Rob van Son

SPECTRO-TEMPORAL FEATURES OF VOWEL SEGMENTS

The series 'Studies in Language and Language use' comprises doctoral dissertations by postgraduate students and monographs and collections produced by researchers of IFOTT.

The theoretical orientation in terms of which IFOTT approaches its research domain can be characterized under the broad term 'funtional'. A functional approach to language and language use implies that the study of language must take account of its primary function as a medium of human communication.

The IFOTT research programme consists of the following:

Argumentation Theory
Augmentation & Alternative Communication
Corpus Linguistics
Creole Linguistics
Discourse Analysis
Functional Grammar
Language Acquisition
Language Change
Language Contact
Language Typology
Language Variation
Pathology of Language & Speech
Sign Language
Speech Processing
Speech Production & Perception

SPECTRO-TEMPORAL FEATURES OF VOWEL SEGMENTS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van Rector Magnificus prof. dr P.W.M. de Meijer,
in het openbaar te verdedigen
in de Aula der Universiteit
(Oude Lutherse Kerk, Singel 411, hoek Spui),
op

vrijdag 3 september 1993 te 13.30 uur

door

Robertus Johannes Joseph Hendricus van Son

geboren te Nijmegen

FACULTEIT DER LETTEREN,
UNIVERSITEIT VAN AMSTERDAM

Promotor: Prof. dr ir L.C.W. Pols

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Son, Robertus Johannes Joseph Hendricus van

Spectro-temporal features of vowel segments / Robertus Johannes Joseph Hendricus van Son. - Amsterdam : IFOTT. - (Studies in language and language use ; 3)
Tevens proefschrift Universiteit van Amsterdam, 1993. - Met index, lit. opg. - Met samenvatting in het Nederlands.
ISBN 90-74698-03-4
NUGI 941
Trefw. : fonetiek ; onderzoek / klinkerreductie / klinkerherkenning.

STUDIES IN LANGUAGE AND LANGUAGE USE
Uitgave IFOTT Amsterdam
Spuistraat 210 1012 VT AMSTERDAM

Printed in the Netherlands by ICG-Printing-Dordrecht

Contents

Foreword

Several years ago, I came to the Intitute of Phonetic Sciences of the University of Amsterdam to apply for a job with only a vague idea about phonetics and its relation to speech. Since then, numerous people have helped me in many different ways to understand phonetics, to perform experiments, to write comprehensible papers, and to complete this thesis. I hope that this book will do justice to their efforts and that they will consider their time well spent.

At this point it is good custom to thank the one who generally shares much of the burden but none of the credit. Sylvia, without you, this would only have been half the book it is now.

1

GENERAL INTRODUCTION

Abstract

This chapter contains a summary of current models on vowel production and perception. The target-undershoot model of vowel production is discussed extensively. Studies that confirm the predictions of this model and those that failed to do so are reviewed. Theories on vowel perception can be divided into those that use information from the consonant-vowel transitions, i.e. dynamic-specification, and those that do not, i.e. target models. Arguments for both types of models are discussed.

In this thesis we present studies on the mechanisms that control vowel production and vowel perception. We test several key predictions made by leading models in these fields of research. In the research in vowel production, the leading model is that of target-undershoot in articulation. The models describing vowel perception can be divided into two "camps". One camp states that all information necessary for recognition is present in the vowel nucleus. The other camp is convinced that the spectro-temporal structure of the consonant-vowel transitions is important for correct vowel identification.

In this chapter, we review the models of vowel production and perception and formulate the problems we want to investigate.

1.1 Target-undershoot in speech production

In natural speech there is a substantial variation in vowel realizations, even when spoken by a single person. Vowels spoken in isolation or in a neutral context, such as /hVd/ in English, are considered to approach the ideal with regard to vowel quality. Such ideal vowel realizations are called canonical realizations. Numerous factors change these canonical realizations to the realizations actually found in natural speech, e.g. speaking style, prosody, context. All these separate influences are generally divided into two groups: coarticulation and reduction (see e.g., the textbooks of O'Shaughnessy, 1987; Clark and Yallop, 1990). Coarticulation causes individual vowel realizations to become more similar to their neighbouring phonemes in the utterance. In an articulatory sense, distinctive features, like place of articulation or rounding, are assimilated. In an acoustic sense, spectral distances between neighbouring phonemes become smaller. Vowel reduction causes realizations of different vowels to become more alike. Reduced vowel realizations are more like the neutral (schwa) vowel. As a result of reduction, the contrast between vowels is smaller (e.g., see Delattre, 1969; Koopmans-van Beinum, 1980 for overviews on reduction). Coarticulation is conventionally described as a result of the immediate context of the vowel (actual neighbours). Differences in the amount of vowel reduction are most evident between stressed and unstressed syllables, but vowel reduction is also reported to occur as a result of differences in speaking style and rate, position in the word, etc..

In practice, it seems often difficult to distinguish between coarticulation and reduction. For many consonantal contexts, the vocalic parts of the consonant-vowel (CV) and vowel-consonant (VC) transitions are "reduced" with respect to the mid-point of the vowel realizations (Schouten and Pols, 1979). In a recent study of the effects of stress, sentence-accent and word-class on vowel reduction, Van Bergem (1993) found that classical reduction could be identical to increased coarticulation. He found that (spectrally) the *non*-lexical schwa vowel, defined as the target of reduction, has no fixed (central) position in the vowel space but is identical to the lexical schwa vowel "... *in the same phonemic context*." (Van Bergem, 1993; p13). In his study, the formant frequencies of reduced /E/-realizations from /wEl/ were not shifted towards the center of the vowel triangle but towards the position of the /ˊ/ vowel from /'Xrywˊl/, which itself was distinctively /O/-like

(F_1=346 Hz, F_2=940 Hz). Results of the work of Koopmans-van Beinum (1992) on schwa vowel realizations can be interpreted to support this idea. This could mean that the schwa is not only the end-point of reduction but also that of coarticulation. In this case, the schwa would be the vowel that is as close to the consonants surrounding it as it possibly could be. If the schwa is variable, and is the most reduced and most coarticulated vowel at the same time, then coarticulation with the context and reduction to the schwa would be identical processes. In this view, the often reported centralization of reduced vowels in vowel space (e.g., Delattre, 1969; Koopmans-van Beinum, 1980) is the result of averaging many different coarticulatory shifts. The center of gravity for a representative sample of consonants seems to be situated in the center of the vowel triangle. More reduction would then mean that the average distance to this center of gravity would be smaller due to more coarticulation. For individual consonant-vowel combinations, the direction of change with reduction could still be different, resulting in a divergence of the formant frequencies of reduced vowel realizations from different contexts (Van Bergem, 1993; especially his figure 7). Only the average change of many different consonant-vowel combinations would be towards centralization.

There is a practical side to the problem of the relation between target-undershoot (e.g. coarticulation and vowel reduction) and prosody, speaking style, and speaking rate. In order to synthesize speech with a natural sounding prosody, variation in the duration of phonemes is necessary. Furthermore, style and rate of synthetic speech should fit the task it is used for. This is important in order to become acceptable for the public. It is therefore important to know how prosody, speaking rate, and speaking style influence the spectro-temporal characteristics of natural speech. Neglecting these changes in synthetic speech may impart naturalness, intelligibility, and, worst of all, acceptation by the intended users.

1.1.1 The classical model of vowel target-undershoot

Coarticulation and reduction are changes in the patterns of movements of the articulators (e.g., tongue, lips, jaw). For vowels, these changes can generally be described as undershoot. The articulators stop before reaching their canonical target position. However, it is very difficult to measure the actual movements of the articulators. Therefore, it are the spectro-temporal features of the uttered sounds that are generally analyzed (e.g., formants). For the study of coarticulation and reduction, both articulatory and formant analysis are expected to give the same results because both are expected to stop short of reaching their canonical targets (e.g., Lindblom, 1963).

Lindblom (1963) found that there was a direct relation between the duration of a vowel realization and the amount of undershoot as determined from the first three formants. He gave a formula linking vowel duration and target-undershoot for each of these formants (equation 1.1).

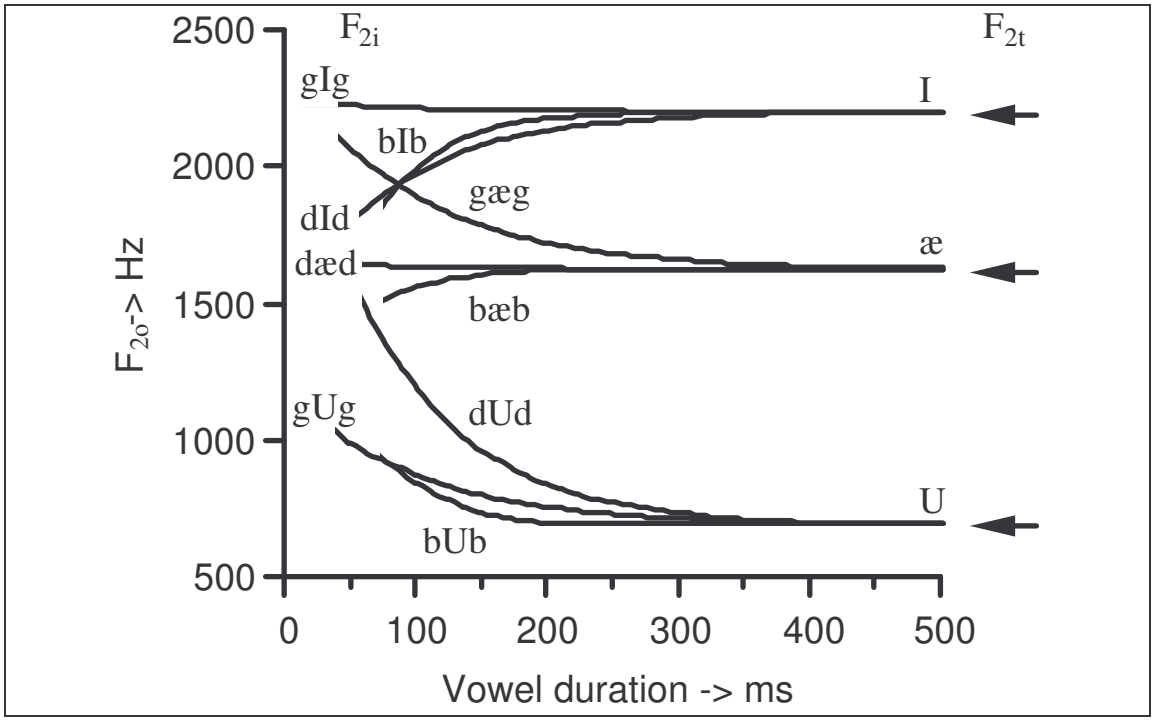

Figure 1.1. The effect of vowel duration on F_2 target-undershoot.
The relation between vowel mid-point value (F_{2o}) and vowel duration, as described by equation 1.1, is illustrated in this example taken from Lindblom (1963). The vowel formant target values (F_{2t}) are indicated by the arrows on the right. Each track starts at the point of "complete assimilation" where vowel mid-point value and vowel onset value are equal, i.e. $F_{2o} = F_{2i}$.

$$F_{no} = k \cdot (F_{ni} - F_{nt}) \cdot e^{-a \cdot DUR} + F_{nt} \qquad [1.1]$$

in which

F_{no} = frequency of formant n (F_n) at vowel mid-point of a CVC
F_{ni} = initial value of F_n at the start of the vowel
F_{nt} = ideal vowel target for F_n
n = formant number (F_1, F_2, or F_3)
DUR = vowel duration, DUR > ln(k)/a
k, a = constants fixed per symmetric consonant environment

Equation 1.1 was derived by Lindblom from vowel realizations with durations between 80 and 300 ms. In this range of durations, undershoot increased considerably from long to short durations. Duration, F_{ni} and F_{no} were measured directly on the spectrograms. For F_{ni}, the average value over all 24 syllables of a certain type was used. The other parameters (i.e., F_{nt}, k, and a) were determined by fitting straight lines through convenient representations of the data points. All in all, equation 1.1 could explain about half of the variance in the data.

We have plotted the function value of equation 1.1 in figure 1.1 for the vowel mid-point value F_{2o}, using parameters determined by Lindblom (1963). The starting point of each line is the point where the vowel mid-point value equals the formant onset value, i.e. DUR = ln(k)/a and $F_{2o} = F_{2i}$. This can be considered to be a hypothetical point of complete assimilation were the consonant completely dominates the spectral structure of the midpoint value. Small changes in durations can have quite large effects on

the vowel mid-point values if vowel durations are already short. If vowel durations would become shorter than those at the (hypothetical) point of complete assimilation (i.e., DUR < ln(k)/a), the vowel mid-point value would "undershoot" the formant onset value (F_{2i}) according to equation 1.1. Therefore, the equation is invalid for these short durations. However, the duration for which this happens (40 - 75 ms, depending on context) is well within the range of possible vowel durations. This is a result of the fact that a fixed value was chosen for the formant onset frequency, F_{2i}. In reality, this onset frequency value changes for short durations (see Broad and Clermont, 1987).

This model of vowel production is called the target-undershoot model because it assumes that the articulators, and therefore the formants, generally fail to reach the canonical target at the vowel mid-point. The formulation of this model was inspired by a damped mass-spring analogy of the articulators (see Lindblom, 1983). In this analogy, undershoot was the result of a power limitation on the movements of the articulators. To reach the same articulatory position in less time would require an increased effort which speakers would not deliver normally. Note that in Lindblom's (1963) interpretation, undershoot is even found for vowel durations longer than 200 ms. This means that articulation speed or effort would be the limiting (and decisive) factor in vowel production even at normal speaking rates.

In Lindblom's (1963) experiment, both consonants in the CVC' syllables were identical plosives (i.e., C=C'). Therefore, the formant onset value in equation 1.1, F_{ni}, could just as well be replaced by the formant offset value (called F_{nf}). Broad and Fertig (1970) found that for /È/, the formant tracks of Consonant-/È/-Consonant' (C/È/C') syllables with mixed consonants could be (re-)constructed by summing independent C/È/ and /È/C' tracks. This was used by Broad and Clermont (1987) to find functions that describe the CV, VC', and CVC' tracks for any combination of consonants and vowel. The vowel on- and offglide formant tracks were modelled by functions akin to equation 1.1. Equation 1.2 gives their complete formant contour as a function of time. We rearranged some terms to give it the same appearance as equation 1.1 (we combined figure 10 and equations 38 and 39 of Broad and Clermont, 1987). Note that equation 1.2 describes the course of a single formant track whereas equation 1.1 describes only the mid-point values. Also, the parameter "k" has different meanings in equations 1 and 2.

$$F_{CVC'}(t) = -k_C \cdot (L_C - T_V) \cdot e^{-B_C \cdot t} + -k'_{C'} \cdot (L'_{C'} - T_V) \cdot e^{B'_{C'} \cdot (t-DUR)} + T_V \qquad [1.2]$$

in which

$F_{CVC'}(t)$	= formant value at time t in a CVC' syllable
T_V	= vowel formant frequency target
L_C, $L'_{C'}$	= initial and final consonant formant locus
k_C, B_C	= initial consonant specific scale factors
$k'_{C'}$, $B'_{C'}$	= final consonant specific scale factors
C, C'	= initial and final consonants respectively
t	= time from start of the vowel, $0 \bullet t \bullet DUR$
DUR	= total vowel duration

To obtain values for the parameters in equation 1.2, $F_{CV}(t)$ and $F_{VC'}(t)$ values were measured for all vowels and consonants. All parameters were estimated by fitting contours to the appropriate data points. Only T_V had also been measured directly, but only for comparison. For equation 1.2 an estimated value of T_V was used. It must be noted that the locus values in equation 1.2 were not considered to be the formant track start- or end-points or extrapolations of the formant tracks. To quote Broad and Clermont (1987, p156): "*... our locus concept generalizes these boundary-oriented definitions [of consonant loci] to involve (1) the whole vowel contour and not just the part near an end-point, and (2) a scaling relation among a set of contours and not just a single contour*". In their approach, for every consonant, a baseline frequency was calculated for which all the formant contours of the various Consonant-Vowel (or Vowel-Consonant) transitions were scaled versions of each other. This baseline frequency was defined as the locus of the consonant. The amount of variance explained by the contours measured for given consonantal loci was not reported. It was only stated that the errors were typical of the order of 1% of the average value.

From the results of Broad and Clermont (1987) it can be inferred that the formant onset frequency (i.e., $F_{CVC'}(0)$) was equal to $T_V - k_C \cdot (L_C - T_V)$, apart from a correction term depending on the vowel duration and the final consonant. With increasing duration, onset frequencies shift due to the waning influence of the final consonant. Using their table VI, shifts of up to 150 Hz can be calculated for the F_2 onset frequencies, when duration increases from 100 to 150 ms (for a /dad/ syllable). This must be contrasted with the assumption, used in equation 1.1, that the formant on- and offset values were fixed. The preceding argument can be made, mutatis mutandis, for the vowel formant offset frequencies.

Broad and Clermont (1987) did not give a formula for the relation between formant-undershoot and duration. However, this formula can be derived in a straightforward manner from equation 1.2 and is given here as equation 1.3 for comparison.

$$T_V - F_{CVC'}(t_{extreme}) = \{k_C \cdot (L_C - T_V) \cdot e^{-B_C \cdot d} + k_{C'} \cdot (L'_{C'} - T_V) \cdot e^{-B'_{C'} \cdot d}\} \cdot e^{-a \cdot DUR} \qquad [1.3]$$

as equation 1.2 but with:

$t_{extreme}$ = the point with $\min(|T_V - F_{CVC'}(t)|), 0 < t_{extreme} < DUR$

$a = B_C B'_{C'} / (B_C + B'_{C'})$

$d = \ln\{(L_C - T_V) \cdot k_C \cdot B_C / ((L'_{C'} - T_V) \cdot k'_{C'} \cdot B'_{C'})\} / (B_C + B'_{C'})$; this factor disappears for symmetric syllables

$DUR > \max(-B_C \cdot d/a, B'_{C'} \cdot d/a)$. The undershoot is determined by the formant on- or offset values for still shorter durations

For equation 1.3, the formant-undershoot is defined as the smallest distance between the formant track and the vowel target value (i.e., $\min(|T_V - F_{CVC'}(t)|)$). Equation 1.3 is only valid if the point where this minimal distance is reached (i.e., $t_{extreme}$) is a global maximum or minimum and is positioned inside the vowel realization, i.e. is not the vowel on- or offset. Equation 1.3 is a more general formulation of equation 1.1; it weights the contributions of different initial and final consonants. The weighting scale

factor "d" depends on the quotient of the formant on- and offset slopes. In a completely symmetrical syllable with identical (apart from sign) on- and offset slope sizes (i.e., d = 0), equation 1.3 reduces to equation 1.1 (with k = k_C+$k'_{C'}$ and a = $B_C B'_{C'}/(B_C+B'_{C'})$). However, equation 1.3 uses estimated consonant-specific locus values (i.e., L_C, $L_{C'}$) instead of the averaged vowel onset values used in equation 1.1 (i.e., F_{ni}). The vowel onset values used by Lindblom depended on both the consonant and the vowel. It is possible to calculate for each set of measurements equivalent syllable scale factors (e.g., $k\cdot(F_{ni}-F_{nt})$ in equation 1.1) and reciprocal duration constants (i.e., "a" in equations 1 and 3). However, the methods with which formant frequencies and duration were determined and the way the estimations of the parameters were optimized differed considerably. Therefore, it is difficult to compare the results of both studies directly.

1.1.2 Interpretations of the target-undershoot model

The choice by Lindblom (1963) of an undershoot function that decays exponentially with duration was inspired on a mechanical analogy for the articulators: a (critically) damped mass-spring system (Lindblom, 1983). Broad and Clermont (1987) set out to test the underlying hypothesis that the formant tracks themselves were also exponential functions of time. If the articulators would behave like a damped mass-spring system, articulator position should indeed show precisely such an exponentially decaying behaviour (see equation 1.2). But if the formant tracks and the articulator position both behave according to such a function, this would indicate a linear relation between the positions of the articulators and the resulting formant frequency. However, there is no evidence for such a linear relation. Therefore, there is no reason to expect that articulators that behave like a (critically) damped mass-spring system will result in formant tracks like those described by equation 1.2.

A damped mass-spring system could in itself be a good model of the articulators. However, at the moment there is no reason to assume that the articulators are critically damped and that they are driven by simple, block-like power functions (as is assumed by Lindblom, 1983). It must be emphasized that, in general, the choice of a function to model a given set of data-points, like the formulations of equation 1.1-1.3, is one of convenience, e.g. a good fit of the data. Such a choice is arbitrary unless it can be validated by an actual understanding of the dynamics of speech. Till then, we must treat equations 1.1-1.3 as descriptive of the data. They cannot be used to explain the process of articulation.

As can be inferred from equation 1.1, Lindblom (1963) concluded that the undershoot of the vowel mid-point values in connected speech could be interpreted as an increase in coarticulation forced by a decrease in duration. It is evident from equations 1.2 and 1.3 that Broad and Clermont (1987) followed him in this. If we abstract from the exact formulations that were chosen in these studies, we can conclude that they both forwarded strong evidence for formant-undershoot that increased exponentially with shorter vowel durations.

In the initial formulation of the target-undershoot model, vowel reduction was interpreted as the combined result of all coarticulatory processes, i.e. vowel reduction is identical to coarticulation (Lindblom, 1963). Other authors disagreed with this interpretation of vowel reduction and vowel reduction itself has been the focus of a lot of studies since (e.g., Delattre, 1967; Koopmans-van Beinum, 1980). Subsequent formulations of the target-undershoot model incorporated some form of overall reduction of vowels as an independent process (Lindblom, 1983).

In a study in which he showed that vowel reduction depends on the language of the speaker, Delattre (1967) pointed out that Lindblom (1963) had only given proof that there exists a relation between vowel duration and coarticulation. He had not presented proof that duration was the independent forcing factor. Still, Lindblom's (1963) conclusion that coarticulation (or reduction as he also called it) is caused by vowel duration was (and is) widely quoted (e.g., Stevens and House, 1963, note 5 on p.123; Öhman, 1966; Verbrugge et al., 1976; Gay, 1981; Miller, 1981a; O'Shaughnessy, 1987, p.113; Duez, 1989; Fox, 1989; Krull, 1989; Nearey, 1989; Strange, 1989a, b). Since its early formulation, the target-undershoot model has been modified by Gay (1981), Lindblom (1983), and Lindblom and Moon (1988) to include speaking effort, articulatory strategies, and speaking style as factors that will modify the effect of duration on the amount of coarticulation and vowel reduction.

The target-undershoot model makes some pertinent and testable predictions. When vowel realizations get shorter, the articulators have less time to complete their movements from one phoneme target to the other. The target-undershoot model assumes (often implicitely) that speaking effort will not be increased enough to compensate for this loss of time. As a result, the articulatory positions that are actually reached in a sequence of phonemes will be drawn closer together, increasing coarticulation. Also, the articulators will travel shorter distances, resulting in levelled-off formant frequency tracks (after normalization for duration), which means that formant frequency excursion sizes diminish. Furthermore, on average, vowel realizations will lie closer to the center of vowel space and vowel realizations will be more reduced (i.e., centralized). However, whether or not centralization is likely depends on the actual distribution of the consonants in the utterance.

1.1.3 Is undershoot the result of articulatory limitations or is it planned?

A multitude of studies have been performed to test the predictions of the target-undershoot model. The results so far are rather ambiguous. The initial idea was that the relation between formant-undershoot and duration could be described using only the distance between the vowel target value and some starting value, i.e. the on- or offset as in equation 1.1 or the consonant locus as in equation 1.3. This starting value is implicitly assumed to be related to the movements of the articulators or to the place of articulation. This idea was supported by the studies of Lindblom (1963), Broad and Fertig (1970), and Broad and Clermont (1987). However, Lisker (1984)

found that high-F_1 vowels (/E œ/) before voiceless stops (/p k/) were shorter than before the corresponding voiced stops (/b g/) and at the same time had higher F_1 values, i.e. shorter realizations showed *less* undershoot than longer ones. If voicing did not change the place of articulation of these stops, this effect would amount to decreasing duration inducing formant-overshoot instead of undershoot. Whalen (1990) challenged the mechanical nature of coarticulation in the target-undershoot model. He presented subjects with words they had to read aloud. Initially, each subject only saw the part of the word up to the vowel of interest. The postvocalic part was only shown after the subject had started to pronounce the vowel. The subjects were able to articulate the words smoothly, but without any anticipatory coarticulation, neither for consonants nor for vowels. He concluded that "*Coarticulation ... is largely a result of planning an utterance rather than an automatic consequence of successfully producing an utterance*" (Whalen, 1990, p.29).

The target-undershoot model linked vowel reduction in unstressed syllables to their short duration. Unstressed vowels proved to be considerably reduced and shorter in most studies (Lindblom, 1963; Gay, 1978; Koopmans-van Beinum, 1980; Engstrand, 1988; Van Bergem, 1993). But some studies found that the duration of unstressed vowels was decreased without an increase in reduction or coarticulation (Den Os, 1988; Fourakis, 1991) or that unstressed vowels were reduced without being shorter (Nord, 1987), for instance in word-final position. This shows that vowel reduction in unstressed syllables can be decoupled from their duration. Therefore it is unlikely that the reduction is completely *caused* by the decrease in duration.

A final test case for the target-undershoot model is the effect of speaking style and rate on coarticulation and reduction. It is known that speaking style strongly affects vowel pronunciation (Koopmans-van Beinum, 1980; Lindblom and Moon, 1988; Moon, 1990). In general, it can be said that the more informal the speaking style, the more reduced and the shorter vowel realizations become (often referred to as sloppy pronunciation). Most studies find that an increase in speaking rate increases undershoot, both articulatory (Gay et al., 1974; Kuehn and Moll, 1976; Flege, 1988) and spectrally (Lindblom, 1963; Den Os, 1980; Gopal and Syrdal, 1988). But the effect proved to be speaker specific (Kuehn and Moll, 1976; Den Os, 1980; Flege, 1988). Some of the subjects in the latter studies did not show an increase in articulatory or formant-undershoot at a fast speaking rate. Other studies did not find any increase in formant-undershoot with speaking rate for their speakers (Gay, 1978; Engstrand, 1988; Fourakis, 1991). The fact that the effects of speaking rate are speaker specific is generally explained as a result of different articulatory strategies (Kuehn and Moll, 1976; Gay, 1981; Lindblom, 1983).

An additional problem with the results of the studies mentioned above might have been the inherent vagueness of the instruction to speak fast. Some speakers might have interpreted it as a request to speak more casual or sloppy, which often would also have been faster. Others might have decided that they should also hyper-articulate. In both cases, apart from speaking rate, style would also be different (e.g. see discussion in Van

Bergem, 1993). In these studies the (carrier) sentences that were used were quite short. Neither the task nor the conditions would have prevented the speakers from pronouncing them in any style they saw fit, from the most casual to clearest of oratorical. In none of the papers were the effects of speaking rate on speaking style explicitly evaluated.

1.1.3.1 *Input-driven versus output-driven control of articulation*

Most studies discussed so far used vowels in only a very limited context. Furthermore, vowels were often embedded in semantically empty syllables or carrier sentences. Such arrangements could influence pronunciation (see discussions in Lindblom and Moon, 1988; Van Bergem, 1993). Context, task, and speaking conditions were generally incompatible between studies. All this makes it very difficult to compare the results of different experiments and to generalize from a restricted environment to natural speech.

The target-undershoot model is based on a simple mechanical analogy. It does not account for the way reduction and durational differences function in normal speech. Word-stress, word-class, and sentence-accent all influence duration and reduction (Koopmans-van Beinum, 1980; Van Bergem, 1993). Word-stress influences word meaning, e.g. the difference between "*to permit*" and "*a permit*" depends on which syllable of the word "*permit*" is stressed. It is known that vowel reduction can change stress assignment on its own (Rietveld and Koopmans-van Beinum, 1987). Sentence-accent is linked to the syntax of the sentence. There is also a difference between words containing "old" information and "new" information (Eefting, 1991) and there could be a relation between the amount of vowel reduction and the frequency of occurrence of a word (as suggested by Van Bergem, 1993). On the other hand, speaking style seems to be related to the intentions of the speaker and to the relation between speaker and audience. A change in speaking style generally indicates a change in these factors. For instance, if a speaker thinks s/he is not understood well, s/he will speak more clearly (Lindblom and Moon, 1988; Moon, 1990).

This complex interplay of factors simultaneously influencing reduction and duration can make that the requirements on vowel reduction and duration clash. This was used by Nord (1987) to produce stressed and unstressed syllables with vowels of equal duration. It is revealing that he found that the degree of reduction in unstressed syllables did not depend on vowel duration. Unstressed vowels were always more reduced than stressed ones. This shows that reduction is linked more to stress than to duration.

Associated with this is the relation between vowel duration and reduction in cases where duration is a part of the vowel identity, as for intrinsically long vowels. In the literature cited above, no reference was made to whether the target-undershoot model also operates on the durational differences found in long-short vowel pairs, i.e. vowel realizations that change identity together with duration. A naive interpretation of the target-undershoot model would predict that realizations of short vowels are more reduced and coarticulated than the corresponding realizations of long vow-

els. However, there was no evidence for this in the studies of Koopmans-van Beinum (1980) and Van Bergem (1993) on Dutch vowels.

The problem about how to explain the variability of vowel realizations in natural speech, centers on how articulation is controlled. The studies discussed above all centered around articulatory and formant-undershoot, incorporating both coarticulation and reduction. Abstracting from all other questions, the models discussed can be interpreted as defining the level of flexibility of articulation and control over articulation. As Whalen (1990) pointed out, the relevant question here is to what extent articulation is planned, and to what extent it is the result of mechanical constraints.

At one extreme there is the position that articulation is organized in programs of fixed patterns of mechanical articulatory actions, more or less like acquired reflexes. These patterns of articulatory actions roughly correspond to phonemes or phoneme transitions. When the programs are triggered, the course of the articulatory actions is fixed and cannot be controlled. In a quick succession of phonemes, the actions start to overlap, i.e. a new program is started before the old one is completed. This leads to undershoot. The extent to which the articulatory actions are completed depends on the time available, i.e. phoneme duration, and the effort invested. To summarize this position, there is no flexibility in the articulation and speakers can only control the global speaking effort and the relative timing of triggering individual patterns, but not their course of action. The articulatory movements are solely determined by the "input" of the articulatory system. Therefore, such a model can be called "input-driven".

The other extreme is that speakers always adapt their articulatory movements to ensure the production of the *intended* sound. In other words, articulation is planned in advance to produce the desired output. There might even be a constant feedback that leads to "on-line" adaptation of articulatory movements. This model is "output-driven", articulatory movements are adapted to produce the desired output.

In the input-driven model, all variation in speech sounds is the predictable result of clashes between articulatory programs. In the output-driven model, the variation in speech sounds is the result of planned differences between realizations. Figure 1.2 describes graphically how duration will or will not influence vowel formant track shape according to the input- and output-driven models. Both extremes are untenable in their pure form and most studies take a middle-stand, only putting more emphasis on the one or the other. The original target-undershoot model (Lindblom, 1963; but also Broad and Clermont, 1987) comes close to a purely input-driven model. Whalen (1990) concluded that coarticulation is to a large extent planned. Delattre (1967) emphasizes the importance of language in the reduction of vowels, suggesting that this reduction is intended and not mechanical. These latter two studies emphasize the output-driven aspects of speaking. In general, studies on coarticulation stress the limitations of the mechanical articulatory process which would lead to a largely input-driven articulatory model. Studies on vowel reduction on the other hand, generally assume implicitly that reduction is somehow intentional, i.e. largely output-driven. Coarticulation and reduction might be different names for

the same process as suggested by Van Bergem (1993), but authors often seem to choose the name according to their conviction about its causes.

The question whether coarticulation and reduction are exclusively input-driven *or* output-driven might be unanswerable. A relation between duration and undershoot that is the result of articulatory constraints at short durations, may have been incorporated in the language and might be reproduced "voluntarily" for longer durations. Such a relation would be planned in longer utterances and mechanically determined in shorter utterances. There could also be other problems. It is possible that whenever mechanical limitations interfere with the desired output, speakers will increase the durations to compensate for it. It will be difficult to demonstrate mechanical limitations unequivocally if the durations always tend to match the desired output.

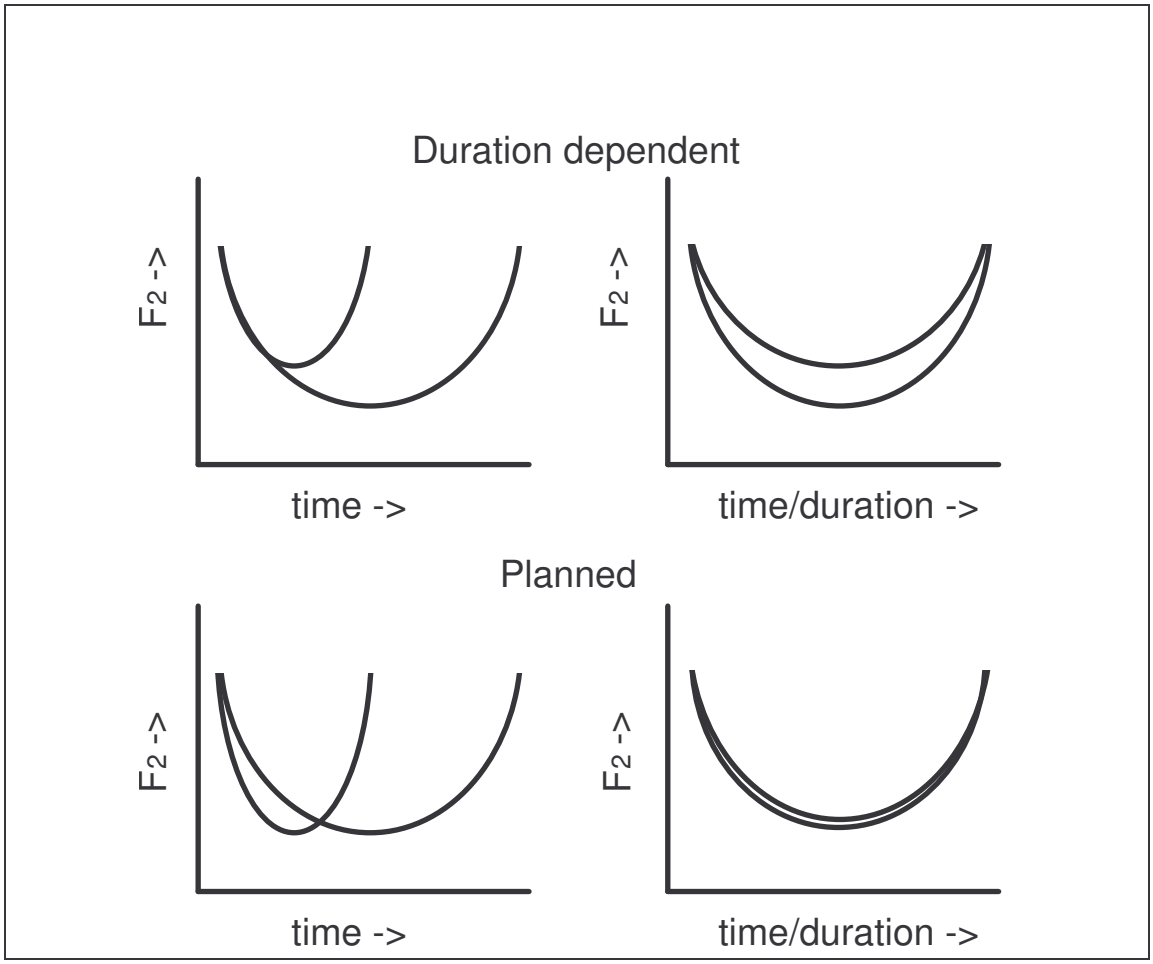

Figure 1.2. The influence of vowel duration on formant track shape. Tracks from two vowels with different durations are depicted for an input-driven model, i.e. duration-dependent undershoot (with excess undershoot, top row), and for an output-driven model (with no excess undershoot, bottom row). The panels on the left give formant tracks in real time (frequency versus time). The panels on the right show the formant tracks when they are normalized for duration (frequency versus time/duration). The two tracks in the lower right panel were displaced a little for clarity. Ideally, they should have been completely identical.

1.1.3.2. *Testing the target-undershoot model*

To study the way vowel duration drives coarticulation and reduction, vowel realizations should differ only in duration. Vowel realizations should be identical in all other respects to prevent different "planning" targets to interfere. The natural variation in vowel duration is strongly coupled to other features of speech that are known to influence vowel spectra, like stress and context. It is difficult to control all these factors and still elicit variation in vowel duration. One possibility is to vary word length or position in the word. An example of the control by way of word length is the initial /È/ that shortens in the sequence will-willing-Willingham (Lindblom and Moon, 1988; Moon, 1990). Examples of control by way of word position are the differences in vowel duration found in word-initial and word-final stressed and unstressed syllables (Nord, 1987). However, the basis of these phenomena is not completely clear and might be a prosodic change that in itself could influence vowel reduction and coarticulation. Furthermore, these methods rely on the construction of special, often artificial, words.

This severely limits the amount of speech that can be used. Using unfamiliar or unknown words might induce an extra clear speaking style. Therefore these methods are not practical if the speech uttered should be close to natural, or at least should be close to normal read speech.

It is much easier to obtain vowel realizations that differ only in duration when different speaking rates are used. The speaker is instructed to speak each utterance with the speaking rate of interest. At the same time care must taken to ensure that speaking style does not change. This keeps context, stress and all other factors nearly identical for every realization of the utterance. As speaking rate in itself does not change the relation between speaker and listener or the circumstances in which the speech is uttered, it should have a minimal effect on any "planned" variation. If a "reading-style" is chosen, a long, normal text can be used. Such a long and normal text will supply vowel realizations from a context that is representative of the language. At the same time, because of its length, a long text will prevent short-term adaptations of articulation strategies to difficult speaking conditions. Such short-term adaptations were suggested to explain the lack of reduction often found in fast rate speech (Kuehn and Moll, 1976; Gay, 1981; Lindblom, 1983). Furthermore, when reading a long text fast, the speakers will be inclined to use a normal reading style. It is difficult to use an unusual speaking style consistently for several minutes when one has also to perform a second task: that of reading. In addition, for a long text, any deviation from normal reading will be obvious to the experimenter. Therefore, it can be ensured that the speaking styles of both readings are (nearly) identical.

Therefore, in our studies we used speaking rate as a variable to determine whether vowel duration is the factor that drives vowel reduction and coarticulation. A long natural text spoken at a fast rate should show more coarticulation when individual vowel-consonant combinations are inspected and should show more centralization of vowel realizations (i.e., more reduction) when averaging over large, representative samples of vowel-consonant combinations.

Reading aloud long texts is a difficult task (see e.g., Eefting, 1991). To be able to read aloud a text twice (at different speaking rates) without too many errors, while keeping stress assignments comparable in both readings, requires a lot of practice. Therefore, we limited our studies to the speech of a single, very experienced, speaker who could accomplish this task. We already know that the articulatory responses to an increase in speaking rate are speaker dependent (Kuehn and Moll, 1976; Den Os, 1988; Flege, 1988). This means that our results cannot be extrapolated to the general population. However, the target-undershoot model (nor any other model of vowel production) does not make reservations regarding the person of the speaker. It claims universal validity and should be applicable to any speaker's utterances. This means that any, non-aberrant, speaker that does not conform to this model could disprove it.

In our experiments, planning of coarticulation and reduction should reveal itself through the fact that, after time normalization, speaking rate has no influence on either of them. Most factors that would otherwise influence coarticulation and reduction other than vowel duration itself, e.g.

stress or speaking style, would now remain unchanged. However, if the mechanical limitations of articulation are more important, the decrease in vowel duration should induce more coarticulation and reduction in fast-rate speech than in normal-rate speech (see figure 1.2).

However, it is theoretically possible that an increase in speaking effort would compensate for the higher speaking rate (e.g., Gay, 1981; Lindblom, 1983; Lindblom and Moon, 1988). From a global increase in speaking effort we would expect either some residual target-undershoot from inadequate compensation or target-overshoot due to hyper-articulation (i.e., over-compensation). If we would not find any formant-undershoot or overshoot in fast-rate speech, this would mean that our speaker had changed his speech to match exactly his intentions, i.e. that his speech is output-driven.

These predictions lead to two potentially independent questions to investigate.

- Is the vowel mid-point or nucleus showing more spectral reduction or coarticulation in fast-rate speech than in normal rate speech?
  This is investigated in chapter 2.
- Are formant tracks of fast-rate vowels more level than those of normal-rate vowels, indicating that articulation movements are shorter in fast-rate speech due to changes in the vowel mid-point and/or on- and off-set positions?
  This is investigated in chapters 3 and 4.

1.2 Perceptual-overshoot and dynamic-specification in vowel identification

In the previous sections we discussed how vowel realizations are influenced by context, prosody and speaking style. We can add to this the variations in pronunciation that exists between individual speakers. Together, these factors induce a high level of variability in vowel pronunciation. This variability could give the impression that vowels are difficult to recognize in normal, connected speech. But, in a normal utterance, vowels are generally identified accurately, whatever the context or speaker characteristics. This raises the question of how listeners accomplish this feat (at the moment, machines cannot). Models of vowel perception try to answer this question by looking for acoustic features in vowel realizations that are invariant to coarticulation, reduction, and speaker identity.

In general, models of vowel perception are tied to models of vowel production. The simple target-undershoot model discussed above inspired the development of a complementary model for vowel perception. In this perceptual model, listeners would compensate for undershoot in production by overshoot in perception. The hypothetical canonical formant target value that was not reached due to target-undershoot could be determined (i.e., calculated) by extrapolating the formant tracks in the Consonant-Vowel (CV) and/or Vowel-Consonant (VC) transition. It is also possible that vowel duration is used together with the "distance" between the vowel realization and its context to factor out the undershoot without a direct recourse to a

dynamical perceptual-overshoot (Nearey, 1989). In this latter case, the listener needs to relate the amount of undershoot to the duration of the vowel.

The perceptual-overshoot theory was first proposed and tested by Lindblom and Studdert-Kennedy (1967). They studied synthetic /wVw/ and /jVj/ syllables with parabolic vowel formant tracks. From subject's responses they derived those F_2 values for which an /U/ percept changed into an /È/ percept (i.e., from a vowel with a low F_2 to one with a high F_2). These F_2 cross-over values were lower in a /wVw/ context with a rising-falling F_2 track than in a /jVj/ context with a falling-rising F_2 track. In short, the targets that were reported by the listeners had markedly overshot the mid-point values that were actually reached in the stimuli (i.e., cross-over value + overshoot = target value).

