947**). *Phonemics*, University of Michigan Press, Ann Arbor.

Pinto, N. B., and Titze, I. R. (**1990**). "Unification of perturbation measures in speech signals," J. Acoust. Soc. Am. **87**, 1278–1289.

Pisoni, D. B. (**1980**). "Variability of vowel formant frequencies and the quantal theory of speech: a first report," Phonetica **37**, 285–305.

Plomp, R. (**1984**). "Perception of speech as a modulated signal," in *Proc. 10th Int. Congress of Phonetic Sciences*, volume edited by M.P.R. van de Broecke and A. Cohen, 29–40, Dordrecht, Floris.

Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.

Pollack, I. (**1971**a). "Amplitude and time jitter thresholds for rectangular–wave trains," J. Acoust. Soc. Am. **50**, 1133–1142.

Pollack, I. (**1971**b). "Spectral basis of auditory "jitter" detection," J. Acoust. Soc. Am. **50**, 556–558.

Pompino-Marshall, B. (**1989**). "On the psychoacoustic nature of the P–center phenomenon," J. Phon. **17**, 175–192.

Pompino-Marshall, B. (**1990**). *Die Silbenprosodie. Ein elementarer Aspekt der Wahrnehmung vor Sprachrhytmus und Sprechtempo*, Linguistische Arbeiten 247, Max Niemeyer Verlag, Tuebingen.

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital processing of speech signals*, Prentice-Hall, Englewood Cliffs, New Jersey.

Rapp, K. (**1971**). "A study of syllable timing," in *Quarterly Progress and Status Report, Speech Transmission Laboratory*, volume STL-QPRS 1/1971, 14–19, Sweden, Stockholm.

Rhode, W. S., and Greenberg, S. (**1992**). "Physiology of the cochlear nuclei," in *The Mammalian Auditory Pathway: Neurophysiology*, volume edited by A.N. Popper and R.R. Fay, New–York, Springer–Verlag.

Rhode, W. S., and Smith, D. H. (**1986**). "Encoding timing and intensity in the VCN of cat," J. Neurophys. **56**, 287–307.

Richards, V. M., Onsan, Z. A., and Green, D. M. (**1989**). "Auditory profile analysis: potential pitch cues," HR **39**, 27–36.

Roads, C. (**1988**). "Introduction to granular synthesis," Comp. Music Journal **12 (2)**, 11–13.

Roucos, W., and Wilgus, A. M. (**1985**). "High quality time–scale modification for speech," in *IEEE ICASSP–85*, volume 2, 493–496.

Rozsypal, J., and Millar, B. F. (**1979**). "Perception of jitter and shimmer in synthetic vowels," Journal of Phonetics **7**, 343–355.

Schoentgen, J., and De Guchteneere, R. (**1997**). "Predictable and random components of jitter," Speech Comm. **21**, 255–272.

Schöne, P. (**1979**). "Mithörschwellen-Tonheitsmuster maskierender Sinustöne," Acustica **43**, 197–204.

Schroeder, M. (**1954**). "Die Statistischen Parameter der Frequenzkurven von grossen Räumen," Acustica **4**, 594–600.

Selkirk, E. (**1982**). "The syllable," in *The Structure of Phonological Representation, Part 2*, volume edited by H. van der Hulst and N. Smith, 337–383, Dordrecht, Floris publications.

Sinex, D. G., and Geisler, C. D. (**1983**). "Responses of auditory–nerve fibres to consonant–vowel syllables," J. Acoust. Soc. Am. **73**, 602–615.

Smith, R. L., and Zwislocki, J. J. (**1975**). "Short–term adaptation and incremental responses of single auditory–nerve fibers," Biol. Cybernet. **17**, 169–182.

Smits, R., and Yegnanarayana, B. (**1995**). "Determination of instants of significant excitation in speech using group delay function," IEEE Trans. Speech Audio Proc. **3**, 325–333.

Sommers, M. S., and Kewley-Port, D. (**1996**). "Modeling formant frequency discrimination of female vowels," J. Acoust. Soc. Am. **99**, 3770–3781.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.

Summerfield, Q., and Culling, J. F. (**1992**). "Auditory segregation of competing voices: absence of effects of FM or AM coherence," Philos. Trans. R. Soc. London. **Ser. B 336**, 357–366.

Sydeserff, H. A., Caley, R. J., Isard, S. D., Jack, M. A., Monaghan, A. I. C., and Verhoeven, J. (**1992**). "Evaluation of speech synthesis techniques in a comprehension task," Speech Comm. **11**, 189–194.

't Hart, H., and Cohen, S. (**1964**). "Gating techniques as an aid in speech analysis," Language and Speech **7**, 22–39.

't Hart, H., and Cohen, S. (**1973**). "Intonation by rule: a perceptual quest," J. Phonetics **1**, 309–327.

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

te Rietmole, P. (**1991**). "Een algoritme ter bepaling van klinkerinzetten en een algoritme ter bepaling van P-centra," IPO report 786, Institute for Perception Research, Eindhoven .

Tekieli, M. E., and Cullinan, W. L. (**1979**). "The perception of temporally segmented vowels and consonant–vowels in syllables," J. Speech Hear. Disord. **22**, 103–121.

Terhardt, E. (**1974**). "On the perception of periodic sound fluctuations (roughness)," Acustica **30**, 201–213.

Treiman, R. (**1986**). "The division between onsets and rhymes in english syllables," J. Memory Lang. **25**, 476–491.