It is known that formant track shape and vowel duration do influence speech perception. These factors are important for the identification of adjacent consonants (e.g., Mack and Blumstein, 1983; Miller and Baer, 1983; Polka and Strange, 1985; Miller, 1981b, 1986; Nossair and Zahorian, 1991; Diehl and Walsh, 1989). Formant track slopes in the nucleus of the realizations also determine the perception of diphthongs (e.g., see O'Shaughnessy, 1987; Peeters, 1991 for overviews). It is therefore natural to expect that these factors will also influence the perception of the vowel realizations themselves. Perceptual-overshoot might be only one of several ways in which formant track shape and vowel duration contribute to vowel identification.

1.2.1 Dynamic-specification versus elaborate target models of vowel perception

In a general fashion, the variability of vowel realizations in speech poses the problem in what way listeners are able to identify these as belonging to the same phoneme. In general, it is assumed that vowel realizations contain invariant acoustical features that allows listeners to resolve their identity. It is maintained that if we could perform the right transformations on the acoustic signal, vowel identity would be unambiguous. Based on whether these invariant features are of a static or dynamic nature, theories on vowel perception can be divided into two "camps" (Strange, 1989a; Andruski and Nearey, 1992).

1) On the one side there are theories that claim that the spectrum at a single cross section in the vowel realization, i.e. the mid-point or nucleus, contains all necessary information that is used to identify it (e.g., Nearey, 1989; Miller, 1989; Andruski and Nearey, 1992). These theories are purely spectral and are called (elaborate) target-models. In these models, the variability in vowel realizations is dealt with by somehow "normalizing" the spectrum to a reference spectrum. The normalizing procedure generally involves combinations of formants and F_0 on a non-linear frequency scale.

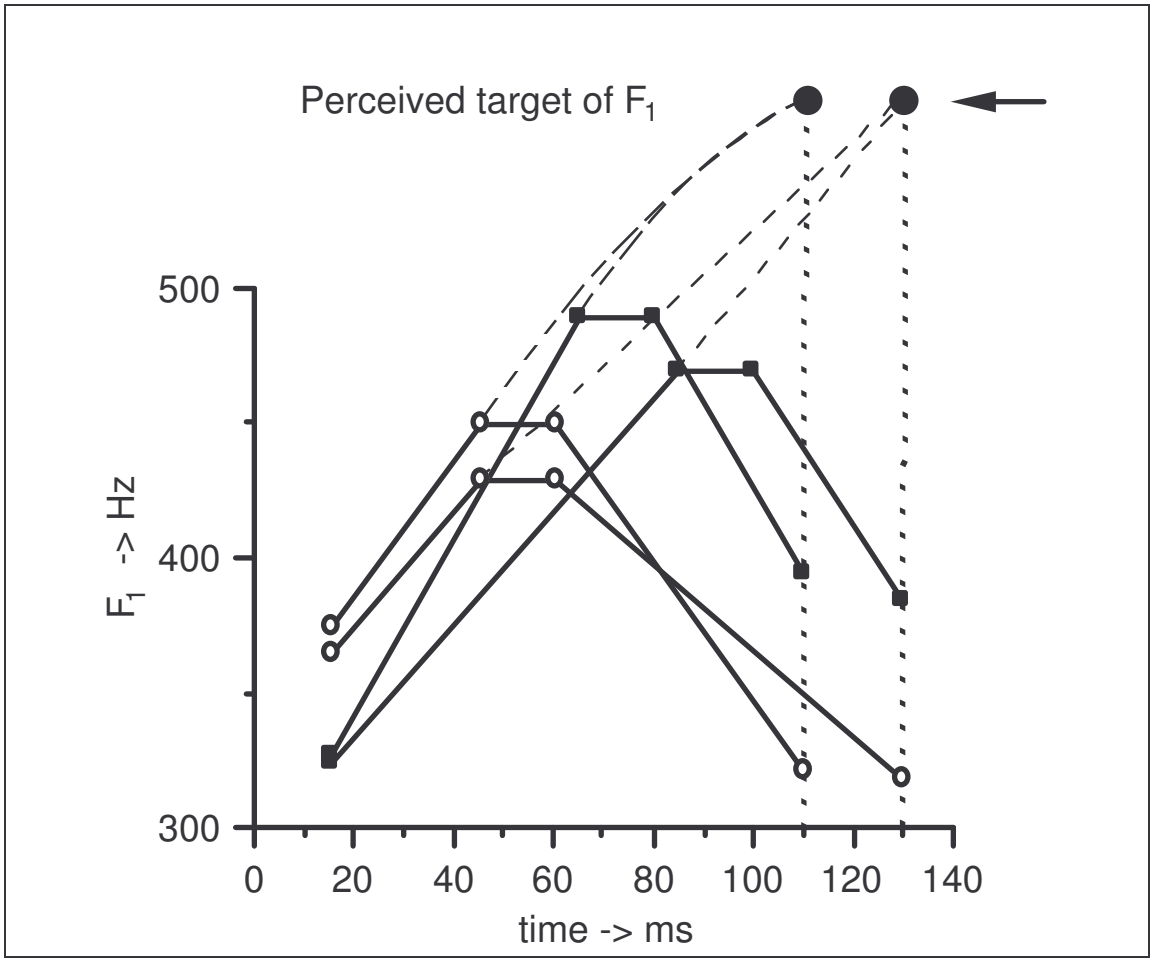

Figure 1.3. Perceptual-overshoot.
The F_1 tracks of four tokens are drawn in a frequency versus time plot. All four tokens lead to the same F_1 "target percept". This target was interpreted to be positioned beyond the maximal values reached in the tokens (indicated by the thin lines). Reproduced from Di Benedetto (1989b, figure 12b).

Vowel-inherent spectral changes, like diphthongization, are modelled by assuming a double, compound, target in the vowel nucleus instead of only a single target (Andruski and Nearey, 1992). Still, the transition parts of the vowel realizations (i.e., the vocalic parts of CV and VC transitions) do not influence vowel recognition according to these theories. Target-undershoot in production would change the spectral contents of the vowel mid-points depending on vowel duration. This could make it necessary to include duration in the normalization procedure in order for this procedure to compensate for the undershoot in production.

2) On the other side there are theories that acknowledge that dynamical information from parts outside the vowel nucleus is also used to disambiguate the information from the vowel nucleus itself (e.g., Lindblom and Studdert-Kennedy, 1967; Huang, 1991, 1992; Di Benedetto, 1989a, b; Fox, 1989; Strange, 1989a, b). These theories are spectro-temporal and rely on "dynamic-specification" to disambiguate the vowel realizations (also called dynamic-cospecification, Andruski and Nearey, 1992). It is assumed that the shape of a vowel formant track is indicative of the direction and amount

of (formant) undershoot. Knowing the amount of undershoot enables a listener to deduce the position of the canonical target of the vowel. A commonly proposed mechanism to achieve this is perceptual-overshoot.

As we already have seen, perceptual-overshoot is a (hypothetical) mechanism by which the listener extrapolates the course of on- or offset transitions into the nucleus of the realization, overshooting the actual mid-point values realized. The listener would perceive a mid-point value closer to the canonical target than the mid-point value actually realized acoustically. This would be a simple mechanism to achieve the aim of undoing the effects of target-undershoot in production. Therefore, it is often incorporated in dynamic-specification theories (e.g., Huang 1991, 1992; Di Benedetto, 1989b; Fox, 1989; Strange, 1989a; Akagi, 1990, 1993). An example of perceptual-overshoot is given in figure 1.3, which was reproduced from Di Benedetto (1989b).

However, it is not always necessary to assume a mechanism of perceptual-overshoot. The shape of the formant tracks (e.g., the slope and excursion size) is in itself informative and could be used to identify a realization. For instance, a large F_1 excursion size and a flat F_2 track could indicate an open vowel (like /a/) without any reference to hypothetical invariant target positions deduced from extrapolating the formant on- and offglide tracks.

1.2.2 Evidence pro and contra dynamic-specification

Evidence for the use of dynamic-specification in vowel recognition comes from several studies. It was noted that coarticulated vowel realizations in a CVC context were identified better, or at least not worse, than vowels spoken in isolation (see discussions in e.g., Strange and Gottfried, 1980; O'Shaughnessy, 1987, p.177; Fox, 1989; Nearey, 1989; Strange, 1989a; Andruski and Nearey, 1992). Also, vowel realizations from which the kernel was removed (silent-center vowels), leaving only the Consonant-Vowel and Vowel-Consonant transitions up to the border of the kernel, were recognized better than the isolated kernel parts alone. Recognition of silent-center vowels was generally only moderately compromised and sometimes recognition was even indistinguishable from that of complete syllables (Strange, 1989b; p.2144). Even when the initial and final transition parts of the silent-center vowels were from speakers of opposite sex, the number of errors remained quite low (Verbrugge and Rakerd, 1986). In all these cases, the vowel mid-point spectrum differed strongly from the canonical case (i.e., vowels pronounced in isolation) or was even absent altogether. This fact did not seem to bother the listeners and as long as the transition parts were present, recognition was hardly compromised. Fox (1989) even found that reducing the transitions in synthetic silent-center realizations to the outermost single pitch period still allowed quite accurate vowel identification.

In a completely different set of experiments, Di Benedetto (1989b) concluded that F_1 transitions and timing were used to distinguish between high (/i È/) and non-high (/e E/) vowels (1989b; her terminology). She discussed perceptual-overshoot as a possible explanation (see figure 1.3) but could not rule out the possibility that her subjects had used a weighted av-

erage of the F_1 contours. Support for dynamic-specification also came from the fact that information about formant track shape could help to distinguish realizations of different vowels with comparable F_1 mid-point or extreme values (Di Benedetto, 1989a; Huang, 1991, 1992).

Andruski and Nearey (1992) interpreted the above evidence in a different way. They concluded that there was no compelling need for dynamic-specification to explain it. Their arguments can be summarized as follows. The initial reports that vowels in context were actually recognized better than isolated realizations could not be confirmed in subsequent studies (e.g., Macchi, 1980; Nearey, 1989; see also discussion in Strange, 1989a). What could be attested was the fact that vowels were recognized equally well in both conditions. But this could also be explained with (compound) target-models. It could also be argued that splicing out the vowel kernel to create silent-center vowels left enough spectral information (e.g., the transition end-points) to identify them without using dynamical information from the CV and VC transitions (this argument was also discussed by Fox, 1989). Finally, the results of Di Benedetto (1989a) about the differences between F_1 transitions in high (/i È/) and non-high (/e E/) vowels from natural speech, can also be interpreted as merely revealing the diphthongized nature of some of these vowels in American-English. The results of her perceptual experiments with synthetic vowels did not distinguish between dynamic-specification and target-models (1989b). Therefore, both studies do not allow to say unambiguously that she has found perceptual-overshoot or dynamic-specification in general.

It is disconcerting to find that an important question as to whether dynamic features of vowels influence their identification cannot be answered unambiguously after so much research. The source of the ambiguity in the results of so many studies has to be known before we will be able to interpret the results of our own experiments (see chapter 5). In chapter 6 we will return to this question and take a closer look at the available literature. We will try to find an answer to the question of what factor(s) in these experiments caused or prevented listeners to compensate for coarticulation or reduction, i.e. in what circumstances we can expect to find perceptual-overshoot and dynamic-specification.

1.2.3 Distinguishing models of vowel perception

A key question in the controversy described above is how vowel identity is affected by vowel duration and formant track shape, if it is affected at all. We could ask whether listeners do compensate for expected undershoot in production and whether they use the information present in the formant transitions to perform this compensation.

In general, dynamic-specification is expected to work in the same direction as perceptual-overshoot. The shape of a formant curve always signifies a vowel with a target on or beyond the mid-point value actually reached. There are no reports of contexts for which the formant mid-point value of any vowel would systematically overshoot the target it reaches when pronounced and sustained in isolation (see section 1.1 above). For example, an open vowel (like /a/) is generally characterized by a strongly curved, rising-

falling F_1 track. The (canonical) F_1 target of this vowel can be found by extrapolating the on- or offglide of this same track. In a first approximation, both the strongly rising-falling curve shape and the target found by extrapolation will indicate an open vowel (i.e., a high F_1-target). Therefore, perceptual-overshoot and dynamic-specification predict the same behaviour of subjects: response targets should overshoot the mid-point values actually present in the tokens. The amount of overshoot should be related to the curvature of the formant tracks and the duration of the tokens.

On the other hand, target-models of vowel perception state that listeners use a cross-section to characterize the complete formant track. In practice, listeners are expected to take the average of some small part of the formant track. This should result either in subject responses that are independent of formant track shape, or alternatively, in some undershoot in strongly curved tracks due to the averaging process. A complicating factor is that listeners could use the wider context of the realization, instead of the formant track shape, to compensate for the *expected* undershoot in production. This would result in an apparent "overshoot" in the responses. However, because this apparent overshoot depends *not* on formant track shape (by definition), the overshoot would *only* depend on context and duration. Therefore, it should be easy to discriminate it from perceptual-overshoot and dynamic-specification.

The differences between models using dynamic-cospecification and target-models seem to hinge on the effect of formant track shape on the responses of the listeners. If the vowel identity is cospecified by the formant track shape, then the targets in the responses should *overshoot* the mid-point values actually present. Furthermore, if there is real perceptual-overshoot, the amount of overshoot should depend indirectly on token *duration*, i.e. a shorter duration with steeper formant slopes should induce more overshoot. However, if formant track shape is not used to specify vowel identity, both formant track shape and duration should have *no influence* on the responses of the listeners, save some *undershoot* due to perceptual averaging and an exchange of long- and short-vowel responses.

In our study we wanted to decide on this question. We investigated how formant track shape and vowel duration influenced vowel identification, i.e. if the responses of the listeners showed perceptual-overshoot or not. Perceptual-overshoot, if it exists, is used to compensate for the effects of coarticulation and reduction. There is a possibility that the listeners will treat vowels presented in isolation quite different from those presented in context. It could be that some change due to coarticulation or reduction must be plausible before listeners will actually use the mechanisms that should compensate for it. Therefore, it is important to check whether the presence of perceptual-overshoot depends on the presence of a non-silent *context*.

In natural speech, the variation in track shapes is limited and linked to other factors that also determine vowel identity. This problem can be controlled in synthetic speech (in this we followed Fox, 1989). Therefore, we opted for synthetic vowel realizations in which we could combine formant track shape, duration, and formant mid-point values in a systematic way.

In chapter 5 we investigate the following three related questions:

- Does a curved formant track shape induce overshoot in the responses of listeners or does it not?
- How does token duration influence vowel identity?
- Are vowel tokens identified differently when presented in simple context than when presented in isolation?

In chapter 6, we will examine the literature on vowel perception to see if we can integrate the results of the experiments presented in chapter 5 with the, often contradictory, results published in the literature. We will also try to find indications in the relevant papers of what might have caused superficially similar experiments to lead to opposing conclusions.

In the General Discussion (chapter 7) we will combine the results of the previous chapters. We will determine whether, for the speech used here, the size of the predicted duration-dependent target-undershoot was large enough to have been detected by the static measurements of vowel formants (chapter 2) and the dynamic point-by-point (chapter 3) and polynomial (chapter 4) analysis. We will weigh the evidence for input-driven and output-driven control of speech. The evidence for the use of dynamic-specification in vowel recognition will be discussed (chapters 5 and 6). Finally, we will try to link the characteristics of vowel production to those of vowel recognition.

2

FORMANT FREQUENCIES OF DUTCH VOWELS IN A TEXT, READ AT NORMAL AND FAST RATE*

Abstract

Speaking rate is thought to affect the spectral features of vowels. Target-undershoot models of vowel production predict more spectral reduction and coarticulation of vowels in fast-rate speech than in normal-rate speech. To test this prediction, a meaningful Dutch text of about 850 words was read twice by an experienced newscaster, once at a normal speaking rate and once as fast as possible. All realizations of seven different vowels and some realizations of the schwa (/ˊ/) were isolated. The first and second formant frequency values of all realizations were measured at five different points, each time by making cross-sections at different points in the vowel realization. The different selections of these points are based on procedures used in literature, such as maximal F_1 or mean formant value. No spectral vowel reduction was found that could be attributed to a faster speaking rate neither was a change in coarticulation found. The only systematic effect was a higher F_1 value in fast-rate speech irrespective of vowel identity. This possibly suggests a generally more open articulation of vowels, speaking louder, or some other general change in speaking style by our speaker when he speaks fast.

*Van Son, R.J.J.H. & Pols, L.C.W. (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", Journal of the Acoustical Society of America 88, 1683-1693.

Introduction

The effects of speaking rate on vowel production have been the objective of many studies (recent examples are e.g., Gopal and Syrdal, 1988; Den Os, 1988; Engstrand, 1988). Speaking rate is thought to affect most, though not solely, coarticulation and spectral reduction (Lindblom, 1963). Both of these are well attested phenomena that play an important role in normal speech (see e.g., the textbooks of O'Shaughnessy, 1987; Clark and Yallop, 1990). The effects of speaking rate on vowels are supposed to be examples of a more general influence of duration on the spectral structure of vowel realizations, an influence described by the target-undershoot model of vowel production, as formulated by Lindblom (1963), Gay (1981), and Lindblom (1983). This model predicts an increase in coarticulation, spectral reduction, or both, in vowels when their realizations shorten.

In its most simple form, the target-undershoot model states that vowels are characterized by their spectrum at a single point in the realization, the vowel target (see also Strange, 1989a). Due to several factors, a vowel realization generally has a target spectrum different from the ideal, or canonical, form. In the target-undershoot model this difference is said to shift the actual target spectrum from the canonical target toward the targets of the neighbouring phonemes (coarticulation) or towards a theoretical neutral vowel (spectral reduction). The articulators are said to miss the ideal target position by undershoot (Lindblom, 1963).

Several factors influencing vowel target spectra are identified and studied, for instance coarticulation (e.g., Pols, 1977; Whalen 1990), speaking style, stress and reduction (e.g., Koopmans-van Beinum, 1980). For the effects of duration on vowel formant frequency targets the results reported are ambiguous. At one hand, several studies support the notion of more target-undershoot with shorter vowel durations (Lindblom, 1963; Broad and Fertig, 1970; Gay et al., 1974; Broad and Clermont, 1987; Lindblom and Moon, 1988). Other studies, however, where unable to detect such an undershoot (Gay, 1978; Nord, 1987; Gopal and Syrdal, 1988; Den Os, 1988; Engstrand, 1988) or found the effect of speaking rate on vowel undershoot to be speaker dependent (Kuehn and Moll, 1976). It can be noted that support for the target-undershoot is mostly found when vowel realizations from only one speaking rate and style are studied, whereas it seems to be difficult to find support when differences between speaking rates or styles are studied. One reason for the ambiguity in the results of these studies might have been the experimental designs used in them. In all studies the speech is uttered under controlled conditions. The level of control often causes the distance between the experimental procedures and natural speech to be large and does not allow the results of these studies to be generalized to more normal modes of speech easily. Most studies used semantically empty words in carrier phrases, and the vowels are often placed in only a limited phoneme context. The experimental procedures used in different studies are often incompatible with one another and comparisons are therefore very difficult.

Three problems especially hamper investigations analyzing the influence of vowel duration on vowel target spectra. First, it is very difficult to elicit

vowel realizations with different durations without altering other factors like context and stress (but note the elegant method used by Lindblom and Moon, 1988), especially if the speech uttered should be close to natural. Second, there seems to be no consensus about how the position of the spectral target in a vowel realization should be determined, different studies use different procedures. For instance, the procedures to determine the point where the target spectrum should be measured of Lindblom (1963), Delattre (1969), Gay (1978), Koopmans-van Beinum (1980), Lisker (1984), Vaissiere (1987), Engstrand (1988), Den Os (1988), and Gopal and Syrdal (1988), all differ largely in definition. Third, there also seems to be a lack of consensus about the representation of the spectral structure of a vowel target. In general, the frequencies of the first two formants are used to characterize a vowel target spectrum. Beside differences in the way these frequencies are measured, there exists a number of ways to represent them (e.g., linear frequencies, logarithmic frequencies, Bark scales) and there is at the moment no reason to prefer one of them over the others for studies testing the target-undershoot model.

To address these problems, an experimental design was selected in which the factors that influence vowel reduction and coarticulation are optimally controlled, and the speech sample was as natural as possible. This was attained by using only a single, experienced, speaker who read a long, meaningful text twice, once at a normal speaking rate and once as fast as possible. A possible drawback of this approach is that stress and vowel context are inherent to that text and are inaccessible to manipulation without losing naturalness. A large collection of vowel realizations was obtained, almost all of which could be used to construct vowel pairs, containing realizations of the same text item at both speaking rates.

The use of only a single speaker could pose a problem if the effects of speaking rate are somehow speaker dependent, as Kuehn and Moll (1976) found. But in this study we are investigating the possibility that a single speaker (this may be any normal speaker) does NOT display any increased articulatory undershoot with an increased speaking rate. If changes in articulatory undershoot are not required in normal speech with an increase in speaking rate, there are profound implications for articulatory theory and research in automatic speech synthesis and recognition.

Vowel target formant values were measured using several procedures in parallel to select the target points in the vowel realizations. This way it is possible to determine whether the detection of durational effects on vowel targets depends on the definition of the targets themselves. The problem of the different representations of the formant frequencies is solved by using statistic tests that are insensitive to the representation of the data. These tests are unlike commonly used statistic tests whose results can be invalidated if, for instance, logarithmic values are substituted for linear values. We therefore will use these distribution-free statistic tests (tests based on rank, see Ferguson, 1981).

In this paper we will investigate whether our speaker produces vowels with more articulatory undershoot (spectral reduction or coarticulation) when he speaks at a fast rate than when he speaks at a normal rate.

2.1 Methods

2.1.1 Speech material

In this study, a long text of about 850 words was used. The text was originally used in a radio broadcast and was informative (concerning economics, see appendix C). The text was read by an experienced, over 60 years old, professional speaker who was selected for his good reading and whose voice was known to give good results with LPC analysis. He speaks the standard form of Dutch (Koopmans-van Beinum, 1980, male speaker #1).

The recorded speech is part of a larger body of speech (in total, 2.5 h of speech recordings) recorded in a 1-day session. The text was read twice. The speaker was instructed to read the text first as he would do for an audience, i.e. at a normal speaking rate. For the second reading, he was instructed to read it as fast as possible. The two readings of this text were done with several hours in between. The speaker was unaware of the specific aims of this project.

The speech was recorded on a commercial Sony PCM recorder, low-pass filtered at 4.5 kHz and digitised at 10 kHz, with 12 bit resolution. Subsequent storage, handling and editing were done in digital form only.

Reading this text took 330 s for the normal speaking rate and 220 s for the fast speaking rate. The overall reduction in duration of the fast-rate realization as compared to the normal-rate realization was one-third when pauses longer than 200 ms were included, and one-fourth when these longer pauses were excluded from both readings.

2.1.2 Segmentation

A waveform editing computer program was used to display the waveform and regenerate the sound of the stored vowels. The waveform and the audio signal were used to identify the boundaries of the vowels (see below). The vowel segments thus identified were copied with a leading and trailing edge of 50 ms of speech to ensure correct spectral analysis at the boundaries of the vowels.

The vowels for this study were selected from the original written text based on their orthographic form. Subsequently, the speech material was searched for realizations of the chosen vowels. Any vowel-like sound that could be attributed to the chosen realization was copied. Only a few vowels were completely absent in the recordings. In some instances, complete words were added to the text. These were used as if they had been in the original text. Both phenomena together resulted in four unpaired vowel realizations. No restriction was imposed on the selection of the vowels except that words and names with a non-Dutch orthography were excluded.

The vowel boundaries were chosen at a zero crossing in the speech waveform. Always, a whole number of pitch periods was used. Any pitch period that could be attributed to the target vowel, and not to the neighbouring phonemes, was considered to be part of that vowel. This included vowel periods that were changed severely by coarticulation. In a plosive-vowel-plosive context this would mean that everything, from the first period following the release burst to (and including) the last discernible period within

the closure, was used (note that Dutch plosives are unaspirated). Some vowels could not be separated from the neighbouring phonemes, especially in vowel-vowel contexts. When this occurred, the whole cluster was used, but the use of these vowel realizations was restricted to formant measuring methods (see below) which are insensitive to segmentation errors.

The read text was labelled for sentence-accent by an experienced phonetician. Labelling for actual phoneme realizations was done by one of the authors. Only standard Dutch phoneme labels were used.

2.1.3 Vowels used

For practical reasons, not all Dutch vowels were used in this experiment. Out of the twelve Dutch monophthongs, only seven were used in this study: the vowels /i y u o A a E/. These vowels were selected on their frequency of use and their representativeness in the vowel space. Five of these are short or half-long vowels (/i y u A E/) and two are long vowels (/o a/). All realizations of these vowels were isolated from the text and used in the analysis. Some realizations differed from their inferred pronunciation and these were labelled according to their actual spoken form. Additionally, some realizations of the schwa, which is a legitimate vowel in Dutch, were selected to serve as a neutral "anchor" in the vowel space. The schwa realizations used came from the words "HET" = /´t/ (English: "THE") and "ER" = /´r/ or /d´r/ (English: "THERE"). In Dutch, these two words are occasionally pronounced with an /E/ instead of with a schwa, but this pronunciation never occurred in the readings of this speaker. In Dutch, the /r/ in "ER" can be an alveolar or a velar consonant (our speaker uses the alveolar variant) and strongly colours vowels towards the /´/ (Pols, 1977). This colouring is expected to change the dynamics of the vowel, but since in this study we only use differences between static features of vowels (i.e. point measurements), this will not pose problems. Some other vowels which were reduced to schwa were included in this group of schwa vowels as well. The schwa in Dutch cannot carry stress in normal (i.e. not contrastive) situations. The various numbers of vowels thus obtained are listed in table 2.1. A grand total of 1178 vowel realizations were isolated existing of 587 pairs of realizations of the same text item at different speaking rates and 4 unpaired realizations. These four unpaired realizations originated from

Table 2.1: Number of vowels occurring in the text that has been analysed in this study. The number of incorrectly segmented vowels is given in parenthesis.

vowel	stressed	unstressed	fast	normal	total
E	59	191 (2)	126	124 (2)	250 (2)
A	58 (2)	181 (6)	116 (2)	123 (6)	239 (8)
a	54	157 (3)	106 (1)	105 (2)	211 (3)
i	52 (1)	132 (7)	92 (1)	92 (7)	184 (8)
o	45	132 (4)	88 (1)	89 (3)	177 (4)
´	0	56 (4)	30 (1)	26 (3)	56 (4)
u	13	19	16	16	32
y	11	4	13	12	25
others	···	···	3	1	4
Total	292 (3)	882 (26)	587 (6)	587 (23)	1178 (29)

vowels inserted by the speaker or deleted from one of the two realizations that were read. Within these 1178 realizations, another four vowels had to be labelled as vowels outside the set studied in this paper. Of the 587 pairs, 17 had different vowel realizations in terms of pronunciation for the two speaking rates and these pairs could not be used in pairwise tests. This leaves us with 570 pairs of realizations that can be used in pair-wise comparisons, as is listed in table 2.2. The 17 vowel pairs with differently labelled phonemes did not show any systematic differences between speaking rates and contained the four vowels labelled outside the set studied in this paper.

2.1.4 Spectral Analysis

A standard software package for speech research was used for LPC analysis (linear predictive code, Vogten, 1986). The vowel segments were analysed with a 10-pole LPC analysis, using a 25 ms Hamming window. The window was shifted in 1 ms steps. This was the basis for formant extraction. The LPC analysis was based on the Split-Levinson algorithm which gives continuous formant tracks (Willems, 1986).

Five different methods were used in parallel to extract five different "target" values from each formant track of each vowel realization. Using the segment boundaries, the value at the mid-point of the realization is read (method Centre), and the (linear) formant frequency average over the complete vowel realization is calculated (method Average). Both these methods were only used on the subset of vowel realizations for which segmentation could be done reliably.

Using a peak (and trough) picking algorithm (a slope segmentator based on Van Son, 1987, see appendix A; see also André-Obrecht, 1988), the point of maximal energy (method Energy) and maximal or minimal value of the appropriate formant (method Formant) were determined to within 3 ms (using a shifting interval one-eighth of the total length of the realization) and the formant frequencies were read at that point. For method Formant, the appropriate formant maximal or minimal value is chosen for each vowel independently, considering its position in the vowel plane. The realizations of the vowels /a A E/ are measured at the point of maximal F_1, the vowels /u o/ at the point of minimal F_2, the vowel /i/ at maximal F_2, and the vowel /y/

Table 2.2: Number of vowel pairs matched on normal versus fast rate. Both realizations in each pair are from the same text item (see text). The number of pairs with incorrectly segmented vowels is given in parenthesis.

vowel	stressed	unstressed	unequal stress	total
E	23	86 (1)	13 (1)	122 (2)
A	25 (2)	82 (3)	8	115 (5)
a	21	72 (2)	11	104 (2)
i	24 (1)	63 (6)	4	91 (7)
o	17	59 (3)	11	87 (3)
´	0	23 (2)	0	23 (2)
u	4	7	5	16
y	5	6	1	12
total	119 (3)	398 (17)	53 (1)	570 (21)

at minimal F_1. With the Formant method, the schwa /´/ was not measured and the values obtained with method Energy are used instead. Peak picking was not perfect, and in about one out of every five formant and energy tracks the "right" peak had to be selected from the suggested alternatives by visual inspection of the tracks. As a fifth method to determine a suitable target point (method Stationary), an automated method for selecting the most stable part of a vowel realization is used (the section with the least variance in the logarithm of the first three formants, Van Bergem, 1988). The last three methods (Energy, Formant, and Stationary) were used on all vowel realizations.

2.2 Results

To determine whether differences in speaking rate introduce differences in vowel formant target values, the properties of vowels realized at normal and at fast rate are compared. It is possible to detect these differences without relying on a specific representation or statistical distribution of the measured values. To decide on statistical significance we used rank-order statistics which is distribution-free. These distribution-free statistical tests are less sensitive and less efficient (Ferguson, 1981) than tests based on a specific distribution (e.g., Normal, Chi-square, or Student's distributions), but they also lack the methodological problems concerning applicability. The range of different stochastic processes for which a distribution-free test can be used is generally much larger than for other statistical tests.

The test results are recalculated to a normal (Gaussian, *z* scores) or Student's (*t* scores) distribution as appropriate, or probabilities are calculated directly (sign-test for small *n*). All tests are derived from Ferguson (1981). Determination of statistical significance is carried out using tables from Abramowitz and Stegun (1965). To obtain a repeated-test result which still has a probability lower than 5% (single test level, indicated by "+") of one or more spurious results that reach the level of significance, a threshold level of 0.1% (10^{-3}, two-tailed, indicated by "++") was used to determine statistical significance in individual tests. In this way, it still is possible to identify the samples that deviate from the H_0 hypothesis out of a large set (up to 50 samples) with an error probability of less than 5%.

2.2.1 Median values

A general way to compare two sets of values is to test for differences in their median values. The standard target-undershoot model predicts a smaller distance between the median formant values of a specific vowel and the schwa for fast-rate speech than for normal-rate speech. This implies lower median formant values for both F_1 of vowels /E a A/ and F_2 of vowels /i E/ for fast-rate speech than for normal-rate speech and higher median formant values for both F_1 of vowels /i y u/ and F_2 of vowels /u o A/. The other values should be more or less the same under both speaking conditions. An analysis of the data per vowel was made, the results of which are shown in table 2.3. In this table, median formant values and a Mann-Whitney U test

were used to test for differences between the distributions of all normal-rate and all fast-rate realizations of one specific vowel in the set.

First, there is a global shortening of vowel duration detectable in fast-rate speech as compared to normal-rate speech, when all vowels are pooled (total row in table 2.3). However, only long vowels, /a/ and /o/, prove to be shorter in fast-rate speech (0.1% level, ++), the other vowels are ambiguous

Table 2.3: Median values for formant frequencies (Hz) and duration (ms).
Statistical significance is determined with a Mann-Whitney U test. Statistical significance is indicated by "++" (at the 0.1% level); a 5% error level for a result is indicated by "+"; other statistically insignificant results are indicated by "ns". Abbreviations of method names: Form.-Formant, Stat.- Stationary, Ener.-Energy, Cent.-Centre, Aver.- Average. In all columns: normal-rate value left (n), fast-rate value right (f). The total mean values and standard deviation of the duration are: normal rate 99 ± 41 ms, fast rate 84 ± 31 ms (correctly segmented vowels only).

Vowel	Form.	Stat.		Ener.		Cent.		Aver.		Duration			
	Rate	n	f	n	f	n	f	n	f	n	f	n	f
E	F_1	554	574	545	565	524	548	544	557	493	520	81	74
		+		+		++		+		++		+	
	F_2	1527	1514	1527	1526	1523	1521	1521	1527	1503	1501		
		ns		ns		ns		ns		ns			
A	F_1	597	618	587	608	581	600	589	609	539	564	81	76
		+		+		+		+		++		+	
	F_2	1151	1153	1112	1133	1128	1133	1119	1131	1133	1129		
		ns		ns		ns		ns		ns			
a	F_1	639	655	631	649	623	637	630	645	579	609	131	97
		+		+		+		+		++		++	
	F_2	1331	1330	1313	1330	1324	1334	1329	1329	1335	1321		
		ns		ns		ns		ns		ns			
i	F_1	312	325	316	332	327	341	313	335	316	333	80	72
		+		+		+		+		+		+	
	F_2	2130	2105	2081	2074	2002	2010	2072	2036	1946	1925		
		ns		ns		ns		ns		ns			
o	F_1	391	413	419	432	412	435	417	439	411	434	121	109
		+		+		+		+		++		++	
	F_2	854	897	930	964	943	972	925	959	995	1029		
		ns		ns		ns		+		ns			
´	F_1	407	440	411	438	407	440	414	434	393	422	52	56
		ns		ns		ns		+		ns		ns	
	F_2	1440	1455	1434	1454	1440	1455	1435	1464	1433	1444		
		ns		ns		ns		ns		ns			
u	F_1	369	368	370	375	376	390	372	373	362	368	83	74
		ns		ns		ns		ns		ns		ns	
	F_2	782	776	800	805	836	821	880	851	947	1012		
		ns		ns		ns		ns		ns			
y	F_1	297	332	313	336	329	364	317	334	316	350	77	76
		ns		ns		+		ns		ns		ns	
	F_2	1452	1416	1576	1442	1624	1566	1590	1476	1582	1504		
		ns		ns		ns		+		+			
Total	F_1	526	553	526	553	498	528	520	535	476	501	89	78
		+		+		+		+		+		++	
	F_2	1339	1351	1341	1347	1343	1361	1334	1357	1345	1360		
		ns		ns		ns		ns		ns			

Figure 2.2. Median values of first and second formant measured with the "Average" method for pairs of realizations of Dutch vowels. Open squares: normal speaking rate values. Filled squares: fast speaking rate values, crosses: /a/, triangles: /i/.

in this respect (at most at the 5% level, +). The averaged shortening of vowel duration due to speaking rate is smaller than the overall shortening of the spoken text (only 15% in vowels versus 25% in the total text, see also section 2.1.1, and 2.3.1 below), but the differences are systematic and present in all but one vowel, the schwa.

The number of vowels, for which significant differences ($p \bullet 0.1\%$, ++) between median formant values at different speaking rates are found, is small. Especially for methods for which inter-vowel spectral distances are large (Formant, Stationary, and Centre) none of the vowels shows a significant difference between speaking rates. The number of (not significant) test results with a low probability ($p \bullet 5\%$, +) is sufficiently high to suggest that there is indeed some difference between speaking rates. The probability to obtain at least 5 out of 8 test results at the 5% level is less than 0.1% (++). For only one method, Average, it is possible to identify the vowels which change with some confidence (at the 0.1% level, ++). Using this measuring method, the vowels /E A a o/ show a statistically significant higher first formant value in fast-rate speech as compared to normal-rate speech (see figure 2.1 and table 2.3). No statistically significant differences between second formant frequencies are found (table 2.3).

Comparing columns in table 2.3, the differences between the different measuring methods are small and seem to be limited to a small reduction in overall size of the vowel triangle going from method Formant to method Average. Although the differences between speaking rates are not always statistically significant, the median values all show the same response to an increase in speaking rate. The differences found here between formant values from vowels spoken at different rates are inconclusive in that for only one method, Average, is it possible to identify statistically significant changes in vowel formant values. Apparently, this kind of statistical analysis is not sensitive enough to show the differences between fast- and normal-rate vowels from unrestricted text reliably. Whether or not a test will show a difference between speaking rates depends on the measuring method used.

2.2.2 Consistency

The consistency with which our speaker reproduces the text in each reading and the ability of our measuring methods to capture the within-speaking-rate variation over different readings must be estimated, before comparisons between the members of vowel realization pairs in both readings can be made. This estimation can be performed by checking the similarity between the measurements in the two readings. The similarity of within-speaking-rate rank order of measurements between different speaking rates is an indicator of the desired consistency. It was measured with a Spearman rank correlation test, the results of which are shown in table 2.4. To illustrate graphically the similarity of rank order, a choice has

been made from the data presented in table 2.4. In figure 2.2, the F_2 frequencies of individual vowel pairs spoken at normal and fast rate, measured with the Average method, are plotted against each other for just three vowels: /o a i/. It can be seen that the formant value pairs are ordered along the diagonal of the plot for /o a/ displaying a fairly monotonic relation between normal-rate and fast-rate F_2 values, and thus a high Spearman rank correlation coefficient. The F_2 values of the /i/ are scattered over a large area, indicating that only a minimal relation exists between normal-rate and fast-rate values of the F_2 for this vowel, and thus only a very small correlation coefficient. As a consequence, the F_2 values measured of /o a/ are consistent over speaking rates whereas the F_2 values of /i/ are not.

In table 2.4, the Spearman rank correlation coefficients of formant values and duration are presented for all methods and vowels used. Except for F_1 of /u/ and F_2 of /i/, all correlation coefficients are above 0.5 for at least some of the methods used. Except for /´/, all durational correlation coefficients are above 0.5. For most vowels, the F_2 formant values correlate with coefficients around 0.7 or well above. These correlation coefficients are comparable in size to those found by Kuehn and Moll (1976) when they compared articulatory velocities from vowel-consonant transitions spoken at different rates. The correlation coefficients show peculiar differences between vowels that are not easily explained without a detailed analysis of the distribution of context features over the different vowels, an analysis that is outside the scope of this paper. The very low correlation of F_2 from /i/ can probably be attributed to problems with the LPC formant analysis of this vowel formant. The F_2 and F_3 values of the /i/ might be too close for the

Table 2.4: Coefficients of a Spearman Rank Correlation test on formant frequency values and durations between the realizations within pairs (normal-rate versus fast-rate) of vowels. For indication of statistical significance and abbreviations see table 2.3.

Vowel	Form.	Stat.	Ener.	Cent.	Aver.	Duration	
E	F_1	0.65 ++	0.56 ++	0.60 ++	0.61 ++	0.61 ++	0.68++
	F_2	0.70 ++	0.64 ++	0.58 ++	0.67 ++	0.71 ++	
A	F_1	0.81 ++	0.74 ++	0.75 ++	0.79 ++	0.79 ++	0.65++
	F_2	0.85 ++	0.86 ++	0.87 ++	0.88 ++	0.89 ++	
a	F_1	0.61 ++	0.57 ++	0.55 ++	0.59 ++	0.65 ++	0.77++
	F_2	0.72 ++	0.73 ++	0.75 ++	0.76 ++	0.84 ++	
i	F_1	0.58 ++	0.53 ++	0.44 ++	0.53 ++	0.58 ++	0.66++
	F_2	0.16 ns	0.13 ns	0.24 +	0.10 ns	0.24 +	
o	F_1	0.78 ++	0.73 ++	0.80 ++	0.86 ++	0.87 ++	0.81++
	F_2	0.79 ++	0.63 ++	0.70 ++	0.69 ++	0.86 ++	
´	F_1	0.70 ++	0.62 +	0.70 ++	0.52 +	0.44 +	-0.06ns
	F_2	0.92 ++	0.89 ++	0.92 ++	0.83 ++	0.91 ++	
u	F_1	0.39 ns	0.31 ns	0.12 ns	0.16 ns	0.27 ns	0.57 +
	F_2	0.63 +	0.62 +	0.53 ns	0.73 +	0.59 ns	
y	F_1	0.01 ns	0.45 ns	0.73 +	0.70 +	0.76 +	0.60 +
	F_2	0.32 ns	0.58 ns	0.69 +	0.54 ns	0.65 +	
Total	F_1	0.94 ++	0.93 ++	0.93 ++	0.94 ++	0.94 ++	0.77++
	F_2	0.96 ++	0.92 ++	0.94 ++	0.93 ++	0.96 ++	

analysis method to resolve the differences between these two formants, resulting in aberrant F_2 values. The total absence of a correlation for the duration of /´/ is to be expected because all pairs of this vowel were taken from only two different, unstressed, high frequency words (/´t/ and /(d)´r/), giving only a very small variation in context.

As before (section 2.2.1), all measuring methods seem to capture the same kind of features with only a difference in sensitivity, and no method behaves at variance with the others. The strong correlations found between values measured for vowels uttered at different speaking rates indicates that whatever systematic differences exist between these vowel realizations, it is conserved by the measurements. This means that a pairwise comparison should indeed be able to discover systematic differences in formant values between speaking rates.

2.2.3 Pairwise changes in formant frequencies and duration

The measured formant and duration values of the vowel pairs were divided into two sets. One set contained all value pairs for which the fast-rate value was higher than the normal-rate value. The other set contained all value pairs for which the fast-rate value was lower than the normal-rate value. Pairs in which both values are equal were omitted. This was done for each of the parameters, F_1, F_2 and duration, and for each method. In table 2.5, the fractions of pairs with a higher fast-rate formant frequency or a lower fast-rate duration are presented as percentages of total. Statistical signifi-

Table 2.5: Percentage of pairs for which the fast-rate realization has a higher formant value than its normal-rate counterpart. Last column (Duration): Percentage of pairs for which the fast-rate realization is shorter than its normal-rate counterpart. Significance is given for a Sign test, ties (fast-rate value = normal-rate value) are omitted. For indication of statistical significance and abbreviations see table 2.3.

Vowel	Form.	Stat.	Ener.	Cent.	Aver.	Duration	
E	F_1	70 ++	73 ++	71 ++	71 ++	80 ++	74++
	F_2	47 ns	48 ns	47 ns	49 ns	44 ns	
A	F_1	70 ++	70 ++	72 ++	70 ++	76 ++	71++
	F_2	64 +	60 +	62 +	64 +	63 +	
a	F_1	62 +	66 +	63 +	68 +	77 ++	92++
	F_2	58 ns	52 ns	51 ns	51 ns	46 ns	
i	F_1	62 +	67 +	64 +	73 ++	71 ++	72++
	F_2	48 ns	55 ns	47 ns	38 +	42 ns	
o	F_1	81 ++	74 ++	81 ++	84 ++	88 ++	81++
	F_2	68 +	64 +	68 +	68 +	71 ++	
´	F_1	61 ns	70 ns	61 ns	76 +	76 +	43ns
	F_2	61 ns	52 ns	61 ns	62 ns	67 ns	
u	F_1	73 ns	55 ns	64 ns	45 ns	55 ns	46ns
	F_2	73 ns	70 ns	64 ns	64 ns	73 ns	
y	F_1	82 ns	73 ns	100 +	82 ns	91 +	73ns
	F_2	36 ns	9 +	17 +	18 ns	9 +	
Total	F_1	69 ++	70 ++	70 ++	72 ++	78 ++	75++
	F_2	57 +	55 +	54 ns	53 ns	53 ns	

cance was determined with a sign-test. Based on the duration figures, most vowels can be said to be shorter when spoken at a fast rate (75%), thus confirming the overall shortening of the vowels in fast-rate speech (section 2.2.1).

With only one exception (i.e., /u/ analysed using the Centre method) the majority (> 50%) of pairs of all vowels with all measuring methods show a fast-rate F_1 value which is higher than the normal-rate formant value. This higher fast-rate F_1 value is found, independent of the identity of the vowel. This means that the first-formant values generally rise with speaking rate, which conforms with the results of the tests using median values (section 2.2.1). This time, however, the differences found are statistically significant (level 0.1%, ++) with all methods used for /E A o/ and vowels pooled (total), and not just for method Average, as was the case when analysing median values (section 2.2.1, see table 2.3). Method Average gives statistical significant differences (level 0.1%, ++) for 5 out of the 8 vowels used (/E A a i o/).

When it comes to vowel formant differences between speaking rates, no clear picture emerges for the second formant. No statistical significant changes can be found except for F_2 of /o/ with the Average method. This averaging method seems to be the most sensitive method for analysis of differences between formant values of vowel realizations, both for F_1 and F_2.

2.2.4 Correlation between formant frequency and duration

The target-undershoot model presupposes a relation between spectral vowel reduction and vowel duration. If vowel formant values move to the schwa value (i.e. show spectral reduction) with shorter vowel durations, there should be a (strong) correlation between vowel duration and vowel formant values. The strength of this correlation, in relation to the correlation between different speaking rates (section 2.2.2), is an indication of the importance of vowel duration in determining the vowel formant value, relative to the other important factors (e.g., stress, context).

The rank correlation between vowel formant values and duration shows very small, but often statistically significant ($p \bullet 0.1\%$, ++), correlation coefficients (table 2.6) which implies that only a very small part of the variation in formant values between vowel realizations can be explained by the differences in duration. This was found for realizations of both speaking rates pooled (table 2.6.a) and for the fast rate realizations (table 2.6.b) and normal rate realization individually (data not shown, they are comparable to those of table 2.6.b). The correlations seem to be stronger when realizations from both speaking rates are used independently instead of pooled together (compare table 2.6.b with table 2.6.a). Of all correlations, only the coefficients of the F_1 values of the vowels /E A a/ are statistically significant.

In contrast, the correlation between formant values of realizations that differ in speaking rate only (table 2.4) is high and statistical significant for both formants and almost all vowels and can thus explain a great part of the variation in formant values. Based on these correlations, it must be concluded that vowel duration has only a marginal power in explaining the vowel formant targets. This small explanatory power holds just as much be-

tween as within speaking rates. The correlation coefficients are so extremely small compared with the pairwise correlations (table 2.4) that it is even possible for these correlations to be the result of a residual correlation stemming from the correlation between both formant target frequency and duration and the stress and context of the vowel.

2.2.5 Influence of phoneme context

Analysis of how the influences of speaking rate depend upon the phonetic context in which the vowels occur (coarticulation) is hampered by the large number of different contextual phonemes per vowel which is inherent to unrestricted (near-natural) text. Consequently, there are so few realizations of any specific vowel-context combination, that a statistical analysis is almost impossible with the amount of text and the statistical methods used in this paper.

As a first attempt, vowels and consonants were pooled on articulatory features. Of all the consonants, the alveolar consonants were most common. In Dutch, the alveolar consonants encompass /n t d s z r l/. Alveolar consonants are articulated very close to the /i/, they can be described as high, closed and fronted phonemes. The vowels were divided into several overlapping sets. A set of closed vowels, /i y u/, versus a set of open vowels, /a A E/, and a set of fronted vowels, /i E/, versus a set of back vowels, /o u/. The vowel realizations in alveolar context were pooled on these groups and the pairwise differences between speaking rates were tested (like in section 2.2.3). Three arrangements are possible: CV*, *VC, and CVC, in which the C is an alveolar consonant and * can be any context. It showed that in, all three arrangements, the same pattern emerged. Because the trailing consonant has the greatest importance in determining stationary vowel spectra (Pols, 1977), and the vowel realizations in this context were most numerous, we only show the *VC results (table 2.7).

It appears that all vowels, grouped on different features, behave identical. The trend of higher F_1 values in fast-rate speech, already found for the individual vowels, without regarding context, emerges again. Also, the lack of significant differences between F_2 values measured at different speaking rates is found again. Despite the fact that open vowels are "distant" in an articulatory sense from the (closed) alveolars, these vowels do not behave different from the more "nearby" closed vowels. The same is found for the distant back vowels and the nearby front vowels. The higher F_1 value in fast-rate speech implies, in these articulatory terms, a more open articulation where a more closed articulation (i.e. lower F_1 values) is expected if a higher speaking rate should result in more coarticulation.

2.2.6 Influence of stress

Thus far, vowels were considered to be comparable when different speaking rates were used. However, the effects of speaking rate could very well be different for stressed and unstressed vowels. This was investigated by comparing the changes between pairs of vowels for the two speaking rates just as in table 2.5, but now for stressed and unstressed vowels separately. Because of the small number of stressed vowel pairs, all vowels were pooled and only these total figures per formant value were used (table 2.8). These total scores indicate a small difference in percentage of pairs changing in one direction for stressed and unstressed vowels. The differences between speaking rates are somewhat less pronounced for the formant values of stressed vowels than for unstressed vowels. The reverse is true for differ-

Table 2.6.a: Similar to table 2.4 but this time the coefficients indicate the Spearman Rank Correlation coefficients between formant values and duration for each vowel realization. Only correctly segmented vowels are used. Normal-rate and fast-rate realizations pooled.

Vowel		Form.	Stat.	Ener.	Cent.	Aver.
E	F_1	0.28 ++	0.19 +	0.21 ++	0.30 ++	0.08 ns
	F_2	0.04 ns	0.03 ns	0.07 ns	0.02 ns	-0.03 ns
A	F_1	0.49 ++	0.42 ++	0.39 ++	0.45 ++	0.31 ++
	F_2	-0.18 +	-0.27 ++	-0.23 ++	-0.26 ++	-0.23 ++
a	F_1	0.41 ++	0.37 ++	0.34 ++	0.33 ++	0.15 +
	F_2	-0.02 ns	-0.02 ns	-0.05 ns	0.02 ns	0.03 ns
i	F_1	0.03 ns	-0.04 ns	0.04 ns	-0.05 ns	-0.02 ns
	F_2	0.33 ++	0.22 +	0.07 ns	0.23 +	0.03 ns
o	F_1	-0.02 ns	0.11 ns	-0.01 ns	0.12 ns	0.12 ns
	F_2	-0.27 ++	-0.12 ns	-0.13 ns	-0.16 +	-0.13 ns
´	F_1	0.17 ns	0.18 ns	0.17 ns	0.10 ns	0.00 ns
	F_2	0.40 +	0.34 +	0.40 +	0.36 +	0.32 +
u	F_1	0.02 ns	0.06 ns	-0.15 ns	-0.09 ns	-0.06 ns
	F_2	0.01 ns	0.17 ns	0.08 ns	-0.05 ns	0.11 ns
y	F_1	0.41 +	0.35 ns	0.26 ns	0.38 ns	0.37 ns
	F_2	-0.11 ns	0.23 ns	0.35 ns	0.24 ns	0.09 ns
Total	F_1	0.23 ++	0.22 ++	0.21 ++	0.23 ++	0.19 ++
	F_2	-0.26 ++	-0.27 ++	-0.27 ++	-0.27 ++	-0.27 ++

ences in duration. This time it matters indeed which method is used to determine the formant frequency. For stressed vowels, methods that are sensitive for the exact shape of the formant track with respect to the vowel boundaries (i.e. Energy, Centre, and Average) indicate more change than do methods that try to catch shape-invariant points of the formants (Formant and Stationary). It is not possible to substantiate this any further with the rather limited set of data used here.

2.3 Discussion

The median formant values found in this study (table 2.3) for normal-rate speech are generally lower than those found by Koopmans-van Beinum (1980, male speaker #1) with speech of the same speaker for stressed and unstressed vowels in read text. Apart from methodological differences in vowel selection and labelling, these differences can be attributed to the differences in spectral analysis (LPC versus spectrographic).

2.3.1 Differences between speaking rates: Duration

Although most fast-rate vowel realizations are shorter than their normal rate counterparts, the differences between these vowel durations are quite small. The global decrease in total duration is about 25%, but the decrease in duration of the vowels studied is less than 15% if the fast-rate reading of the text is compared to the normal-rate reading. The exception is the vowel /a/, which seems to shorten by approximately 25% (*median* values from table 2.4, see also section 3.2.1).

Table 2.6.b: As table 2.6.a. Vowel realizations from fast-rate reading only.

Vowel		Form.	Stat.	Ener.	Cent.	Aver.
E	F_1	0.34 ++	0.27 +	0.32 ++	0.38 ++	0.18 +
	F_2	0.02 ns	-0.02 ns	0.04 ns	0.01 ns	-0.03 ns
A	F_1	0.49 ++	0.38 ++	0.31 ++	0.42 ++	0.27 +
	F_2	-0.16 +	-0.27 +	-0.22 +	-0.26 +	-0.22 +
a	F_1	0.57 ++	0.55 ++	0.50 ++	0.53 ++	0.36 ++
	F_2	-0.04 ns	-0.05 ns	-0.06 ns	-0.04 ns	-0.06 ns
i	F_1	0.00 ns	-0.05 ns	-0.03 ns	-0.10 ns	-0.10 ns
	F_2	0.23 +	0.18 ns	0.04 ns	0.16 ns	-0.04 ns
o	F_1	-0.05 ns	0.19 ns	0.08 ns	0.25 +	0.22 +
	F_2	-0.34 +	-0.14 ns	-0.14 ns	-0.18 ns	-0.19 ns
´	F_1	0.18 ns	0.06 ns	0.18 ns	0.06 ns	-0.10 ns
	F_2	0.38 +	0.37 +	0.38 +	0.37 +	0.40 +
u	F_1	-0.05 ns	0.05 ns	0.14 ns	0.07 ns	0.05 ns
	F_2	0.07 ns	0.00 ns	0.20 ns	0.00 ns	0.19 ns
y	F_1	0.57 +	0.15 ns	0.34 ns	0.36 ns	0.45 ns
	F_2	0.44 ns	0.58 +	0.42 ns	0.51 ns	0.38 ns
Total	F_1	0.23 ++	0.23 ++	0.22 ++	0.24 ++	0.20 ++
	F_2	-0.27 ++	-0.27 ++	-0.28 ++	-0.28 ++	-0.28 ++

Different explanations are possible. At one hand, we may have overestimated the global decrease in duration by including too much of the silent parts (pauses shorter than 200 ms, section 2.1.1). These silent parts could be the elements that absorb the shortening. At the other hand, our segmentation may have been biased toward longer fast-rate vowels by including more pitch periods in fast-rate vowel realizations than in normal-rate realizations. This kind of bias is difficult to detect if the context from which the vowel realizations are obtained is as diverse as in this study.

Apart from these methodological problems, another reason for the small difference in vowel duration between speaking rates may be the fact that the normal rate vowel realizations themselves already are quite short. A normal, and pleasant, speaking rate for reading a long text will be faster than the speaking rate used for isolated sentences in a citation style of speaking. The attainable durational differences between speaking rates for vowel realizations in studies using that kind of speech may be higher than what is found in the present study.

Whatever the explanation of the rather small size of the differences in vowel duration between speaking rates, these differences are highly systematic. Therefore, the fast-rate vowel realizations should nevertheless show the differences in target values associated with speaking rate differences, but actually did not.

2.3.2 Differences between speaking rates: Formant frequencies

Considering the material and methods used here, it is not possible to uncover the cause of the higher F_1 values found in all vowels with a higher speaking rate. An explanation for this higher formant value might be that, given the fact that F_1 is related to the openness of vowels, our experienced speaker lowers his jaw somewhat more in fast-rate speech than in normal-rate speech. This could be the result of overcompensation or overshoot when the speaker accommodates for the high speaking rate. An alternative explanation might be that our speaker reads the fast-rate realization with a louder voice than the normal-rate realization. It is known that differences in speech effort can change the articulation (Schulman, 1989) and the formant values of vowels (Traunmüller, 1988). A louder voice might also be partly responsible for the relatively long vowel durations in fast rate speech

*Table 2.7: Similar to table 2.5 but this time only vowels uttered in *VC context are used, for which the C is an alveolar consonant (one of* /n t d s z l r/*) and * can be any context. The vowel pairs are pooled on the features [+Closed] (*/i y u/*, n=60), [+Open] (*/E A a/*, n=255), [+Front] (*/i E/*, n=141), and [+Back] (*/u o/*, n=46).*

Vowel	Form.	Stat.	Ener.	Cent.	Aver.	Dur.	
Closed	F_1	53 ns	62 ns	64 +	63 ns	67 +	73++
	F_2	45 ns	51 ns	38 ns	38 ns	43 ns	
Open	F_1	66 ++	69 ++	68 ++	70 ++	79 ++	79++
	F_2	54 ns	53 ns	52 ns	52 ns	52 ns	
Front	F_1	62 +	70 ++	67 ++	68 ++	75 ++	75++
	F_2	45 ns	48 ns	44 ns	43 ns	45 ns	
Back	F_1	80 ++	73 +	82 ++	83 ++	87 ++	85++
	F_2	64 ns	61 ns	67 ns	67 +	70 +	

(Schulman, 1989; c.f., section 2.3.1). Because we did not calibrate our recordings for loudness, we are not able to check this. The difference between the F_1 values at different speaking rates is, however, very small and its perceptual relevance is questionable.

These results show that a different style of speaking, fast-rate versus normal-rate reading of a text, can change the duration of the vowels without changing the vowel formant values or can change the vowel formant target values in unexpected ways. Even when using vowels in identical context, a simple correlation between vowel formant target values and vowel duration cannot be extended over different speaking styles. Indications for speaking-style specific correlations between F_1 and duration were also found by Lindblom and Moon (1988) when they compared clear and citation form speech. Also the explanatory power of duration when predicting vowel target values must be judged marginal if compared to other (contextual) factors.

It is known that articulatory adaptation to a fast speaking rate can be speaker dependent (Kuehn and Moll, 1976) and it is to be expected that the ability to read aloud at a fast rate, and still pronounce correctly, depends on experience and training. The speaker used in this experiment has had a very long career as a professional speaker and newscaster, so his capabilities are not likely to be shared by naive, untrained, subjects. The results are nevertheless important for general theories on articulation and the design of systems for automatic speech recognition and synthesis. The experience of the speakers used should also be considered seriously when designing an experiment regarding the effects of speaking rate on speech sounds.

2.3.3 Differences between measuring methods

In this paper different methods to measure vowel formant values in a given formant track were used. Averaging the formant values over the complete vowel is the method most sensitive to speaking rate changes; at the same time this method also produces formant frequencies that deviate most from the values reported in literature (e.g., Pols, 1977; Koopmans-van Beinum, 1980). However, the differences between the various methods used are in most respects marginal and all methods used essentially give the same outcome. When studying vowel targets, the method that is most convenient can be used.

Probably all points in a vowel segment change in concert when speaking rate changes, so it may not be crucially important which cross-section in

Table 2.8: Similar to table 2.5 but this time with all vowel pairs pooled on stress, first row: unstressed; second row: stressed; last row: stressed and unstressed combined. Only pairs with equal stress realization on both readings are used.

VowelForm.		Stat.	Ener.	Cent.	Aver.	Dur.	
no stress	F_1	72 ++	72 ++	71 ++	74 ++	78 ++	73++
	F_2	58 +	56 +	55 ns	54 ns	54 ns	
stress	F_1	57 ns	63 +	69 ++	66 ++	76 ++	84++
	F_2	53 ns	52 ns	50 ns	50 ns	51 ns	
Total	F_1	69 ++	70 ++	70 ++	72 ++	78 ++	75++
	F_2	57 +	55 +	54 ns	53 ns	53 ns	

the realization is actually used to measure the difference. Such a model of vowel dynamics can only be checked with a detailed analysis of the total dynamic shape of vowel formant tracks, not by using point measurements as was done here. This dynamic description of formant tracks is the subject of the next two chapters (see also, Van Son and Pols, 1989, 1991a, 1992).

2.4 Conclusions

With the restriction that speech of only one speaker was used and that the speech was constrained to two readings of one text, our analysis reveals that neither excess vowel reduction (in terms of vowel targets) nor excess coarticulation accompanies a higher speaking rate. The only change in vowel formant frequency that could be detected was a higher value of the first formant frequency in fast-rate speech as compared to normal-rate speech, irrespective of the vowel identity. This shift in formant frequency may be linked to a more open articulation of the vowels or an increase in loudness of the speech. No difference due to stress or consonantal context was found that could explain this behaviour, neither was there an effect of the method with which the target points within the vowel realizations were determined.

3

FORMANT MOVEMENTS OF DUTCH VOWELS IN A TEXT, READ AT NORMAL AND FAST RATE*

Abstract

Speaking rate in general, and vowel duration more specifically, is thought to affect the dynamic structure of vowel formant tracks. To test this, a single, professional speaker read a long text at two different speaking rates, fast and normal. The present project investigated the extent to which the first and second formant tracks of 8 Dutch vowels varied under the two different speaking rate conditions. A total of 549 pairs of vowel realizations from various contexts were selected for analysis. The formant track shape was assessed on a point-by-point basis, using 16 samples at the same relative positions in the vowels. Differences in speech rate only resulted in a uniform change in F_1 frequency. Within each speaking rate, there was only evidence of a weak leveling off of the F_1 tracks of the open vowels /A a/ *with shorter durations. When considering sentence-stress or vowel realizations from a more uniform, alveolar-vowel-alveolar context, these same conclusions were reached. These results indicate a much more active adaptation to speaking rate than implied by the target-undershoot model.*

*Van Son, R.J.J.H. & Pols, L.C.W. (1992). "Formant movements of Dutch vowels in a text, read at normal and fast rate", *Journal of the Acoustical Society of America* 92, 121-127.

Introduction

In the target-undershoot model of vowel articulation, vowel duration is considered an important parameter in determining the actual realization of the vowel formants (e.g., Lindblom, 1963; Broad and Fertig, 1970; Gay, 1978; Gay, 1981; Lindblom, 1983; Broad and Clermont, 1987; Di Benedetto, 1989a; Lindblom and Moon, 1988; Moon, 1990). Vowel duration is important both for the formant frequency inside the vowel nucleus (for use of "vowel nucleus", see Krull, 1989) as well as for the shape of the complete formant tracks. The target-undershoot model predicts more spectral reduction when vowels become shorter, i.e. more schwa-like formant values in the vowel nucleus and more level, less curved, formant tracks.

Inside the vowel nucleus, the formant frequencies appeared to be correlated to vowel duration in the way predicted by the target-undershoot model, at least when speaking style was held constant (Broad and Fertig, 1970; Broad and Clermont, 1987; Lindblom and Moon, 1988; Moon, 1990). In contrast, formant frequencies were only weakly correlated to vowel duration, or not at all, when the speaking style differed (e.g., clear speech versus citation form speech, Lindblom and Moon, 1988; Moon, 1990; fast-rate speech versus normal-rate speech, Van Son and Pols, 1990). Several studies did not find speaking-rate dependent differences between formant frequencies that were in any way connected to vowel identity (e.g., Gay, 1978; Gopal and Syrdal, 1988; Den Os, 1988; Engstrand, 1988; Van Son and Pols, 1990; Fourakis, 1991). Van Son and Pols (1990) did find a systematic higher F_1 in fast-rate speech, but this difference occurred in all vowels (even /a/). This rise in F_1 cannot be interpreted as vowel reduction in the sense of the target-undershoot model. These studies suggest that there are two kinds of durational differences between vowel realizations. The first type of durational differences are those found between vowels spoken in the same speaking style and at the same rate. These differences in vowel duration are (cor-)related to spectral differences as predicted by the target-undershoot model. The other type of durational differences are the differences between vowels spoken in different speaking styles or at different rates. These latter differences in duration are not related to spectral differences between vowels.

Relatively few studies have considered the relation between vowel formant dynamics and duration (e.g., Broad and Fertig, 1970; Broad and Clermont, 1987; Di Benedetto, 1989a; Van Son and Pols, 1989) and these were limited to only one speaking style. Studies that did use different speaking styles or different speaking rates generally only measured formant frequencies within the vowel nucleus. Therefore, it is not clear whether fast-rate speech is just "speeded-up" normal-rate speech, or whether different articulation strategies (as proposed by Gay, 1981) or a higher speaking effort (Lindblom, 1983) are used. Differences in articulation or speaking effort should result in different shapes of the formant tracks, e.g. a levelling-off of the formant movements in fast-rate speech.

Formant track shape is generally characterized by the lengths and slopes of vowel on- and off-glide which are measured using two to four points from each formant track (Di Benedetto, 1989a; Strange, 1989a, b;

Duez, 1989; Krull, 1989). However, it is very difficult to determine the boundaries of the stationary part (Benguerel and McFadden, 1989) and to measure formant track slopes accurately. Therefore, another method to characterize formant track shapes was chosen. We performed a point-by-point analysis on sampled vowel formant tracks (16 points, adapted from Broad and Fertig, 1970) and compared the formant frequencies on comparable, relative, positions in the vowel realizations.

Differences between speaking rates are best studied by using vowel realizations that differ *only* in speaking rate. In order to obtain a large and varied inventory of such vowel pairs, a long text was read twice by a single professional speaker (a well known newscaster), once at a normal rate and once at a fast rate (Van Son and Pols, 1990). With these vowels, we have tested whether vowel formant track shape depends on vowel duration and speaking rate and how this relation can be modelled. Also the effects of stress and vowel context were taken into account.

Using a single, professional speaker will make it difficult to generalize the results of this study to other, more “naive”, speakers. However, the way an experienced newscaster, who speaks standard Dutch and whose pronunciation is perceived as “correct”, reacts to speaking rate differences will be very likely an “accepted” way of doing so. General theories of articulation do not consider personal skill or experience as a factor of importance. Therefore, if our speaker does not utter vowels in the way predicted then we have, for non-aberrant speech, a counter example to the general theories of articulation. We do acknowledge that large sections of the population might react in a different way to changes in speaking rate. Our experiment should be viewed only as a test on the predictive power of articulation theories on the effects of speaking rate.

3.1 Methods

The present project investigated a subset of the material used in our previous study (chapter 2; Van Son and Pols, 1990). Here, we will only summarize the procedures used.

3.1.1 Speech material and segmentation

A meaningful text of 844 words (1440 syllables) was read twice by an experienced speaker, once as fast as possible, once at a normal rate (i.e., as for an audience). The speech was recorded on a commercial Sony PCM-recorder, low-pass filtered at 4.5 kHz and digitized at 10 kHz, with 12 bit resolution. Subsequent storage, handling and editing were done in digital form only. Reading the text took 330 s for the normal speaking rate and 220 s for the fast speaking rate (4.4 and 6.6 syll./s including pauses, cf. Koopmans-van Beinum, 1990). The overall reduction in duration of the fast-rate as compared to the normal-rate realization was one-third when pauses longer than 200 ms were included, and one-fourth when these longer pauses were excluded. A subjective evaluation did not reveal differences in reading style between speaking rates.

Based on the orthographic form of the original text, we selected putative realizations of the vowels we wanted to study. These vowel realizations were localized in the speech recordings and the segment boundaries were placed with the help of a visual display of the waveform and auditory feedback. The vowel boundaries were chosen at a zero crossing in the speech waveform. A whole number of pitch periods was used. Any pitch period that could be attributed to the target vowel, and not to the neighbouring phonemes, was considered to be part of that vowel realization. The segments were copied with a leading and trailing edge of 50 ms of speech. Vowel realizations that could not be separated from their context with confidence were not used, contrary to what was done in chapter 2 (Van Son and Pols, 1990). The tokens were labeled for sentence accent and actual phoneme realization. Stress and phoneme labels at the two rates were not always identical but the differences between the speaking rates were not systematic.

3.1.2 Vowels used

Seven of the twelve Dutch monophthongs were used: /i y u o a A E/. These vowels were selected because of their rather high frequency of use in Dutch and their representativeness in the vowel space. Five of the vowels used are short or half-long vowels (/i y u A E/) and two are long vowels (/o a/).

As a neutral "anchor" in the vowel space, a small number of realizations of the schwa was selected as well. These schwa realizations came from the words "HET" = /´t/ (English: "THE") and "ER" = /´r/ or /d´r/ (English: "THERE"). Some other vowels which were reduced to schwa, were included in this group of schwa vowels as well.

The various numbers of vowels thus obtained are listed in table 3.1. Out of 1178 isolated tokens, only equally paired tokens that could be segmented with confidence were used in this study, leaving 549 pairs of tokens.

To assess the importance of stress and vowel context, more homogeneous subsets of realizations of the vowels /E A a i o/ were selected from the total set of tokens and analysed separately: We used tokens with and without sentence-stress and those tokens that occurred in a CVC context in which

Table 3.1: Number of vowel pairs matched on normal- versus fast-rate. Both tokens in a pair are from the same text item. Only pairs with comparable vowel realizations that could be reliably segmented are presented, 38 pairs from the original material were not used and are not included in this Table (see text). The schwa is never stressed. In the last column the number of tokens in an alveolar-vowel-alveolar context is added between parenthesis for some vowels (Dutch alveolar consonants are /n t d s z l r/*, see text).*

vowel	stressed	unstressed	unequal stress	Total	
E	23	85	12	120	(21)
A	23	79	8	110	(33)
a	21	70	11	102	(27)
i	23	57	4	84	(38)
o	17	56	11	84	(16)
´	0	21	0	21	
u	4	7	5	16	
y	5	6	1	12	
Total	116	381	52	549	(135)

both C's were alveolar consonants (i.e., one of /n t d s z l r/, table 3.1). Alveolar consonants can be considered to be closed and fronted phonemes, from an articulatory viewpoint close to the vowel /i/. The target-undershoot model predicts the largest influence of duration when the articulatory distance between consonant and vowel is largest. Therefore, we would expect the largest coarticulatory effects on the F_1 tracks of the open vowels /E A a/ and the F_2 tracks of the back vowel /o/. There were not enough tokens in another (non-alveolar) homogeneous context to merit analysis.

Of the three other vowels, /´ u y/, there were too few stressed tokens or realizations in an alveolar context to enable analysis.

3.1.3 Spectral analysis and formant track sampling method

The vowel segments were analyzed with a 10-pole LPC analysis, using a 25.0 ms Hamming window, which shifted in 1 ms steps (Vogten, 1986). The formant analysis was based on the Split-Levinson algorithm, which gives continuous formant tracks (Willems, 1986).

The formant tracks obtained from the different vowels were sampled at 16 equidistant points, including both boundaries. The linear formant frequency, in Hz, was used. Two tokens (both /i/) were shorter than 16 ms and thus gave less than 16 different frames in a track. From these we doubled some frames to obtain the 16 desired values. Symmetry was preserved by the doubling.

3.2 Results

The formant values and vowel durations were compared for the two speaking rates. Comparisons were done between pairs of tokens taken from readings of the same text items at different speaking rates.

All statistical tests are from Ferguson (1981), all statistical tables from Abramowitz and Stegun (1965, pp.966-990). Correlation coefficients were recalculated to a Student's *t*-test to determine significance. To prevent repeated-test results from containing spurious errors, a two-tailed threshold level for statistical significance of p•0.01% was chosen for testing the point-by-point formant data (16 points per formant per vowel) and a threshold level of p•0.1% was chosen for testing differences in duration (1 value per vowel). When the two speaking rates were tested in parallel, i.e. not pooled, only results that were statistically significant at both speaking rates were considered, because the methods used were not well qualified to distinguish between speaking rates.

3.2.1 Duration

Mean differences of duration between speaking rates were tested (table 3.2). As was to be expected, the fast-rate tokens were shorter than the normal-rate tokens. The difference was around 15% for all vowels combined, intrinsic long vowels (/a, o/) showed a mean shortening of around 20% at a higher speaking rate.

Mean duration was statistically significantly (p•0.1%) shorter for fast-rate tokens than for normal-rate tokens for the vowels /E A a i o/ (table 3.2). Realizations of the schwa did not differ in length between speaking rates. This could be explained by the fact that they were already extremely short. The vowels /u y/ showed no significant differences, probably because of their small numbers (see table 3.1). From the results presented in table 3.2 it was found that the mean duration of the long vowels (V:) was related to the mean duration of the short and half-long vowels (V, excluding /´/) as: V: = a·V - d, in which 'a' and 'd' are speaking-style independent constants (Fant and Kruckenberg, 1989; Koopmans-van Beinum, 1990; they found V: = 1.9·V - 45 ms and V: = 2.05·V - 38 ms, respectively). As only two speaking conditions were available, the coefficient 'a' could not be determined reliably from our data and was chosen to lie between the two published values, i.e. 'a'= 2. The constant 'd' was found to be 45 ms in normal-rate speech and 47 ms in fast-rate speech.

The correlation between vowel duration values of tokens spoken at normal and fast rate was significant for all vowels tested, except for the vowel /´/, and correlation coefficients were larger than 0.71 for all vowels except for the vowels /i ´/ (table 3.2). This meant that the within-speaking-rate variation in duration is preserved between different speaking rates. The lack of correlation between durations of the schwa at fast and normal-rate, could possibly be attributed to the restricted contexts from which these tokens were extracted and the lack of differences between realizations at the two speaking rates.

3.2.2 Effects of speaking rate on formant frequencies

Speaking rate differences resulted in differences in vowel durations and probably also in formant values. Mean formant frequency differences between speaking rates proved to be rather small. In figure 3.1, the differences in formant values between speaking rates are displayed as the

Table 3.2: Mean duration (in ms) of tokens for both speaking rates, and mean difference in duration between speaking rates. The mean duration of short vowels (/E A i u y/*, all tokens pooled) was 86 ms (normal-rate) and 76 ms (fast-rate). The mean duration of long vowels (*/a o/*, all tokens pooled) was 128 ms (normal-rate) and 104 ms (fast-rate). Last column: Correlation coefficient of vowel duration between tokens of the same text item at both speaking rates. Statistical significance is tested with a Student's t-test on difference. Correlation coefficients were recalculated to a Student's t-test variable before testing. Statistical significant differences and correlation coefficients are underlined (level p≤0.1%, last two columns), the others are not significant.*

vowel	normal	fast	normal-fast	Corr. coeff.
E	85	74	11	0.78
A	87	77	10	0.74
a	127	102	26	0.79
i	86	74	13	0.64
o	129	107	23	0.80
´	56	54	2	-0.02
u	89	82	8	0.89
y	92	81	11	0.86
Total	99	84	15	0.82

normal-rate formant frequency subtracted from the corresponding fast-rate formant value, so any deviation from a straight line at 0 value might be interesting. For each vowel, the differences between tokens spoken at different rates, corresponding to a certain point in the token (points 1 through 16), were averaged and the statistical significance was determined by a Student's *t*-test on difference. Statistical significance for individual points was indicated in the legend of figure 3.1.

For F_1, the differences were statistically significant in more than half of the vowel segment (more than 8 points) for the vowels /E A a o/ and in less than half of the vowel segment in /i/ (see figure 3.1 upper panel). The differences in F_1 were small, on the average 20 Hz. The parts showing significant differences did not correspond to a certain position within the vowel. Thus, fast-rate tokens showed a slightly higher F_1 value than normal-rate tokens in all parts of the vowel, irrespective of vowel identity.

Despite quite large differences between mean F_2 values (figure 3.1 lower panel), statistically significant differences were only found in a small part in the second half of /A/. Thus, no consistent differences in frequency were found between F_2 values from vowels spoken at a fast rate as compared with those spoken at a normal rate. This result suggests that there were no large, systematic effects of speaking rate on the shape of the second formant track.

3.2.3 Correlation between speaking rates

The two readings resulted in two correlated sets of formant measurements. The context of each text item was identical in both readings so the formant frequency values measured in tokens of the same text item at different speaking rates might very well be correlated. The correlation coefficient over pairs of tokens of the same vowel is then a measure of the amount of context dependent variance captured with the measurements (see also chapter 2; Van Son and Pols, 1990). These correlations were calculated for each point in the vowels and the resulting correlation coefficients were plotted in figure 3.2.

The values measured at both speaking rates from the same text item, indeed showed high correlation coefficients. The correlations were statistically significant for F_1 in all parts of the vowels /E A a i o/ (figure 3.2 upper panel). For F_1, the correlation coefficients surpassed 0.71 (more than 50% of variance explained) in most parts of the vowels /A o/ and were larger than 0.5 (more than 25% of variance explained) in the vowels /E a i/, i.e. in those vowels that showed significant correlations ($p \bullet 0.01\%$) between F_1 values. The vowels /´ u y/ did not show significant correlations between speaking rates, despite some fairly high correlation coefficients (e.g., for /y/ tokens).

For the second formant (F_2, see figure 3.2 lower panel), the tokens of /E A a o ´/ showed significant correlations between speaking rates ($p \bullet 0.01\%$) in all or most parts of the vowels, the vowels /i u y/ only in small parts. Except for the vowel /i/, the values of the statistically significant correlation coefficients were almost all above 0.71 and thus explained more than half of the variance in most parts of the vowels. Note that the correlation

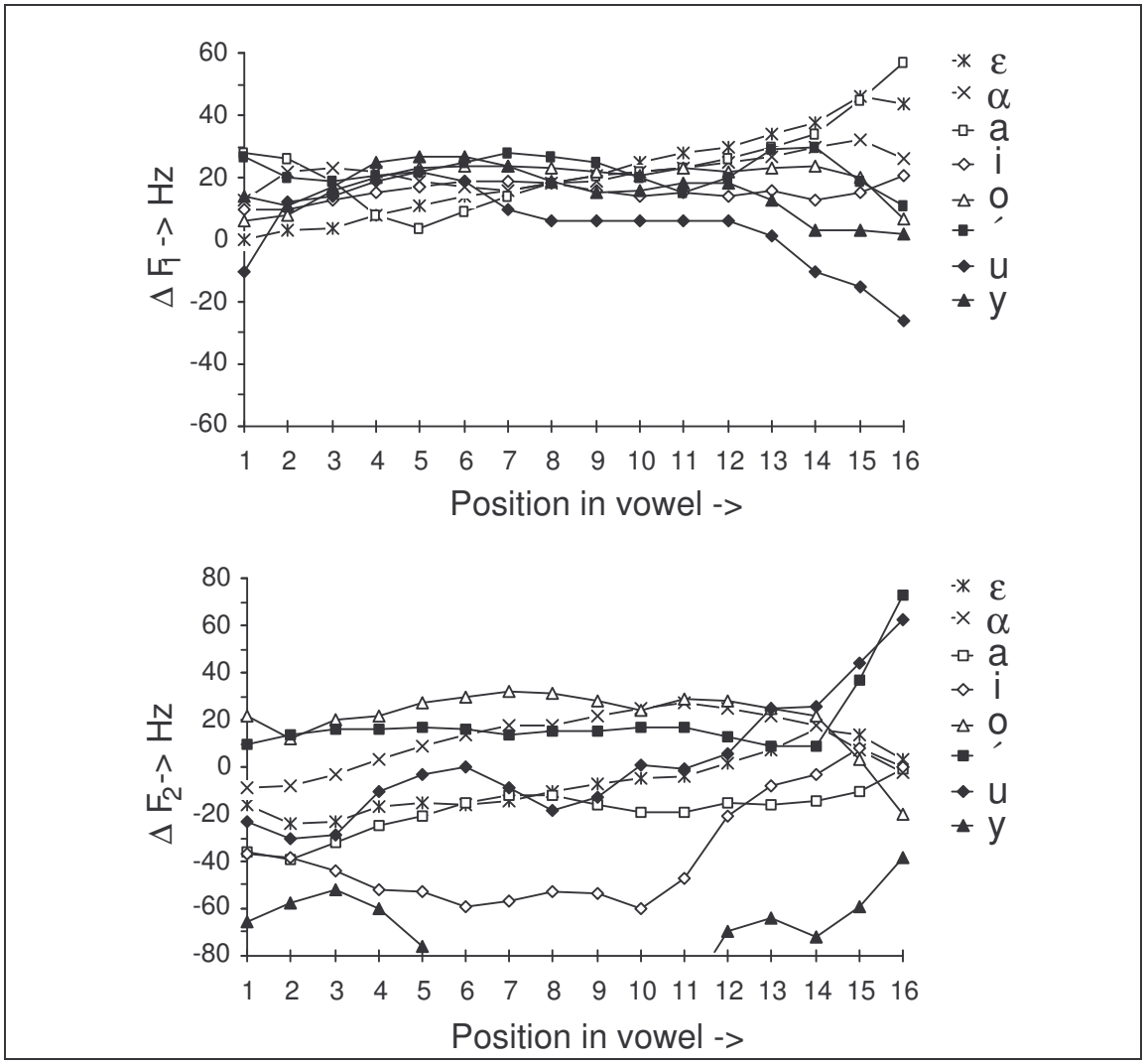

Figure 3.1. Mean differences in formant frequency values in Hz (fast-rate value minus normal-rate value) for all 16 points within the vowels. Statistical significance is determined by a Student's t-test on difference (p≤0.01%). Upper panel: First formant (F_1). The differences are significant at the points /E/*: 7-16;* /A/*: 2-15;* /a/*: 2, 8-16;* /i/*: 3-8;* /o/*: 3-14. Lower panel: Second formant (F_2). The differences are significant at the points* /A/*: 10-12.*

coefficients between the formant values of vowels spoken at normal and fast rate (figure 3.2) were often larger than the corresponding correlation coefficients between vowel durations (table 3.2).

These results indicate that a large fraction of the variation in vowel formant values within each speaking rate was indeed systematic and reproduced when the text was reread.

3.2.4 Effects of duration on formant frequencies

Because durations differed between speaking rates (c.f. section 3.2.1) and F_2 values did not seem to (c.f. section 3.2.2), it would not have been prudent to pool tokens from both speaking rates to calculate correlation coefficients between vowel duration and vowel formant frequency. Therefore, correlation coefficients between formant values and vowel durations were calculated for each speaking rate independently (not shown). The strength of the

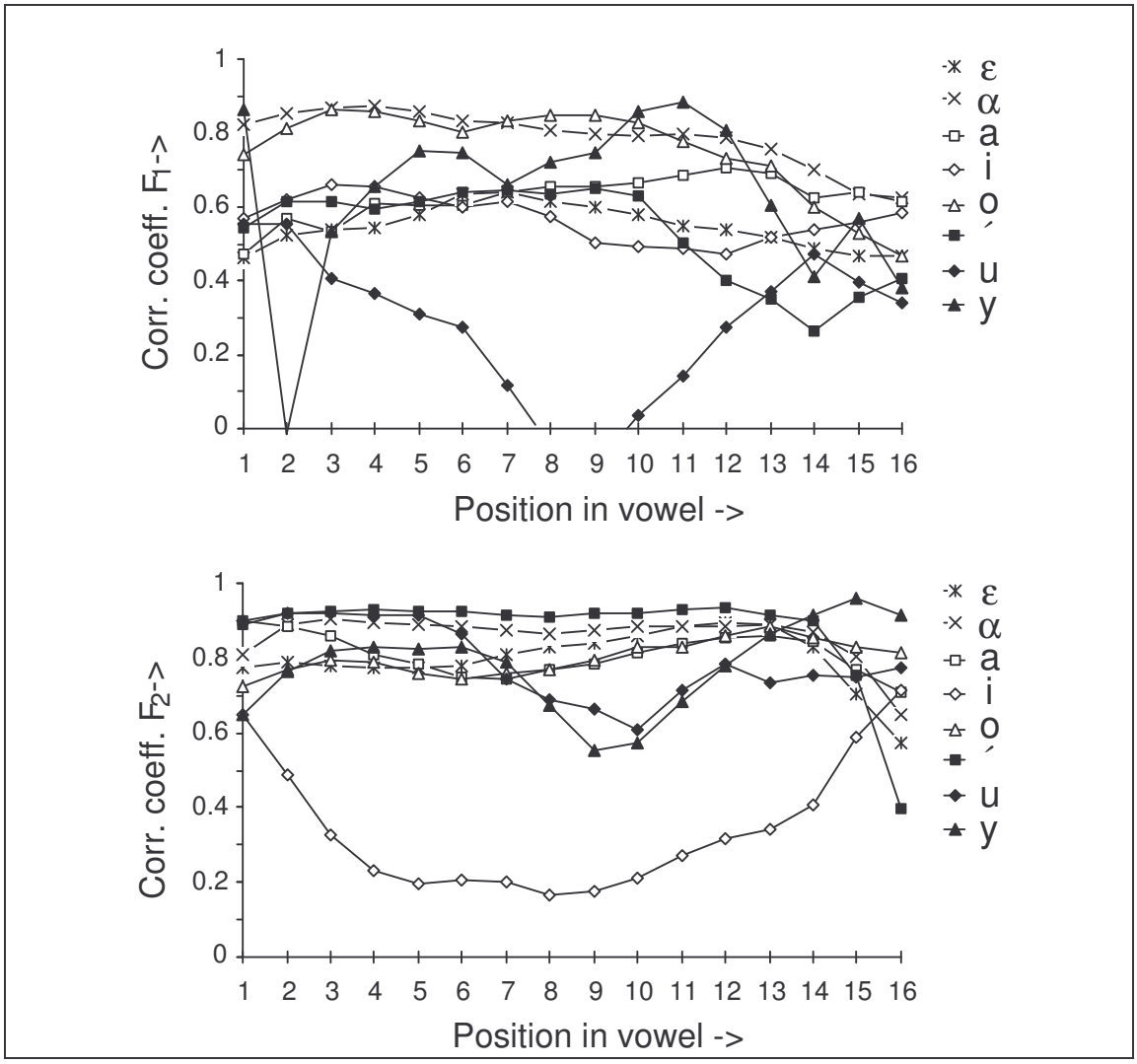

Figure 3.2. Correlation coefficients between formant frequency values measured in fast-rate tokens and the corresponding values measured in normal-rate tokens for all 16 points within the vowels. Statistical significance is determined by recalculating the correlation coefficients to a Student's t-test (p≤0.01%). Upper panel: First formant (F_1). The correlations are significant at all 16 points within the vowels /E A a i o/*. Lower panel: Second formant (F_2). The correlations are significant at all 16 points within the vowels /E A a o/, and at the points* /i/*: 1, 2, 15, 16;* /´/*: 1-15;* /u/*: 1-6;* /y/*: 14-16.*

correlation between formant frequency values and vowel duration denotes the importance of the duration in determining vowel formant frequency (and vice versa). The stronger the relation between formant frequencies and vowel duration, the higher the absolute value of the correlation coefficient between both values. It must be remembered that a lot of variance could be explained due to the strong correlation between speaking rates, both for duration (section 3.2.1) and formant values (section 3.2.2).