Turner, C. W., and Van Tasell, D. J. (**1984**). "Sensorineural hearing loss and the discrimination of vowel-like stimuli," J. Acoust. Soc. Am. **75**, 562–565.

van Bezooijen, R., and Pols, L. C. W. (**1990**). "Evaluating text–to–speech systems," Speech Comm. **9**, 263–270.

Wakefield, G. H., and Viemeister, N. F. (**1990**). "Discrimination of modulation depth of sinusoidal amplitude modulation (SAM) noise," J. Acoust. Soc. Am. **88**, 1367–1373.

Zera, J., Onsan, Z. A., Nguyen, Q. T., and Green, D. M. (**1993**). "Auditory profile analysis of harmonic signals," J. Acoust. Soc. Am. **93**, 3431–3441.

Zwicker, E. (**1952**). "Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones," Acustica **2**, Akustische Beihefte AB125–AB133.

Summary

This thesis presents two studies aimed at investigating the perception of features of natural and algorithmically modified speech. Speech can be regarded as complex acoustic information characterized by both temporal and spectral variations (Plomp, 1984). Despite the complexity and the dynamical nature of speech, the human auditory system is generally able to process this acoustic information in a robust manner. In this respect, robustness refers for instance to reliable processing of speech signals of different speakers, having different speaking styles. Even though the acoustical signals of these speakers differ considerably, the intended messages of the individual speakers are conveyed without great difficulty. Robustness also applies to perceiving speech under adverse acoustical conditions, e.g., in the presence of many interfering noises and voices (the "cocktail party effect").

A conventional way of labeling the successive stages of human speech processing in response to acoustical stimulation is: *hearing, understanding, and comprehension.* The *hearing* stage refers to the (presumably) language-independent processing of acoustical information. In the *understanding* stage, the processed acoustic information is transformed into a sequence of labeled speech sounds (phonemes) and words. Finally, in the *comprehension* stage, meaning is assigned to such a sequence. A generally accepted notion is that each of these stages possesses mechanisms to achieve the robustness mentioned above.

This thesis focuses on the *hearing* stage in the context of perception of I) rhythm and intonation in natural speech, and II) of modified speech signals in speech manipulation and synthesis. In Part I (Chapter 2), the speech features under investigation are vowel onsets which are conceived of as "landmarks" in the speech signal (Hermes, 1990). These landmarks are thought to play an important role in the perception of accentuation and rhythm, and may therefore contribute to the robustness of speech perception. The study presented in this thesis deals with the automatic detection of vowel onsets. In Part II (Chapter 3-5), the speech features are signal degradations introduced by time-domain

modification of speech signals for speech-modification and speech-synthesis purposes. The study focuses on the audibility of these degradations and aims at understanding this audibility from a psychoacoustical perspective. In other words, the study investigates the robustness of the auditory system to particular (speech) signal degradations.

The vowel-onset detection study presented in **Chapter 2** builds on a study by Hermes (1990) who proposed an algorithm for vowel-onset detection. The algorithm mimicks some processing features of the human auditory system, such as frequency selectivity, in a functional way. The performance of the algorithm was judged to be insufficient, especially in the case of application to the speech of profoundly hearing-impaired children, for which it had been developed. Hermes (1990) expressed the assumption that cell responses to the onset of stimulation of the auditory system ("onset responses") may signal acoustic events to the brain, and that this may also hold for vowel onsets. The study presented in this thesis addresses the hypothesis that, given the apparent importance of vowel onsets for the perception of natural speech, vowel onsets may be distinctively coded in cell responses in some stage of the human auditory pathway. To this end, computational models of the first two neural stages of the auditory pathway, the auditory nerve (Meddis, 1986, 1988) and the cochlear nucleus (Meyer, 1993a), are used. The study focuses on *firing rate information*, i.e. the amount of spikes per unit of time in response to stimulation, rather than to synchronized activity, i.e. the firing of neurons in phase with a particular frequency present in the stimulating sound. The two questions addressed in Chapter 2 are formulated as follows: "Does the physiological literature on cell responses in the auditory nerve and the cochlear nucleus support the hypothesis that vowel onsets are coded distinctively in these two stages of the auditory system? Is it possible to automatically detect vowel onsets on the basis of simulations of such cell responses using computational models?"

For the first question, a literature survey did not yield physiological data in support of some distinct coding of vowel onsets in the auditory nerve or the cochlear nucleus. Because physiological measurements of cell responses to stimuli as complex as consonant-vowel (CV) syllables are somewhat sparse, however, this finding should not be regarded as a conclusive answer. Nevertheless, on the basis of characteristics of cell responses reported in the literature it is argued in Chapter 2 that the so-called Chop-T responses, observed in the anteroventral cochlear nucleus, presumably provide the most stable neural representation for vowel-onset detection. The reason for this is that cells showing such responses receive input from auditory nerve fibers which, individually, have limited dynamic ranges of about 20-30 dB. Because these fibers have various firing thresholds, combination of their excitatory activity results in a stable firing-rate coding of, for instance, the stimulus power spectrum, over a dynamic range much larger than that of individual auditory nerve fibers (e.g., Blackburn and Sachs, 1990).