The correlation coefficient values between F_1 frequency and vowel duration generally were positive in the center and smaller or negative in the on- and offglide for the open vowels /E A a/ (not shown). This means that realizations of these high F_1 vowels that have a longer duration also have higher F_1 frequencies in the center and equal or lower F_1 frequencies in the on- and offglide part. This can also be described as a decrease in the differ-

ence between the center and the on- and offset frequencies of the F_1 track with a decrease of duration. This indicates a leveling of the formant track with a shorter duration. However, significant correlation strengths between F_1 values and duration were reached for both normal-rate and fast-rate for the vowels /A a/ only (not shown), and there only in a small part (2-8 points) in the center of the vowels. Only in fast-rate tokens of the vowel /a/ did the correlation coefficient surpass 0.5, but then for three sample points only ($|r| \bullet 0.55$). This indicated that the amount of variance explained this way (i.e., less than 25%) was small but could still be of importance.

For F_2, none of the vowels showed a statistically significant correlation between formant values and vowel duration for both speaking rates (not shown). There was no measurable relation between vowel duration and F_2 frequency values.

3.2.5 Effects of context

The tokens of the vowels /E A a i o/ in an all alveolar CVC context (C is one of /n t d s z r l/) were also analysed. The number of tokens per vowel available in an alveolar context was quite small (n = 16-38, table 3.1). For small numbers, the estimated parameter values will have a large error. Therefore, we concentrated on the relation between the tokens in the subset and those of the parent set and not on the actual sizes of the differences between the two sets. For this analysis, a threshold level of significance of $p \bullet 0.1\%$, reached at two or more points within a vowel, was sufficient.

The fast-rate tokens of this subset had a uniform higher mean F_1 frequency than the normal-rate tokens but the difference was not statistically significant ($p>0.1\%$ at all points). The between-speaking-rate correlation coefficients of the formant frequencies were high for both F_1 and F_2, often higher than those for the parent set. The trends were the same as in the parent set of tokens.

The correlation coefficients between formant frequencies and vowel duration were generally higher in the subset of tokens in alveolar context than in the parent set, especially for F_1 of /A a/. Still, only few correlation coefficients were statistically significant (F_1 in the center of /A/, $p \bullet 0.1\%$ for more than 2 points).

These results show that the tokens from the subset of vowels in alveolar context were not different from the complete parent set of vowel tokens.

3.2.6 Effects of stress

The previous analyses were repeated on token pairs of the vowels /E A a i o/ for which both tokens were stressed or unstressed (data not shown). This was done to check whether sentence-stress might be significant with respect to the effects of differences in speaking rate or duration.

Stressed tokens were 30% longer than the unstressed ones for both speaking rates ($p \bullet 0.1\%$). The differences in vowel duration between speaking rates were comparable for stressed and unstressed tokens (i.e., 15%). The mean duration of the long vowels (V:) was related to that of the short

vowels (V) as V: • 2·V - 54 ms in stressed tokens and V: • 2·V - 43 ms in unstressed tokens (cf. 3.2.1).

For the F_1, formant frequencies of the stressed tokens were generally higher than those of the unstressed tokens at both rates. This difference was largest for the high F_1-target vowels (p•0.01% in the center of /A/ for both speaking rates). The vowel space of the stressed tokens was larger, i.e. less reduced, in the F_1 direction (/i/ to /a/) than that of the unstressed tokens. There was no indication that, compared to stressed tokens, unstressed tokens were spectrally reduced with respect to the F_2. The fast-rate stressed and unstressed tokens had a uniform higher F_1 than the normal-rate tokens. For unstressed tokens the difference was statistically significant (p•0.01%). For stressed tokens the difference was smaller than for unstressed tokens and not statistically significant (p>0.1%).

Correlation coefficients between speaking rates were higher in stressed tokens than in unstressed tokens and statistically significant for both (p•0.01%). The reverse was found for the correlation between formant values and vowel duration. For both stressed and unstressed tokens the correlation between formant values and vowel duration was never statistically significant (p>0.1%) for both speaking rates. As far as could be checked, the results obtained from all tokens pooled were equally valid for both subsets of tokens individually.

3.3 Discussion

3.3.1 Effects of speaking rate

The difference in vowel duration between tokens spoken at normal and fast rate was small but consistent. In fact, the difference was only half of what would have been expected from the overall difference in duration of both readings, which was 25% (see section 3.1.1). For both readings the mean duration of long vowels (V:) was twice the mean duration of short vowels (V) minus a constant, i.e. V: • 2·V - 46 ms. From this relation it follows that the absolute difference in vowel duration between speaking rates should have been approximately twice as large for long vowels than for short vowels. But this relation does not explain why the overall differences were so small. A possible explanation could be that vowels are more resistant to durational compression than other phonemes. Indeed, this was found by Eefting (1991) using the same speaker.

In other studies, larger differences in vowel duration were found between speaking styles and rates (e.g., Lindblom and Moon, 1988) than in the present study. These studies used speech which contained longer vowel realizations than did our speech material. Starting with (much) shorter vowel realizations from a long read text, the small reductions in vowel duration found in this study were likely to strain the articulatory capabilities of our speaker more than did the much larger reductions of vowel duration in studies which used isolated words or sentences. As the articulatory models discussed before emphasize articulatory effort as an important factor influencing vowel formant tracks, even this relatively small reduction should have had a measurable effect on vowel formant tracks.

Despite the fact that the fast-rate vowel realizations were generally (and consistently) shorter than the normal-rate realizations, there was hardly a difference between the formant frequency values measured at different speaking rates. This means that a difference in speaking rate did not result in systematic differences in formant values. Only the F_1 frequency is higher in vowels spoken at a fast rate compared to vowels spoken at a normal-rate. This rate-dependent rise in F_1 frequency was present irrespective of vowel identity and it was uniform (independent of the position inside the vowel). This means that the equivalent results found in chapter 2 (Van Son and Pols, 1990) for vowel nucleus measurements cannot be attributed to a change in formant track shape due to speaking rate. It also indicates that our speaker increased articulation speed when he spoke faster. This increase in articulation speed matched the decrease in vowel duration.

3.3.2 Effects of duration on formant tracks

A simple, one-way, relation between vowel formant tracks and vowel duration would result in a clear-cut, and strong, correlation between these two. However, correlation coefficients between formant frequencies and vowel duration were only significant for the F_1 tracks of the high-F_1 target vowels (/A a/). The correlations implied a leveling off of the F_1 tracks with shorter durations of the tokens. This is predicted by the target-undershoot model. However, the correlation coefficients were rather small in all cases. The correlation between formant frequency and vowel duration hardly explains more than 30% of the variance in formant frequencies ($|r| \bullet 0.55$, section 3.2.3). Between-speaking-rate correlations for these three vowels, which measure the context dependent variation captured by the measurements, sometimes explained up to 70% of the variance in F_1 formant frequencies ($|r| \bullet 0.85$, figure 3.2 upper panel). This difference in correlation indicated that duration is not a major determinant of overall vowel formant track shape in read speech.

F_2 formant tracks did not show any sizeable correlation between formant track frequency and vowel duration.

3.3.3 Effects of context and stress

The context in which a vowel is spoken might be of importance for changes in speaking rate (or changes in duration). We compared the results for stressed with those for unstressed token pairs and also the results for tokens from an alveolar context with those from all tokens pooled.

Stressed vowel tokens were generally longer than the unstressed tokens and spectrally less reduced (at least for F_1). No differences between stressed and unstressed tokens were found when the effects of changes in speaking rate or duration were considered. The difference in duration between stressed and unstressed tokens was twice the difference between speaking rates. There was a difference in F_1 formant frequency between stressed and unstressed tokens but stressed and unstressed tokens did not differ in the way speaking rate affected their formant frequencies, i.e. F_1 was higher in fast-rate speech, although the size of the effect of speaking rate might have been smaller in stressed tokens than in unstressed tokens.

All this indicates that vowel duration alone is not enough to explain the differences between stressed and unstressed vowel realizations. This confirms the results of Nord (1987).

For tokens from an alveolar CVC context, the same uniform higher F_1 frequency in the fast-rate tokens was found as in the parent set. There was the same lack of effect of either speaking rate or duration on the F_2. These results indicate that if coarticulation from an all-alveolar context was stronger in fast-rate speech than in normal-rate speech, the difference was too small to be measured by the methods used in this paper. We were only able to test a subset of Dutch vowels and consonants. It is still possible that other CVC combinations are more strongly affected by speaking rate changes.

To summarize, the trends observed in vowel realizations in our parent set were also present in the stressed and unstressed realizations and in the realizations from an alveolar-vowel-alveolar context. Therefore, we conclude that the variation of these textual factors in our data did not influence the results we obtained.

3.4 Conclusions

This study was limited in that only one speaker was used who read aloud a single text. From the results we conclude that this speaker did not behave as predicted by the target-undershoot model. Even the refined versions of the target-undershoot model that incorporate alternative articulation strategies (Gay, 1981) and increased effort (Lindblom, 1983) would predict some measurable differences in formant frequency values between speaking rates. That these differences were not found indicates that these theories are not universally valid for all speakers using continuous read speech. We found evidence that they might explain some aspects of the relation between vowel duration and formants within a single speaking style. However, our study indicates that their explanatory powers are limited and probably speaker specific.

The results presented here indicate that the articulatory effects of differences in vowel duration *between* speaking rates (and probably speaking styles) are not the same as the effects of differences in vowel duration *within* a single speaking rate (or style). This difference should be addressed by articulation theories based on the target-undershoot model. It is also clear that our speaker was readily able to actively adapt his articulation to a fast speaking rate. It is therefore unlikely that articulation speed is a limiting factor in his vowel pronunciation as is implied by the target-undershoot model.

4

THE INFLUENCE OF SPEAKING RATE ON VOWEL FORMANT TRACK SHAPE AS MODELED BY LEGENDRE POLYNOMIALS*

Abstract

Speaking rate in general, and vowel duration more specifically, is thought to affect the dynamic structure of vowel formant tracks. To test this, a single, professional speaker read a long text at two different speaking rates, fast and normal. The present project investigated the extent to which the first and second formant tracks of eight Dutch vowels varied under the two different speaking rate conditions. A total of 549 pairs of vowel realizations from various contexts were selected for analysis. Legendre polynomial functions were used to model and quantify the shape of normalized formant tracks. No differences in normalized formant track shapes were found that could be attributed to differences in speaking rate. But a higher F_1 frequency in fast-rate speech relative to normal-rate speech was found that can be explained as the result of a uniform change in frequency. These results indicate a much more active adaptation to speaking rate than implied by the target-undershoot model. Within each speaking rate, there was only evidence of a weak leveling off of the F_1 tracks of the open vowels /E A a/ *with shorter durations. These same conclusions were reached when sentence-stress was taken into consideration and when vowel realizations from a more uniform, alveolar-vowel-alveolar, context were examined separately. In the alveolar context, a small rise in F_2 of the vowel* /o/ *might indicate more coarticulation in fast-rate speech.*

*Adapted from: Van Son, R.J.J.H. & Pols, L.C.W. (1991). "The influence of speaking rate on vowel formant track shape as modeled by Legendre polynomials", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 15, 43-59.

Introduction

Vowel duration is generally considered an important parameter in determining the pronunciation of vowels and therefore of vowel formant tracks (e.g., Lindblom, 1963; Broad and Fertig, 1970; Gay, 1978, 1981; Lindblom, 1983; Broad and Clermont, 1987; Di Benedetto, 1989a; Lindblom and Moon, 1988; Moon, 1990). Vowel duration is important for the shape of the overall formant tracks. The target-undershoot model (Lindblom, 1963, 1983) is often cited to explain vowel formant behaviour under different speaking conditions. It predicts more coarticulation when vowels become shorter. In a large sample of normal speech, with typical utterances, this averages out to more spectral reduction, i.e. more schwa-like formant values in the vowel nucleus and more level (less curved) formant tracks (cf. Koopmans-van Beinum, 1980; Van Bergem, 1993). In a previous study we found that there was no evidence for an increased reduction or more coarticulation in fast-rate speech of a highly experienced speaker (chapter 1; Van Son and Pols, 1990), at least not in the vowel nucleus.

Relatively few studies have considered the relation between vowel formant dynamics and duration (exceptions are Broad and Fertig, 1970; Broad and Clermont, 1987; Di Benedetto, 1989a) and these were limited to only one speaking style. Studies that did use different speaking styles or different speaking rates generally only measured formant frequencies within the vowel nucleus (but see chapter 3; Van Son and Pols, 1989, 1992). Therefore, it is not clear whether fast-rate speech is just "speeded-up" normal-rate speech, or whether different articulation strategies (as proposed by Gay, 1981) or a higher speaking effort (Lindblom, 1983) are used. Differences in articulation or speaking effort should result in different shapes of the formant tracks, e.g. a levelling-off or, conversely, an amplification of the formant movements in fast-rate speech.

Formant track shape is generally characterized by the lengths and slopes of vowel on- and off-glide which are conventionally measured using two to four points from each formant track (Di Benedetto, 1989a; Strange, 1989 a, b; Duez, 1989; Krull, 1989). However, it is very difficult to determine the boundaries of the stationary part (Benguerel and McFadden, 1989) and to measure formant track slopes accurately. Therefore, another method was developed to characterize formant track shapes. First vowel formant tracks were sampled (16 points, adapted from Broad and Fertig, 1970). Second, the global "shape" of the sampled formant tracks was modeled with Legendre polynomials of order 0-4 (see section 4.1.1). This modeling approach was used to investigate the effects of speaking rate on vowel formant track shape. In chapter 3, this problem was studied using the 16 equidistant points directly (cf. Van Son and Pols, 1992).

Differences between speaking rates are best studied by using vowel realizations that differ *only* in speaking rate. In order to obtain a large and varied inventory of such vowel pairs, a long text was read twice by a single professional speaker, once at a normal rate and once at a fast rate (Van Son and Pols, 1990). With these vowels, we have tested whether vowel formant track shape depends on vowel duration and speaking rate and how this re-

lation can be modeled. The effects of stress and vowel context were also taken into account.

4.1 Methods

The work presented in this chapter used the sampled formant track values obtained in chapter 3. We refer to that chapter for a description of the vowel segments and the methods used to obtain the sampled formant tracks. For convenience we reproduce the table with the number of vowel realizations used (table 4.1, which is identical to table 3.1).

4.1.1 Measuring differences between formant tracks

Legendre polynomial coefficients of order 0-4 were used as measures of formant track shape, see table 4.2 and figure 4.1 (appendix B; Churchhouse, 1981; Abramowitz and Stegun, 1965, pp.773-802). The Legendre polynomials are the simplest set of orthogonal polynomials and are generally easier to use than other sets. For practical reasons, we used the shifted Legendre polynomials which are defined on the base [0,1] instead of [-1,1].

An analysis using Legendre polynomials is a kind of regression analysis. The Legendre polynomial coefficients are calculated as a linear combination of the formant track sample points (see appendix B). Therefore, when the data points have a Gaussian distribution, all the coefficients also have a Gaussian distribution and the corresponding statistics can be used to test for differences between Legendre coefficients. The coefficients include the mean value (order 0) and linear regression slope (order 1). The second-order coefficient measures the parabolic excursion within a vowel realization, independent of the overall slope of the formant track. The third- and fourth-order coefficients measure, among other things, the amount of "stability" in the central part of the vowel (c.f. figure 4.1). The Legendre polynomials are orthogonal, meaning that the Legendre polynomial coefficients that describe track shape are mathematically independent. Because the zeroth-

Table 4.1: Number of vowel pairs matched on normal versus fast rate. Both tokens in a pair are from the same text item. Only pairs with comparable vowel realizations that could be reliably segmented are presented, 38 pairs from the original material were not used and are not included in this table (see text). The schwa is never stressed. In the last column the number of tokens in an alveolar-vowel-alveolar context is added between parenthesis for some vowels (Dutch alveolar consonants are /n t d s z l r/*, see text).*

vowel	stressed	unstressed	unequal stress	total	
E	23	85	12	120	(21)
A	23	79	8	110	(33)
a	21	70	11	102	(27)
i	23	57	4	84	(38)
o	17	56	11	84	(16)
´	0	21	0	21	
u	4	7	5	16	
y	5	6	1	12	
total	116	381	52	549	(135)

Figure 4.1.a: The first five Legendre polynomials, L_0-L_4. The polynomials are drawn with different Legendre coefficients P_i (actually the function $P_i \cdot L_i$ is drawn): P_0=1, P_1=P_2= -0.5, P_3=P_4= -0.25.

order measures the mean formant frequency, the results for this order should be identical to those found with the Averaging method in Van Son and Pols (1990) which uses the same speech data (see chapter 2).

Calculation of the Legendre polynomial coefficients was done by integration of the product of the sampled formant track and the appropriate Legendre polynomial function. We used the closed-type Newton-Cotes formulas to perform the numerical integration (Abramowitz and Stegun, 1965 p.886; appendix B). Because no 15th-order version of the Newton-Cotes formulas was available, we integrated the 15 intervals between the 16 track samples in two parts with the Legendre functions. The first part with the leading eight intervals (eighth-order Newton-Cotes formula) and the second part with the trailing seven intervals (seventh-order Newton-Cotes formula).

Legendre polynomials are used to model data points. The remaining variance after the fit is calculated by subtracting the variances of the various order polynomials, defined as $P_i \cdot P_i/\{1+2 \cdot i\}$ (P_i is the Legendre polynomial coefficient and i the order, Abramowitz and Stegun, 1965 pp.773-802; Churchhouse, 1981), from the original variance of the function. The remaining error (i.e., the RMS error) is the square-root of the remaining variance. The precision of the coefficients, especially the higher order ones, is limited by the precision of the calculations and the incomplete equivalence between the integration of continuous functions and the numerically integration of sampled data. However, this proved to be no problem.

Table 4.2: First five shifted Legendre polynomials and their slope at three points.
The polynomials, $L(\tau)$, are defined between 0 and 1 (inclusive). Next to the expressions the slope values of the polynomials are given for three points in the first half of the interval. The relative time τ is defined as time/duration ($0 \le \tau \le 1$). $L_i(0) = 1$ for even-order polynomials and $L_i(0) = -1$ for odd-order polynomials, $L_i(1) = 1$ for all polynomials. Even-order polynomials are symmetrical and odd-order polynomials are anti-symmetrical, i.e. if $-0.5 \le \varepsilon \le 0.5$ and $L_i' = dL_i/d\tau$ then $L_i(0.5+\varepsilon) = L_i(0.5-\varepsilon)$ and $L_i'(0.5+\varepsilon) = -L_i'(0.5-\varepsilon)$ if i is even and $L_i(0.5+\varepsilon) = -L_i(0.5-\varepsilon)$ and $L_i'(0.5+\varepsilon) = L_i'(0.5-\varepsilon)$ if i is odd (Adapted from Abramowitz and Stegun, 1965).

order	L_i ($0 \le t \le 1$)	$L_i'(0)$	$L_i'(0.25)$	$L_i'(0.5)$
0	1	0	0	0
1	$2 \cdot \tau - 1$	2	2	2
2	$6 \cdot \tau^2 - 6 \cdot \tau + 1$	-6	-3	0
3	$20 \cdot \tau^3 - 30 \cdot \tau^2 + 12 \cdot \tau - 1$	12	0.75	-3
4	$70 \cdot \tau^4 - 140 \cdot \tau^3 + 90 \cdot \tau^2 - 20 \cdot \tau + 1$	-20	3.125	0

Figure 4.1.b: Example of Legendre polynomials and their use in modeling functions. Tracks composed of different Legendre polynomials, using the same coefficient as in 4.1.a. Top: $1L_0$ - $0.5L_1$ - $0.25L_3$, bottom: $1L_0$ - $0.5L_2$ - $0.25L_4$. When formant frequency tracks are modeled, the horizontal axis represents the normalized time and the vertical axis the formant frequency in Hz. Note that tracks are shaped like formant tracks.

4.2 Results

The formant tracks were compared for the two speaking rates. Comparisons were done between pairs of tokens taken from readings of the same text items at different speaking rates.

All statistical tests are from Ferguson (1981), all statistical tables from Abramowitz and Stegun (1965 pp.966-990). Correlation coefficients were recalculated to a Student's *t* (Ferguson, 1981) to determine significance. To prevent repeated-test results from containing spurious errors, a two tailed threshold level for statistical significance of $p \bullet 0.1\%$ was chosen for testing Legendre polynomial coefficients (five values per formant per vowel). When the two speaking rates were tested in parallel, i.e. not pooled, only results that were statistically significant at both speaking rates were considered, because the low numbers of realizations prevented us from distinguishing between speaking rates.

Vowel tokens spoken at a fast rate were 15% shorter (on average) than tokens spoken at a normal rate. The difference was consistent for all vowels except /´/ and statistically significant for /E A a i o/ ($p \bullet 0.1\%$). The correlation between vowel durations at different speaking rates was high and statistically significant ($p \bullet 0.1\%$, $0.64 \bullet r \bullet 0.89$ except for /´/).

4.2.1 Goodness of fit

The Legendre polynomials were meant to model formant track shape. It was therefore important to know how well they fit the formant tracks and how much each order contributes to the overall fit (see section 4.1.1). In table 4.3, the proportion of variance (in percent), explained by each component was calculated for individual tokens and then averaged over all tokens. The contribution of the zeroth-order component (the mean formant frequency) represents the variance around zero frequency, which is not instructive for models of formant track *shape*. Therefore, the zeroth order component was left out: the variance was calculated around the mean frequency. Also, the remaining part of the variance left after the fit (the RMS error) was calculated.

In table 4.3 it can be seen that the bulk of the variance in the individual formant tracks could be explained by the first- and the second-order polynomials (65% - 93%). The remaining variance, left after fitting all Legendre polynomials up to order 4, was between 1% and 12%. The proportion of the variance that remained after the fit, tended to be higher when there was less movement in the formant tracks, i.e. when there was only a small variance to explain (e.g., F_1 of /u o y i/). For most vowel formant tracks, the amount of variance explained decreases with the order of the Legendre coefficient. Exceptions are the F_1 tracks of the vowels /E A a/, and the F_2 track of the vowel /i/. For these formant tracks the second-order coefficient explains most of the variance (up to 66%, table 4.3), making it the determining factor of track shape.

Table 4.3: Mean percentage of formant track variance around the mean formant frequency (i.e., excluding the zeroth-order Legendre coefficient) explained by the higher order Legendre polynomials (order 1-4) for each vowel. In the last column (rest), the mean percentage of the remaining (i.e., not explained) variance is given. Tokens from both speaking rates are pooled.

vowel		1	2	3	4	rest
E	F_1	39	54	3	2	2
	F_2	51	32	9	4	4
A	F_1	31	61	5	2	2
	F_2	67	17	8	3	5
a	F_1	25	66	4	2	3
	F_2	62	23	7	4	5
i	F_1	51	21	15	6	7
	F_2	38	42	7	5	7
o	F_1	40	29	17	5	9
	F_2	47	32	10	7	5
´	F_1	58	32	6	3	1
	F_2	56	26	9	5	4
u	F_1	47	18	14	9	12
	F_2	60	31	4	3	2
y	F_1	37	37	14	6	6
	F_2	82	10	3	2	3

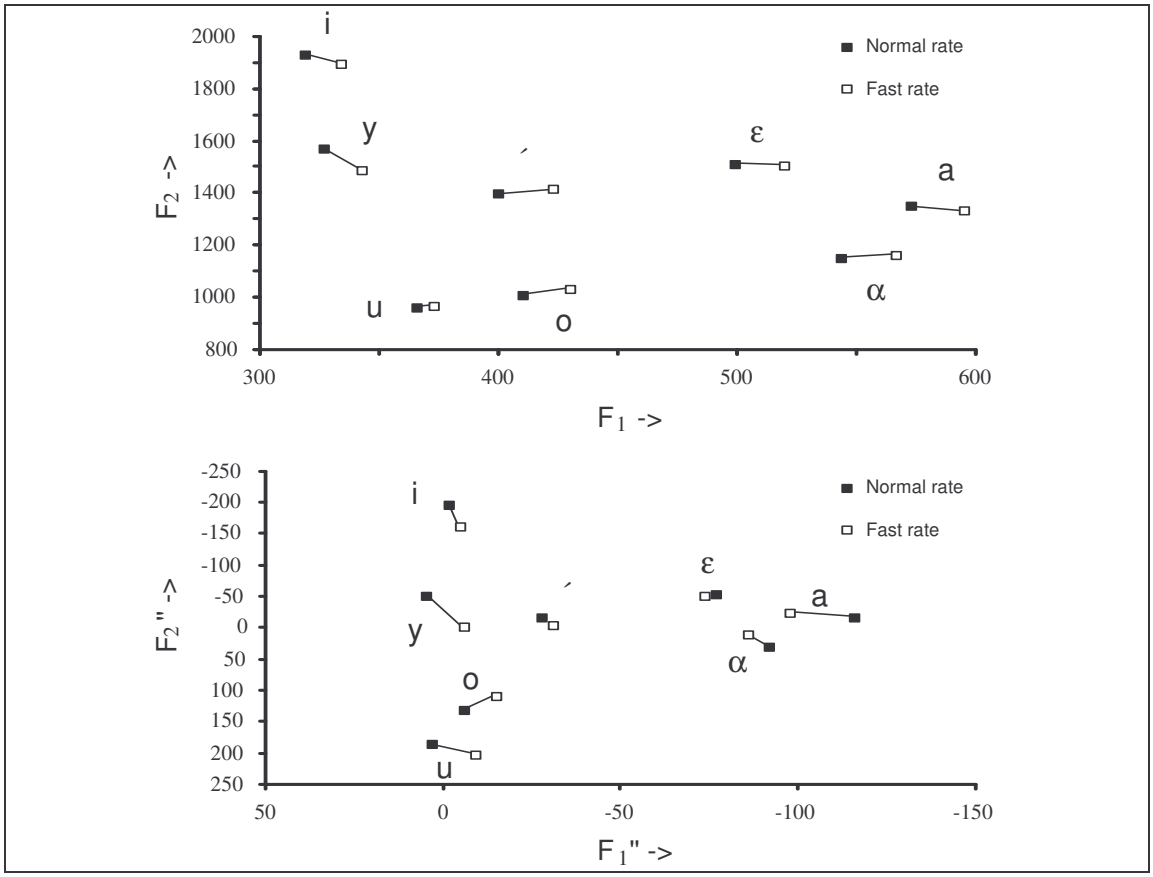

Figure 4.2: Vowel space (F_1/F_2 space) constructed by plotting mean Legendre polynomial coefficient values for the second formant frequency against the mean coefficient values for the first formant frequency for all vowels used. Filled squares: normal-rate tokens, open squares: fast-rate tokens. Upper panel: Zeroth-order Legendre polynomial coefficients P_0 (i.e., mean formant frequency within the realization). This plot results in the normal vowel triangle. Lower panel: Second-order Legendre polynomial coefficients P_2 (F_n", note reverse axes).

4.2.2 Legendre polynomial coefficients and their interpretation

In table 4.4 the mean values of the Legendre coefficients are presented for the orders 0-2. Of all polynomial coefficients, only the zeroth- and second-order coefficient values differed systematically (i.e., statistically significant for both speaking rates) from zero. Almost all mean first-order coefficient values were negative but only a few values were statistically significantly different from zero for both speaking rates (F_2 of /A/). Therefore, the first order polynomial coefficient, which corresponds to the linear regression slope, was important for describing the shape of each individual formant track (see previous section), but the sign of the coefficient (i.e., the slope) was not determined for any vowel.

The zeroth-order coefficient corresponds to the mean formant frequency. It is known that the value of the mean formant frequency is a strong cue to vowel identity (e.g., see chapter 1; Van Son and Pols, 1990). The value of the second-order coefficient can be interpreted as an excursion size relative to a straight line, i.e. the difference between maximum and minimum value of the second-order polynomial.

From the formulae in table 4.2 it follows that this excursion size is 1.5 times the value of the second-order coefficient (in Hz). For F_1, the values of the mean second-order coefficient were between 5 and -116 (table 4.4.a), which amounts to excursion sizes of between 0 and about 180 Hz. For F_2, the mean second-order coefficient values were between -196 and +203 (table 4.4.b), which corresponded to excursion sizes (absolute values) between 0 and approximately 300 Hz. These values are in line with the differences between formant values of vowel onset and nucleus found by Di Benedetto (1989a) for F_1, and Krull (1989) and Weismer et al. (1988) for F_2. These studies also show that much larger excursion sizes are found when speaking styles other than reading a text are involved (reference speech in Krull, 1989), or with certain consonant-vowel combinations that were hardly or not at all present in the speech material used here (e.g., /w/ context in Weismer et al., 1988; /u/ in Krull, 1989). The fact that in a variable context the mean excursion size of some vowels was systematically, and substantially, different from zero indicates that formant excursion size could be used to determine vowel identity (see below).

The mean third- and fourth-order coefficient values were not statistically significantly different from zero, except the fourth-order coefficient values of F_1: 9, 16 for /a/ and F_2: 32, 37 for /o/, normal and fast respectively (data not shown). Also, the contribution of the third- and fourth-order polynomials to the total fit were small and often negligible (table 4.3). Therefore, we will not discuss them in the remaining part of this paper. We did use them to estimate the slope values (see below).

From the polynomial coefficients, the normalized slope at each point in

Table 4.4.a: Mean values of first formant (F_1) in Hz.Legendre polynomial coefficients (order 0-2) and calculated mean value of normalized slope at τ = 1/4 and τ = 3/4 (SL 1/4 and SL 3/4 in Hz/segment, see table 4.2) Mean values that are statistically different from zero are underlined (Student's t-test, p≤0.1%). Whenever the fast-rate value differs significantly from the normal-rate value, this is indicated with a "" (Student's t-test on difference, p≤0.1%). Normal-rate: top row (N), fast-rate: bottom row (F).*

vowel		0	1	2	SL 1/4	SL 3/4
E	N	499	-33	-77	-161	-297
	F	* 520	* -9	-74	199	-241
A	N	544	-21	-92	236	-324
	F	* 567	-15	-86	213	-280
a	N	573	-24	-116	252	-338
	F	* 595	-10	* -98	249	-287
i	N	319	-12	-2	-21	-21
	F	* 334	-11	-5	-6	-24
o	N	410	-14	-6	18	-62
	F	* 430	-10	-15	41	-65
´	N	400	-32	-28	16	-139
	F	423	-33	-31	18	-144
u	N	366	-11	-3	14	-57
	F	373	-26	-9	-15	-82
y	N	327	13	5	5	54
	F	343	5	-6	12	23

the original formant tracks was approximated by summing the values of the slopes of the individual Legendre polynomials at these points (table 4.2), multiplied by the corresponding Legendre coefficient. We calculated the normalized slopes at points at one-fourth (SL1/4) and three-fourths (SL3/4) of the normalized duration of each vowel and averaged them just like the Legendre coefficients (table 4.4, last two columns). These two points are positioned to lie in the on-and off-glide of the vowels, except for the long vowels, /a o/, where they may occasionally lie in the vowel nucleus.

The slopes in the on- and off-glide parts of the vowels, as estimated from all five Legendre polynomials, differed in a systematic way from zero for many vowels but were nevertheless difficult to interpret. Often the absolute values of the slopes on the onglide of the tokens were very different from those on the offglide (table 4.4). This difference showed that vowel formant track shapes were generally asymmetric.

The differences in slope of the formant tracks between fast- and normal-rate tokens (after time-normalization) were never statistically significant and thus did not help us to determine the effects of speaking rate on formant track dynamics.

4.2.3 Relations between polynomial components

The mean values of the zeroth- and second-order coefficients were linked together: higher zeroth-order coefficient values were accompanied by lower (more negative) second-order coefficients. Negative second-order coefficients imply a maximum in the formant track, positive coefficients imply a minimum. This correlation was statistically significant for all vowels pooled ($|r| = 0.6$, $p \bullet 0.1\%$). In the upper panel of figure 4.2, the mean zeroth-order coefficient values are plotted, F_2 against F_1, for both speaking rates (compare figure 2.1; Van Son and Pols, 1990). In the lower panel, the second-order coefficients are presented. For both orders, the mean coefficient values of the individual vowels form the familiar vowel

Table 4.4.b: As table 4.4.a. Second formant (F_2)

vowel		0	1	2	SL 1/4	SL 3/4
E	N	1507	-55	-53	23	-249
	F	1500	-35	-49	41	-192
A	N	1146	-51	31	-160	-31
	F	1159	-40	* 11	* -89	-69
a	N	1349	-38	-16	-65	-117
	F	1329	-26	-23	2	-121
i	N	1929	-67	-196	447	-724
	F	1892	-40	-162	358	-528
o	N	1009	-30	132	-339	221
	F	1031	-35	111	-305	156
´	N	1396	-7	-15	55	-85
	F	1414	1	-4	88	-60
u	N	960	-35	187	-605	432
	F	962	2	203	-603	597
y	N	1568	-157	-49	-145	-471
	F	1487	-157	-1	-388	-219

triangle. For the zeroth-order coefficient values this was expected, for the second-order coefficient values this was new. Presupposing random ordering, the probability of just this constellation for the mean second-order coefficients is less than 0.1% (in the upper panel /i y u o A a E ´/ are ordered in a spiral, the probability of just such a spiral in the lower panel is 4·8/8!•0.0008, allowing for the freedom to choose the signs of the axes (2·2=4) and the ambiguity of the order of a single pair (/u o/: 8)).

Figure 4.2 suggests that in the F_1 direction the second-order coefficient values could be interpreted as a measure of openness: closed has value zero, e.g. the vowels /u y i/. In the F_2 direction it could be interpreted as a measure of front- versus back-articulation: schwa has value zero (i.e., flat), /u/ is positive (i.e., a minimum) and /i/ is negative (i.e., a maximum). Based on the second-order polynomial coefficient and the vowels used here, the vowels could be grouped in distinguishable sets. This meant that the vowel-sets /u o/, /y/, /i/, /E A a/ and /´/ could be distinguished from each other with statistical significance (p•0.1%, Students-*t* test on means of F_1 or F_2), by only using the value of the second-order coefficient of individual vowel realizations. This fact and the large contribution to the overall shape of the formant tracks (especially F_1, see section 4.2.3) suggested that the second-order coefficient could be an important cue of the relation between vowel identity and vowel formant track shape.

The correlation between zeroth- and second-order Legendre coefficients was not statistically significant for the tokens of any *single* vowel (|r| • 0.15 none significant, not shown), contrary to what was found when all vowel realizations were pooled. Therefore, zeroth- and second-order Legendre coefficient values can be considered to be independent apart from being both related to the vowel identity.

Correlations between different orders of Legendre polynomial coefficients were not always small. Of all correlations between all different order coefficient values from tokens of the same vowel, approximately 7% was statistically significant (p•0.01% each). However, we could not find any pattern in these correlations (data not shown). From this we inferred that the contributions of polynomials of different orders were indeed independent from each other, but that extraneous (e.g., textual) factors could have caused correlations between polynomial coefficients of different orders that depended on the distribution of these factors in the text.

4.2.4 Effects of speaking rate

The zeroth-order component (i.e., mean formant value) of F_1 from the vowels /E A a o/ (table 4.4.a) showed a higher fast-rate value compared to the normal-rate value. The other, higher order, components rarely showed statistically significant differences between speaking rates, only first-order F_1 of the vowel /E/, and second-order F_1 of the vowel /a/ and F_2 of the vowel /A/ (table 4.4.a, b). From this we can conclude that the F_1 frequency of fast spoken vowels is higher than the F_1 frequency of tokens spoken at a normal rate. The difference is uniform and irrespective of vowel identity.

Correlations between speaking rates of the zeroth-order (mean value) component were high and statistically significant (p•0.1%, table 4.5). First-

order coefficient values showed significant correlations between speaking rates, but generally with lower correlation coefficients than those of the zeroth-order components. Second-, third- and fourth-order components often showed statistically significant correlations between speaking rates, especially for F_2 (table 4.5, only second-order is shown). The correlation coefficients of F_2 were higher than those of F_1 in most vowels. The correlation coefficients decreased with increasing order but still remained quite high (up to r=0.74 for /o/, third-order F_2, not shown). These results led to the conclusion that higher order components of formant tracks contained information that was preserved between speaking rates. All different order components could be used to investigate the effects of duration on vowel formant shape.

Generally, there was no extra information to extract from the on- and off-glide slopes. Between-speaking-rate correlation coefficients of the slope values were almost always lower than those of the first-order component.

4.2.5 Relation between polynomial coefficients and vowel duration

The polynomial coefficient values found for the formant tracks were correlated with vowel duration. This correlation was performed for both speaking rates independently (not shown). Generally, the correlation coefficients between Legendre coefficient values and vowel duration were small and statistically not significant for both speaking rates. An exception were the second-order Legendre coefficients of the F_1 of the vowels /E A a/ (r•0.33-0.52, p•0.1%). These coefficient values were almost as high as the between-speaking-rate correlation coefficients (cf. table 4.5). The correlations between duration and second-order components of F_1 implied a decrease in

Table 4.5: Correlation coefficients between speaking rates of Legendre polynomial coefficients (order 0-2) and of calculated mean values of normalized slope at τ = 1/4 and τ = 3/4 (SL 1/4 and SL 3/4, see Table 4.2). Correlation coefficients that are statistically different from zero are underlined (coefficients recalculated for Student's t-test, p≤0.1%).

vowel		0	1	2	SL 1/4	SL 3/4
E	F_1	0.62	0.47	0.47	0.46	0.41
	F_2	0.87	0.76	0.54	0.69	0.44
A	F_1	0.86	0.67	0.46	0.64	0.49
	F_2	0.91	0.86	0.68	0.81	0.61
a	F_1	0.71	0.59	0.55	0.47	0.52
	F_2	0.85	0.85	0.67	0.85	0.56
i	F_1	0.57	0.69	0.46	0.42	0.51
	F_2	0.32	0.50	0.25	0.29	0.04
o	F_1	0.85	0.69	0.70	0.66	0.60
	F_2	0.87	0.78	0.76	0.68	0.75
´	F_1	0.55	0.36	0.40	0.74	0.28
	F_2	0.95	0.83	0.19	0.66	0.55
u	F_1	0.04	0.75	0.26	0.06	0.58
	F_2	0.73	0.86	0.73	0.83	0.75
y	F_1	0.73	0.62	0.54	0.39	0.19
	F_2	0.84	0.88	0.72	0.32	0.81

curvature (or excursion size) for shorter durations, i.e. shorter vowels had more level formant tracks.

The correlation coefficients between on- and offglide slopes and vowel duration that were statistically significant were all comparable in size to those between the second-order coefficients and vowel duration. The former relation can most likely be explained from the latter. All other correlation coefficients were small and not statistically significant for either speaking rate.

4.2.6 Effects of context

A subset of the tokens of the most numerous vowels /E A a i o/ in an all alveolar CVC context was analysed separately (i.e., C is one of /n t d s z r l/). For each vowel, the number of tokens available in an alveolar context was quite small (between 16 and 38, see table 4.1). For small numbers, the estimated parameter values will have a large error. Therefore, we concentrated on the relation between the tokens in the subset and those of the parent set and not on the actual sizes of the differences between the two sets.

The mean values of the Legendre polynomial coefficients (order 0-2) and the estimated slope at 1/4 and 3/4 of the vowel did not differ much from those found for the tokens of the parent set (table 4.4). The second-order Legendre coefficients of the F_1 tracks of the vowels /E A a/ might be an exception. The tokens of these three high F_1 target vowels had a somewhat higher (up to 20%) mean second-order coefficient value for both speaking rates and the slopes at both points inside the tokens were somewhat steeper.

The fast-rate tokens of this subset had a uniformly higher F_1 than the normal-rate tokens ($p \bullet 0.1\%$ for /A o/, zeroth-order). The vowel /o/ also showed a slightly higher F_2 in the fast-rate tokens (42 Hz $p \bullet 0.1\%$, zeroth-order). The between-speaking-rate correlation coefficients of the Legendre coefficients were high for both F_1 and F_2, often higher than those for the parent set. The trends were the same as in the parent set of tokens (table 4.5).

The correlation coefficients between Legendre polynomial coefficients or slope and vowel duration were generally higher in the subset of tokens in alveolar context than in the parent set (section 4.2.2). Still, only few correlation coefficients were statistically significant ($p \bullet 0.1\%$, fast-rate F_1: second-order coefficient of /E A a/ and slope at 1/4 of /E/) or larger than the corresponding correlation between speaking rates (c.f. table 4.5). An exception was the second-order Legendre coefficients of the F_1 tracks of the fast-rate tokens of the vowels /E A a/. Here the correlation coefficients were higher ($|r| \bullet 0.60\text{-}0.75$, $p \bullet 0.1\%$) than the coefficients obtained from the corresponding correlation between the two speaking rates.

These results show that the tokens from the subset of vowels in alveolar context were not different from the complete parent set of vowel tokens.

4.2.7 Effects of stress

The previous analyses were repeated on token-pairs of the vowels /E A a i o/ for which both tokens were stressed or unstressed (data not shown). This

was done to check whether sentence-stress might be significant with respect to the effects of differences in speaking rate or duration. Stressed tokens were 30% longer than the unstressed ones for both speaking rates ($p \bullet 0.1\%$). The differences in vowel duration between speaking rates were comparable for stressed and unstressed tokens (i.e., 15%).

For the F_1, zeroth- and (negative) second-order Legendre coefficient values of the stressed tokens of the high F_1-target vowels /E A a/ were higher than those of the unstressed tokens at both rates ($p \bullet 1\%$ for vowels pooled). The vowel space of the stressed tokens was larger, i.e. less reduced, in the F_1 direction (/i/ to /a/) than that of the unstressed tokens, both for zeroth-order (5%) and second-order coefficients (25%). The slopes of the F_1 tracks of stressed tokens were generally steeper than those of unstressed tokens. Both the fast-rate stressed and unstressed tokens had a uniformly higher F_1 than the normal-rate tokens (zeroth-order, $p \bullet 0.1\%$, stressed /E a/, unstressed all individual vowels).

Due to the lower number of realizations, the second-order coefficient values and track slopes of the F_2, were often not statistically significantly different from zero for the stressed tokens of vowels that did show significant values for the unstressed tokens. There was no indication that, compared to stressed tokens, unstressed tokens are spectrally reduced with respect to the F_2.

Generally, correlation coefficients, both for vowel duration and formants between speaking rates and between formants and vowel duration, were higher in stressed tokens than in unstressed tokens. The comparison was difficult because results for the stressed tokens were often statistically not significant due to the small number of stressed tokens. No other difference between stressed and unstressed tokens was found. As far as could be checked, the results obtained from all tokens pooled were equally valid for both of these subsets of tokens.

4.3 Discussion

The results found here generally are in agreement with those found using a more conventional type of analysis based on a direct comparison of the 16 equidistant points per vowel segment. These latter results are discussed in chapter 3 (see also Van Son and Pols, 1989, 1992). In this chapter we discuss specifically coordinated, whole track differences between speaking rates, instead of "local" point-by-point differences.

4.3.1 Effects of speaking rate

Despite the fact that the fast-rate vowel realizations are generally (and consistently) shorter than the normal-rate realizations, there is hardly a difference between the formant track shape parameters measured at different speaking rates. This means that, after normalization for duration, a difference in speaking rate did not result in systematic differences in formant track shape. Only the F_1 frequency is higher in vowels spoken at a fast rate than in vowels spoken at a normal rate, see figure 4.2. This rate-dependent rise in F_1 frequency was found irrespective of vowel identity. It was also limited to the zeroth-order Legendre polynomial (i.e., mean formant value).

This means that the F_1 frequencies in all parts of the fast-rate tracks were raised by roughly the same amount. This means that the equivalent results found by Van Son and Pols (1990; see chapter 2) for "static" measurements, in which method Average is identical to using the zeroth-order coefficient, must be attributed to an uniform increase in formant frequency over the whole F_1 track in fast-rate speech. It cannot be attributed to an increase in only the vowel nucleus or only the transition parts, which would also have changed the *shape* of the formant tracks (i.e., higher order Legendre coefficient values).

4.3.2 Effects of duration on formant tracks

A simple, one-way, relation between vowel formant tracks and vowel duration would result in a clear-cut, and strong, correlation between these two. This means that duration should explain a significant part of the variance in formant track parameters (i.e., the variance in track parameters would be systematic and linked to the variance in duration). However, correlation coefficients between formant frequencies and vowel duration were only significant for the F_1 tracks of the high F_1 target vowels (/E A a/), see section 4.2.5. The correlations implied a leveling off of the F_1 tracks with shorter durations of the tokens. This is predicted by the target-undershoot model. However, the correlation coefficients were rather small in all cases. The correlation between formant frequency and vowel duration hardly explains more than 30% of the variance in second-order Legendre coefficients ($0.33 \bullet |r| \bullet 0.52$). Between-speaking-rate correlations for these three vowels sometimes explained up to 70% of the variance in F_1 formant track parameters (zeroth order, $|r| \bullet 0.86$, table 4.5). This indicates that a very large part of the variance in formant track parameters is indeed systematic and reproduced for each "reading" of the text, independent of speaking-rate. The fact that the correlation between formant track parameters and vowel duration is much weaker than the between-speaking-rate correlation indicated that duration is not a major determinant of overall vowel formant track shape in read speech.

There is one area where the correlation between formant track parameters and vowel duration is as strong as the between-speaking-rate correlation and where duration might indeed explain much of the systematic variance. For the second-order Legendre polynomial coefficients of the F_1, the between-speaking-rate correlation coefficients were not larger (i.e., $0.46 \bullet |r| \bullet 0.55$, table 4.5) than those between Legendre coefficients and duration. This indicates that much of the *systematic* variance of the second-order Legendre coefficients of the F_1, as measured by the between-speaking-rate correlation, might indeed have been determined by vowel duration. The correlation between second-order Legendre coefficients and vowel duration was as predicted by the target-undershoot model, i.e. shorter duration were combined with more level formant tracks. But again, the absolute size of the effect of duration on track shape is minimal, generally explaining less than a quarter of the *total* variance observed.

F_2 formant tracks do not show any sizeable correlation between track parameters and vowel duration.

4.3.3 Effects of context and stress

The context in which a vowel is spoken might be important for the effects produced by changes in speaking rate (or changes in duration). We compared the differences in duration and in formant track shape between speaking rates for stressed with the differences for unstressed token-pairs and also the differences between speaking rates for tokens from an alveolar context with those from all tokens pooled.

Stressed vowel tokens were generally longer than the unstressed tokens and less reduced spectrally (at least for F_1). No differences between stressed and unstressed tokens were found when changes in speaking rate or duration were considered. The difference in duration between stressed and unstressed tokens was twice the difference between speaking rates. There was a difference in F_1 formant frequency between stressed and unstressed tokens but no difference between speaking rates. This indicates that the vowel duration alone is not enough to explain the differences between stressed and unstressed vowel realizations, confirming the results of Nord (1987).

For tokens from an alveolar CVC context, we would expect the largest effects on the open vowels /E A a/ for the F_1 tracks and on the back vowel /o/ for the F_2 tracks (see section 3.1.2 of chapter 3). For fast-rate tokens we found an increase in the correlation between the second-order Legendre coefficient of the F_1 tracks of the vowels /E A a/ and vowel duration. This suggests that the constraints on F_1 formant movements might have been tighter for vowel realizations spoken at fast rate than for realizations spoken at normal rate in this extreme consonant context, i.e. closed-open-closed. The same uniformly higher F_1 frequency in the fast-rate tokens was found as in the parent set. For vowels in an alveolar context we found the same lack of effect of either speaking rate or duration on the F_2, except that in this context the F_2 of the vowel /o/ showed a small, uniform, increase in fast-rate speech. Therefore, there might have been more coarticulation or "target-undershoot" in the F_2 in this extreme context (alveolar-/o/-alveolar). But because only one vowel was affected it is difficult to interpret the change.

The trends observed in vowel realizations in our parent set were also present in the stressed and unstressed realizations and in the realizations from an alveolar-vowel-alveolar context. This shows that the effects of speaking rate on vowel realizations is to a large extent independent of sentence-stress and (alveolar) context.

4.4 Conclusions

This study was limited in that only one speaker was used who read aloud a single text. From the results we conclude that this speaker did not behave as predicted by the target-undershoot model, which predicts more reduction (both static and dynamic) in vowel articulation with a faster speaking rate, especially when vowel durations are quite short to begin with. Even the refined versions of the target-undershoot model that incorporate alternative

articulation strategies (Gay, 1981) and increased effort (Lindblom, 1983) on a global level, would predict some measurable differences in formant track shape or frequency values between speaking rates. That neither was found indicates that these theories are not universally valid for all speakers using continuous read speech. We cannot rule out the possibility that these theories might explain some aspects of the relation between vowel duration and formants within a single speaking style or when strong coarticulation is predicted. However, our study indicates that their explanatory power is limited and probably speaker specific. Based on these results, articulation models are needed that acknowledge a much more active behaviour of the speaker in adapting to a high speaking rate.

5

THE INFLUENCE OF FORMANT TRACK SHAPE ON THE IDENTIFICATION OF SYNTHETIC VOWELS

Abstract

Synthetic vowels were used to investigate whether listeners use vowel duration and formant track shape to determine vowel identity. The synthetic vowels had level- or parabolically-shaped formant tracks and variable durations. They were presented in isolation as well as in synthetic Consonant-Vowel-Consonant context. There was no evidence of perceptual compensation for expected target-undershoot due to token duration or context. The only asserted effects of duration and context were in the number of long- and short-vowel responses. There was also no evidence that the listeners used the formant track shape or slopes independently to identify the synthetic vowel tokens. Tokens with curved formant tracks were generally identified near their formant offset frequencies.

Introduction

There is an ongoing discussion about how listeners identify vowel realizations. Two types of models can be distinguished: target-models and models using dynamic-specification (see Strange, 1989a). In target-models, the identity of a vowel is determined by the spectral contents of the vowel kernel, or even of a single cross-section through the realization (e.g., Nearey, 1989; Andruski and Nearey, 1992). In models using dynamic-specification, the identity of a vowel is to a large extent determined by the spectral dynamics in the vowel on- and offglide, such as formant track slopes (e.g., Di Benedetto, 1989a, b; Strange, 1989a, b).

A related problem is that of vowel "target-undershoot" in articulation. This occurs when vowels spoken in connected speech are pronounced with less contrast than canonical realizations (e.g., Lindblom, 1963, 1983; Lindblom and Moon, 1988; Gay, 1978, 1981). It is suggested that listeners would compensate for this undershoot in pronunciation by "overshooting" the target in perception (see discussion in Strange, 1989a).

Central to the "dynamic-specification" and "target-undershoot" models is the question of how formant track shape, vowel duration, and context together affect vowel identification. Identification experiments have shown that vowel realizations with the stable vowel kernel removed, leaving only the vowel on- and offglide, can be identified quite well by listeners ("silent-center" realizations, e.g., Strange, 1989a, b). This suggests that the formant track slopes in the on- and offglide of vowel realizations carries sufficient information about vowel identity. In section 4.2.3 we did indeed find that a related measure, the formant track excursion size, correlated with mid-point vowel formant frequencies in connected speech (Van Son and Pols, 1991a). The question remains whether this information is actually used by listeners.

In the present study we investigated whether listeners use information from the formant track shape to decide on the vowel identity and whether vowel duration and context influence this decision. Especially, it was investigated whether in situations where target-undershoot in production is expected, listeners automatically compensated for this *expected* undershoot in production by perceptual mid-point overshoot. In an attempt to answer these questions we concentrated our investigation on the effects of formant target frequency, vowel duration, and formant track shape on vowel identity. These three factors were varied independently to determine their relative contribution to identification. This cannot be done using natural speech, we therefore used synthetic vowels.

We chose the (parabolic) excursion size to represent formant track shape instead of the more commonly used track slope. This was done because the definition of formant track slope is linked to the duration of the (stationary) vowel nucleus, which is notoriously difficult to determine in natural speech (Benguerel and McFadden, 1989). This would make it difficult to obtain plausible values of formant track slopes and transition durations for vowel synthesis at all durations. The formant tracks of vowels can be approximated very well by a parabolic function as long as the vowel duration is not too long (chapter 4; Van Son and Pols, 1991a). It is easy to synthesize vow-

els with plausible parabolic formant tracks for which the excursion sizes are determined from natural speech (e.g., Van Son and Pols, 1991b).

Vowels pronounced in context are expected to show more target-undershoot than those pronounced in isolation. To investigate whether this leads to compensation in the perception (i.e., perceptual-overshoot), an experiment was performed in which vowel tokens were presented in isolation as well as in context (CVC, CV, and VC), using two simple synthetic consonants (/n/-like and /f/-like). Using the same consonant-tokens in both pre-vocalic and post-vocalic position enabled us to determine the influence of the position of a consonant token in the syllable on the identification of the associated vowel and the consonant token itself.

5.1 Methods

5.1.1 Isolated vowels

5.1.1.1 Token synthesis

All tokens were synthesized using an LPC-10 synthesizer with a pre-emphasis of 0.9. The synthesis parameters were: F_0 = 159, F_3 = 2490, F_4 = 3500, and F_5 = 4500 (Hz) and variable F_1 and F_2. All bandwidths were 50 Hz. This is equivalent to a cascade formant synthesizer using five formants and a pulse source. Synthesis was done at 10 kHz sampling rate and 12 bit resolution. We used a low-pass filter cut-off of 4.5 kHz for digital-to-analog conversion. The pitch was fixed at F_0 = 159 Hz, which corresponds to a period of 63 samples (6.3 ms), to prevent the introduction of a perceptive change in formant frequency due to the interaction between F_0 declination and higher formants.

Before waveform samples were actually generated, the synthesizer had run for four pitch periods with the values of the first synthesis frame. This procedure was necessary to damp onset transients in the responses of the synthesizer filters. The source amplitude was constant and was chosen at 75% of the maximum to prevent clipping of the waveform. We did not use autoscaling of the amplitude because it can produce widely fluctuating sound levels for the tokens. Synthesizing the /o/-like target pair (F_1=450 Hz, F_2=900 Hz, see below) with an excursion size of ΔF_2 = 375 Hz still resulted in a clipped waveform. This was alleviated by lowering the source pulse-amplitude for this combination to 30% of the maximum. The resulting four tokens sounded less loud than the other vowel tokens. The boundaries of all tokens were smoothed with a Hanning window of (2 times) 2 ms duration before recording, to remove click sounds. These vowel signals were D/A-converted and recorded on one audio channel of a VCR-tape on a Panasonic NV-F70HQ VHS stereo video cassette recorder that was also used for stimulus presentation.

5.1.1.2 Token construction

Nine formant mid-point value pairs (F_1, F_2) were defined using published values for Dutch vowels (Koopmans-van Beinum, 1980). These formant fre-

quency pairs corresponded approximately to the vowels /i u y È o E A a π/ in terms of vowel quality (see table 5.1, note that /È/ corresponds to /I/), but not in terms of duration. Using these mid-point F_1/F_2 pairs with fixed values for F_0 and F_3-F_5 results in tokens that do not quite sound like the original vowels from which the F_1/F_2 pairs were extracted. This was alleviated by tuning the formant mid-point values until the resulting tokens sounded well. When the mid-point values gave good vowel percepts, we changed them a little towards those of a neighbouring vowel to make the token label somewhat ambiguous (see table 5.1 and figure 5.1). Using tokens with somewhat ambiguous vowel quality will make our listening tests more sensitive to shifts in the perception of the tokens.

For these nine targets, formant tracks were constructed for F_1 and F_2 that were either level or parabolic curves according to equation 5.1.

$$F_n(t) = \text{Target} - \Delta F_n(4(t/\text{Duration})^2 - 4t/\text{Duration} + 1) \quad [5.1]$$

$$dF_n(t)/dt = -\Delta F_n(8t/\text{Duration} - 4)/\text{Duration}$$

in which:

$F_n(t)$: the value of formant n (i.e., F_1 or F_2) at time t.
$dF_n(t)/dt$: the slope of the formant track at time t.
ΔF_n : F_n(mid-point) - F_n(on/offset), i.e. the excursion size.
t : the time-point inside the token, 0 • t • Duration.
Target : the formant target frequency.
Duration : the total token duration.

For non-zero excursion sizes, formant tracks shaped according to equation 5.1 are symmetric and actually have no completely flat steady state part (e.g., figure 5.2). The target value is the maximum or minimum value of the formant track, depending on whether the excursion size is positive or negative, respectively. At the vowel on- and offset the formant-track slope is plus or minus $4\Delta F_n$/Duration, respectively (see equation 5.1). This means that, for a fixed token duration, the formant-track slope is a linear function of the excursion size.

Formant tracks were defined at 125 "sample" or frame points within the duration of the vowel-token (e.g., figure 5.2). All 125 frame values were used for synthesizing a token which meant that the synthesizer parameters were updated several times within each pitch period, i.e. more often than pitch synchronous. Different durations were obtained by varying the number of synthesized speech samples from 2 to 12 per frame point. This resulted in tokens with durations of 25 - 150 ms. Tokens with durations shorter than 25 ms were obtained by using only part of a longer token. All vowel tokens had an integer number of pitch periods.

For each target, tokens with level formant tracks ($\Delta F_1 = 0$, $\Delta F_2 = 0$) were synthesized with durations of 150 ms, 100 ms, 50 ms, 25 ms, 12.5 ms, and

Table 5.1: Vowel formant target frequencies (Hz) with the approximate Dutch vowel label in the top row.

V	i	u	y	È	o	E	A	a	π
F_1	300	300	300	450	450	650	700	750	450
F_2	2450	750	1900	2200	900	1950	1100	1300	1550

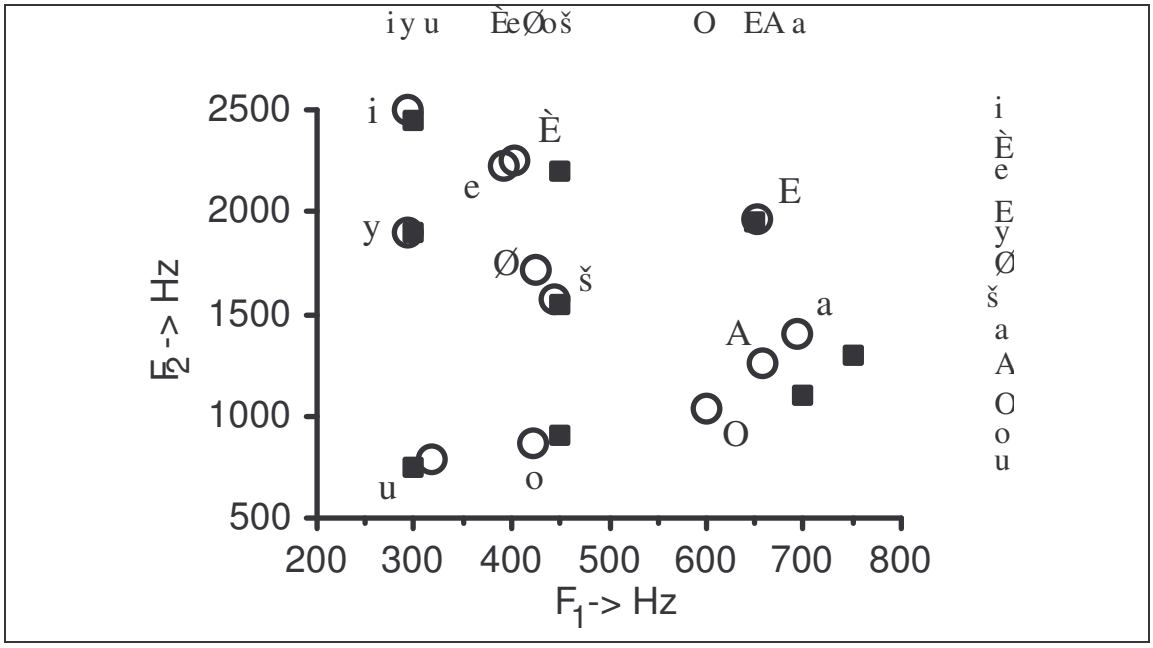

Figure 5.1: Formant frequencies of a single set of Dutch vowels pronounced in isolation by a male speaker (open circles, data taken from Koopmans-van Beinum, 1980). Above and to the right are the vowel labels in the established order along the F_1 and F_2 axis respectively (combined from 8 sets, see section 5.2). The nine formant target positions used to synthesize vowel tokens for the listening experiments are indicated by squares.

6.3 ms. Tokens with durations of 6.3 ms and 12.5 ms were generated by using only the first, or first two, pitch periods of the corresponding 25 ms tokens. For fear of suppressing too much of the shortest, 6.3 ms, tokens with the Hanning-window, we also recorded them without smoothing. However, the responses to these unsmoothed tokens were too erratic and they will not be used here. F_1 and F_2 formant tracks were constructed according to equation 5.1 for all tokens synthesized (F_0 and F_3-F_5 were always level). Table 5.2 describes all tokens that were recorded.

Almost all targets were used to synthesize tokens with excursion sizes of ΔF_1 = 225 Hz (see figure 5.2 for an example), ΔF_1 = -225 Hz, ΔF_2 = 375 Hz, and ΔF_2= -375 Hz (see table 5.2). Except for ΔF_1 = -225 Hz, these excursion sizes can be found in natural speech for /a i u/ respectively (see appendix D, ΔF_1 and ΔF_2 ; cf. chapter 4; Van Son and Pols, 1991a). The unusual F_1 tracks with ΔF_1 = -225 can be compared to that of a medial closed vowel flanked by more open vowels in a three-vowel sequence (see appendix D). Although the above excursion sizes were fixed in Hz, expressed in the perceptually more relevant semitones the variation was large, due to the variation in target frequency. Excursion sizes varied from 5-12 semitones in the F_1 direction and 3-9 semitones in the F_2 direction.

Tracks that crossed any other formant track (e.g., F_3 = 2490 Hz) or the F_0 (159 Hz) were not synthesized. This is indicated by the dashes in table 5.2. Tokens with curved formant tracks ($\Delta F_1 \bullet 0$ and/or $\Delta F_2 \bullet 0$) were synthesized with durations of 150 ms, 100 ms, 50 ms, and 25 ms and complete, symmetric formant tracks. From these tokens, the onglide and offglide parts (first and second half respectively) were also used separately. The on- and offglide-only tokens were half the length of their "parent" tokens (i.e., from 12.5 ms to 75 ms).

Figure 5.2: The waveform and vowel-token formant tracks of an example /naf/ token. This was a filler-token from the second experiment but test-tokens were constructed likewise (section 5.1.2). The vowel part was also used in the first experiment (section 5.1.1). The vowel-token was synthesized with an /a/-like formant target (F_1 = 750 Hz, F_2 = 1300 Hz), ΔF_1 = 225 Hz, ΔF_2 = 0 and a duration of 100 ms. The corresponding formant tracks of the vowel part are displayed in the upper part of the figure. The lower part displays the waveform.

Next to synthesizing vowel tokens with these rather extreme and fixed /a u i/-like excursion sizes, for each of the other vowel formant targets, tokens were synthesized according to the same principles but with excursion sizes that were more realistic according to the specific vowel targets. For vowels other than /a u i/, these tokens are displayed in the right hand side of table 5.2 (column 7-12, all values taken from chapter 4; Van Son and Pols, 1991a). The total number of stimuli is also specified in this table.

5.1.1.3 *Presentation*

All 495 synthetic vowel realizations were written to a VCR-tape in a pseudo-random order. Tokens were presented in blocks of ten with a 3.5 second inter-stimulus interval. At the start of the sequence and at the end of each block a 1000 Hz beep of 500 ms was sounded. The time between blocks was 6 seconds. After 240 stimuli a short (1 minute) break was inserted. The stimuli were presented binaurally at a comfortable sound level over open headphones (Sennheiser HD441) to up to six subjects at a time. Before the actual presentation of the test stimuli, the subjects heard a set of 10 stimuli of 200 ms duration to get accustomed to the signals and the procedure. These 10 stimuli were not present in the test sequence and no feedback was given.

The subjects were instructed to mark the vowel they heard on an answering sheet. They could choose from orthographic representations of all twelve Dutch monophthongal vowels, i.e. /Ø π E e È i y u o O A a/ (presented as "EU U E EE I IE UU OE OO O A AA"). For the Dutch language, presentation in orthographic form causes no ambiguities and no training was required. When no response at all was given to a certain vowel token, the subject was asked to identify this single token afterwards in an isolated presentation (i.e., only the missing token was presented and it was presented only once). However, this situation was very rare. In total only 6 out of nearly 14,000 responses were missing.

5.1.1.4 Subjects

29 Dutch subjects participated in the experiment (15 male, 14 female). Participation was voluntary and no rewards of any kind were offered. The subjects varied in their previous experience in phonetics from naive, i.e. no previous contact (5 subjects) or only limited contact (3 subjects), to undergraduate students (8 subjects), and postgraduate students and senior phoneticians (13 subjects). Age varied between twenty and sixty years. None of the subjects reported hearing problems.

None of the subjects had heard the presentation before and none was acquainted with its composition or the construction of the stimuli. Only one subject had any knowledge about the general aims of the experiment.

5.1.2 Presentation in synthetic CVC syllables

5.1.2.1 Consonants

With a Dutch speech synthesizer of Nijmegen University (Kerkhoff et al., 1986) a single realization each of the /n/ and /f/ sounds were generated (table 5.3) using a modified Klatt synthesizer (Klatt, 1980; cascade filters) with a periodic pulse-source for the /n/, or a noise source for the /f/, and the amplitude gradually rising and falling (see figure 5.2). The duration was 95 ms for both phonemes. Both separately generated synthetic realizations were chosen because they could be added easily before and/or after the synthetic vowel tokens into convincing pseudo-syllables. These two specific consonant realizations were used in *all* stimuli whenever they were specified with one or both of these consonants.

Table 5.2: Number of vowel tokens synthesized.
For each target (first column) and formant excursion size (first two rows, values in Hz) the number of tokens synthesized for each duration is indicated. Dashes indicate items that could not be synthesized (see text). 1: complete tracks only, 3: complete tracks, onglide-only and offglide-only, the latter two with half the duration of the former (see text). Total: Row and column totals of tokens for each duration. #Dur.: number of durations for which tokens were synthesized (see text). Tokens: Number of tokens, i.e. the product of #Dur. and Total.

ΔF_1	0	225	-225	0	0	0	75	75	150	150	75		
ΔF_2	0	0	0	375	-375	225	225	-225	150	-75	75	Total	Tokens
a	1	3	3	3	3							13	55
i	1	-	3	3	-							7	31
u	1	-	3	-	3							7	31
y	1	-	3	3	3	3						13	55
Ê	1	3	3	3	-		3					13	55
O	1	3	3	3	3			3				16	67
E	1	3	3	3	3				3			16	67
A	1	3	3	3	3					3		16	67
/n/ π	1	3	3	3	3						3	16	67
Total	9	18	27	24	21	3	3	3	3	3	3	117	-
/f/#Dur.	4	4	4	4	4	4	4	4	4	4	4	-	250
Tokens	36	72	108	96	84	12	12	12	12	12	12	-	495

Table 5.3: Parameters used to synthesize /n/ and /f/ segments. All values in Hertz. F_1-F_4: formants, NP: Nasal Pole, NZ: Nasal Zero, Z_2: Zero 2 (Zero 1 was not used).

	F_1	F_2	F_3	F_4	NP	NZ	Z_2
/n/	Frequency	180	1600	2600	3800	320	240	-
	Bandwidth	75	300	100	150	100	100	-
/f/	Frequency	4300	1600	2500	3500	-	-	250
	Bandwidth	50	500	1200	600	-	-	495

5.1.2.2 Vowel segments and syllable construction

Testing all 495 vowel tokens under all context conditions would have strained the endurance of our subjects too much. Therefore, we only used a subset of the available vowel tokens. The subset tested in context consisted of tokens with a duration of 50 and 100 ms and targets corresponding to /A È E o/ (table 5.1 and 5.2). These targets were expected to be the most sensitive to formant track shape. Furthermore, the 50 ms tokens would elicit primarily short-vowel responses, diminishing the problems with long-short confusions. The 100 ms tokens were included for comparison. The test-set corresponded to the tokens from the first five columns of table 5.2 ($\Delta F_1 = \Delta F_2 = 0$; $\Delta F_1 = 225$ or -225 Hz and $\Delta F_2 = 0$; $\Delta F_1 = 0$ and $\Delta F_2 = 375$ or -375 Hz). For a duration of 50 ms, tokens with complete formant tracks and on- and offglide-only tokens were used (the latter were half of the 100 ms tokens). For a duration of 100 ms, only vowel segments with a complete, symmetric formant track were used. All these vowel segments were presented in isolation and as part of synthetic syllables (table 5.4). For the four vowel targets used in the test set this added up to 68 tokens with isolated vowel segments (V) and 152 syllable tokens (CVC, CV, VC), for a total of 220 test tokens (table 5.4).

Additionally, realizations of the /a i u y π/ targets were used as fillers both with level formant tracks and with realistic formant excursion sizes (e.g., figure 5.2). For those realizations of the fillers that had curved formant tracks, the on- and offglide parts were also used separately. These filler realizations were used to prevent the subjects from homing in on the test set which would have limited their responses. The filler tokens were combined with the synthetic consonant realizations in a similar way to give 50 different filler tokens. Each filler token was used twice so in the test there were 100 filler tokens and 220 test tokens.

5.1.2.3 Presentation and subjects

We changed the procedure for the presentation of the tokens to be able to assess the consistency of the responses of our listeners. The 320 tokens were written in a pseudo-random order to a VCR-tape in *two* different orders. Each sequence was preceded by the same leader of 10 practice tokens. The practice tokens were selected from the filler tokens and were representative of the total. Each sequence of tokens was presented binaurally to each subject individually over closed headphones (Sennheiser HD220) in a small, quiet room. Each presentation lasted for about 25 minutes; no pause was inserted. Between the presentations of the two sequences to each subject there was a time-interval of approximately a month (42 days median, 7 days shortest) to ensure that the particulars of the first sequence were forgotten.

The tokens were more complex in this second experiment than in the first experiment. Therefore, we decided to use an open response paradigm in this experiment. The subjects were instructed to write down orthographically, as a sequence of single sounds, whatever they heard. They were informed that the tokens could deviate from Dutch phonotactics. Because the orthographic form might be influenced by existing Dutch words, the subjects were especially instructed to use isolated-character forms to write down sounds, e.g. "G" and "AA" instead of the Dutch orthographic form "ga" for the sequence /xa:/. This transcription procedure is not intuitive. Therefore, testing only started after we were confident that the subject indeed had understood the task. After the ten practice stimuli, it was checked whether the stimuli were transcribed as prescribed. If necessary, additional explanations were provided. At the end of the experiment, none of the subjects reported difficulties with this task (note that all subjects had a background in phonetics). In total, only a single response was missed by the subjects (out of 9600 responses). The subject involved identified the missed stimulus in a second, isolated presentation (the same procedure as was used in the first experiment).

15 Subjects participated in this experiment. All but one of them had also participated in the previous experiment. The subjects were under- and post-graduate (language) students and senior phoneticians. Each subject participated twice, responding to each sequence of tokens once.

5.2 Results

We were more interested in differences between responses to tokens that differed in duration or formant track shape than in the absolute responses for each duration or formant track shape. Therefore, we mainly tested differences in the responses to different tokens on a within-subject basis, i.e. responses from each subject were compared separately. All subjects did recognize the test tokens as vowels without problems. However, the vowel stimuli were not always "natural" because of the sometimes unnatural formant track shapes and very short durations and it was often difficult to

*Table 5.4: Syllables constructed from vowel and consonant segments. The vowel tokens used were a subset of those described in table 5.2. First two rows - formant excursion size in Hz; first column - vowel token duration; second column - syllable structures. Entries give the formant track parts (cmp - complete, on - onglide only, off - offglide only) and the number of targets for which syllables were constructed: 4 - /È E A o/, 3 - /E A o/. Each vowel segment was used in only two syllables, one for each context (see second column). Stationary tokens with a duration of 50 ms (indicated by an asterisk *) were used in six syllables, the same one for each context. All these 68 unique vowel segments were also presented in isolation. This brings the total number of stimulus tokens to 220 (152+68).*

		ΔF_1	0	225	-225	0	0	
Duration	Syllable	ΔF_2	0	0	0	375	-375	Total
50 ms	nVf, fVn		*cmp 4	cmp 4	cmp 4	cmp 4	cmp 3	19
	nV, fV		*cmp 4	on 4	on 4	on 4	on 3	19
	Vn, Vf		*cmp 4	off 4	off 4	off 4	on 3	19
100 ms	nVf, fVn		cmp 4	cmp 4	cmp 4	cmp 4	cmp 3	19
	Total		16	16	16	16	12	76
	Tokens		32	32	32	32	24	152

decide which specific vowel was heard.

The vowel symbols used can be positioned in the vowel triangle in a two-dimensional formant space. We were interested in which direction the responses would change as a result of movements in the first and second formant of the tokens. We therefore decided to rank-order the vowel labels along the two dimensions of vowel space (e.g., figure 5.1). Changes in the responses were investigated by performing a sign-test on the differences in the label rank-orders in one or both dimensions. We used a threshold level of significance of p•0.1% (i.e. p•0.001) to prevent repeated tests from producing spurious results.

"Ideal" rank-order numbers for Dutch vowels were determined by assigning rank numbers separately to the F_1 and F_2 values of eight sets of formant measurements taken from Koopmans-van Beinum (1980; two female and two male speakers, vowels and monosyllabic words uttered in isolation, cf. figure 5.1 for one specific set). The order of the vowel labels was not identical for each of these eight vowel sequences. However, the discrepancies were small and individual discrepancies could be resolved by using the ordering present in the majority of the sequences. Along the F_1 the (ascending) rank-order established was /i y u È e Ø o π O E A a/, along the F_2 it was /u o O A a π Ø y E e È i/ (cf. figure 5.1).

With the rank-order established, we were able to sort the labels from "low" to "high" F_1 or F_2. Counting responses upwards from the low side of this sorted set of labels, we could determine the label that was used halfway (50%) of the responses. This was called the median response label. Also, for every pair of responses from a certain subject to a certain pair of tokens (say tokens with the same duration but with level and curved formant tracks), we could determine whether the response to the second token was "higher" (+) or "lower" (-) than the response to the first. This enabled us to perform sign-tests on the responses to different tokens and thus to determine the direction of change brought about by any difference between the tokens. The results of these operations are independent of the metrics of the formant space, e.g. whether formants are measured in Hz, semitones, or Bark.

5.2.1 Isolated vowel presentation

5.2.1.1 Effects of duration on tokens with level formant tracks

Our stimuli were sometimes rather artificial and we did not know beforehand how our tokens would be labeled. We also did not know whether the responses of our subjects to individual tokens would tend to converge to a single label. It would have been possible that certain tokens were so unnatural that the responses to them would be erratic. Also it is important to know whether and how token duration influenced identification, especially for short durations. Therefore, we first determined the median responses to tokens with level formant tracks and the influence of token duration.

Theoretically, the median response can be different along both formant directions, but this only occured once (see table 5.5). In table 5.5 the median response is given for each target and for each token duration (tokens with level formant tracks only). For tokens with a duration of 25 ms and up the subjects were very consistent. Twenty (or more) out of our 29 subjects (> 67%) either used the same label for tokens with the same target (/u y i E/) or chose one of a long/short vowel pair (for targets /a A o È π/). The only discrepancy between the responses of the subjects was whether some tokens represented long or short vowels.

For tokens with durations of 6.3 ms and 12.5 ms there was more confusion. The number of /È/ responses increased dramatically compared with those to longer tokens. The number of /O/ responses increased to a lesser extent. Together, the /È O/ labels account for almost half (31% /È/ and 15% /O/) of all responses for tokens with a duration of 6.3 ms, but only one out of every five responses (21%) for 25 ms tokens. The /È/ responses were concentrated on the tokens with "neighbouring" targets, i.e. /i È y/ and to a lesser extent to /A a E π/. The /O/ responses were distributed more widely, i.e. to /u O A E π/-like tokens. The number of /π/ responses remained approximately equal between tokens of 25 ms and shorter tokens. For all other labels the share in the responses declined with shorter durations.

Except for the drive to mid-F_1 vowel labels (i.e., /È O/ and less for /π/) in the responses, the subjects tended to confuse neighbouring vowel labels in short duration tokens (not shown). Still, even for tokens with a duration of 6.3 ms, at least 14 out of the 29 subjects used the same label in their responses to each token (this label was /È/ only for token targets /i È y/). Albeit that it was not always the same label as used for longer tokens with the same target.

Dutch has four vowel pairs with a durational opposition: /A a:/, /O o:/, /È e:/, and /π Ø:/ (the ':' mark is most of the time omitted). The other vowels can be considered to be short or half-long (i.e., /E i y u/). For the members of these four long/short vowel-pairs the total number of long (i.e., /a o e Ø/) and corresponding short (i.e., /A O È π/) vowel responses are displayed in figure 5.3.a as a function of token duration.

As is to be expected, the number of long-vowel responses increased with token duration while the number of corresponding short-vowel responses decreased at the same time. Without the /È/ and /e/ responses, i.e. ignoring

Table 5.5: Median vowel responses to individual tokens with level formant tracks.
Columns correspond to individual targets (vowel labels on the top row, see table 5.1). Rows correspond to tokens of a single duration (in ms). Median vowel responses were identical when determined along the F_1 and F_2, except for the token marked with "" for which the F_1 is mentioned first. Whenever the median response was used by 20 or more out of 29 subjects (2/3) it is underlined.*

Duration	a	A	o	u	y	i	È	E	π
6.3	A	A	O	u	È	i,È*	<u>È</u>	E	π
12.5	A	<u>A</u>	O	<u>u</u>	È	i	<u>È</u>	<u>E</u>	<u>π</u>
25	A	<u>A</u>	O	<u>u</u>	<u>y</u>	<u>i</u>	<u>È</u>	<u>E</u>	<u>π</u>
50	A	<u>A</u>	o	<u>u</u>	<u>y</u>	<u>i</u>	<u>È</u>	<u>E</u>	<u>π</u>
100	a	A	<u>o</u>	<u>u</u>	<u>y</u>	<u>i</u>	È	<u>E</u>	<u>π</u>
150	<u>a</u>	A	<u>o</u>	<u>u</u>	<u>y</u>	<u>i</u>	e	<u>E</u>	Ø

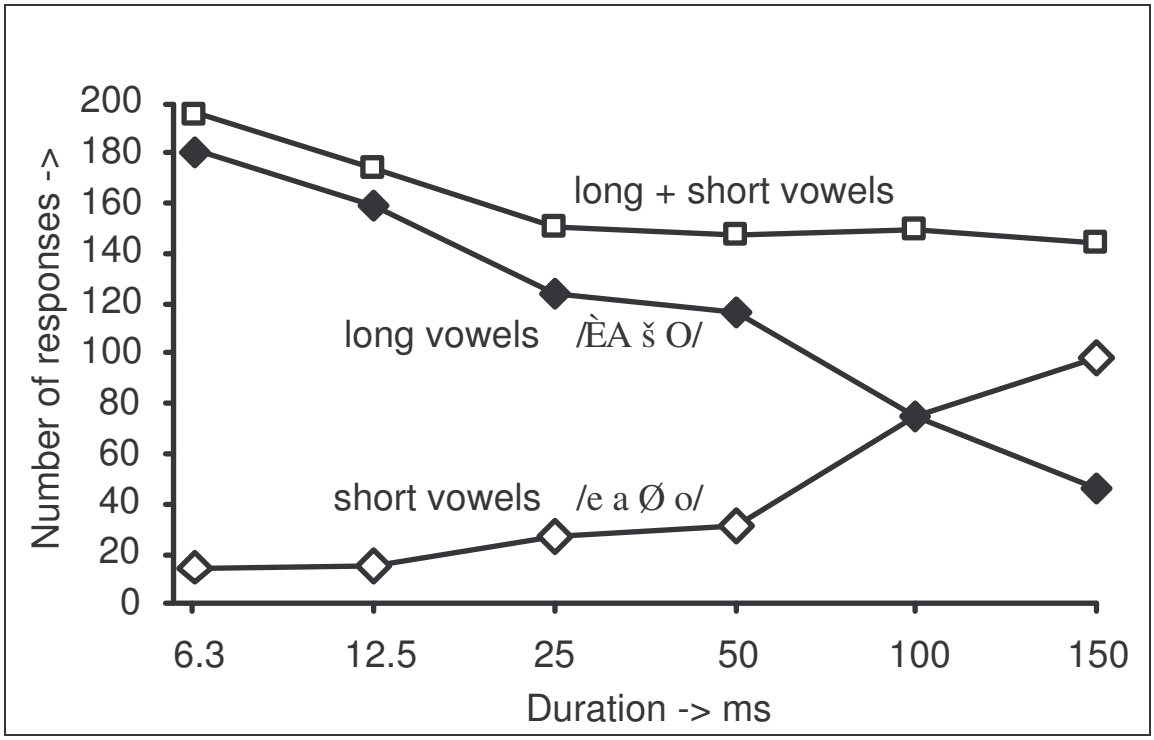

Figure 5.3.a: Absolute number of long- and short-vowel responses (/e a Ø o/ *and* /È A π O/, *respectively) for different durations. The total number of responses per duration was 261.*

the large /È/ response bias at short durations, the sum of the short- and long-vowel responses remained almost constant with token duration. This indicates that there is an exchange between long and short responses to the same targets for different durations. This exchange between a short response at one duration and the corresponding long response, by the same subject to the same target, at the next higher duration (and vice versa) is displayed in figure 5.3.b.

For durations below 50 ms there was a low level of random changes between long and short labels. But when going from 50 ms to 100 ms and to 150 ms tokens, over half of all differences between the responses of our subjects can be attributed to changes from short-vowel labels for shorter tokens to the corresponding long-vowel labels for the longer tokens. There were only few changes the other way round (70 versus 9, p•0.1%, sign-test).

For token durations of 6.3 and 12.5 ms the responses were dominated by /È O/ labels and a general confusion between labels. We will therefore concentrate on the longer tokens. When the responses to 25 ms duration tokens were directly compared with those to 150 ms duration tokens (for each subject), 104 responses out of a total of 261 (9 targets times 29 subjects = 261 pairs with different durations) differed between durations. Of these, 60 pairs (58%) of different responses could be described as short-to-long vowel transitions or vice versa. This left only 44 pairs of differing responses (42%) to be explained by "other" effects of duration (cf. "other" in figure 5.3.b). No systematic trends could be found in these remaining differences. As 20 of these pairs (19%) had an /È O/ response for the short token, part of these differences might have been the result of the increased number of /È O/ responses for short duration tokens which can still be found at 25 ms. When this analysis between tokens of 25 ms and tokens of 150 ms duration was

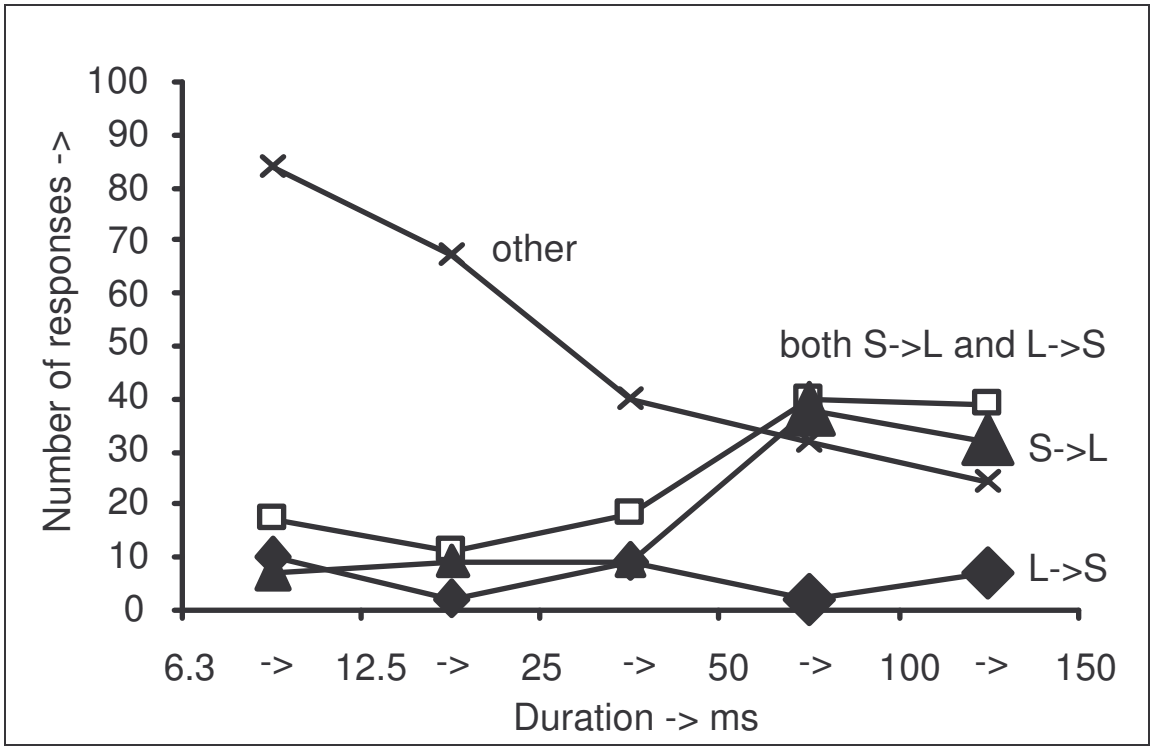

Figure 5.3.b: The number of short-to-long vowel label exchanges when tokens become longer (max. 261). Large symbols indicate cases where S->L and L->S differed significantly (p≤0.1%, two tailed sign-test). S->L: Short-vowel labels that are exchanged for the corresponding long-vowel labels (/A/->/a/, /O/->/o/, /È/->/e/, or /π/->/Ø/). L->S: The reverse of S->L. Both: The sum of all long/short exchanges. Other: All other changes.

repeated for other pairs of durations the results were the same (not shown). For tokens with level formant tracks, no systematic effects of duration could be found in the responses other than an exchange between long- and short-vowel labels and an increase in /È O/ labels for durations shorter than 25 ms.