For addressing the second question ("Is it possible to automatically detect vowel onsets in simulated cell responses?"), a model of Chop-T responses developed by Meyer (1993a) was employed. The simulated Chop-T responses were subjected to a vowel-onset detection scheme that basically integrates the simulated activities into two, rather broad, frequency channels associated with the first and second formant. A sudden onset of (integrated) activity in both channels signals a vowel onset, provided a number of additional conditions is met, such as an appropriate ratio of activity in both channels. The scheme was tested on a large database of read speech, uttered by a large number of speakers. Its performance was somewhat worse than the algorithm proposed by Hermes (1990), which was reevaluated using this database, in the sense that the number of missed onsets and false alarms were both slightly higher. In addition, the performance was found to be dependent of the presentation level of the sentences in the database: missed-onset rates were found to decrease with increasing level, but false-alarm rates were found to increase. By extending the scheme with rough voiced/unvoiced information extracted from the simulated responses, the number of false alarms decreased, although the number of missed onsets remained constant.

In addition to the use of a scheme operating on simulated Chop-T responses, a conventional pattern-recognition technique was employed, namely multi-layer perceptrons (MLPs). Basically these networks were trained to signal the presence of a vowel. Vowel onsets were detected by finding the instances of sudden increase in the network output. Training these MLPs with the simulated Chop-T responses showed a level dependency in detection performance, and moderate detection rates of missed onsets and false alarms. Better results were obtained using mel-scaled spectra as input to the MLPs. In such cases, performance was comparable to the algorithm developed by Hermes (1990).

By comparing the detection performances of the algorithm designed by Hermes (1990), the scheme based on Chop-T response simulation, and the MPLs, the importance of several physical characteristics of the speech signal for automatic vowel-onset detection was derived. It was found that rough spectral content, such as a two-channel representation, and intensity information contribute most to detection. Harmonicity information, such as a voiced/unvoiced labeling, can be regarded as a secondary source of information to reduce the number of false alarms. An analysis of the phonemic context in which false alarms and missed onsets occurred revealed a high degree of conformity for the three detection methods, despite substantial differences in detection strategies. Missed onsets predominantly occured in /ə/, /i/, and to a lesser extent /I/ vowel contexts. False alarms were mainly scored in long vowels, in /r/ contexts, and for plosives.

Chapters 3 to 5 of this thesis focus on the audibility of signal degradations due to time-domain speech-waveform manipulation. As a well-known and typical example, the Pitch

Synchronous OverLap and Add (PSOLA; Charpentier and Moulines, 1989; Moulines and Laroche, 1995) technique is evaluated. PSOLA is known to often yield satisfactory quality of the modified speech, although unpredictable artefacts ar known to occur, such as roughness and hoarseness. The basic question of this study is why the human auditory system often is insensitive, or tolerant to waveform modifications of speech. In this study, PSOLA application is restricted to modification of the fundamental frequency ($F0$), of both natural and synthetic speech signals.

In **Chapter 3** the discriminability of PSOLA-modified from unmodified synthetic single-formant signals (Klatt, 1980) was determined as a function of the shift in $F0$. Discrimination performance of subjects was typically found to vary in an almost sinusoidal fashion as a function of the shift in $F0$. Such a shift in $F0$ results, apart from the desired shift in the frequencies of the harmonics, in changes of both the harmonics levels and the (relative) harmonic phases. The largest changes occur around the formant frequency. In auditory discrimination between modified and unmodified signals, subjects may for instance detect these level changes. In such case listeners are said to attend to spectral cues for discrimination. Such cues are expected to be particularly available in the case of *resolved* harmonics around the formant frequency, i.e., if each harmonic falls within a separate auditory filter. In the case of *unresolved* harmonics, i.e., if several harmonics fall within a single filter, listeners may attend to differences in modulations ("beats"). In such case, listeners are said to attend to temporal cues. The listening experiments were paralleled by psychoacoustical modeling efforts in order to clarify whether subjects used spectral or temporal cues. The subjects' discrimination performance for stimuli with a low formant, which likely results in resolved harmonics around the formant, could be modeled best by using an excitation-pattern model based on spectral cues. In the case of a high formant, where harmonics are expected to be unresolved, subjects' performance could be best described using a modulation-discrimination model. The modeling results were thus in agreement with current psychoacoustical understanding of auditory processing of resolved and unresolved harmonics.

In general, the influence on discrimination performance of randomly varying the overall level of the stimuli was small. In other words, the variability in level observed in natural speech presumably does not explain any perceptual tolerance to signal degradations induced by PSOLA. Randomly varying the formant frequency, however, did drastically deteriorate discrimination sensitivity in the case of low formants, but not in the case of high formants.

Chapter 3 also addresses the perceptual sensitivity to the accuracy of positioning the so-called "pitch markers". Pitch markers indicate boundaries of local pitch periods and serve to align the time-domain PSOLA operations. In these experiments, all pitch

markers were shifted in time by equal amounts. The excitation instances of the formant filters were regarded as reference pitch-marker positions. The results show that (uniform) shifts of 10-25 % of the fundamental period are below the discrimination threshold. In other words, the auditory system seems to be somewhat tolerant to pitch-marker positioning, at least for uniform shifts. For such shifts, the excitation pattern model mentioned above could accurately describe the subjects' discrimination behavior.

In **Chapter 4** more speech-like, and thus more complex, signals were subjected to PSOLA modification, and discrimination performance was determined. First, double-formant signals were modified which were composed of combinations of the single formants used in Chapter 3. The discrimination performance was found to be dominated by the first formant, possibly because of its higher level relative to the second formant. After equalization of the formant levels, subjects could use the information from the second formant region in some, but not all conditions. In the case of PSOLA-modified natural vowels, which have more than one formant, cues for detecting PSOLA-induced changes thus are more likely to originate in the first formant region than in higher regions. As was shown in Chapter 3, such cues may vanish if the (low) formant frequency varies, which is likely to occur in natural speech. Therefore, the tolerance of the auditory system in perceiving $F0$-shifted natural vowels is expected to be high.