When the subjects were grouped with respect to previous experience in phonetics, there were no obvious differences in the distribution of long- and short-vowel labels, neither were there any obvious differences with regard to the consistency of the responses to tokens.

5.2.1.2 *Effects of extreme formant excursion sizes on token identification*

Our primary interest was how token identification was influenced by formant track shape. It was to be expected that effects would be most dramatic for large formant excursion sizes. Therefore, we examined first in this section the effects on identification of the more extreme excursion sizes found in natural speech (i.e., column 3-6 of table 5.2; section 5.1.1.2; Van Son and Pols, 1991a).

Responses to tokens with the above mentioned excursion sizes were compared with the corresponding responses to tokens with level formant tracks (same target, duration, and subject). For each response to a token with a *curved* formant track it was noted whether the vowel label had a lower or higher rank number than the label used in the corresponding response to the token with *level* formant tracks. The vowel labels were rank

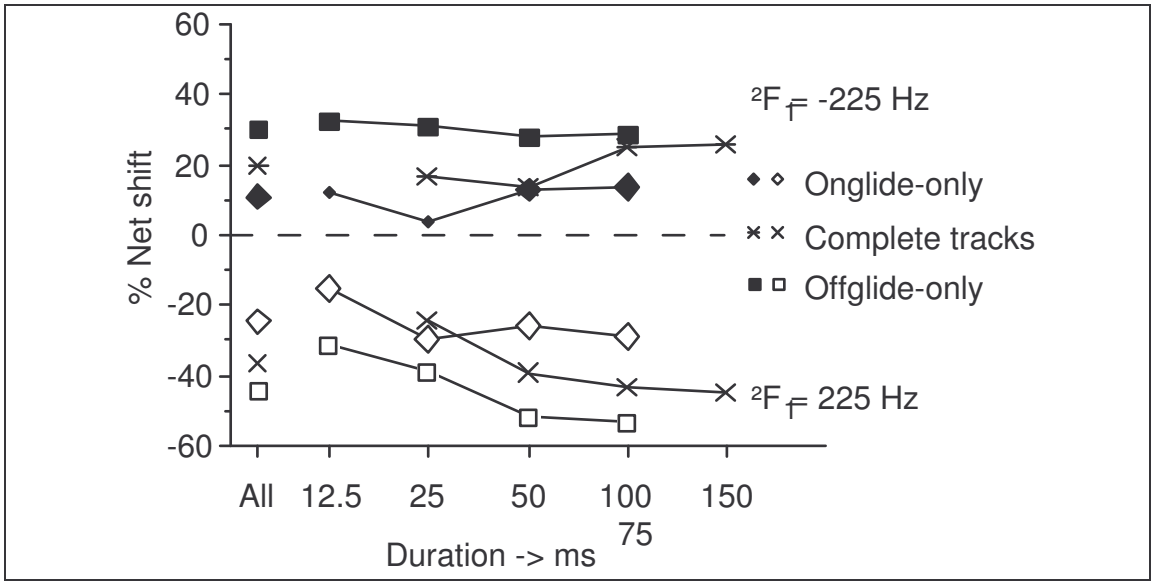

Figure 5.4.a: Net shifts in responses to isolated vowel tokens as a result of formant track curvature of the F_1.
All values in percent of total number of responses (see text). Positive shifts are towards higher formant frequencies, negative shifts towards lower frequencies. Large symbols indicate statistically significant shifts (p≤0.1%, two-tailed sign-test). All: all four durations pooled. Open symbols: ΔF_1 = 225 Hz (n=696), filled symbols: ΔF_1 = -225 Hz (n=1044). The second formant is level (i.e., ΔF_2 = 0) for both.

ordered along that formant dimension which was the curved formant in the stimulus. All response pairs where pooled over subjects and the net shift towards a lower or higher rank-number was calculated as a percentage of the total number of responses. Statistical significance of the differences was determined with a two-tailed sign-test (level of significance p•0.1%). The (signed) net shifts between the responses have the advantage that, for large numbers of responses, they can be added, e.g. the net shift between set A and set C is the approximately the sum of the shifts between the sets A and B, and B and C.

A numerical example will further clarify the approach used. The responses to tokens with an excursion size of ΔF_1= 225 Hz, a complete formant track, and a duration of 50 ms were compared with responses to tokens with level formant tracks and the same duration of 50 ms. According to the third column in table 5.2, six tokens, with targets corresponding to /È π o E A a/, were synthesized with this excursion size. The responses to tokens for these six different targets were pooled (this practice is discussed below). In total there were 6 (targets) times 29 (subjects) or 174 responses to these tokens with curved formant tracks. These were compared with the 174 responses to the corresponding tokens with level formant tracks. Of these 174 response pairs (each time same subject and target), 74 (43%) had different labels for the "curved" and "level" tokens. With vowel labels ranked along the F_1, 70 (40%) responses to the token with a curved F_1 had a lower rank number and 4 (2%) had a higher rank number than the corresponding responses to tokens with level formant tracks. The remaining 100 (58%) responses had identical rank numbers. The difference between the number of responses with a lower and those with a higher rank number indicated a net shift towards a lower rank number in 70 - 4 = 66 (38%) of the

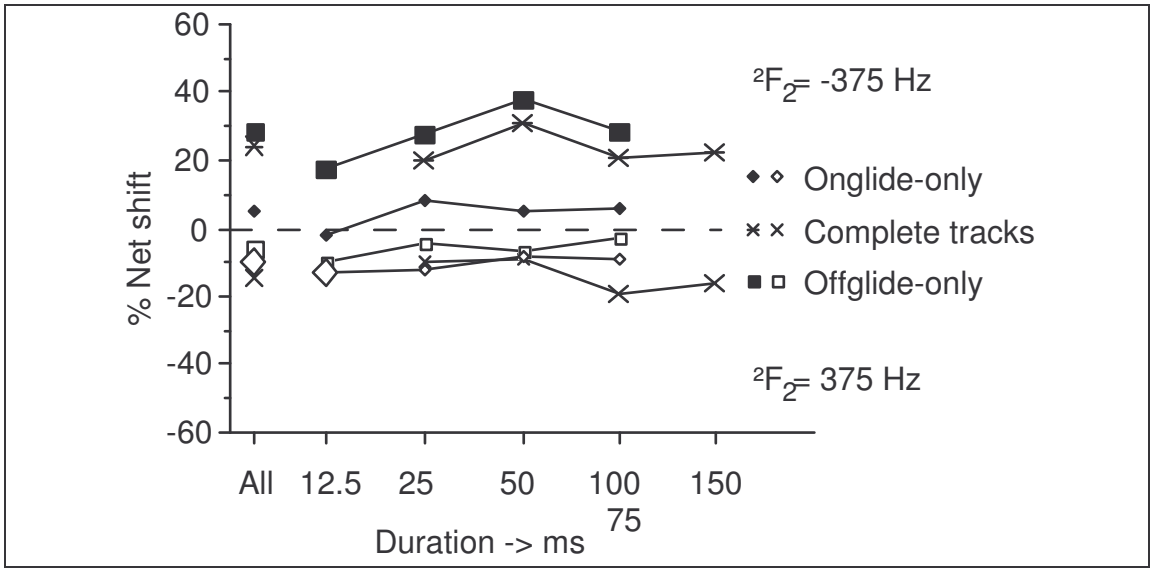

Figure 5.4.b: As figure 5.4.a. Net shifts in responses to isolated vowel tokens as a result of formant track curvature of the F_2. Open symbols: ΔF_2 = 375 Hz (n=928), filled symbols: ΔF_2 = -375 Hz (n=812). The first formant is level (i.e., ΔF_1 = 0) for both.

responses. Because the shift was towards a lower F_1 rank number, the number was entered as a negative number (-38%) in figure 5.4.a (the second entry from below at 50 ms). The remaining 74 - 66 = 8 (5%) responses are "undirected" differences (i.e., four each way) and can be considered as a base level of confusion (or noise) in the responses.

When 70 out of 74 responses have a lower rank number this is statistically significant at the p•0.1% level (two-tailed sign-test, it is customary to use only the pairs that differ for a sign test). The same procedure was applied to every combination of token duration, formant excursion size, and formant track part (on- or offglide-only). The responses to on- and offglide-only tokens with a duration of 75 ms were compared with the responses to tokens with level formant tracks with a duration of 100 ms since there were no corresponding 75 ms tokens in the set. The results of all comparisons are displayed in figure 5.4.a and 5.4.b.

Token duration had only a slight effect on the net shift in the responses. Only for curved F_1 tracks did shorter tokens have a statistically significant smaller net-shift (see figure 5.4.a). When we compared responses to tokens with a duration of 150 ms with those to tokens of 50 ms, on average only 10% of the responses were shifted towards the on/offset of the F_1 formant (not counting long-short differences, significant for complete formant tracks, ΔF_1 = 225 Hz and -225 Hz pooled, p•0.1%, sign-test). This means that a significant increase in token duration induced only a small shift in the responses (see figure 5.4.a). For tokens with excursions in the F_2 tracks no effect of duration could be found (see figure 5.4.b).

A consistent picture emerges from figure 5.4. For each formant track excursion size and each duration, the subjects responded with labels that where shifted towards the on/offset of the (parabolic) formant tracks. This shift was more pronounced for curved F_1 tracks than for curved F_2 tracks. There are marked differences in responses between tokens with complete, symmetrical formant tracks and those with only the on- or offglide parts.

Except for tokens with ΔF_2 = 375 Hz where differences generally were not statistically significant, the offglide part alone elicited the greatest shift in responses, followed by tokens with the complete, symmetrical formant track and the smallest shift was found for tokens with the onglide only. This is most apparent when all shifts are pooled on duration (column "All" in figure 5.4).

The differences between the responses to offglide- and onglide-only tokens were statistically significant for all durations (tokens pooled on excursion size, p•0.1%, sign-test) and all excursion sizes (tokens pooled on duration, p•0.1%, sign-test) except for ΔF_2 = 375 Hz. For tokens with a duration of 12.5 ms the differences between the responses to *off*- and *on*glide-only tokens were statistically significant for three out of four excursion sizes (p•0.1%, sign-test, not for ΔF_2 = 375 Hz). The differences between responses to *off*glide-only tokens and the corresponding responses to tokens with *complete* formant tracks were statistically significant (p•0.1%, sign-test, all relevant durations pooled). The differences between responses to onglide-only tokens and responses to tokens with complete formant tracks were only statistically significant for ΔF_2 = -375 Hz (p•0.1%, sign-test, all relevant durations pooled).

Whether the perception of a vowel-token will actually shift by increasing the excursion size, also depends on the position of its target in Dutch vowel space. Especially, the perceptual distance to the nearest boundary in vowel space would matter and whether the excursion would draw the "target" away from it or towards it. These perceptual distances varied widely for our targets. Therefore, it is not surprising that the sizes of all the reported net shifts in the responses, i.e. the proportion of the responses that differed, were very sensitive to the particular token target frequencies (not shown). However, we *never* found that the *direction* of the shift was different for different token targets. Therefore, we present any shift in the responses as pooled over all targets.

The differences between the responses that remained after the net shift was subtracted from the total number of different responses could be regarded as confusions and errors by the subjects. In the numerical example given above (with ΔF_1=225 Hz and 50 ms tokens), this rate of confusions and errors was fairly low, only 8 out of 174 responses (5%). For other excursion sizes and durations the rate of confusions was generally higher. Averaged over all token responses, the rate of confusion was 18%. This rate was fairly independent of token duration except for the tokens of 75 and 100 ms durations where it peaked at 22% due to a larger number of long/short confusions.

5.2.1.3 *Effects of realistic formant excursion sizes on token identification*

The fixed excursion sizes used in the previous section were rather unnatural for most token targets and most certainly so for ΔF_1 = -225 Hz. This might have resulted in "unnatural" responses from our subjects. Therefore, we also presented tokens with more realistic (i.e., more natural) formant track excursion sizes matched to the formant target (see table 5.2). Each target vowel was synthesized with target-specific formant track excursion

sizes. The *number* of responses that differed from those to tokens with level formant tracks was very sensitive to the position of the target (but not the *direction* of the shift). The net shifts as a fraction of the total number of responses do not show the systematic differences between the tokens. Therefore, we will present the net shifts in the responses as a fraction of only the responses that actually differed.

From a total of 1044 responses, 307 differed from those to stationary tokens with level formant tracks. Of these, 254 had labels with lower F_1 rank numbers and 53 had labels with higher F_1 rank numbers. The net shift towards a lower F_1 rank number was 201 responses, which is 65% from a total of 307 responses that differed. This means that most responses to tokens with curved formant tracks that differed from those to stationary tokens were shifted towards the on/offset of the F_1 tracks. The size of the shift, i.e. the net proportion of differing responses that was shifted towards the on/offset of the F_1 tracks, was related to the excursion size of the token. The net size of the shift was from 91% (for /A a/) to below 2.5% (for /È/) in the order /A a E π o È/ (significant for /A a E π/, p•0.1%, sign-test). Except for /A a/, whose net shifts were almost equal, this order corresponded to a decreasing F_1 target frequency and decreasing formant track excursion size (see table 5.1 and 5.2).

For the realistic F_2 excursion sizes, we could not find a relation between excursion size and the size of the shift (not shown). Anyway, for this formant the differences between the responses to tokens were not significant for any of the targets (p>0.1%, sign-test) except for the /π/-like target, for which it can be explained as interference from the F_1 track shape.

To summarize the results: It appears that realistic excursion sizes in the F_1 tracks elicited graded (size-dependent) shifts in the responses. No effects were found of realistic excursions in the F_2 tracks, even for tokens that had level F_1 tracks (i.e., /i y u/ targets).

5.2.2 Presentation of vowels in context

In the above experiment, vowel tokens were presented in isolation, i.e. with silence preceding and following the vowel segment. This might have induced our subjects to focus their attention on features that are specific to isolated, sustained vowels. To investigate the effects of the token context on vowel identification we performed an experiment with vowel tokens presented in isolation mixed with identical tokens presented in a CV, VC, and CVC context. We wanted to compare the responses to identical vowel-tokens under different conditions (isolated and in context). This prevented us from using smooth, natural-like consonant-vowel transitions to construct the syllables.

In this experiment, our subjects heard a number of vowel tokens of either 50 ms or 100 ms in isolation and in context. The subjects were asked to write down what they heard, but at least they should respond with a vowel or a diphthong. The subjects were instructed to use a question mark when they could not decide on the identity of a heard consonant. Diphthong and triphthong answers were considered to consist of two or three vowel-labels. However, only one, monophthong, label was used to represent each multi-

vowel response. In this we gave the subject the benefit of the doubt. When the "target" response was present, it was used as the monophthong label for the whole response. If the target label was not present, the first vowel label in the response was used. For instance, the response /π/-/y/ (i.e., Dutch "ui") was considered to be an /y/ when the target of the token was /y/-like, else it was considered to be an /π/. This way, we could reduce diphthong and triphthong labels to monophthong labels without unduly amplifying (or even producing) the dominance of the vowel-token offset as found in the first experiment.

A consonant-token in the stimulus was considered to be recognized when a consonant label of the same class was used in the response, i.e. any nasal for the synthetic /n/-sound and any fricative for the synthetic /f/-sound. Transcription errors of the subjects regarding the order of the vowel and consonants were ignored. This way we can investigate what the effect is of the *presence* of a consonant (but not the effect of the conscious *perception* of a consonant).

5.2.2.1 *Consistency in responses to synthetic vowels*

Each subject responded twice to each test-token, once in each session, and four times to each filler token, twice in each session. With these responses we were able to check the consistency with which our subjects responded to identical tokens, both within sessions and between sessions.

Between the two sessions, the vowel-labels differed in 19% of the responses to the test tokens. Within both sessions the vowel-labels differed in 17% of the responses to the filler-tokens (cf. section 5.2.1.2). When long-short confusions were discarded, the differences dropped to 12% for the test-tokens (between sessions) and 14% for the filler-tokens (within sessions). Without long-short confusions the number of differences between the responses depended mainly on the formant target frequencies and less on the formant excursion size. The differences ranged from 2% (/E/-target) to 19% (/o/-target) not counting long-short confusions.

Diphthongs or triphthongs were heard 4% of the time on a total of 6600 responses (30·220), both to vowel segments in isolation and in context. Most of the multi-vowel responses were given for 100 ms tokens (8% of 1710 responses) and when the excursion size of the F_1 was not zero (all tokens pooled, 6% for ΔF_1 = 225 Hz and 10% for ΔF_1 = -225 Hz, both of 1440 responses). For the 100 ms tokens with curved (i.e., non-stationary) formant tracks, diphthong responses were over 10 times more frequent for vowel-tokens presented in isolation (V) than for those presented in context (CVC; 31% of 450 and 2% of 900 responses respectively).

5.2.2.2 *The responses to synthetic consonants and their influence on vowel identification*

Artificial syllables were used to be able to investigate how the consonantal context in which a vowel-token was presented influenced its identification, and vice versa. To understand how the consonantal context influences the identification of vowels it is necessary to investigate how these consonant-

tokens themselves were "perceived". For instance, it is not clear how consonants that are "missed" by the subjects will influence the identification of the neighbouring vowel. The responses to individual consonant segments in different conditions (different position and vowel segments) could be compared because each individual consonant segment occurred in every position (syllable initial or final) in every syllable used (CV, VC, and CVC).

The synthetic /f/-sound was considered to be identified correctly when it was labelled as a fricative, the /n/-sound when it was labelled as a nasal. The prime factor that influenced consonant recognition proved to be the position in the "syllable". In token-initial position 70% of the synthetic consonants was recognized and 9% was heard but not identified, i.e. a question mark was responded. In token-final position 98% was recognized and less than 0.5% unidentified. This difference was significant ($p \bullet 0.1\%$, sign-test). The synthetic /n/-sound was slightly better recognized than the /f/-sound in both positions.

The identity of the vowel token following or preceding the consonant did influence recognition but much less so than did its position in the syllable. Recognition was worst in both positions when the consonant preceded a vowel-token with $\Delta F_1 = -225$ Hz or when the formant track was level (both very unnatural for a CV or VC transition). When preceding such a vowel-token, only 60% of the consonant-tokens were recognized (91% in token-final position).

Beside the "induced" consonant labels, i.e. fricatives for the /f/-sound and nasals for the /n/-sound, the subjects also responded with other consonant labels indicating that they perceived consonants that were not present in the tokens as independent sound segments. Such an additional consonant was indicated to precede the vowel in more than 6% (overall) of the responses, 16% when there was a token-initial /n/. An additional consonant was reported to follow the vowel in less than 2% of the responses, less than 0.5% when a consonant was actually present in that position. Over half of the additional consonants reported to have been heard, were /b/ (pre-vocalic) and /p/ (post-vocalic). In all contexts, an excursion size of $\Delta F_1 = 225$ Hz almost doubled the number of added consonants heard with respect to other excursion sizes.

The responses to a vowel token were influenced by the context in which it was presented, silence (i.e., in isolation) or synthetic consonants. When a vowel token was followed by a consonantal sound (VC and CVC, C one of /f/ or /n/) there was, on average, a decline by half (to 50%) in the number of long-vowel responses compared to when it was presented in isolation (statistically significant for each context, $p \bullet 0.1\%$, sign-test). The long-vowel responses were not only replaced by the corresponding short-vowel responses but also by other "nearby" vowels. In contrast, when the vowel token was only preceded by a consonant (CV condition, i.e., an "open syllable") the number of long-vowel responses increased (by 50% for /f/, less for /n/) compared to when presented in isolation (statistically significant for /f/V only, $p \bullet 0.1\%$, sign-test). Generally, the presence of a synthetic /n/-sound lead to less long-vowel responses than the presence of an /f/-sound in the same position (statistically significant for all contexts pooled, $p \bullet 0.1\%$, sign-test) especially in the CV and CVC condition (the effect was almost absent

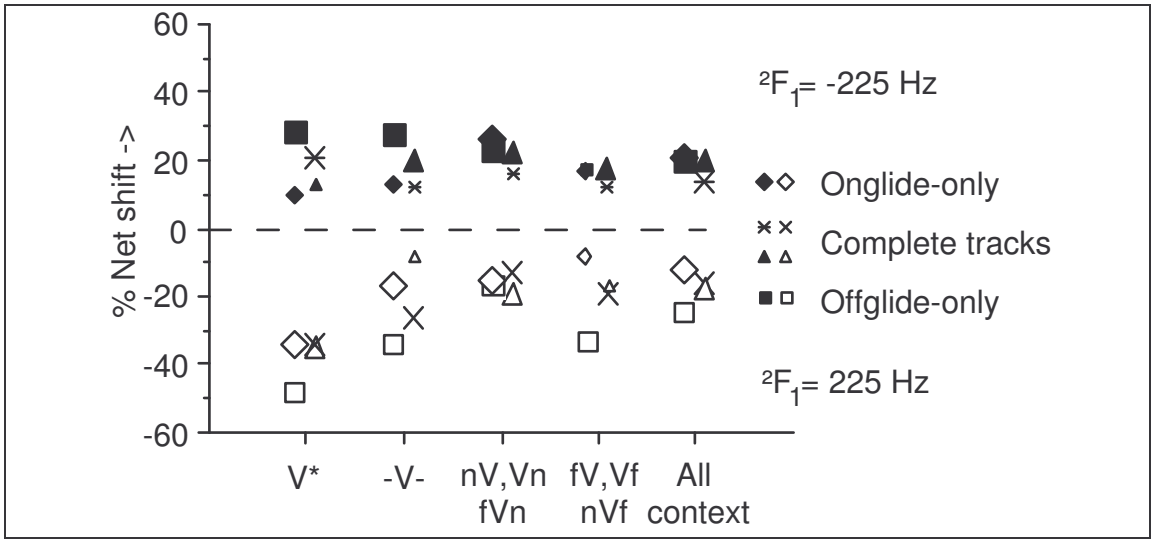

Figure 5.5.a: Net shifts in responses as a result of curvature of the F_1 both for isolated vowel tokens and vowel tokens in context. All values in percent of total number of responses (see text). Large symbols indicate statistically significant shifts (p≤0.1%, two-tailed sign-test). V: results from the first listening experiments (n=116 or n=87). -V-: isolated vowel tokens. nV, fV: onglide-only tokens. Vn, Vf: offglide-only tokens. nVf, fVn: tokens with complete tracks. All context: all tokens in context pooled. Net shifts are grouped in columns according to the context of the corresponding tokens. In each column, the symbols have been displaced horizontally for clarity. Triangles: 100 ms tokens, all other tokens have a duration of 50 ms. Open symbols: ΔF_1 = 225 Hz (n=120), filled symbols: ΔF_1 = -225 Hz (n=120). The second formant is level (i.e., ΔF_2 = 0) for both.*

in the VC condition). Therefore, it is possible that the difference in the number of long-vowel responses, found between the /n/ and /f/ sounds, was the results of the effect of the pre-vocalic consonant only. Diphthong responses always decreased when vowel-tokens were presented in context. No other systematic effect of context could be attested.

To summarize these results: The relative position of a consonant in the synthetic syllable was found to be the major determinant influencing its identification. The only other systematic effect found was a position-dependent change in perceived vowel length due to context.

5.2.2.3 *The influence of formant excursion size on vowel identification*

In figure 5.5 the net shifts in the responses as a result of vowel token formant excursion size are presented for different contexts (i.e., V, CV, VC, and CVC tokens). The results of the second experiment cannot be compared immediately with those of the first experiment (presented in figure 5.4) because in the second experiment only a subset of the tokens (targets) was used. For comparison, we extracted the responses to an identical subset of vowel tokens from the first experiment and included the net shifts of these in figure 5.5 (the column labelled V*).

The responses to the vowel tokens presented in *isolation* (V condition, second experiment) were influenced by the design of the experiment and the task of the subjects, but only the sizes of the net shifts were affected. The sizes of the shifts in the second experiment were clearly smaller but, for each duration, the pattern was more or less the same (V* and -V-

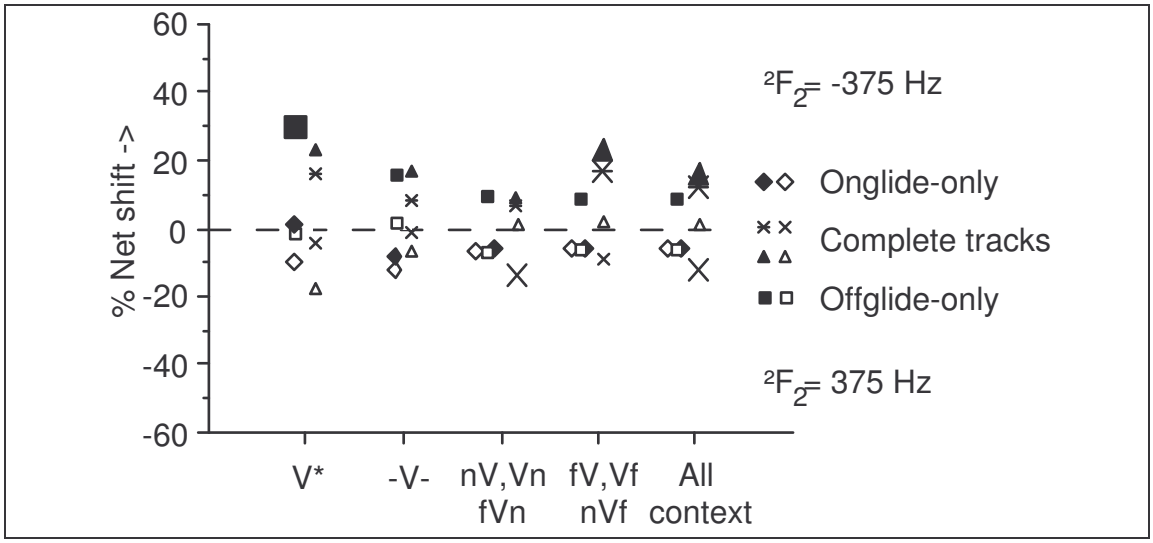

Figure 5.5.b: As figure 5.5.a. Net shifts in responses as a result of curvature of the F_2 for isolated vowel tokens and vowel tokens in context. Open symbols: ΔF_2 = 375 Hz (n=120), filled symbols: ΔF_2 = -375 Hz (n=90). The first formant is level (i.e., ΔF_1 = 0) for both.

columns in figure 5.5, compare to figure 5.4). The use of diphthong and triphthong labels in the second experiment could probably explain this difference. The "target" label would be more often present in these compound labels than in forced-choice monophthongal responses. In the second experiment we granted the listeners the benefit of the doubt. This could have reduced the number of responses that were shifted towards the vowel formant track offset frequencies.

In general, the responses to the offglide-only tokens were shifted more towards the formant offset than responses to tokens with complete formant tracks, and these responses in turn were shifted more than the responses to onglide-only tokens (in reverse order for ΔF_2 = 375 Hz).

In *context*, the responses to the vowel tokens were essentially the same as in *isolation*. Most of the differences between the responses to vowel tokens presented in context and those presented in isolation were related to the differences in the number of long vowel, diphthong and triphthong responses. Also, the pattern of shifts caused by formant track shape was similar for vowel tokens presented in context and those presented in isolation. Generally, for each duration and type of presentation (i.e., isolation and context), the largest shifts were found for the responses to the offglide-only tokens. However, the differences between the responses to onglide-only, offglide-only, and complete tokens were rather small in the second experiment and statistically not significant.

5.3 Discussion

In this chapter we will limit our discussion to the results of our own experiments. In chapter 6 we will give a detailed account of the literature in relation to the results presented in this chapter. A final discussion follows in chapter 7.

5.3.1 The effects of duration

For our synthetic vowel tokens with level formant tracks, token duration had no influence on the identification apart from the obvious exchange between long- and short-vowel labels (see figure 5.3). Only when the token duration was below 25 ms did other effects become important. Next to the general confusion of neighbouring vowel labels, there was a relative increase in mid-F_1 vowel labels (i.e., /È O π/, in that order). This might have been caused by a loss of spectral resolution at very short durations, which induces a linear averaging (i.e., Δfreq. • 1/duration • 170 Hz). For tokens with curved formant tracks, the net shift in the responses that resulted from curvature of the F_1-track was smaller in shorter tokens. It is very reasonable to assume that formant on- and offglides were less well resolved spectrally and therefore perceptually less pronounced as they became shorter and steeper.

It is apparent from the identification results that subjects were quite capable of labelling extremely short tokens consistently down to a single pitch period (i.e., 6.3 ms), although the error-rate became quite high for some tokens. This consistency in labelling short tokens was also found by Van der Kamp and Pols (1971) and Fox (1989). For both tokens with level and curved formant tracks, there was no indication that our subjects in any way compensated with perceptual-overshoot for expected target-undershoot as a result of duration or context.

5.3.2 The effects of formant excursion size

In all respects, curvature of the second formant had less influence on the responses of the listeners than curvature of the first formant. This can at least partly be explained by a difference in the perceptual size of the excursions of the F_2 and F_1 tracks. Expressed in semitones, the excursion sizes used for the second formant were considerably smaller than those used for the first formant (3-9 versus 5-12 semitones, respectively).

From the results of both listening experiments it is clear that our subjects used the "information" that was present in the formant dynamics only to determine the formant offset-frequencies, or some average near this point. They did not use the curvature, on-, or offset slopes independently of the actual formant frequencies. The large formant excursion sizes, used for all targets, suggested realizations of high-F_1 vowels (ΔF_1 = 225 Hz), high-F_2 vowels (ΔF_2 = 375 Hz), low-F_2 vowels (ΔF_2 = -375 Hz), or even a completely unusual phoneme sequence (ΔF_1 = -225 Hz). If our subjects had used the information present in the shape independently from the actual formant frequencies, they would have shown some "perceptual-overshoot". Perceptual-overshoot would have resulted in labels representing vowel targets beyond the target actually "reached" in the token. When this had occurred, the sign of the net shifts in the responses would have been the same as that of the formant excursions. But actually net shifts in the responses were of opposite sign. Therefore, compared with tokens with level formant tracks, the responses to tokens with curved formant tracks were shifted towards the on/offset of the formant tracks and away from the actual targets. This can be described as "perceptual-undershoot". Target

vowels were also combined with excursion sizes that were more realistic for these specific vowels. This induced the same shift in a graded (i.e., size dependent) way along the F_1 direction, no shifts in the responses were found along the F_2 direction.

The "perceptual-undershoot" found in both experiments suggests some kind of averaging of the formant frequency inside the tokens. The largest shifts were found in offglide-only tokens, followed by tokens with complete formant tracks (with equal duration and therefore steeper slopes). The onglide-only tokens induced the smallest shifts. A simple, linear average of the formant track frequency is identical for each of these formant track parts (i.e., complete, onglide and offglide). Therefore, the "perceptual averaging" apparently was not symmetric. However, any averaging method that attaches the greatest weight to formant frequencies in the final part of the tokens would reproduce the relation between formant track shape and vowel identification. This suggests some kind of dominant perceptual recency effect in the responses of the subjects (i.e., last heard is best remembered).

5.3.3. The effects of context

Synthetic consonants presented in pre-vocalic position were identified less well than those presented in post-vocalic position. Furthermore, an /n/ sound preceding the vowel-token induced more /b/ percepts. Both these findings suggest that the conflicting cues from the artificial CV transitions influenced consonant identification more than the equally artificial cues from the VC transition. The fact that the perception of a consonant is more affected by a vowel following it than by one preceding it was also found by Mann and Soli (1991). Also, according to their results, the post-vocalic consonant would be more important for the identification of the vowel than the pre-vocalic consonant. As the (synthetic) post-vocalic consonant was "identified" quite well (• 98% "correct") in our experiment, we can, in a first approximation, act as if all post-vocalic consonantal tokens were indeed recognized as such. Again according to Mann and Soli, we can expect that the impact of the rather large number of "missed" pre-vocalic consonantal tokens (• 30% missing) on vowel identification was small.

When vowel tokens were presented with a consonant following it (i.e., VC and CVC tokens), our subjects responded with less long-vowel labels. Presented with a consonant preceding it (i.e., CV tokens) they responded with more long-vowel labels. In Dutch, short vowels (/A O π E È/) are not allowed in open syllables, apart from some exclamations. So this last effect could be the result of phonotactic constraints. Between the two synthetic consonants used, the difference seems to be that an /n/ sound preceding the vowel token results in less long-vowel responses than an /f/ in the same position. No other systematic effects of context were found.

When followed by silence (i.e., a pause), the final part of a vowel can be considered the most reliable part, i.e. the part least affected by coarticulation. Therefore, it would have been advantageous if our subjects would have focused on the formant offset frequencies to identify the isolated vowel tokens. In closed syllables, the central part of the vowels is the most reliable.

To give an impression of closed syllables we presented the vowel tokens surrounded by synthetic consonants. If listeners did indeed use the "most reliable" part (as defined here) to identify vowel realizations, they should have shifted their attention to the central part of the vowel tokens from closed pseudo-syllables.

However, with vowel-tokens presented in context, the responses of our subjects showed the same shift of the labels in the direction of the on/offset frequency of the formant tracks as found when presented in isolation. The differences between vowel tokens with curved and level formant tracks hardly changed when vowels were presented in context instead of in isolation. Therefore, the possibility that the sheer presence of other (not integrated) speech sounds would focus the attention of the subjects away from the offset of the vowel tokens and towards the center can be rejected for our type of stimulus. Also, no evidence was found that our subjects compensated for the *expected* coarticulation that normally would have resulted from context in natural speech (coarticulation that was not really present). However, the decline in diphthong responses in context might in some way be related to such a compensation.

5.3.4 Relevance for natural speech

Manipulating synthetic speech is a powerful method for studying speech perception. But it is always necessary to confirm whether the results can be applied to natural speech. On first sight, the results of our experiments seem to be inconsistent with common experience. In natural speech, vowels are generally recognized well in context (cf. Strange, 1989a) where formant excursion sizes can exceed those used in these experiments. Furthermore, vowel formant excursion size is correlated with vowel identity (chapter 4; Van Son and Pols, 1991a). It is surprising to find that this, extra, information was not used by our listeners.

The experiments described here are only a single step towards solving the question of how the temporal features of vowels influence vowel identity. We isolated only two factors, formant curvature and duration, that were expected to be important for vowel identification, and ignored all others (e.g., sound level, tracks of other formants, integrated context). The influence of formant curvature and duration was investigated using synthetic tokens. From the results it can be concluded that duration had negligible effects, except for long/short vowel identification. Furthermore, the formant curvature, i.e. on- and offglide slope, was not used as an independent marker of vowel identity. On the contrary, the presence of steep formant slopes made the identification of a hypothetical target value (as defined by target-undershoot models of production) less likely.

There are several reasons why our results cannot be directly extrapolated to natural speech. In our tokens, the sound level was kept constant for the duration of the vowel token, whereas in natural speech it peaks in the center of the vowel. It is possible that this would cause the formant frequencies there to be more prominent and more important for identification. Furthermore, we deliberately tried to obtain formant target frequencies that were close to the perceptual borders between vowels. For some targets

we succeeded and their identification was very prone to shifts in perception. Other targets were apparently still located more peripherally with respect to the center of the perceptual area and were very resistant to shifts (e.g., the /E/ target). The partially ambiguous formant targets, together with the fixed values for the higher formants (F_3-F_5), might have made our tokens much more sensitive to formant movements than vowel targets in natural speech would have been. The fact that we found that users used a weighted average of the formant tracks agrees with the results of Di Benedetto (1989b; see also section 6.1.3.3), except that in her case most weight was placed on the *on*set parts. Still, this finding contradicts existing theories on vowel perception.

To be able to compare responses in different context, the individual token segments had to be identical. Therefore, we had to refrain from integrating the synthetic consonants with the synthetic vowel tokens into realistic syllables. Coarticulation between consonants and vowels was deliberately not modelled. This might have induced our subjects to process the syllables as sequences of unconnected sounds, viewing the vowel part still as an isolated sound. In natural speech the integrated movements of all formants, the pitch, and the loudness might induce listeners to focus on other parts of the vowel segment than in our synthetic tokens.

5.4 Conclusions

Bearing in mind that there is still a gap between our synthetic tokens and natural speech, it is possible to draw some general conclusions from our experiments. First, token duration did not influence vowel identification, except for obvious long/short-vowel exchanges. For durations of 25 ms and longer, no evidence for any duration-dependent perceptual over- or undershoot was found. Below 25 ms the number of general confusions increased as did the number of mid-F_1 vowel responses (i.e., /È O π/). Adding synthetic consonants to the tokens, creating CV, VC and CVC syllables, only changed the number of long-vowel and multi-vowel responses. No compensation for articulatory target-undershoot in the form of perceptual over- or undershoot could be attested.

Second, formant excursion size, and therefore formant track slope (cf. equation 5.1), was not used independently by our listeners to identify vowels. Our subjects identified the vowel-tokens using formant frequencies primarily from the final part of the token as the "target", irrespective of the formant track slope at that point. This result was not influenced by presenting the vowel tokens in a context of synthetic consonants, i.e. in "pseudo-" syllables.

We conclude that our results with synthetic vowels agree more with a *modified target-model of vowel perception* than with a model that uses dynamic-specification (cf. Strange, 1989a). However, our results indicate that the target was located near the offset of the vowel-tokens and not in the nucleus, so the target-model should at least be modified to supply an explanation for this behaviour. No evidence was found that the listeners compensated in any way for token duration.

If indeed a target-model is the better description of the identification of vowels in natural speech, the question remains whether the listeners select the location of the target in natural speech in the same way. However, dynamical features could very well be indicative for vowel identity, as many studies have concluded (see chapter 6). If they are used, our own results imply that the use of these dynamical features must depend crucially on factors other than the shape of the first and second formant track alone.

6

VOWEL PERCEPTION: A CLOSER LOOK AT THE LITERATURE

Abstract

The literature on vowel perception contains contradictory claims concerning the use of information from the consonant-vowel and vowel-consonant transitions in vowel recognition. Some studies claim to have found that listeners use formant track shape to compensate for changes in production brought about by coarticulation. Others claim that no evidence for such a compensation could be found. Our own experiments show that the information in the formant track shape of synthetic vowels is not always used in a way that would have benefited recognition of comparable natural vowels. A re-evaluation of the literature shows that evidence for compensatory processes, i.e. perceptual-overshoot and dynamic-specification, was only found when vowel realizations were presented in an appropriate context. Some studies show that vowel recognition deteriorates when vowel segments are presented out of context. These facts suggest that the presence of an appropriate context is essential for any perceptual compensation of coarticulatory changes.

Introduction

In chapter 1 we signalled a disagreement in the literature with regard to the role of Consonant-Vowel (CV) transitions in vowel recognition (see Strange, 1989a; Andruski and Nearey, 1992). Several studies lead to the conclusion that dynamic features, and especially formant transitions, are used to identify vowel realizations. However, no evidence for such a mechanism could be found in other studies. The evidence that was presented in favour of perceptual-overshoot and dynamic-specification could also be interpreted against it (Andruski and Nearey, 1992). In chapter 5, we too could find no evidence of dynamic-specification or perceptual-overshoot. On the contrary, we found that non-level formant tracks would lead subjects away from the mid-point values towards perceptual-undershoot. This means that, instead of alleviating the effects of coarticulation, curved formant tracks would aggravate them. The cause of all these contradictory results remains unknown.

The experiments we have done cannot answer this question. Only new experiments might be able to solve it. To see in what direction the answer might be found, we will re-evaluate the existing literature in the light of our own results. We will try to indicate what factors might have been responsible for the presence or absence of dynamic-specification and perceptual-undershoot in different experiments. We will have to reinterpret existing publications to find such factors. These new interpretations are bound to remain speculative, at least in as far as we will stretch the published data beyond the scope given to them by the authors of the original papers. Only new experiments could prove the validity of any such new interpretations.

In this chapter we will weigh the evidence for perceptual-overshoot and dynamic-specification put forward in the literature. We will consider dynamic-(co)specification to designate any model that assumes that listeners use spectro-temporal information from the CV- or VC-transitions to compensate for the effects of coarticulation or reduction. Perceptual-overshoot is one such model. Any effect of the formant track shape inside the CV- and VC-transitions that increases vowel recognition is evidence for dynamic-specification.

Perceptual-overshoot will be considered an automatic, peripheral process which moves the perceived vowel formant mid-point, or extreme, value beyond the value actually reached in the acoustic signal. The perceived formant track should be an *extrapolation* of the vowel on- and/or offset formant transitions (CV and/or VC; see chapter 1, Figure 1.3). Therefore, we only speak of perceptual-overshoot when the size of the difference between the perceived formant value and the value actually present in the acoustic signal depends on the slope and extent of the CV or VC formant transition. This means that a positive, but not necessarily linear, correlation must have been established between the amount of overshoot and the slope and/or extent of the formant transition before we can speak of perceptual-overshoot as a special form of dynamic-specification.