Second, the influence of vocal perturbation, i.e. jitter and shimmer, on discrimination performance was determined. In a baseline experiment, detection *thresholds* for shimmer and jitter in single-formant signals were measured. Thresholds for jitter, and to a lesser extent for shimmer, could be replicated by using the modulation-discrimination model mentioned above. The influence of vocal perturbation on subjects' discrimination sensitivity of PSOLA-modified and unmodified stimuli was found to be small. This finding implies that irregularities in $F0$ or intensity do not deteriorate the detectability of PSOLA-induced changes, and thus do not account for any perceptual tolerance in perceiving PSOLA modified natural speech. On the other hand, such irregularities do not either give rise to additional PSOLA-induced discrimination cues.

Finally, in **Chapter 5** it was verified whether some of the findings of Chapters 3 and 4 for synthetic signals can be generalized to PSOLA-modification of sustained natural vowels. In particular, the experiments reported in this chapter dealt with pitch-marker shifts. The results obtained for uniform shifts of the pitch markers were in good agreement with those reported in Chapter 3 for synthetic signals. The excitation-pattern model could reasonably well describe the subjects' discrimination performance, except for the above-threshold performance in some conditions. Additionally, thresholds were measured for detection of single pitch-marker shifts and random shifts of all pitch markers within the vowel. For the single shifts, thresholds depended on the position of the pitch marker

within the vowel: thresholds for shifts in the middle in the vowel were lowest. For the random shifts, thresholds were in reasonable agreement with the jitter thresholds reported in Chapter 4. In summary, the findings of Chapters 3 and 4 on pitch-marker positioning and the role of jitter appear to the generalizable to natural speech, at least in the case of F0 modification of sustained vowels.

Samenvatting

Dit proefschrift beschrijft twee studies waarin onderzocht is hoe een aantal kenmerken (Engels: "features") van natuurlijke en gemodificeerde spraak waargenomen wordt. Spraak kan beschouwd worden als akoestische informatie die, als gevolg van zowel temporele als spektrale variatie, gekarakteriseerd wordt door een grote complexiteit (Plomp, 1984). Ondanks de complexiteit en het dynamische karakter van spraak is het menselijke gehoorsysteem over het algemeen goed in staat deze akoestische informatie op een robuuste manier te verwerken. Bij robuustheid wordt hier bijvoorbeeld gedacht aan het adequaat kunnen verwerken van spraaksignalen van verschillende sprekers, elk met een andere spreekstijl. De beoogde boodschappen van de individuele sprekers kunnen vrijwel probleemloos overgedragen worden, ook al verschillen de spraaksignalen van deze sprekers aanzienlijk. Bij robuustheid kan ook gedacht worden aan het kunnen waarnemen van spraak in, akoestisch gezien, ongunstige omstandigheden, zoals bijvoorbeeld in de aanwezigheid van vele interferende stemmen en andere achtergrondgeluiden (het "cocktail party effect").

Een conventionele indeling van de opeenvolgende stappen in het proces van menselijke spraakverwerking is: *horen*, *verstaan* en *begrijpen*. Op het niveau van *horen* wordt de akoestische informatie op een, zo wordt verondersteld, taalonafhankelijke manier verwerkt. Op het niveau van *verstaan* wordt deze verwerkte informatie omgezet in een sequentie van geclassificeerde spraakklanken (fonemen) en woorden. Uiteindelijk wordt aan een dergelijke sequentie betekenis toegekend op het niveau van *begrijpen*. Bij deze indeling wordt er over het algemeen vanuit gegaan dat elk van deze drie niveau's beschikt over mechanismen om de bovengenoemde robuustheid te bereiken.

Dit proefschrift richt zich op het niveau van *horen*, in de contekst van waarneming van I) ritme en intonatie in natuurlijke spraak en van II) gemodificeerde spraaksignalen in het kader van spraakmanipulatie en spraaksynthese. In Deel I (Hoofdstuk 2) gaat het bij de onderzochte spraakkenmerken ("features") om klinkerinzetten. Klinkerinzetten worden beschouwd als ankerpunten in het spraaksignaal (Hermes, 1990) waarvan wordt aange-

nomen dat zij een belangrijke rol spelen in de waarneming van accenten en van ritme. Op die manier spelen klinkerinzetten mogelijk een rol bij de robuustheid van menselijke spraakwaarneming. In de studie die in dit proefschrift gepresenteerd wordt zijn algoritmes voor het automatisch detecteren van klinkerinzetten ontwikkeld en geëvalueerd. In Deel II (Hoofdstukken 3-5) gaat het bij de onderzochte spraakkenmerken om vervormingen (distorties) van het spraaksignaal. Met name worden distorties onderzocht die geïntroduceerd worden wanneer spraaksignalen ten behoeve van spraaksynthese in het tijddomein gemodificeerd worden. De studie richt zich op de hoorbaarheid van deze vervormingen waarbij het doel is om deze hoorbaarheid te begrijpen op basis van kennis uit de psychoakoestiek. Anders gezegd, de studie behandelt de vraag in hoeverre het menselijk gehoor gevoelig is voor bepaalde vervormingen van het spraaksignaal.

De studie naar klinkerinzetdetectie van Hoofdstuk 2 bouwt voort op een studie van Hermes (1990) waarin een algoritme voor klinkerinzetdetectie uitgewerkt wordt. Het algoritme bootst op een functionele wijze een aantal eigenschappen van verwerking in het menselijk auditief systeem na, zoals frequentieselectiviteit. De resultaten verkregen met dit algoritme werden ontoereikend bevonden, in het bijzonder bij toepassing op de spraak van slechthorende kinderen waarvoor het algoritme ontwikkeld was. Hermes (1990) stelde dat neurale responsies aan het begin van stimulatie (Engels: "onset responses") mogelijkerwijs dienen om nieuwe akoestische gebeurtenissen (Engels: "events") te signaleren aan het brein, hetgeen ook van toepassing kan zijn op klinkerinzetten. De in dit proefschrift beschreven studie behandelt de hypothese dat, gegeven de klaarblijkelijke rol die klinkerinzetten spelen bij de waarneming van natuurlijke spraak, klinkerinzetten op aparte wijze gecodeerd worden in zenuwcelresponsies in het auditief systeem. Hiertoe zijn computationele modellen gebruikt van de eerste twee stadia van de neurale verwerking van auditieve informatie, namelijk van de gehoorzenuw (Meddis, 1986, 1988) en van de nucleus cochlearis (Meyer, 1993b). Het onderzoek heeft zich beperkt tot het onderzoeken van de vuringsactiviteit van zenuwcellen, d.w.z. het aantal vuringen per tijdseenheid ten gevolge van akoestische stimulatie. Als alternatieve codering wordt in de fysiologische literatuur de gesynchroniseerde vuringsactiviteit genoemd, d.w.z. of het vuringspatroon in fase verloopt met bepaalde frequenctiecomponenten van de stimulatie. De twee vragen die in Hoofdstuk 2 behandeld zijn kunnen als volgt geformuleerd worden: "Wordt er in de fysiologische literatuur bewijs geleverd voor een aparte neurale codering van klinkerinzetten in de gehoorzenuw of the nucleus cochlearis? Is het mogelijk om klinkerinzetten automatisch te detecteren op basis van responsies die gesimuleerd worden met computationele modellen?"

Wat betreft de eerste vraag werden er in een literatuuronderzoek geen fysiologische meetgegevens gevonden die zo'n aparte codering van klinkerinzetten in de gehoorzenuw of de nucleus cochlearis aannemelijk maakten. Fysiologische metingen van celreponsies op

stimuli met de complexiteit van lettergrepen, zoals bijvoorbeeld een consonant-klinker (CV) struktuur, zijn echter maar in beperkte mate gepubliceerd. Daarom geeft de uitkomst van die literatuurstudie geen definitief antwoord op de gestelde vraag. Op basis van in de literatuur beschreven karakteristieken van celresponsies worden in Hoofdstuk 2 wel een aantal argumenten gegeven om te veronderstellen dat de zogenaamde Chop-T responsies, die gemeten zijn in het anteroventrale gedeelte van de nucleus cochlearis, de meest bruikbare neurale respresentatie van spraaksignalen leveren voor klinkerinzetdetectie. Cellen die zulke responsies vertonen worden namelijk geëxciteerd door meerdere cellen uit de gehoorzenuw, die elk afzonderlijk slechts een beperkt dynamisch bereik van 20-30 dB hebben. Doordat deze gehoorzenuw-cellen verschillende vuurdrempels bezitten, resulteert de combinatie van hun exciterende activiteit in een stabiele representatie van het vermogensspectrum van een akoestische stimulus, over een dynamisch bereik dat veel groter is dan het bereik van afzonderlijke cellen uit de gehoorzenuw (zie bijvoorbeeld Blackburn and Sachs, 1990).

Om de tweede vraag te beantwoorden ("Is het mogelijk om klinkerinzetten automatisch te detecteren op basis van gesimuleerde celresponsies?") is een model gebruikt voor de simulatie van Chop-T responsies dat ontwikkeld is door Meyer (1993b). In een klinkerinzetdetectie-algoritme dat ontwikkeld is voor dit proefschrift worden deze gesimuleerde responsies verdeeld over twee brede frequentiebanden, die geassocieerd kunnen worden met de eerste en tweede formant. De activiteit van de verschillende responsies wordt gesommeerd in beide banden. Een abrupte stijging van (gesommeerde) activiteit in beide banden signaleert een klinkerinzet, op voorwaarde dat aan een aantal criteria wordt voldaan, zoals bijvoorbeeld een geschikte verhouding van de activiteit in de twee banden. Dit algoritme is geëvalueerd op een grote database met voorgelezen spraak van een groot aantal sprekers. De resultaten waren iets slechter dan de resultaten van het algoritme ontwikkeld door Hermes (1990) dat eveneens op basis van deze database geëvalueerd is: zowel het aantal gemisde klinkerinzetten als het aantal ten onrechte aangemerkte inzetten was iets hoger. Bovendien bleek het nieuw ontwikkelde algoritme gevoelig te zijn voor het ingangsniveau van de databasezinnen: het aantal gemisde inzetten bleek af te nemen naarmate het ingangsniveau hoger werd, maar tegelijkertijd steeg dan ook het aantal ten onrechte aangemerkte inzetten. Door een simpele "stemhebbend/stemloos beslissing" aan het algoritme toe te voegen, bepaald op basis van de gesimuleerde responsies, kon het aantal onterechte detecties teruggebracht worden, terwijl het aantal gemisde inzetten constant bleef.