6.1 An evaluation of the relevant literature

The results of our experiments seemed to disagree with at least some that were reported in the literature (see chapters 1 and 5). In this chapter we will interpret our results in the light of results reported in the literature. We will first discuss two questions that are related to the question of whether dynamic information is used to identify vowels. First, is there dynamic information in the spectro-temporal structure of vowel segments that could be used to identify vowel realizations (section 6.1.1). Second, is the ambiguity found in the responses to synthetic stimuli also found in natural speech or are natural vowels always recognized well (section 6.1.2). The remainder of section 6.1 will be dedicated to findings that are directly related to the question of whether listeners use dynamic information from consonant-vowel (or vowel-vowel) transitions to identify vowel realizations. We divided the experiments reported in the literature into two groups:

1. Experiments using synthetic speech (section 6.1.3)
2. Experiments using natural speech (section 6.1.4)

6.1.1 Information present in formant dynamics

Several studies have tried to determine whether vowel realizations contain dynamic information that could be used to identify them. In chapter 4 we found that excursion size could be used to distinguish vowels with high F_1- or F_2-targets from vowels with low target values for either of these formants (see figure 4.2). The relation between excursion size and vowel formant target frequencies indicated that vowel formants started and ended, on average, from a closed (low-F_1) and non-high/non-low (mid-F_2) position. Stressing the fact that these starting and ending points are averages, this seems not to be unreasonable from an articulatory point of view. Furthermore, the strong correspondence between formant spaces constructed from "excursion size" and "mid-point" values (cf. figure 4.2) indicates that the link found between formant excursion size and vowel identity is unlikely to be an artefact of the low number of realizations used.

Examining natural speech, Di Benedetto (1989a) found that she could use the time at which the maximum in the F_1 was reached to distinguish realizations of the vowels /È E/. Huang (1991, 1992) reported that characterizing a vowel formant track with three points (at 25%, 50%, and 75% of duration) instead of only at a single point, could increase the recognition score of a Gaussian classifier. This shows that information on formant track shape could help classification. Akagi (1990, 1993) also concludes that information from spectral dynamics could be used to improve automatic vowel classification in natural speech. Both Huang and Akagi suggested that a mid-point "overshoot" mechanism that compensates for coarticulatory undershoot could do the job.

These studies show that the spectral dynamics of vowel realizations can be used to help classify vowel realizations automatically. This was found using several different methods to measure these dynamic features. The systematic nature of the relation between formant track shape and vowel identity suggested the possibility that human listeners would use this information too. However, our own study has shown that the matter is not

that simple (chapter 5). It is clear that some conditions must be met before listeners will actually use the dynamic information present in vowel realizations.

6.1.2 Natural versus synthetic speech

In our experiments, we used synthetic stimuli with simplified formant contours. The formant trajectories in our vowel tokens were in a sense quite unnatural, moving mostly along one formant at a time. It could be that, for each *natural* vowel realization, the combined trajectory of the formants in formant space (i.e., F_1/F_2 space) would spend most of its time within the boundaries of the perceptual area of that vowel. This way it would not matter on which part of a natural vowel realization its identity was determined. In most experiments using synthetic speech, it is tried to make the trajectories in formant space similar to those in natural speech (c.f. Lindblom and Studdert-Kennedy, 1968; Fox, 1989). However, it is known that reduced vowels and vowels excised from their context are identified less well than vowels spoken in isolation (Koopmans-van Beinum, 1980; Van Bergem, 1993; see also section 6.1.4). From this we can conclude that in natural speech too, formant trajectories seem to leave the perceptual area of the vowel, just as in our experiments. Therefore, some other mechanism seems to ensure correct identification.

It is important to note that even for our extreme formant excursion sizes, the changes in the responses often were quite small. For example, the /E/ target we used was almost incorruptible and the high and low F_2-target tokens (i.e., those with /i È u/-like mid-points) did hardly show any change in responses due to curvature of the F_2. However, responses to some other targets, e.g. /o/, were easily shifted in all directions. This indicates that the vowel mid-point formant values determined the sensitivity of subjects to formant track shape.

Formant excursion sizes in natural speech are generally smaller than the extreme excursion sizes used in our listening experiments (compare chapter 4 and 5). We found that the corresponding shifts in responses were also smaller when we used smaller and more realistic excursion sizes. It is to be expected that vowel realizations from natural speech, with "good" mid-point formant frequencies and moderate formant excursion sizes, will generally be identified correctly. This might in part explain the generally high recognition scores for natural vowel realizations uttered in context (see discussions in Strange, 1989a; Nearey, 1989; Andruski and Nearey, 1992). However, this fact cannot explain everything, because of the above mentioned fact that vowel realizations from natural speech are identified much worse when presented out of context.

6.1.3 Experiments using synthetic speech

The strongest claims for the existence of perceptual-overshoot were based on experiments using synthetic vowel tokens with well defined formant tracks. The oldest and most cited paper that reported perceptual-overshoot is the study of Lindblom and Studdert-Kennedy (1967). This study contrasts with our own study in which we did find the opposite results: clear

perceptual-undershoot (chapter 5). Their stimuli were similar to ours and it certainly requires some explanation why the results of both studies disagreed. We will therefore discuss their experiments extensively. We will also discuss several other papers.

A preliminary remark must be made about an important difference between the experiments discussed below and that of our own (chapter 5). All experiments discussed in this section, 6.1.3, used a forced choice paradigm for the responses. Listeners were always asked to respond with only one of a limited set of possibilities, often only two labels were available, irrespective of what they actually heard. In our experiments we either asked our listeners to respond with any of the Dutch monophthongs (forced choice) or they were asked to respond whatever they heard (open response). In chapter 5 we saw that restricting the response categories to all Dutch monophthongs, therefore excluding diphthongs and triphthongs, already increased the size of the perceptual-undershoot found. Restricting the response categories still further to only two labels (e.g., /U È/ or /È E/) will result in even more dramatic changes in the outcome of the experiments. Essentially, in the experiments discussed below, the *listeners* were forced to place their responses on a single continuum. In our experiments, *we* constructed these continua ourselves by rank-ordering the response labels along the F_1 and F_2 directions. It is certain that these two different procedures for ordering responses along a continuum will give different results. However, it is very *un*likely that this methodological difference will change perceptual-overshoot in the responses into perceptual-undershoot and therefore we will not elaborate on this difference. The number and quality of response categories might, however, have a very strong effect on the sizes of the over- or undershoot found. Therefore, between-paper comparison of results can only be done in a qualitative way, not in a quantitative way.

6.1.3.1 *The paper of Lindblom and Studdert-Kennedy (1967)*

Lindblom and Studdert-Kennedy (1967) used vowel tokens in a well defined and integrated context. Vowel token mid-point values spanned a continuum in the range between /U È/ (F_1 = 350 Hz, F_2 = 1-2 kHz, F_3 = 2.3-2.8 kHz). Vowel tokens were presented to subjects in isolation with level formant tracks and in /wVw/ and /jVj/ syllables with parabolically shaped formant tracks. The vowel on- and offset frequencies were F_1 = 250 Hz, F_2 = 800 Hz, F_3 = 2200 Hz in /wVw/ context and F_1 = 250 Hz, F_2 = 2200 Hz, F_3 = 2900 Hz in /jVj/ context. The consonants were synthesized as two stationary 20 ms sounds with formant frequencies that were identical to the vowel formant on- and offset frequencies. The responses of the subjects were limited to only two categories: /U/ and /È/. Stimuli of different durations and with or without context were presented in a blocked fashion. Ten native speakers of American English participated in the experiments. Four were tested in Sweden (KTH, Stockholm) and six in the USA (Haskins Laboratories, New York). Pseudo-random sequences of tokens of each duration in context and in isolation were presented on separate days (four blocks, /wVw/ and /jVj/ together versus #V# for each duration, i.e. 200 ms and 100 ms).

Next to the similarities in stimuli, several important differences with our experiments are apparent (cf. chapter 5). Spectral changes from consonants to vowels and vice versa were continuous in the experiment of Lindblom and Studdert-Kennedy (1967). The formant tracks of the vowel parts always started and ended at the values used for the consonants. Furthermore, their consonants were synthesized as "vowel-like" sounds. The consonants and vowels in the Consonant-Vowel-Consonant (CVC) syllables were therefore well integrated. Next, the F_2 excursion sizes were often larger than those used in our experiments, up to 1200 Hz (compared to a maximum of 375 Hz in chapter 5). With our relatively small excursion sizes we already induced a sizeable amount of diphthong responses. It is to be expected that the stimuli of Lindblom and Studdert-Kennedy induced an even stronger perception of diphthongs than our own. This might have influenced the responses of the subjects in ways unaccounted for in their experiments.

As a last difference, the subjects were asked specifically to identify the vowel token in a known context and in a two-alternatives forced-choice paradigm. The difference in the response paradigms between both studies is unlikely to have produced the perceptual-overshoot versus -undershoot difference in the responses. However, the fact that Lindblom and Studdert-Kennedy excluded all responses except /U È/ can have hidden other important differences between tokens, e.g. the perception of diphthongs and glides (the importance of diphthong perception for their study was discussed by Lindblom and Studdert-Kennedy).

Lindblom and Studdert-Kennedy reported a definite overshoot in the responses to /wVw/ and /jVj/ context when these responses were compared to the responses of the corresponding tokens presented in isolation (i.e., #V# stimuli). However, the responses to tokens presented in context and those presented in isolation were collected on separate occasions. Furthermore, there is a significant difference between the responses to the 200 ms and 100 ms #V# tokens, which too were presented on different days. Therefore, it would be more prudent to compare the responses to /wVw/ and /jVj/ tokens collected within one session directly, i.e. the "combined" overshoot. This approach will be used here. For two subjects, no perceptual boundary between /U/ and /È/ could be determined for the /jVj/ syllables. Therefore, we can only use the responses of eight of the ten subjects.

The median difference between the F_2 mid-point values in /wVw/ context and in /jVj/ context for which /U/ changed into /È/ responses, i.e. the cross-over point in the responses, was 180 Hz for 200 ms vowel tokens and 274 Hz for 100 ms tokens. The cross-over point for /jVj/ syllables had a higher F_2 value than that for /wVw/ syllables, showing clear perceptual-overshoot. However, three out of the eight subjects showed consistent perceptual-undershoot instead of overshoot (all three tested in Sweden). If only the responses of the five subjects showing consistent overshoot were used, the median differences in F_2 mid-point value between /wVw/ and /jVj/ context, i.e. the combined perceptual-overshoot, became 289 Hz and 363 Hz (200 ms and 100 ms tokens respectively). This is a considerable amount of overshoot, approximately 30% of the combined excursion sizes (by defini-

tion: combined excursion sizes + combined overshoot = /jVj/ onset - /wVw/ onset = 1400 Hz for this experiment).

Lindblom and Studdert-Kennedy used the position of the cross-over point for vowel tokens presented in isolation to estimate the overshoot. From their numbers it followed that around two-thirds of the combined overshoot could be attributed to the /wVw/ context and one-third to the /jVj/ context. The amount of perceptual-overshoot (i.e., the difference between the cross-over points of the corresponding CVC and #V# tokens) proved to be unrelated to the excursion size (i.e., the difference between the onset and cross-over frequency) of the /wVw/ and /jVj/ tokens at the cross-over point or was even negatively correlated. The /wVw/ context induced much more overshoot than the /jVj/ context with only moderately larger excursion sizes. This was even found when only the data of the subjects showing consistent overshoot were used. In this experiment, formant on/offset track slope was directly related to formant excursion size. Therefore, when perceptual-overshoot was not related to the formant excursion size, it was also not related to formant track slope. It might have been related to the /w/ and /j/ context itself (see section 6.1.3.6).

Lindblom and Studdert-Kennedy also reported that a shorter duration (100 ms) increased the amount of perceptual-overshoot in the /wVw/ syllables for 9 out of 10 subjects (median increase in F_2 overshoot was 68 Hz, all ten subjects completed the answers for the /wVw/ tokens). However, when the significant effect of token duration on the responses to the isolated vowel tokens was taken into account, the increase in perceptual-overshoot in the /wVw/ syllables was found only for 6 out of 10 subjects (median increase in F_2 overshoot was 32 Hz). For the short duration too there was no relation between formant-overshoot and formant excursion size. When we combined their results for 200 and 100 ms tokens there was a strong negative correlation between excursion size and perceptual-overshoot for the /wVw/ tokens ($r \bullet -0.93$, $p \bullet 1\%$) and no correlation at all for the /jVj/ tokens.

The negative correlation between perceptual-overshoot and formant excursion size can undoubtedly be traced back to the design of the experiment. Because the on- and offset formant frequencies were fixed, the perceptual-overshoot can be defined as the #V# cross-over point minus the excursion size at the corresponding CVC cross-over point. The minus sign in this dependency creates a strong bias for a negative correlation. Nonetheless, if there had been a perceptual "target", calculated from the actual F_2 mid-point value and an extrapolation of the F_2 tracks, then there should have been a positive correlation between F_2 excursion size and perceptual-overshoot. The lack of any correlation between formant excursion size and perceptual-overshoot for the /jVj/ tokens could be the result of the smaller distance between the F_2 onset and cross-over frequencies and the small number of responses (no cross-over points were available for two of the subjects).

Lindblom and Studdert-Kennedy related their results to the overshoot found in diphthong perception. They discussed the fact that in diphthongs, generally only one of the two targets is actually realized. The presence of the other target is only suggested by the movements of the formants. "*Thus, an articulatory movement [Ae] or [AE] is heard as [Ai] by the naive*

listener" (quote from Lindblom and Studdert-Kennedy, 1967, p.842). From our results, described in chapter 5, we could infer that the tokens used in their experiments were indeed long enough, and had sufficiently large excursion sizes, to induce diphthong responses. Nearey (1989, p.2103) reported that stimuli with a similar formant track shape produced glide-like percepts. The fact that vowel-like consonants (i.e., /w/ and /j/) were added would only have strengthened this tendency. If their subjects would have interpreted their tokens as diphthongs, this would explain the overshoot in identification found. Subjects would have used the extent of the "glide" part as a co-specification to diphthong or glide identity. The design of the tokens then would cause a negative correlation between formant excursion size and "perceptual-overshoot". Diphthong or glide perception could also make more understandable the large differences between subjects. For some subjects the threshold for glide-perception might be so large that the F_2 track would "overshoot" the #V# cross-over F_2 frequency. In our experiments we also found that the number of diphthong responses varied widely between subjects. But we did not find any variation in the "direction" of the responses (i.e., perceptual over- or undershoot) between subjects when responses to formant curvature in general were examined.

Lindblom and Studdert-Kennedy (1967) concluded that vowel perception in context was influenced by perceptual-overshoot. When we consider the fact that their tokens strongly resembled glides or diphthongs (or even triphthongs), we might conclude instead, that they have only proven perceptual-overshoot for glides and diphthongs. When their tokens were interpreted as diphthongs, this might also explain the variation in behaviour between the subjects.

6.1.3.2 *The paper of Nearey (1989)*

Nearey (1989) repeated the experiments of Lindblom and Studdert-Kennedy (1967) with isolated vowels, /bVb/ and /dVd/ syllables, the latter two replacing respectively /wVw/ and /jVj/. Isolated vowels were synthesized with stationary formants. Instead of a parabolic formant track for the vowels in context, Nearey used a sixth order polynomial (i.e., $F(t) = F_{target} + (F_{initial} - F_{target}) \cdot (2 \cdot t/Duration - 1)^6$). Preliminary tests had shown that polynomials of lower orders did not give convincing stop-like percepts. The parabolic shape used by Lindblom and Studdert-Kennedy (1967) gave glide-like percepts.

The mid-point values of F_1, F_3, and F_4 were fixed at 700, 2400, and 4000 Hz, respectively. The F_2 mid-point value was varied in 20 steps from 900 to 1800 Hz. The vowel tokens were 100 ms long and had an F_0 of 120 Hz. The on-/offset values for F_1, F_2, and F_3 were 150, 2000, and 3000 Hz for /dVd/ and 150, 700, and 2100 Hz for /bVb/, respectively. In principle, this would have given F_2 excursion sizes ranging from 200 to 1100 Hz for both /dVd/ and /bVb/ tokens. However, due to the low F_1 on/offset frequencies, the F_2 amplitude was very low at the formant on- and offset points. The real F_2 on- and offset frequencies were measured at the -20 dB point and ranged from about 800 to 1170 Hz for /bVb/ tokens and from

about 1510 to 1920 Hz for /dVd/ tokens. This gives F_2 excursion sizes ranging from 100 to 630 Hz and from 120 to 610 Hz for /bVb/ and /dVd/ tokens respectively.

Subjects heard the tokens in blocked sessions, i.e. only one of #V#, bVb, or dVd per session, as well as in a mixed presentation, containing all three token types. They were asked to label the vowel stimuli as /Å/, /U/, or /E/. From the responses the cross-over F_2 mid-point values were determined where /Å/-/U/ and /U/-/E/ labels change. There was a clear effect of formant track shape on these cross-over points (i.e., silence, /dVd/, or /bVb/ context) indicating perceptual-overshoot. For the mixed condition, the overshoot was from 108 to 125 Hz with a single low value of 11 Hz for the /U/-/E/ boundary in the /bVb/ syllables (the former overshoot values were significant, the latter was not). The overshoot in the blocked condition was lower, from 36 to 88 Hz and 15 Hz respectively. The excursion sizes at the cross-over points were approximately from 160 to 430 Hz (/bVb/) and from 120 to 340 Hz (/dVd/).

Both when expressed in Hertz and in semitones, there seemed to be a negative correlation between F_2 excursion size (and therefore F_2 slope) and size of the overshoot ($r \bullet -0.7$), or no relation at all. The largest F_2 excursion size (430 Hz) resulted in the smallest overshoot (11 Hz) and vice versa (120 Hz excursion size and 125 Hz overshoot respectively). The excursion sizes of the /dVd/ tokens at the cross-over points were all smaller than those of the /bVb/ stimuli (both in Hz and in semitones). However, the perceptual-overshoot was always larger in /dVd/ tokens (in Hz, all but one in semitones). So, as in the work of Lindblom and Studdert-Kennedy (1967), there seems to be a context dependent co-specification of the vowels by F_2 track shape (e.g., excursion size).

Nearey compared the perceptual-overshoot he found with the amount necessary to compensate for the target-undershoot predicted by Lindblom (1963) and Broad and Clermont (1987). It was clear that the amount of perceptual-overshoot found in his listening experiments (11 to 125 Hz) was insufficient to compensate for the expected amount of target-undershoot (140 to 260 Hz). Again, there even seemed to be a negative correlation between the expected amount of target-undershoot and the amount of perceptual-overshoot actually found, or no relation at all. Considering the fact that 75% of the formant change was confined to only 20% of the total duration (compared to 50% of duration in Lindblom and Studdert-Kennedy, 1967), it is remarkable that any effect of formant track shape could be detected at all. The fact that these short transitions of the vowel have such a large effect on vowel identity suggests that the "perceptual-overshoot" found in this experiment is not caused by formant track shape itself but by the perception of the context it caused. This would mean that the context, and not the vowel realization, triggers the compensation for coarticulation. Such a mechanism would induce perceptual-overshoot in any vowel realizations presented in the proper context. This mechanism could be tested by presenting stationary tokens in the same context as "correct" and "incorrect" dynamic tokens. However, it is difficult to elicit good stop consonant percepts without the proper formant movements. This means that experi-

ments using stop consonants as a context could not readily distinguish between vowel inherent effects and context effects on perception.

Nearey concludes that his experiments have shown the existence of perceptual compensation effects for formant-undershoot in production. The amount of compensation found is quite small and seems to be unrelated to the formant excursion size or the formant track on- and offglide slopes. There also seems to be no relation with the amount of expected formant-undershoot in production. Therefore, the "overshoot" found could have been the result of some high level compensation for coarticulation instead of a low level "perceptual" overshoot.

6.1.3.3 *The paper of Di Benedetto (1989b)*

Di Benedetto (1989b) also found evidence that the shape of the F_1 formant tracks did influence vowel identification. She presented vowel tokens in a /dVd/ syllable with linear on- and offglides and a plateau of 15 ms in F_1 (see chapter 1, figure 1.3). The F_1 maximum varied between 330-500 Hz in 10 steps, the F_1 excursion size varied between 26-170 Hz (1.4-7.2 semitones). The F_2 changed symmetrically from 2593 to 2800 Hz and back. Her seven subjects had different language backgrounds, i.e. American English (4), Italian (2), and Japanese (1). Subjects were asked to label the tokens as /È i/ (high, closed) or /e E/ (non-high, open) depending on native language (using her terminology).

For all seven subjects, tokens with an onglide of 30 ms and an offglide of 70 ms were perceived as more open and less high than identical tokens with a time-reversed F_1 track (total token duration always 115 ms). The same was found when the long, 70 ms glide was shortened to 50 ms (total duration 95 ms). However, for the shorter tokens the cross-over F_1 frequency between /i È/ and /e E/ responses was always higher than for the longer tokens (for all subjects and for both stimulus types). Di Benedetto explained this effect from the intrinsically shorter duration of /È/ and /i/ in all languages involved. In a separate experiment she presented subjects with vowel tokens with different F_1 track shapes. From the results of this experiment she concluded that her subjects used the complete formant tracks to identify vowels.

Di Benedetto did not include control tokens with level F_1 contours. Therefore, she could not decide whether her subjects used perceptual-overshoot of the onglide or a weighted formant time average to identify the tokens. For the long tokens (115 ms), the cross-over points for the tokens with short and long onglides had almost identical onglide slopes. The fact that the same onglide slope could lead to less overshoot for longer onglides argues against perceptual-overshoot, but not against co-specification of vowel identity by onglide slope. For the shorter tokens (95 ms), the cross-over points of the long-onglide tokens had an almost 50% steeper slope than those of the short-onglide tokens. Still, some co-specification of vowel identity by F_1 onglide slope cannot be ruled out.

However, when we compared her results with those presented in chapter 5 we are inclined to conclude that the use of a weighted formant time average by the subjects is the more likely explanation. A conclusion that was

also favoured by Di Benedetto herself. With her data we made a (very) crude estimate of the relative weights attached to the first and second half of each of her tokens. The relative weights of the first and second half showed to be around 8:1 in favour of the first half (both durations, all subjects). This contrasts sharply with our own results that showed that the final half was most important for identification (chapter 5). This might mean that there was an effect of formant track slope after all. It is possible that the perception of the initial /d/ interfered with the weighting of the formant tracks. We might speculate that the curious effect of formant onset slope on cross-over frequencies mentioned above might be linked to a shift in the perception of the pre-vocalic consonant, which again might have induced a stronger perceptual compensation in the form of overshoot. This could be tested by presenting the tokens from Di Benedetto's experiment in isolation as well as in context.

6.1.3.4 *The paper of Fox (1989)*

Fox (1989) performed silent-center experiments with synthetic stimuli using a 7-step /bÈb/-/bEb/ continuum. Next to the mid-point values, his tokens also modelled the "natural" movements of F_1-F_3 with linear line segments. The total duration of the tokens was 300 ms. The duration of the vowel parts of the tokens was 255 ms, they consisted of symmetrical linear on- and offglides of 30 ms each and a stationary medial part of 195 ms. Listeners were asked to identify these tokens as either /bÈb/ or /bEb/, or to discriminate pairs of tokens to be the same or different. He presented listeners with the full tokens, silent-center tokens, and with medial vowel tokens. The silent-center tokens consisted of only the first and last 4 pitch periods of each vowel token (35 ms and 38 ms respectively) with a silent gap in between. The stationary tokens only contained the stationary medial vowel part (185 ms). The on-/offset to mid-point excursion sizes in the 7 tokens were in the range (maximal-minimal formant frequency), F_1: 30-95 Hz (1.3-3.5 semitones), F_2: 306-265 Hz (3.2-3.0 semitones), and F_3: 177-128 Hz (1.2-0.9 semitones). The formant track excursions in this continuum were such that a higher F_1 excursion size and a lower F_2 or F_3 excursion size indicated a more /E/-like vowel. It would therefore be difficult to distinguish perceptual over- and undershoot of formant mid-point values. Evidence for perceptual-overshoot from one formant would point to perceptual-undershoot for another formant.

In a set of discrimination experiments Fox was able to show that the silent-center tokens were perceived differently from the stationary medial vowel tokens. In separate experiments he presented the silent-center tokens also with only the outer 1, 2, 3, or the full 4 pitch periods of the on- and offset transitions, i.e. removing respectively 3, 2, 1, or no pitch periods from the inside of the original silent-center tokens. It appeared that the number of pitch periods present in the tokens influenced the identification scores. In general, the more pitch periods were present in a token, the more /E/-responses it got. This result could be explained by assuming that subjects identified the tokens on the transition end-point formant frequencies.

From the results of this last experiment it could be inferred that the F_1 frequency was the most important clue to token identity with the F_2 frequency as a good second (compare his table 4 with his figure 7, note that the F_2 end-point frequencies in this table 4 are incorrect). To test the hypothesis that tokens were identified on their transition end-point frequencies, Fox synthesized 200 ms vowel tokens with stationary formants with exactly these transition end-point frequencies. Listeners were asked to identify these tokens as either /È/ or /E/. The results clearly showed that the silent-center tokens were perceived as different from the stationary tokens with identical "medial" formant values.

Fox interpreted his results as evidence for dynamic-specification without discussing the direction of the perceptual difference between stationary and transition-only stimuli. However, from his figures 8 and 9, it followed that his results could be explained by assuming perceptual-undershoot of the F_2 or perceptual-overshoot of the F_1. For low F_1 values, there is little difference between token responses. At higher F_1 frequencies there is a steady excess of /È/ responses for the stationary tokens. This finding is consistent with both perceptual-undershoot of the F_2 and perceptual-overshoot of the F_1 in the silent-center tokens. However, the excess /È/ responses in the experiments of Fox do remind us of the same excess /È/ responses we found in our own experiments (see chapter 5). In our experiments the increase in the number of /È/ responses at short token durations was indiscriminate and could not be traced to any kind of under- or overshoot. This raises the possibility that the increase of /È/ responses in both experiments might have been caused by some factor unrelated to formant track shape. We will not pursue this matter further because at the moment this possibility cannot be substantiated.

To decide which explanation is more likely, perceptual-undershoot of the F_2 or overshoot of the F_1, we must estimate which would have the most effect. From our own results we would have expected the effects of F_1 movements to be more important than those of the F_2. However, in the experiments of Fox, the F_2 excursion sizes in the /È/-/E/ continuum were much larger than the F_1 excursion sizes, even when expressed in semitones. In our own experiments, the corresponding F_2 excursion sizes were comparatively smaller. Expressed in semitones, the F_1 excursions of our tokens were even larger than the F_2 excursions (cf. chapter 5). Furthermore, in the experiments of Fox, the parallel F_3 excursions are likely to have strengthened the perceptual prominence of the F_2 movements. All this might have made the F_2 movements more salient in the stimuli of Fox. From the fact that the F_2 movements were likely to be perceptually more salient than the F_1 movements, we are inclined to conclude that perceptual-undershoot of the F_2 (and F_3) formant tracks is the more likely explanation for his results.

The fact that Fox (1989) obtained consistent identification scores for single pitch period stimuli confirms our results with double pitch period stimuli. We too found that "transition-only" stimuli with a duration of 12.5 ms could be used reliably to find small shifts in the responses of listeners (see also Van der Kamp and Pols, 1971).

From the work of Fox (1989) we can conclude that transition-only silent-center stimuli are perceived differently from the corresponding stationary medial stimuli, i.e. the excised centers from the silent-center stimuli. From the experiment with short and very short transitions we can conclude that there was strong evidence for perceptual-undershoot of the F_2.

6.1.3.5 *The paper of Akagi (1993)*

As part of a larger effort to model coarticulation, Akagi (1993) studied vowel formant boundary shifts in perception (see also Akagi, 1992; and the review of this work by Repp, 1993). In his experiment, two Japanese subjects were asked to identify synthetic vowels as either /u/ or /a/. The stimuli in this experiment were stationary, five formant, vowel tokens with a duration of 50 ms. They were preceded by a stationary single formant anchor of 50 ms that was separated from the vowel token by a variable silent gap. The F_1 of the vowel tokens varied in such a way as to form a continuum from /u/ to /a/. The formant frequency of the anchor token preceding the vowel varied from below the lowest F_1 frequency to over the F_5 frequency. The duration of the silent gap, separating the anchor from the vowel token, varied from 0-300 ms in 25 ms steps. The results of his experiments showed that the F_1 values for which /u/ responses changed into /a/ responses depended on both the formant frequency of the anchor and the duration of the silent gap. Akagi concluded that there was an assimilation effect when the duration of the silent gap was below 70 ms (i.e., perceptual-undershoot) and a contrast effect when the duration of the silent gap was longer (i.e., perceptual-overshoot). This means that the presence of perceptual under- or overshoot was determined by the duration of the silent gap. Therefore, it seems that it was the temporal structure of the context that influenced the perception of the vowel more than the spectral difference between anchor and vowel token. This points towards an important role for context in the process of vowel identification. It also shows that perceptual-overshoot is not limited to "natural" stimuli.

6.1.3.6 *What factor could induce perceptual-overshoot?*

Akagi's (1993) study indicates that the structure of the vowel context might be crucial to the existence of perceptual-overshoot, or dynamic-cospecification in general (see also Brady et al., 1961). When we compare the results of our own study to that of Lindblom and Studdert-Kennedy (1967), we see that it is exactly there that the major differences are located (leaving aside the differences in response categories). They supplied a convincing and contrasting context to their vowel tokens, we did not. Nearey (1989) also ensured that the formant track slopes at the consonant-vowel transitions were as acute as those found in natural speech. He described the percepts of the plosives as convincing. Both Di Benedetto (1989b) and Fox (1989) used linear line segments to model plosive-vowel transitions. The quite long and gradual vowel formant on- and offset transitions used by Di Benedetto and Fox cannot be expected to have added much to the perception of the plosive context (compare these with the very acute on- and offsets of Nearey, 1989). What is more important, in these latter

studies all vowel tokens were presented in the same context so any effect of context would have gone unnoticed. It seems therefore, that the presence of perceptual-overshoot depends more on the perception of the context than on the actual formant track shape, i.e. formant excursion size, inside the vowel token itself (see also Tohkura et al., 1992; Repp, 1993; for related studies on context effects). This is supported by the fact that in none of the experiments the size of perceptual-overshoot of formant mid-point values was positively correlated with formant excursion sizes or formant track slopes. Without the perception of a proper context, subjects seemed to have reverted to the use of a weighted formant average to identify the vowel tokens.

6.1.4 Experiments using natural speech

With regard to the question of how vowels are identified by listeners, experiments using natural speech can be divided into two groups. One group investigates how vowel intelligibility is influenced by the context in which they are uttered. The other group compares the importance of the consonant-vowel transitions and the, more or less stationary, medial vowel part (i.e., the vowel kernel) for vowel recognition.

6.1.4.1 The influence of context on vowel intelligibility

Vowels spoken in consonantal context have mid-point spectra that differ from spectra taken from canonical realizations, i.e. vowels spoken in isolation (e.g., Stevens and House, 1963; Lindblom, 1963). It is therefore logical to suspect that vowels spoken in context are less well understood than those uttered in isolation. Initial experiments comparing vowel recognition in context with recognition of isolated vowels claimed that vowels in context were actually recognized *better* than those spoken in isolation (10% versus 30% errors, e.g., Strange et al., 1976; Gottfried and Strange, 1980; Strange and Gottfried, 1980). However, by taking more care on various methodological aspects such as dialect background and response procedure, Macchi (1980) found no difference between the intelligibility of isolated vowels and vowels in context (errors around 2%, see also the extensive reviews of Strange, 1989a; Nearey, 1989). Koopmans-van Beinum (1980) found that vowels excised from one-syllable words uttered in isolation were recognized worse than vowels spoken in isolation (16% versus 10% errors, p•0.01%, her tables 7.2 and 7.4). Most of the errors in the responses to her isolated vowels were caused by the problems of identifying the realizations of the short vowels /O A È π/ spoken in isolation because of their relatively long durations. Removing responses to these four tokens made the differences even more dramatic (13% versus 3% errors respectively). This shows that the difficulties with the duration cannot explain the differences in identification scores. Unstressed vowels from free conversation, which were severely reduced, performed extremely poorly (77% *errors*). As these unstressed and reduced realizations were very short, the errors were now concentrated in the responses to the long vowels (/a e/ received only 4.2% *correct* responses). However, even the four short vowels mentioned before were identified incorrectly in more than half of the responses (54% *errors*).

The differences in recognition rates reported can probably be explained by noting that the studies discussed by Strange (1989a) and Nearey (1989) primarily used plosive-vowel-plosive context and presented subjects with complete syllables. Koopmans-van Beinum (1980) used a mixed context of which plosives constituted only 25% and presented the vowels separated from their context, but with as much of the transitions as possible. This could indicate that the presence of the context itself would boost the identification of the vowels. This notion received support from the work of Huang (1991, 1992) and Kuwabara (1985).

Huang presented consonant-vowel-consonant syllables to subjects as well as the excised vowels from these syllables (i.e., without the consonants). The recognition rate for the full syllables was more than 8% higher than that for the excised vowels alone (79% versus 71%, $p \bullet 0.1\%$, Huang, 1991; calculated from her tables 4.4-4.11). Kuwabara found an even more dramatic effect of context. He used Japanese three-vowel sequences, taken from sentences. The medial vowel of each sequence was presented both in context and separately in isolation (i.e., without the two flanking vowels). Recognition of the medial vowel in isolation was much worse than in context (recognition rates of 80% and 96% respectively). However, it was not clear how much of the Vowel-Vowel transitions was included with the medial vowels when they were presented in isolation. It is therefore difficult to assess the significance of his results.

Next to the presence of the context, the nature of the context might also influence vowel recognition (as was also found by Gottfried and Strange, 1980). The results of Koopmans-van Beinum, Huang and Kuwabara show that the conclusion that vowels in context are recognized as well as vowels spoken in isolation (Strange, 1989a; Nearey, 1989) does not hold for vowel realizations presented without their proper context.

6.1.4.2 *The importance of the transition for vowel recognition*

Experiments that try to determine the importance of consonant-vowel transitions in vowel recognition, generally use the silent-center paradigm. Simple syllables, mostly of the stop-vowel-stop type (e.g., /bVb/) are recorded in carrier sentences. The vocalic part of the target syllables are divided into three parts: an initial part which contains all of the consonant-vowel transition (e.g., /bV/), a final part, which contains all of the vowel-consonant transition (e.g., /Vb/), and a medial part which contains the more or less stationary vowel kernel. Generally, care is taken to include only the transitions in the initial and final parts and to exclude parts of the vowel kernel. Then two new kinds of syllables are constructed, one containing only the medial part and one containing only the initial and final transition parts with silence substituted for the medial part. The original as well as the new syllables are then presented to listeners and the number of recognition errors is noted.

Several variations of the basic design of silent-center experiments are in use. The length of the syllables, either the medial vowel kernels or the silent centers, can be manipulated to exclude the original durational information from the tokens. The initial and final parts of the vowels used to

create the silent-center syllables can be taken from different realizations or even from different speakers (with opposite sex). Finally, the initial and final parts can also be presented separately in isolation. Sometimes, vowels spoken in isolation are also added for comparison.

Several studies using the silent-center paradigm are reported in the literature (e.g., Strange et al., 1983; Verbrugge and Rakerd, 1986; Strange, 1989b; Andruski and Nearey, 1992). Verbrugge and Rakerd asked listeners to identify /bVb/ syllables. The vowel could be one of /È i E œ U A U ɯ/. They heard the original syllables, silent-center syllables (with the medial 60% removed), hybrid silent-center syllables whose initial and final part were from different speakers (of opposite sex), and the initial and final parts separately. The pattern of recognition errors was typical for experiments with silent-center syllables. The error rate of the labelling was: whole syllables 8%, silent-centers 20%, hybrid silent-centers 26%, initial parts 48%, and final parts 66% errors. The error rate was much lower when short-long vowel errors were removed. All differences were significant, except for the differences between the two types of silent-center syllable. Others found that the centers-only were recognized as well as the silent-center syllables (Strange et al, 1983; Strange, 1989b). From these latter studies it could also be deduced that removing durational information almost doubled the error rate.

Verbrugge and Rakerd tried to device a way to predict the silent-center recognition scores from the individual recognition scores of the initial and final parts. In general, combining the recognition scores of the initial and final parts severely overestimated the recognition errors for the silent-center syllables, even when short-long errors were not counted. This was even so under the unlikely assumption that the recognition would be incorrect only when both parts were not recognized correctly. The same difference between recognition of individual parts and complete silent-center syllables was found in the other studies (Strange et al., 1983; Strange, 1989b). Both Verbrugge and Rakerd (1986) and Strange (1989b) found that the initial parts were recognized significantly better than the final parts. Strange also found that there was no difference in the error rate between the centers and the initial parts when durational information was removed from the centers. This result is similar to our own results. In chapter 5 we found that the difference in responses between onglide-only tokens and stationary tokens was small. Both differed markedly from the offglide-only tokens. The apparent difference in "error rate" in silent-center experiments and our own experiments (chapter 5) can be attributed to methodological differences (type of speech, language). Furthermore, it is difficult to define an error rate for our synthetic stimuli ("net shift" is not synonymous to error rate) as we do not know what the "correct" response should be.

What is striking in most of these studies is the small difference in recognition rate between the original syllables and the silent-center syllables. The 12% difference found by Verbrugge and Rakerd (8% versus 20% errors) was the largest of the studies discussed here. Strange et al. (1983) and Strange (1989b) found no significant difference at all between these two types of syllables. Verbrugge and Rakerd found that combining the initial part of a man's vowel realization with the final part of a female's, and vice versa, did not significantly affect the recognition of these hybrid silent-

center syllables. The results of the latter study indicate that the recognition of the vowel "target" frequency could not have been the result of a simple extrapolation of the formant tracks into the silent center. It strongly suggests that both parts were processed separately and that the resulting vowel "targets" were abstracted in such a way that they could be combined into a single, more dependable target.

In general, the results from these silent-center studies support our own results. We saw that the responses to the offglide transition of a vowel were generally shifted (i.e., caused more "errors") from those to the onglide and stationary medial parts. We also saw that there is at most only a small difference between responses to the onglide transition part and to the stationary medial part (Strange et al., 1983; Strange, 1989b). A large difference between our study and these silent-center studies was found when the different parts of the vowel realizations were assembled into a syllable. In our study we found that the combined on- and offglide tokens performed inbetween onglide-only and offglide-only tokens, i.e. these synthetic "syllables" did not perform any "better" than any one part alone. Literature shows that recognition of complete silent-center syllables from natural speech even outperformed the most optimistic predictions of errors made by combining recognition errors for the individual parts. Clearly, combining the on- and offglide transitions into a silent-center syllable added something that helped the subjects in recognizing the vowels. When fixed length syllables were used, recognition of silent-center syllables consistently outperformed recognition of the medial vowel part (recognition rates reached a ceiling when the original duration was preserved). This shows that the combined initial and final parts were not just used to reconstruct the missing medial part of the vowel because then they could never have been recognized better than the medial part alone.

6.2 Integration of the available results

When we combine the results of the silent-center studies with the studies using synthetic speech (most notably Lindblom and Studdert-Kennedy, 1967; Nearey, 1989;) a possible explanation emerges. In the studies using synthetic speech we saw that the effects of coarticulation were compensated in well integrated syllables and could be demonstrated when different consonants were contrasted. Such compensation (e.g., perceptual-overshoot) was absent in our own, non-integrated syllables and could not be proven in the several other studies (Di Benedetto, 1989; Fox, 1989). These latter studies have in common that less pain was taken to produce convincing consonant-vowel transitions in contrasting arrangements. When compensation for coarticulation was found in experiments using natural speech, e.g. with silent-center syllables, the original context (such as the release bursts) was always present with most, if not all, of the consonant-vowel transitions (e.g., Strange et al., 1982; Verbrugge and Rakerd, 1986; Strange, 1989b). So we might very well assume that the original context was indeed perceived as such.

We can now hypothesize that there is a mechanism to compensate for vowel formant target-undershoot in production due to coarticulation. This

mechanism does not work on the spectro-temporal shape in the vowel itself. Instead, it works at the level of the syllable and beyond. It will compensate vowel formant target-undershoot using the syllabic or wider context. The evidence so far available indicates that dynamic information from the transition parts of the vowel is used for compensation, but only when it contains sufficient information about the context. This mechanism would explain a lot of the results discussed so far.

It is not surprising that the vowels-with-context in silent-center syllables will not be recognized any better than vowel realizations spoken in isolation, as Andruski and Nearey (1992) found. A vowel spoken in isolation will contain all information necessary to be recognized in its original context, i.e. silence. Any compensation for context in silent-center syllables can hardly be expected to improve that. However, it will be clear that silent-center vowels will be better recognized than the isolated medial vowel parts because these medial parts do not contain the information necessary to compensate for coarticulation. The initial and final parts, when presented separately, do contain this information but are not perceived as syllables and therefore, no compensation is performed.

In our own experiments (chapter 5) we wanted to compare identical vowel realizations in different context (including presentation in isolation). We wanted to test the effects of the presence of a context *an sich* on the identification of vowel tokens. To achieve this, we deliberately did not change the formant track shape to match the context in which the vowel token was presented. Therefore, the vowels in the /nVf/ and /fVn/ pseudo-syllables we used might have been perceived as still being "pronounced" in isolation and not in well integrated syllables. Furthermore, we do not know whether /n/ and /f/ are capable of inducing a detectable amount of compensation even in natural speech. In neither case, any compensation would have been found in our experiments.

Another serious problem in our experiments might be the effect of context on perceived duration. In our experiments, any consonantal context changed the number of long-vowel and diphthong responses. As a consequence, any comparison of responses to identical vowel tokens presented in isolation and in different contexts immediately faltered on exchanges of long- and short-vowel responses. After removing these long-short exchanges, there were not enough changed responses left to give meaningfull results. Therefore, the results of our experiment could only be used to show that vowel-inherent (dynamical) cues are not enough to induce compensation for coarticulation. Our results could not be used to decide whether the vowel context can induce such compensation.

If the compensation for coarticulation is performed only after the context is "reconstructed" by the listener, this would also explain the good results for hybrid silent-center syllables. Both parts in a hybrid silent-center syllable give the same (hypothetical) "proto-targets" for the vowel and context. These would then have been combined and the compensation would have been determined from the combination of these elements. What information is actually used to determine the compensation is not clear at this moment. The results of the experiments using synthetic speech do point towards dynamic information, specifying formant movements. But in these experi-

ments, the dynamic information strongly correlated with the "locus" values of the consonants in the context. This still leaves the possibility that, in these experiments too, the listeners used the *identity* of the perceived consonants to help identify the vowel and not the formant track shape itself. It is therefore not really possible to distinguish between these two possibilities at the moment.

6.3 Conclusions

We can summarize the evidence presented in section 6.1 and 6.2 as follows. The shape of formant tracks carries information that could be used to compensate for coarticulatory formant-undershoot in production and so could help to improve vowel identification (section 6.1.3.1). Experiments with synthetic speech indicated that, when tokens were presented in an appropriate context, subject did use the formant track shape in a way that would have compensated for the effects of coarticulation in that context. Without such a context, this dynamic information was not used by subjects and was even detrimental to "identifying" any canonical target, assumed to correspond to the given formant track shape (section 6.1.3.3). Experiments with natural speech indicated that (parts of) vowel realizations were identified better in their original context than when excised from it and presented in isolation (section 6.1.4.1). In their original context, vowel realizations were equally intelligible as vowels spoken in isolation.