Naast het algoritme dat opereerde op de gesimuleerde Chop-T responsies is een conventionele patroonherkenningstechniek gebruikt, namelijk multi-layer perceptrons (MLPs). Deze netwerken werden getraind om de aanwezigheid van een klinker in het spraaksignaal aan te geven. Klinkerinzetten werden gedetecteerd door die momenten te traceren

waarop de uitgang van het netwerk ("aanwezigheid van een klinker") snel toenam. Bij het trainen van deze netwerken op de gesimuleerde Chop-T representaties werd eveneens een niveau-afhankelijkheid gevonden, met matige aantallen van gemisde klinkerinzetten en onterechte detecties. Betere resultaten werden verkregen indien mel-geschaalde spectra als input dienden voor de MLPs. In dat geval waren de resultaten vergelijkbaar met de resultaten voor het algoritme beschreven door Hermes (1990).

Door de detectieresultaten van het algoritme van Hermes (1990), het algoritme gebaseerd op de simulatie van Chop-T responsies en de MLPs te vergelijken kon het belang voor automatische klinkerinzetdetectie van een aantal fysische karakteristieken van het spraaksignaal afgeleid worden. De globale spectrale inhoud, zoals een representatie in twee kanalen, en informatie over de intensiteit bleken het meest bij te dragen aan de detectie. Informatie over de harmoniciteit, zoals een stemhebbend/stemloos classificatie, kan gezien worden als een secundaire informatiebron om het aantal onterechte inzetten te reduceren. Uit een analyse van de foneemcontext waarin onterechte en gemisde klinkerinzetten voorkwamen bleek dat er een hoge graad van overeenstemming was tussen de drie detectiemethoden, ondanks wezenlijke verschillen in de afzonderlijke detectiestrategieën. Gemisde inzetten kwamen vooral voor bij de klinkers /ə/, /i/ en, in mindere mate, /I/. Onterechte inzetten werden vooral gevonden in lange klinkers, in de contekst van de medeklinker /r/ en bij plosieven.

In de Hoofdstukken 3 t/m 5 van dit proefschrift wordt ingegaan op de hoorbaarheid van signaalvervormingen ten gevolge van golfvormmanipulaties van spraak in het tijddomein. Als een bekend en typerend voorbeeld is de Pitch Synchronous OverLap and Add (PSOLA; Charpentier and Moulines, 1989; Moulines and Laroche, 1995) geëvalueerd. Deze techniek blijkt, met verrassend eenvoudige bewerkingen, gemodificeerde spraak met veelal hoge kwaliteit op te kunnen leveren. Op onvoorspelbare momenten treden er echter ook hoorbare vervormingen op die bijvoorbeeld leiden tot ruwheid en heesheid van de gemodificeerde spraak. Het uitgangspunt voor deze studie is de vraag waarom het menselijk auditief systeem veelal ongevoelig, of tolerant, is voor golfvormmanipulaties van spraak. In dit onderzoek is de toepassing van PSOLA beperkt tot het modificeren van de fundamentele frequentie ($F0$) van zowel natuurlijke als synthetische spraaksignalen.

In Hoofdstuk 3 is bepaald in hoeverre PSOLA-gemodificeerde, synthetische signalen met één enkele formant auditief onderscheiden kunnen worden van niet-gemodificeerde signalen, als functie van de verschuiving in $F0$. De resultaten van luisterexperimenten vertoonden over het algemeen een bijna sinusoïdale variatie in de onderscheidbaarheid als functie van de verschuiving in $F0$. Zo'n $F0$ verschuiving heeft een verandering van de (complexe) spektrale omhullende tot gevolg hetgeen leidt tot verandering van zowel

de niveau's als de fasen van de harmonischen van het signaal. De grootste veranderingen vinden plaats rondom de formant. Bij het auditief onderscheiden van gemodificeerde en ongemodificeerde signalen kunnen proefpersonen bijvoorbeeld luisteren naar de niveauveranderingen. In dat geval spreekt men van spektrale distinctieve kenmerken (Engels: "spectral cues"). Van zulke cues kan verwacht worden dat ze vooral gebruikt kunnen worden indien de harmonischen rond de formant *opgelost* worden, d.w.z. als elke harmonische in een afzonderlijk auditief filter valt. In het geval dat de harmonischen *onopgelost* zijn, d.w.z. als meerdere harmonischen binnen een enkel auditief filter vallen, kunnen proefpersonen luisteren naar verschillen in modulatie ("zwevingen"). In zo'n geval is er sprake van temporele distinctieve kenmerken (Engels: "temporal cues"). Naast het uitvoeren van luisterexperimenten is er getracht om met behulp van psychoakoestische modellen vast te stellen of luisteraars gebruik maken van spektrale of temporele cues. De resultaten van de experimenten waarin gebruik gemaakt werd van een formant met lage frequentie, waarbij de harmonischen rond de formant grotendeels opgelost zijn, konden het best gemodeleerd worden met een excitatiepatroonmodel dat gebaseerd is op spektrale cues. In het geval van een hoge formant, waarbij deze harmonischen maar zeer ten dele opgelost worden, konden de resultaten het best beschreven worden met een model gebaseerd op het onderscheiden van modulatiediepte. De resultaten van het modeleren waren dus in overeenstemming met het huidige psychoakoestische inzicht in de auditieve verwerking van opgeloste en onopgeloste harmonischen.

In één van de experimentele condities werd het geluidsniveau random gevarieerd. Hierdoor werden de resultaten van het auditief onderscheiden van gemodificeerde signalen echter maar in beperkte mate beïnvloed. Met andere woorden, de variabiliteit in geluidsniveau zoals die voorkomt in natuurlijke spraak verklaart waarschijnlijk niet waarom het auditief systeem veelal tolerant is voor de signaalvervormingen veroorzaakt door PSOLA. Daarentegen deed het random variëren van de formant frequentie het onderscheidingsvermogen aanzienlijk verminderen in het geval van de lage formanten, echter niet in het geval van de hoge formanten.

In Hoofdstuk 3 wordt ook onderzocht hoe gevoelig het menselijk gehoor is voor de nauwkeurigheid waarmee de zogenaamde "pitch markers" geplaatst worden. Pitch markers geven de grenzen van lokale toonhoogteperiodes aan en dienen voor het in de pas laten lopen van de PSOLA operaties met het spraaksignaal. In de bijbehorende experimenten werden alle pitch markers over eenzelfde tijdsduur verschoven. Daarbij werden de momenten van excitatie van de formant filters als referentiepunten genomen. De resultaten laten zien dat (uniforme) verschuivingen van 10-25 % van de fundamentele periode niet onderscheiden kunnen worden. Met andere woorden, het auditief systeem lijkt in enige mate tolerant te zijn voor de pitch-marker positionering, tenminste in het geval van uniforme verschuivingen. De resultaten van proefpersonen konden voor dergelijke

verschuivingen tot in redelijke mate van detail gemodeleerd worden met behulp van het bovengenoemde excitatiepatroonmodel.

In Hoofdstuk 4 zijn meer spraak-achtige, en dus ook complexere, signalen gemodificieerd met behulp van PSOLA, waarbij wederom de onderscheidbaarheid t.o.v. niet-gemodificeerde signalen bepaald werd. Allereerst zijn dubbele-formantsignalen gemodificeerd die samengesteld waren uit combinaties van formanten zoals gebruikt in Hoofdstuk 3. Het onderscheidingsvermogen van proefpersonen bleek gedomineerd te worden door informatie van de eerste formant, mogelijkerwijs vanwege het feit dat het niveau van deze formant hoger was dan dat van de tweede formant. Als de niveaus van beide formanten gelijkgetrokken werden bleken proefpersonen in staat om de informatie uit het gebied van de tweede formant te kunnen gebruiken in sommige, maar niet in alle condities. In het geval van PSOLA-gemodificeerde natuurlijke klinkers, die meer dan één formant bezitten, zullen de cues voor het onderscheiden van gemodificeerde en niet-gemodificeerde signalen dus waarschijnlijk uit het spectrum rond de eerste formant komen, en niet rond hogere formanten. Zoals aangetoond is in Hoofdstuk 3 kunnen de cues voor het onderscheiden verdwijnen als de (lage) formant frequentie varieert, hetgeen over het algemeen het geval is bij natuurlijke spraak. De tolerantie van het auditieve systeem voor het verschuiven van $F0$ is daarom naar verwachting hoog.

Als tweede aspekt is de invloed van stemperturbatie, d.w.z. jitter en shimmer, op het onderscheidingsvermogen van proefpersonen bepaald. Als basisexperiment zijn de waarnemingsdrempels voor jitter en shimmer in enkelvoudige-formantsignalen gemeten. De gemeten drempels voor jitter, en in mindere mate voor shimmer, konden gesimuleerd worden door gebruik te maken van het bovengenoemde model gebaseerd op het onderscheiden van modulatie. De invloed van stemperturbatie op het auditief onderscheiden van PSOLA-gemodificeerde en niet-gemodificeerde signalen bleek klein te zijn. Dit gegeven doet veronderstellen dat onregelmatigheden in $F0$ of in intensiteit de waarneembaarheid van veranderingen ten gevolge van PSOLA niet verslechteren. Daarom ligt de eventuele aanwezigheid van stemperturbaties niet ten grondslag aan de perceptieve tolerantie bij het waarnemen van gemodificeerde, natuurlijke spraak. Anderzijds worden door de aanwezigheid van dergelijke onregelmatigheden geen nieuwe, door PSOLA veroorzaakte cues ter onderscheiding geïntroduceerd.

Tot slot is in Hoofdstuk 5 nagegaan of een aantal van de uitkomsten van Hoofdstuk 3 en 4 voor synthetische signalen gegeneralizeerd kunnen worden naar PSOLA-bewerking van aangehouden natuurlijke klinkers. Met name is de rol van de plaatsing van de pitch markers onderzocht. De resultaten voor uniforme verschuivingen van de pitch markers kwamen goed overeen met de resultaten die in Hoofdstuk 3 beschreven zijn voor synthetische signalen. Het excitatie-patroon model was in staat om een redelijk goede

beschrijving te geven van de resultaten, behalve in enkele condities voor bovendrempelige resultaten. Daarnaast zijn drempels gemeten voor de waarneming van de verschuiving van één enkele pitch marker en van random verschuivingen van alle pitch markers. In het geval van enkelvoudige verschuivingen bleken de drempels af te hangen van de positie in de klinker: drempels voor verschuivingen in het midden van de klinker waren het laagst. In het geval van random verschuivingen waren de drempels in redelijk goede overeenstemming met de drempels voor jitter zoals besproken in Hoofdstuk 4. Concluderend kan gesteld worden dat de uitkomsten van de Hoofdstukken 3 en 4 met betrekking tot pitch marker plaatsing en de rol van jitter gegeneralizeerd kunnen worden naar natuurlijke spraak, in elk geval naar aangehouden klinkers.

Curriculum vitae

9 februari 1968	geboren te Velsen
1980 - 1986	ongedeeld VWO aan de Haaksbergse Scholengemeenschap
1986 - 1992	studie Alpha-Informatica aan de Universiteit van Amsterdam.
	specialisatie: spraakcommunicatie aan het Instituut voor Fonetische Wetenschappen (IFA) te Amsterdam.
	afgestudeerd (cum laude) op een onderzoek naar "cellular automata and speech recognition"
juni 1992 - sept 1993	vervangende dienstplicht op het Instituut voor Perceptie Onderzoek (IPO) in Eindhoven.
nov 1993 - okt 1997	assistent in opleiding (AIO) op het IPO.

Dankwoord

Aan het slot van dit proefschrift wil ik graag nog een aantal mensen met name noemen die in belangrijke mate hebben bijgedragen aan de inspirerende tijd waarin ik gewerkt heb op het IPO. Dankzij hen leek niet alleen de tijd zich geregeld te versnellen, maar werd de harmonieuze werksfeer frequent met een paar tonen verhoogd. Met andere woorden, deze mensen hebben de PSOLA techniek in de vingers en brengen haar nog in de praktijk ook.

Ik wil graag Dik bedanken voor zijn inspirerende begeleiding bij het "vangen" en determineren van de klinkerinzetten, en het wekken van mijn interesse in de fysiologie van het gehoorsysteem. The same goes for Georg, also for trying to convince me that guinee pigs are not particularely interested in vowel onsets. Armin is de laatste jaren een begeleider geweest zoals het bij de wet vastgelegd zou moeten worden: geïnteresseerd, altijd op de hoogte en de grote lijnen in de gaten houdend. Bovendien, Armin schroomt er niet voor om, buiten het "veilige" vakgebied, een "baantje rechtsom te schaatsen" en desondanks een goede tijd neer te zetten. Rene wil ik bedanken voor de raad en het advies aan het begin (en eind) van mijn projekt, Aad voor het "overnemen van de fakkel" in het tweede gedeelte. Steven, mijn kamergenoot sinds jaar en dag, wil ik bedanken voor alle humoristische en rake analyses van de beurskoersen, het IPO, J.S. Bach, paars, en alle andere, steeds weer verrassende onderwerpen. Natuurlijk ook dank aan de Horen groep door de jaren heen (Marcel, Wout, Ralf, Andy, Joyce en Jeroen). En dank aan de geduldige proefpersonen, wie Het Waarom van de experimenten soms ontging. Speciale dank aan subject KM wier "behaviour" iedere voorspelling telkens weer overtreft.

Stellingen

behorend bij het proefschrift

Physiological and psychoacoustical correlates of perceiving natural and modified speech

van Reinier Kortekaas

1. Voor het automatisch detecteren van klinkerinzetten zijn globale spectrale inhoud en intensiteit, beide als functie van de tijd, de belangrijkste spraaksignaalkarakteristieken. Harmoniciteit is hierbij een secundaire bron van informatie.

   Dit proefschrift, hoofdstuk 1.

2. Aan de perceptieve tolerantie ten opzicht van PSOLA golfvormmodificatie van natuurlijke spraak ligt de spektrale variabiliteit van die spraak ten grondslag, en niet haar irregulariteit in grondfrequentie of in intensiteit.

   Dit proefschrift, hoofdstukken 3 en 4

3. Meetsystemen ter klinische beoordeling van pathologische stemmen dienen gebaseerd te zijn op modellen van, bijvoorbeeld, waargenomen ruwheid en niet op meting van de irregulariteit van het spraaksignaal.

   naar: Rabinov et al. (1995) *J. Speech Hear. Res.* **38**, 26-32.

4. Het streven naar een hoge mate van detail in de modelering van de glottale excitatie voor spraaksynthese ontbeert vaak perceptieve relevantie.

   Veldhuis (1996) *IPO annual progress report* **31**, 100-108.

5. Waar Hillenbrand (1988) in diens Figuur 2 ten onrechte spreekt van autocorrelatie, wordt autocovariantie bedoeld; in het geval van autocorrelatie zou de spreiding in periodeduur of stemamplitude vele malen groter zijn dan de bijbehorende gemiddelden.

Hillenbrand (1988) *J. Acoust. Soc. Am.* **83**, 2361-2371.

6. Ondanks de aanwezigheid van meer raakvlakken dan verschillen, is het voor veel spraak- en gehoorwetenschappers nog steeds moeilijk om dezelfde taal te spreken en goed naar elkaar te luisteren.

7. Geluidskwaliteitsonderzoek voor mobiele telefoons is slechts van secundair belang omdat het gros der gebruikers alleen erin praat, en niet luistert.

8. Het gegeven dat de zin "Ze hebben een nieuwe auto gekocht" al meer dan 30 jaar gebruikt kan worden voor prosodisch spraakonderzoek naar de expressie van bijvoorbeeld blijdschap, stemt niet optimistisch over de ontwikkeling van de auto-mobiliteit.

9. AIO's zullen binnen een aantal jaren uitgroeien tot Nederland's grootste exportprodukt.

10. De toenemende invloed van de media op de aandachtspunten van het Nederlandse parlement vergroot, in weerwil van de persvrijheid, het belang van democratische verkiezingen van die media.

11. Nog steeds is het vaak zo dat degene die de lakens uitdeelt, ze niet zelf wast.