Together the above facts strongly suggest that the information in formant dynamics is used only when vowels are heard in an appropriate context. It might even mean that it was the context, and not the formant dynamics, that determined how vowel realizations were identified, e.g. whether there was some "perceptual" compensation for coarticulation.

7

GENERAL DISCUSSION

Abstract

In previous chapters we have found that an increase in speaking rate did not result in an increase of the amount in formant-undershoot in the vowel nucleus. It also did not change the time-normalized formant track shape. In this final chapter we discuss several possible alternative explanations for this lack of effect of speaking rate on formant-undershoot. We demonstrate that our methods were sensitive enough to detect the predicted amount of excess undershoot. We also show that the context from which our vowel realizations were taken should have induced a sizeable amount of excess reduction if a higher speaking rate indeed increases formant-undershoot. From this we conclude that our speaker has read the text faster without an increase in formant-undershoot (i.e., coarticulation and reduction). This means that target-undershoot is not the results of articulatory limitations but is most probably planned. Our perceptual experiments showed that listeners did not compensate for vowel target-undershoot unconditionally. A large excursion size in the formant tracks of synthetic vowel tokens induced perceptual-undershoot instead of perceptual-overshoot, at least when these tokens were presented in isolation or in a non-integrated /nVf/ *context. Our subjects tended to identity the vowel tokens on their offset formant values. These results disagree with current models of vowel perception. A close inspection of the relevant literature showed that the role of the context in vowel recognition is probably underrated in current theories.*

Introduction

In the previous chapters, we investigated some aspects of the production and perception of vowels. We tested predictions that were obtained from current theories on vowel articulation and perception (see chapter 1). On all accounts, the results of our experiments disagreed with some of the leading models about vowel production and perception. If vowel target-undershoot is defined as a shift of the formant values in the vowel nucleus away from the canonical target, then our study showed that speaking rate did not influence target-undershoot (i.e., coarticulation and reduction). It also did not change the distance between vowel formant on/offset and nucleus values (i.e., the excursion size). Together, this means that, after time-normalization, articulation was not affected by speaking rate. Also, vowel identification was impaired, instead of supported, by the presence of non-level formant tracks.

Naturally, these results raised new questions. How could they be reconciled with the results presented in the literature? Would it be possible to incorporate all the contradictory reports from the literature, and our own results, into a model of how vowels are used in speech? In the following sections we will discuss these questions and we will try to answer them.

7.1 Target-undershoot in production

In the production part of the present study we determined whether speaking rate had an effect on the production of vowels by an experienced newscaster. This way we investigated the question whether formant-undershoot is planned or whether it is caused by the mechanical limitations of the articulators (i.e., jaw, tongue, lips). If mechanical limitations were the cause of the vowel formant target-undershoot found in normal, connected speech, we would have found excess undershoot, i.e. even more coarticulation and reduction, when our speaker spoke at a fast rate. If mechanical limitations were not the cause of target-undershoot, then our speaker would have been able to adapt to a higher speaking rate without any excess undershoot, for instance, by increasing speaking effort.

Comparing vowel realizations uttered at a fast and a normal speaking rate, we were not able to detect any differences in the amount of spectral reduction or coarticulation between them. This implied that when speaking fast, our speaker reproduced all formant movements that he also produced when speaking at a normal rate, but now using less time.

In this section, we will discuss our findings in the light of the prevalent target-undershoot model. We will try to determine whether our results can indeed be used to distinguish between undershoot caused by articulatory limitations and undershoot as a pre-planned process, i.e. between input-driven and output-driven undershoot. We will do this by addressing the question whether the target-undershoot model predicts a detectable difference in formant-undershoot for the two speaking rates used.

There were several factors that could have prevented us from finding any excess target-undershoot due to an increased speaking rate, such as:

1. The method of formant analysis was inadequate.
2. The durational difference between speaking rates was too small.
3. The undershoot had already reached a ceiling (or floor) in normal-rate speech.
4. Contextual variation had averaged out any change.
5. The differences between vowel target and on/offset were too small (i.e., not enough coarticulation).
6. Other articulation strategies were used in fast-rate speech.

Below we will discuss them all.

7.1.1 Quasi-stationary formant analysis might give inaccurate values

In our study we determined formant frequencies. This was done by using an LPC-10 analysis procedure with a shifting 25 ms window (1 ms step-size). This method basically assumes that the signal is stationary within the 25 ms window, hence the phrase "quasi-stationary". Speech is of course not stationary. Consequently, the analysis will give results that are some kind of average over the 25 ms of the window. As a result of averaging, shorter realizations will tend to show some "undershoot" compared to longer realizations. However, most vowel realizations were well over 50 ms long and the central part of these vowel realizations tended to be rather stationary. Therefore, we think that our vowel formant nucleus frequencies were not influenced much by this spectral averaging. Furthermore, in a recent study, the accuracy of LPC-10 analysis in capturing formant track shape was assessed to be quite good (Smits, submitted). Therefore we do not think that this problem really corrupted our measurements. This conclusion is supported by the fact that we did not measure any duration-related undershoot. Had we found any undershoot, the averaging might have been a problem. Because we did not, the argument seems to remain rather academic.

When determining formant track shapes, the effect of averaging by using an analysis window would be a levelling of the tracks. This levelling would have increased with decreasing durations. We used whole vowel modelling of formant tracks with polynomials of a low order (only up to fourth order Legendre polynomials, see chapter 4). This in itself already constitutes a smoothing of the formant tracks. We think that this smoothing is stronger than that produced by the window in an LPC analysis. Again, we did not find any solid evidence for a duration dependent levelling of the formant tracks. The averaging effects of the analysis window seemed not to have caused any problems.

There is one area where the window-size does cause problems. At the vowel on- and offset boundaries, half of the analysis window will sample the context of the vowel realization instead of the vowel itself. As formant frequencies tend to be ill-defined in consonants or in rapidly changing consonant/vowel boundaries, formant frequencies measured here might be atypical (Smits, submitted). This is not to say that the (possibly incorrect) formant frequencies at the on- and offset boundaries behave in an irregular way. Correlations between speaking rates for formant frequencies at the

boundaries were as high as for those in the middle part (figure 3.2, chapter 3).

To summarize this discussion: Using a quasi-stationary method for formant measurements could have introduced the duration-dependent undershoot we were looking for. Because we did not find any duration-dependent undershoot, these fears remained unsubstantiated.

7.1.2 Too small a difference between normal- and fast-rate speech

The most obvious explanation for not finding any excess target-undershoot is that the differences between the two speaking rates were too small to cause any detectable difference in formant-undershoot. Indeed, the difference in vowel duration was on average only 15% (short vowels 12%, long vowels 19%, schwa none). This difference is quite small compared to the differences reported in other papers (e.g., Lindblom, 1963; Lindblom and Moon, 1988; but see Den Os, 1988, p.66). The difference in vowel duration between stressed and unstressed vowels was double that between speaking rates (30% versus 15% in our data, cf. Den Os, 1988, p.71). However, it must be remembered that undershoot is expected to increase exponentially at shorter durations. The durations of our vowel realizations were on the lower edge (and beyond) of those used by Lindblom (1963). Small changes in duration should exert large changes in undershoot at these already rather short durations.

The question of whether the differences in vowel duration between speaking rates were too small to induce a measurable increase in undershoot, depends on the sensitivity of our tests. Assessing the sensitivity of our method on an a priori basis was difficult. The sensitivity depended on the number of realizations and on how systematic the differences between speaking rates were. Not enough is known about the differences between speaking rates to assess their impact on the sensitivity of our methods. However, we can do an a posteriori assessment of sensitivity by determining the smallest differences that were found to be significant. For both F_1 and F_2, the smallest differences that could be positively identified between speaking rates were only 20 Hz (chapters 2-4), with an occasional outlier down to 15 Hz. So we must conclude that only if an overall decrease in vowel duration of 15% had induced a systematic increase in formant-undershoot of less than 20 Hz, we would have been unable to detect this excess undershoot. For the F_1 values that we presented in chapter 2, there is no question of whether excess undershoot could have been detected or not. If these F_1 frequencies in fast-rate speech showed anything, it was overshoot instead of undershoot. However, for the F_2 values, no apparent differences between speaking rates were found. To know whether this lack of a difference in F_2 could have been due to the small difference in duration it is necessary to estimate the expected amount of excess undershoot.

We used the model and data of Lindblom (1963) and the mean vowel durations from chapter 3 of the present study to estimate the size of the expected excess undershoot in F_2 due to speaking rate in our own data (see figure 1.1, chapter 1). This was done for the three different contexts that Lindblom had used (i.e., b_b, d_d, g_g) and the vowels that were closest to

ours (/È œ Ø a O U/ in his study). Of the three values of formant-undershoot predicted for each of our vowels (one for each /b/, /d/, or /g/ context), we used only the median value. Using the median value is more realistic than using the extreme values because of the diverse context in our samples which would tend to average out the excess undershoot. For our realizations of the vowels /A a o u i/, the expected amount of excess undershoot due to a higher speaking rate was in the range of 30-40 Hz. This value is larger than the threshold of detection determined earlier.

We used primarily a sign-test to detect differences between speaking rates. Therefore, the size of the difference in F_2 values between normal- and fast-rate might have been less important. It was the systematic nature of the excess undershoot that would have counted. Fast-rate vowel realizations were measurably, and systematically, shorter than the corresponding normal-rate realizations for instances of the vowels /A a o i/ (chapter 2 and 3). If a shorter duration had invariably resulted in more centralization (i.e., reduction), this excess undershoot should have been detected just as readily as the shorter duration. This is especially so for any excess undershoot in the back vowels /A o u/. For these back vowels, excess F_2 undershoot should have been towards higher F_2 values in (almost) every context. Therefore, any excess undershoot in realizations of these three vowels due to speaking rate should have been highly systematic.

We conclude that the amount of undershoot predicted from the literature would have been large enough to have been detected by the methods used in this study. However, we did not find any systematic increase in formant-undershoot due to an increased speaking rate. This indicates that the increase was either not systematic or much smaller than previously expected from a purely passive model with all parameters fixed.

7.1.3 A ceiling (floor) in undershoot was already reached

It could be that there is a maximum amount of formant-undershoot. At the most extreme case, undershoot could not exceed (nearly) complete assimilation if the remaining sound should still be a vowel. When this "minimal" vowel is reached and the vowel realization has completely blended with its context, a further decrease in duration would not lead to an increase in undershoot. If this ceiling for undershoot had already been reached in normal-rate speech, no extra undershoot should have been expected when speaking rate was increased. If this is true, the target-undershoot model seems to be of limited use for explaining variation in vowel realizations in normal speech.

However, we did find differences between stressed and unstressed vowels at both speaking rates (chapter 3, 4). Speaking-rate-related differences in duration were comparable for stressed and unstressed vowels. Therefore, there seemed to be enough room for additional formant-undershoot in the stressed vowels at a normal speaking rate. This potential extra formant-undershoot was not found with a faster speaking rate.

7.1.4 Variation in context has averaged out any difference between speaking rates

Coarticulation and reduction cause vowel formant mid-point values to shift towards the formant on- and offset frequencies (e.g., see Van Bergem, 1993). For some consonants this might result in a shift away from the center of vowel space, for others it might result in a shift towards the center of vowel space. As a result, the shift of formant mid-point values for vowel realizations taken from a mixture of contexts might average out to zero (i.e., no shift at all).

In this study we used an existing text. The text had been used in a radio broadcast and discussed economics (see appendix C). The text was used unaltered and no provisions were made for the occurrence of vowels, consonants, or words. Therefore, this text can be considered to be a typical example of modern Dutch. From this text, we used all realizations of seven vowels. In table 7.1 we present for each vowel the frequency of pre- and post-vocalic context. From the study of Pols and Schouten (1979) it can be concluded that for the back vowels /a A O o u/ and the high vowel /i/, the vowel formant on- and offset values will lie in the inner parts (i.e., away from the edges) of the vowel triangle for the most important consonantal contexts (i.e., /n d t r/). We must also consider the fact that /n d t s z/ have very similar "loci" and therefore will cause formant-undershoot in approximately the same direction. Therefore, the conclusion that more reduction equals more centralization can be extended to all five consonants. Together with the /r/, these consonants make up half of the context of our vowel realizations (cf. table 7.1). As a result, we would expect the vowel formant on- and offset frequencies to be, on average, more central than the vowel mid-point frequencies. So there is no reason to expect that an increase in formant-undershoot due to an increase in speaking rate should not have

Table 7.1.a: Context preceding the vowels.
For each vowel the number of occurrences of the 10 most frequent context items are displayed. Context is given without regard for syllable, word, or sentence boundaries. However, a perceptual silence or pause was considered a distinct item and is indicated by the symbol "#". Phonemes from voiced/voiceless oppositions were pooled, as were preceding vowels. The last column but one contains the total as a percentage of all realizations. The last column (labelled KvB) contains the corresponding percentage taken from Koopmans-van Beinum (1980) for free conversation averaged over four speakers (her tables 2.2 and 2.3). Consonant contexts that were also investigate by Van Bergem (1993) are underlined.

Contex	E	A	a	i	o	´	u	y	total	%	KvB %
d/t	16	40	15	22	11	5	1	5	115	19.6	18.1
n	6	0	12	30	9	6	2	2	67	11.4	6.0
s/z	12	1	10	13	14	3	2	3	58	9.9	6.0
m	19	3	23	5	1	2	2	0	55	9.4	6.2
v/f	2	28	5	0	15	0	0	1	51	8.7	5.2
#	34	4	3	0	5	3	0	0	49	8.3	15.0
r	3	6	6	9	6	3	0	0	33	5.6	6.8
Vowel	4	12	3	2	3	4	1	0	29	4.9	0.5
w	10	9	3	1	1	0	0	0	24	4.1	5.7
X	2	3	5	3	6	0	4	1	24	4.1	0.9
Others	16	17	20	7	18	0	4	0	82	14.0	26.3
total	124	123	105	92	89	26	16	12	587		

shown up as more centralization.

The previous arguments were rather theoretical. We would like more solid evidence that a sample of vowel realizations like ours, indeed showed centralization with increased reduction. In previous studies, it was found that reduction means more centralization when samples of vowels from normal utterances were used (Koopmans-van Beinum, 1980; Krull, 1989; Van Bergem, 1993). For Dutch, both Koopmans-van Beinum (1980) and Van Bergem (1993) found reduction to be almost synonymous to centralization for large samples of vowel realizations. It is therefore interesting to compare the distributions of context for their vowel realizations with ours. We included the corresponding numbers from the study of Koopmans-van Beinum (1980) in table 7.1, and also indicated which consonants were used by Van Bergem (1993). We can see that, compared to the study of Koopmans-van Beinum, our sample of vowels was not biased towards rare or unusual contexts. Most consonants used by Van Bergem were also dominant in our sample. Both the study of Koopmans-van Beinum and that of Van Bergem showed that reduction in a typical sample of Dutch vowels averages out to formant-undershoot towards the center of the vowel triangle (i.e., centralization). As a consequence of the similar distribution of consonants over the context of our sample of vowel realizations, an increase in vowel reduction due to speaking rate should also have resulted in increased centralization of our vowel realizations. Therefore, the fact that we did not find more centralization in our sample of vowels means that there was no increase in formant-undershoot due to an increase in speaking rate.

From the previous discussion it could be concluded that, on average, vowel formant on- and offset frequencies were centralized with regard to the vowel nucleus. This was tested for our speech material. For this test, we determined the average excursion size for each vowel. The formant excursion size was calculated from the Legendre polynomial coefficients (estimated as $\Delta F = -3/2\ P_2 - 5/8\ P_4$, see chapter 4).

As expected, we found that mean excursion sizes were definitely different from zero for all but the closed vowels (/u y i/ for F_1 excursion sizes) and the mid-F_2 vowels (/y ´ a/ for F_2 excursion sizes). For these latter vowels,

TABLE 7.1.b: As 7.1.a Context following the vowels.

Context	E	A	a	i	o	´	u	y	total	%	KvB %
n	52	41	18	10	4	0	0	1	126	21.5	15.4
t/d	12	35	17	23	2	17	3	1	110	18.7	12.8
r	12	2	37	3	30	8	0	3	95	16.2	18.4
l	15	16	9	0	4	0	0	0	44	7.5	6.4
k	3	7	5	5	6	0	6	1	33	5.6	9.3
s/z	4	4	3	17	5	0	0	1	34	5.8	9.9
X	7	9	5	1	6	0	0	0	28	4.8	3.4
v/f	2	0	1	2	18	0	2	0	25	4.3	2.5
b/p	5	2	1	3	5	0	0	3	19	3.2	3.8
w*	1	0	1	10	0	0	0	2	14	2.4	0.7
Others	11	7	8	18	9	1	5	0	59	10.0	17.4
total	124	123	105	92	89	26	16	12	587		

** /w/ was limited almost completely to the vowel /i/. Therefore, we present it here, although the more evenly distributed /m/ was somewhat more frequent (16 versus 14 times).*

the F_1 or F_2 excursions indeed averaged out. For all others, the average excursion sizes were significantly different from zero and the variations due to context clearly did not cancel out (cf. P_2 values of chapter 4, figure 4.2). Indeed, the average formant excursion sizes all indicated that formant on- and offset frequencies were centralized with respect to the F_2 values at mid-point and closed with respect to F_1 values (i.e., towards low values for F_1). This test too lead to the conclusion that more formant-undershoot should on average result in more centralized vowel realizations.

To summarize this discussion: if we compare the context from which we excised our vowel realizations with that used in other studies, we can conclude that more formant-undershoot due to an increased speaking rate is expected to result in a centralization of formant values. When we actually analyzed the formant excursion sizes we again saw that, on average, an increase in formant-undershoot due to speaking rate should have resulted in more centralization. In neither case was there any evidence that context variation could have averaged out changes in the amount of formant-undershoot due to differences in speaking rate.

7.1.5 Coarticulation was not strong enough to require extra undershoot

Target-undershoot depends on the difference between vowel formant on- and offset frequencies and the canonical target frequency, the latter being the theoretical mid-point value of very long realizations. The vowel formant on- and offset frequencies in turn depend on the consonants in the context. Not all consonants induce strong coarticulatory effects in the vowels. It all depends on the "articulatory distance" between vowels and their flanking consonants. If we had used vowel segments from a more or less neutral context, e.g. hVd in American English (Stevens and House, 1963), no additional undershoot would have been expected.

In our study we used a normal text (see discussion in section 7.1.4). We used *all* realizations of the chosen vowels, irrespective of context. The chosen vowels were distributed over the vowel triangle. Therefore, our set of vowel realizations can be considered to sample the natural range of contexts in Dutch (see table 7.1). We must acknowledge that some highly coarticulating consonants, like /w j/, were rare. But that was because these consonants are rare in Dutch. If an increase in speaking rate only induces a detectable amount of additional undershoot in these rare, highly coarticulating contexts, then duration is obviously not a major determinant of variability in vowel realizations.

We did test whether a larger "articulatory distance" between vowels and context would have changed our results. To do this, we selected vowel realizations from an alveolar context (i.e., one of the consonants /n s z t d r l/). These consonants would restrict the tongue to a high and fronted position, i.e. close to the position it takes for the vowel /i/. The articulatory distance between the consonants of the context and the high, fronted vowels (i.e., /i E y/ from our sample) would be relatively small. The distance with the low, back vowels (i.e., /u o A a/ from our sample) would be comparatively large and should therefore induce a sizeable amount of excess formant-

undershoot with an increase in speaking rate (c.f. Gopal and Syrdal, 1988). But even this subset of realizations with a large articulatory distance between vowels and context did not show any excess formant-undershoot at a fast speaking rate.

7.1.6 Alternative articulating strategies

A reorganization of articulatory movements is often forwarded as an explanation of a lack of undershoot (e.g., Kuehn and Moll, 1976; Gay, 1981; Lindblom, 1983; Engstrand, 1988). If this really is the explanation, it is not clear what triggered the change in articulation strategy in the speakers of these studies. Especially because this change seemed to be very speaker specific (see also Flege, 1988). In our experiment, we did make sure that our speaker used a regular "reading" style of speaking. The text was long, it was only one of a whole collection of texts that had to be read on a single day, and there were several hours between both readings of the same text. Therefore, the style of speaking must have been "normal", apart from speaking rate itself, for both readings or else our speaker would not have been able to maintain this style throughout the day. Informal listening did not reveal any conspicuous difference in speaking style, except for speaking rate.

Any change in articulatory strategies, including a change in articulatory effort, that is not just a uniform acceleration of articulatory movements should result in a change in formant track shape after time-normalization. We did not find any evidence for such a change in strategy. The results of chapter 3 and 4 all point to a uniform increase in articulation speed.

7.1.7 Does duration control vowel target-undershoot?

We must conclude that our speaker indeed did read the same text faster without an increase in formant-undershoot. This means that duration in itself does not determine formant-undershoot. Together with the results obtained by other studies (Engstrand, 1988; Lindblom and Moon, 1988; Fourakis, 1991), this leads to the conclusion that the relation between vowel duration and formant-undershoot is specific for each speaking style and rate. Speakers were generally able to adapt their speech to any articulatory rate.

It has been shown that reduction in unstressed syllables can be independent of duration (Den Os, 1988; Nord, 1988; Fourakis, 1991). Whalen (1990) showed that, at least sometimes, coarticulation is planned (i.e., output-driven). It is also known that spectral vowel reduction depends strongly on speaking style (Koopmans-van Beinum, 1980) and even language (Delattre, 1969). Therefore, we must conclude that, whatever the cause of formant-undershoot (coarticulation and reduction), it is not the mechanical limitations of the human articulators, i.e. it is not input-driven. Considering the evidence discussed above, we follow Whalen (1990) in that it is more likely that undershoot is to a large extent planned.

7.2 Perceptual-overshoot, dynamic-specification, and target models of perception

If we conclude that the variation in vowel realizations that result in coarticulation and reduction are introduced on purpose (i.e., planned), the question of how listeners cope with this variation becomes even more complex. If the variation in vowel realizations would have been systematic and the result of physiological factors, listeners could compensate for it at the level of the individual segment. Such perceptual compensation could be automatic and "low-level". However, if the variation in vowel realization is wilfully introduced (and possibly language dependent), its presence cannot always be relied upon or be deduced from the vowel segment alone. Therefore, this variation cannot be neutralized automatically by the listener using only clues from the vowel segment itself.

As a consequence of the putative planned nature of coarticulation and reduction, there are only two ways of compensating for the variation that results from it. First, vowel realizations could contain invariant clues that are not affected by coarticulation and reduction. These could be used to compensate for variability or circumvent it altogether. The other possibility is that the presence of a likely "cause" of changes in a realization would be deduced first (e.g., coarticulation with a certain consonant). This knowledge could then be used to undo the expected changes in the vowel realization. The former approach is the basis for most theories on human vowel recognition. A limited version of the latter approach is used successfully in automatic speech recognition where phonemes are classified in context only, e.g. when using triphone models and Multi-Layered-Perceptrons (for an overview, see e.g., O'Shaughnessy, 1987).

To sort out those acoustic features that listeners use to identify vowel realizations is a difficult job. Natural speech is very complex. Even though vowels are comparatively simple sounds, they are characterized by the temporal course of many variables (e.g., F_1-F_3, intrinsic F_0 and duration, loudness). All these variables are also context sensitive. As most of these parameters are strongly correlated in natural speech, it is not generally possible to determine what variable caused what effect in perception. This leads to a dilemma in the study of speech between using natural and synthetic speech. The more natural the speech used in an experiment is, the less clear it will be which acoustic feature caused what perceptual response. However, the more individual variables are isolated and controlled in synthetic speech, the more likely it is that relevant features have been removed with the uncontrolled variation. In the former case we are not sure of what has actually been measured. In latter case, it is difficult to ensure that what has been measured is relevant to natural speech too. The result of this dilemma is a dependency between experiments with natural and synthetic speech. Experiments with natural speech are necessary to suggest which parameters might be of importance in perception. Experiments with synthetic speech are needed to prove that the suggested parameter is indeed capable of inducing the perceptual effect. After which a new round of experiments is needed to check whether there are more acoustic features that could induce the same percept.

For this reason we cannot interpret our results without taking into account other studies using natural and synthetic speech. In the next sections we will summarize the results of our experiments with synthetic speech and try to integrate them with the existing literature which was evaluated in chapter 6. Finally we will try to decide whether and how static and dynamic features of vowel realizations influence vowel recognition.

7.2.1 Recapitulation of our vowel identification results

In chapter 5 we found a consistent perceptual-undershoot in the responses of our subjects (see also Pols and Van Son, 1993). We concluded that our listeners used mostly formant values from the final part of each token to identify it. This was found for all durations and both in isolation as well as in pseudo-syllables with /n/ and /f/. The perceptual-undershoot was consistently found for all four track shapes, i.e. concave downward and upward, both for F_1 and F_2. However, the predominance of the final part of the tokens in the responses could not be found for concave downward tracks in the F_2. The size of the shift in the responses depended on the size of the F_1 excursion. The shift was larger for larger excursion sizes (a dose-response relation).

The size of the shift in responses due to perceptual-undershoot was almost insensitive to duration. There were only minor differences between the responses to tokens of 25 ms and 150 ms, apart from the obvious differences in the number of long-vowel responses. Furthermore, listeners did not use the exact offset point for identification. If they had done so, the onglide-only tokens would not have shown any shift in responses. However, onglide-only tokens did induce a small but consistent amount of perceptual-undershoot. Therefore, as duration did not matter, it appeared that listeners used either a fixed fraction of the total duration or a weighted average of each formant track, scaled for token duration. In both cases, most emphasis was laid on the final half of the vowel tokens.

From a practical point of view, it makes sense to use the final part of an isolated vowel realization to identify it. In speech, short, isolated vowels would come closest to their canonical target at their offset. But, we also found this tendency when we surrounded our tokens with synthetic consonants. Here, one would have expected that listeners would use the part furthest from the consonants to identify a vowel token. But this specific context did barely influence their responses.

The shape of vowel formant tracks also influenced the identification of the surrounding /n/ and/or /f/ segments. These consonants were most often *mis*-identified around vowel tokens with level formant tracks or an "unconsonantal" concave upward F_1 track (i.e., ΔF1 = -225 Hz). Furthermore, the probability of reporting "extra" consonants, i.e. those not explicitly inserted in the signal, also depended on the formant track shape. It was highest with a concave downward F1 track shape (i.e., ΔF1 = 225 Hz).

The fact that we found that a relatively small part of each token was used to identify it would be in agreement with (compound) target-models (Strange, 1989a; Andruski and Nearey, 1992). However, compound target-models assume that listeners use the vowel kernel or nucleus to identify it.

In our study listeners used the offset part. The relevant literature does not supply data on how listeners detect the vowel kernel in natural speech. It is generally assumed that listeners somehow use the vowel mid-point or the part with the least spectral change. Both these strategies can be ruled out for our tokens.

Other options are the point inside the vowel realization with maximal loudness or furthest from the context in an integrated syllable. Our tokens were synthesized with constant source power. Therefore, the importance of the loudness envelope could not be checked with our data.

To determine the role of the context in determining the perceptual "target"-point, we presented vowel tokens also in pseudo-syllables (i.e., nVf or fVn). This did not change the responses markedly. From this we can conclude that the sheer presence of speech surrounding a vowel will not induce compensation for coarticulation, nor will it shift the "identification" point of the token towards the mid-point. It is still possible that such a compensation or shift will occur only in more integrated contexts and that our tokens in the peculiar n/f context were still perceived as isolated vowels. However, this would mean that a listener would first have to identify the context, detect the coarticulation and only then would pick a point inside the realization to identify it.

Whatever the reasons for our unexpected results, they do show that current models of vowel perception are incomplete. If dynamic-specification is important in normal speech perception, factors other than the mere shape of the first and second formant track are of crucial importance. If listeners use a (compound) target, determining its position inside the vowel might be a non-trivial problem.

7.2.2 Results from the literature

In chapter 6 we have looked at the relevant literature to see whether we could find a reason for the differences between our results (i.e., perceptual-undershoot) and reports from others who found perceptual-overshoot or evidence of dynamic-specification. There is no doubt about the fact that the spectro-temporal structure of vowel segments contains information about their identity (Huang, 1991, 1992; Akagi, 1993; see also chapter 3 and 4). This information can be used to enhance the automatic classification of vowel segments. However, we demonstrated in chapter 5 that human listeners will not use this information unconditionally, as some other studies suggested (e.g., Lindblom and Studdert-Kennedy, 1967; Nearey, 1989).

The condition under which listeners would compensate for target-undershoot in production (i.e., coarticulation or reduction) is not known. However, it seems that the major difference in experimental method between studies that did report this "perceptual compensation" and studies that did not, is the use of complete syllables in contrasting arrangements. We also saw that, in general, vowel segments were identified less well when presented out of context. Together, the above facts suggested that the information in formant dynamics was used only when vowels were heard in an appropriate context. It might even mean that it was the context, and not the formant dynamics, that determined how vowel realizations were iden-

tified, e.g. whether there was some "perceptual compensation" for formant target-undershoot.

7.3 Target-undershoot and vowel perception

In this study we have looked at two aspects of vowel formant dynamics. With respect to vowel production, we tested how formant track shape was influenced by vowel duration. With respect to vowel perception, we examined how the formant track shape affected identification. The underlying question was how well the produced sounds corresponded to the intended vowels. Were vowel sounds produced as intended or were they corrupted by the limitations of the articulatory system? Was dynamic information used to determine vowel identity and did it improve recognition or was it simply ignored or even detrimental to recognition?

The process of articulation has indisputable "mechanical" aspects. The articulators are bodies with a mass, stiffness, and damping. They have to be moved around in synchrony using muscles with limited power. These mechanical aspects will certainly affect articulation and shape the sounds uttered. The simple damped mass-spring model of Lindblom (1983) is just an illustration of this principle. However, to conclude that this mechanical side to articulation dominates vowel production is too one-sided. After all, speaking is a conscious act, and in general, people have very good control over their voluntary actions, especially after some practice. If anything, speaking is practised a lot.

The mechanical aspects of articulation imply that a reduction in duration means either less movement or more force. If a speaker has to complete all articulatory movements in a shorter time, s/he must increase speaking effort. If the force of articulation cannot be controlled at the level of the syllable, a decrease of duration would result in undershoot. The target-undershoot model implicitly states that the force of articulation can only be controlled at the level of sentences or higher, if it can be controlled at all. But if a complex process, like stress, can be applied on individual syllables, it is entirely conceivable that the force of articulation can also be controlled at this level. Our study showed that a speaker can reproduce a long stretch of speech at a different rate consistently. Stress, durational, and formant patterns were quite faithfully replicated. In short, his control over his speech was excellent. When we include all other evidence, we can conclude that it is quite likely that, in general, speakers are able to control vowel undershoot and duration at will and independently.

Still, a strong case can be made for a relation between duration and formant-undershoot, as exemplified by equations 1.1-1.3 of chapter 1. It is clear that in natural speech, a shorter vowel duration will generally occur together with more formant-undershoot. On the other hand, undershoot seems to be under the control of the speaker, i.e. is planned or output-driven. If undershoot is intentional, the question of its function in normal speech is raised.

With the available evidence, two functions for undershoot suggest themselves. As context determines undershoot, the context could in principle be reconstructed from the undershoot. This means that coarticulation would help in identifying consonants. The importance of vowel formant track shape for consonant recognition has been discussed extensively in the literature (to name only a few: Pols and Schouten, 1978; Pols, 1979; Mack and Blumstein, 1983; Polka and Strange, 1985; Miller, 1986; Klatt, 1987; Nossair and Zahorian, 1991). Not surprisingly, we also found that formant track shape influenced the number and identity of the consonants in the responses of our subjects.

In addition to the impact of the immediate context, undershoot is also implicated with the perception of prosody (Rietveld and Koopmans-van Beinum, 1987) and word frequency (e.g., Van Bergem, 1993). Reduction could increase and decrease with the speakers estimation of how well the audience will understand individual words or syllables. In this way, reduction could be used to signify unstressed syllables and high-frequency function words. Listeners could then focus their attention on stressed syllables and low-frequency content words.

We can summarize these two putative functions of target-undershoot by concluding that the identification of context and prosodic structures is facilitated by coarticulation and reduction. In other words, the prominence of vowels is actively manipulated, and vowel intelligibility is sacrificed, to enhance syllable and word intelligibility.

The results of our experiments on vowel perception indicated that information, relevant to the compensation for the effects of context (i.e., formant track shape), was not used unconditionally to support vowel identification, at least not in the context we used. An evaluation of the existing literature showed that the results, as published, did suggest a crucial role for the syllabic or word context in vowel recognition (see chapter 6). This would mean that the information present in the vowel segment itself would only be used properly if the segment is heard as part of an appropriate syllable or word. So, we have seen first that vowel realizations were changed to fit in particular syllables when uttered. Now we have seen that when they have to be recognized, the whole syllable or word might help to identify them.

At present, target and dynamical models of vowel perception highlight different aspects of the process of vowel recognition. But they concentrate completely on information from the vowel segment itself. Now there are strong indications that listeners might also use the context (syllables or words) when trying to identify individual vowel segments. For a better understanding of vowel perception, this syllabic and word context should be taken into account.

7.4 Conclusions

We can summarize the preceding discussion by saying that:

- For our speaker, speaking rate, and therefore duration *an sich*, did not influence vowel formant-undershoot or (time-normalized) track shape.

- Our listeners did not use perceptual-overshoot or dynamic-specification in identifying synthetic vowel tokens. Neither did they use the vowel mid-point.

This lead us to conclude that the amount of vowel formant-undershoot is planned by the speaker. Listeners do not automatically compensate for this undershoot at the level of the individual vowel token.

7.5 Suggestions for future research

In this thesis we concluded that vowel target-undershoot, i.e. coarticulation and reduction, is largely planned or output-driven. It could be that the function of coarticulation in speech is different from that of reduction. Studies of vowel articulation generally concentrate on either coarticulation or vowel reduction. Few studies address the relation between these two phenomena. As a result, it is not known how coarticulation and reduction interact. Some studies suggest that they might be different aspects of the same process, e.g. vowel reduction could be a measure of the average amount of coarticulation. A quantitative study of the relation of the contrast between vowels (i.e., reduction) and the amount of formant-undershoot due to coarticulation should resolve this issue.

Vowel articulation is influenced by context, prosody, and speaking style. The effects of prosody and speaking style on vowel realizations are generally referred to as vowel reduction. In this thesis we only studied vowel realizations. In a future project we will investigate whether the spectro-temporal features of consonant realizations change under the influence of prosody and speaking style in ways that could be described as "consonant reduction" (Van Son and Pols, 1993).

The possibility that it is the context that induces compensation in perception could be checked by presenting synthetic vowels like those used in chapter 5 with and without a convincing context of other vowels, i.e. inside three-vowel or vowel-glide sequences. As vowel-vowel sequences are strongly coarticulated in natural speech and are easy to synthesize, it must be possible to decide whether it is the context or the formant movements that induce "perceptual-overshoot" in the listener. Preparations for such an experiment are currently under way at our institute.

References

Abramowitz, M. & Stegun, I.A. (**1965**). *Handbook of mathematical functions* (Dover Publications, Inc., New York NY, 9th printing).

Akagi, M. (**1990**). "Evaluation of a spectrum target prediction model in speech perception", *Journal of the Acoustical Society of America* **87**, 858-865.

Akagi, M. (**1992**). "Psychoacoustic evidence for contextual effect models" in *Speech perception, production and linguistic structure*, edited by Y.Tohkura, E.Vatikiotis-Bateson & Y.Sagisaka (Ohmsha, Tokyo; IOS Press, Amsterdam), 63-78.

Akagi, M. (**1993**). "Modeling of contextual effects based on spectral peak interaction", *Journal of the Acoustical Society of America* **93**, 1076-1086.

André-Obrecht, R. (**1988**). "A new statistical approach for the automatic segmentation of continuous speech signals", *IEEE Transactions on Acoustics Speech and Signal Processing* **36**, 29-40.

Andruski, J.E. & Nearey, T.M. (**1992**). "On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables", *Journal of the Acoustical Society of America* **91**, 390-410.

Benguerel, A-P & McFadden, T.U. (**1989**). "The effect of coarticulation on the role of transitions in vowel perception", *Phonetica* **46**, 880-896.

Brady, P.T., House, A.S., Stevens, K.N. (**1961**). "Perception of sounds characterized by a rapidly changing resonant frequency", *Journal of the Acoustical Society of America* **33**, 1357-1362.

Broad, D.J. & Clermont, F. (**1987**). "A methodology for modelling vowel formant contours in CVC context", *Journal of the Acoustical Society of America* **81**, 155-165.

Broad, D.J. & Fertig, R.H. (**1970**). "Formant-frequency trajectories in selected CVC-syllable nuclei", *Journal of the Acoustical Society of America* **47**, 1572-1582.

Churchhouse, R.F. (editor) (**1981**). *Handbook of applicable mathematics*, *vol. III: Numerical methods* (John Wiley & Sons), 194-201.

Clark, J. & Yallop, C. (**1990**). *An Introduction to phonetics and phonology* (Basil Blackwell, Oxford, UK), 116-151.

Delattre, P. (**1969**). "An acoustic and articulatory study of vowel reduction in four languages", *International Review of Applied Linguistics and Language Teaching (IRAL)* **VII**, 295-325.

Den Os, E.A. (**1988**). "Rhythm and tempo of Dutch and Italian; a contrastive study" Ph.D. Thesis, University of Utrecht, The Netherlands.

Di Benedetto, M.G. (**1989a**). "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency", *Journal of the Acoustical Society of America* **86**, 55-66.

Di Benedetto, M.G. (**1989b**). "Frequency and time variations of the first formant: Properties relevant to the perception of vowel height", *Journal of the Acoustical Society of America* **86**, 67-77.

Diehl, R.L. & Walsh, M.A. (**1989**). "An auditory basis for the stimulus-length effect in the perception of stops and glides", *Journal of the Acoustical Society of America* **85**, 2154-2164.

Duez, D. (**1989**). "Second formant locus-nucleus patterns in spontaneous speech: some preliminary results on French", *Phonetic Experimental Research Institute of Linguistics University of Stockholm (PERILUS)* **X**, 109-114.

Eefting, W. (**1991**). "The effect of information value and accentuation on the duration of Dutch words, syllables and segments", *Journal of the Acoustical Society of America* **89**, 412-424.

Engstrand, O. (**1988**). "Articulatory correlates of stress and speaking rate in Swedish VCV utterances", *Journal of the Acoustical Society of America* **85**, 1863-1875.

Fant, G. & Kruckenberg, A. (**1989**). "Prelimenaries to the study of Swedish prose reading and reading style", *KTH Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)* **2**/1989, 1-83.

Ferguson, G.A. (**1981**). *Statistical analysis in psychology and education. International student edition* (McGraw-Hill, 2nd printing), 381-406.

Flege, J.E. (**1988**). "Effects of speaking rate on tongue position and velocity of movement in vowel production", *Journal of the Acoustical Society of America* **84**, 901-916.

Fourakis, M. (**1991**). "Tempo, stress, and vowel reduction in American English", *Journal of the Acoustical Society of America* **90**, 1816-1827.

Fox, R.A. (**1989**). "Dynamic information in the identification and discrimination of vowels", *Phonetica* **46**, 97-116.

Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements", *Journal of the Acoustical Society of America* **63**, 223-230.

Gay, T. (**1981**). "Mechanisms in the control of speech rate", *Phonetica* **38**, 148-158.

Gay, T., Ushijima, T., Hirose, H. & Cooper, F.S. (**1974**). "Effect of speaking rate on labial consonant-vowel articulation", *Journal of Phonetics* **2**, 47-63.

Gopal, H.S. & Syrdal, A.K. (**1988**). "Effects of speaking rate on temporal and spectral characteristics of American English vowels", *Speech Communications Group Working Papers* **VI**, Research Laboratory of Electronics MIT, 162-180.

Gottfried, T.L. & Strange, W. (**1980**). "Identification of coarticulated vowels", *Journal of the Acoustical Society of America* **68**, 1626-1635.

Huang, C.B. (**1991**). "An acoustic and perceptual study of vowel formant trajectories in American English", Ph.D. Thesis, Massachusetts Institute of Technology, USA (Research Laboratories of Electronics, Technical report no. 563, Cambridge, MA 02139).

Huang, C.B. (**1992**). "Modelling human vowel identification using aspects of formant trajectory and context" in *Speech perception, production and linguistic structure*, edited by Y.Tohkura, E.Vatikiotis-Bateson & Y.Sagisaka (Ohmsha, Tokyo; IOS Press, Amsterdam), 43-61.

Kerkhoff, J., Loman, H. & Boves, L. (**1986**). "Fonpars2: A compiler for the implementation of phonetic rules", *Proceedings Institute of Phonetics, Catholic University of Nijmegen*. **10**, 75-80.

Klatt, D.H. (**1980**). "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America* **67**, 971-995

Klatt, D.H. (**1987**). "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America* **82**, 737-793.

Koopmans-van Beinum, F.J. (**1980**). "Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions", Ph.D. Thesis, University of Amsterdam, The Netherlands.

Koopmans-van Beinum, F.J. (**1990**). "Spectro-temporal reduction and expansion in spontaneous speech and read text: the role of focus words", *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* 90 Vol. **1**, 21-24.

Koopmans-van Beinum, F.J. (**1992**). "What's in a schwa?", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **16**, 53-64.

Krull, D. (**1989**). "Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech", *Phonetic Experimental Research Institute of Linguistics University of Stockholm (PERILUS)* **X**, 87-108.

Kuehn, D.P. & Moll, K.L. (**1976**). "A cineradiographic study of VC and CV articulatory velocities", *Journal of Phonetics* **4**, 303-320.

Kuwabara, H. (**1985**). "An approach to normalization of coarticulation effects for vowels in connected speech", *Journal of the Acoustical Society of America* **77**, 686-694.

Lindblom, B. (**1963**). "Spectrographic study of vowel reduction", *Journal of the Acoustical Society of America* **35**, 1773-1781.

Lindblom, B. (**1983**). "Economy of speech gestures" in *The production of speech*, edited by P.F.MacNeilage (Springer-verlag, New York N.Y.), 217-246.

Lindblom, B. & Moon, S.-J. (**1988**). "Formant undershoot in clear and citation-form speech", *Phonetic Experimental Research Institute of Linguistics University of Stockholm (PERILUS)*. **VIII**, 21-33.

Lindblom, B. & Studdert-Kennedy, M. (**1967**). "On the role of formant transitions in vowel recognition", *Journal of the Acoustical Society of America* **42**, 830-843.

Lisker, L. (**1984**). "On reconciling monophthongal vowel percepts and continuously varying F patterns", *Haskins Laboratories: Status Report on Speech Research* ***SR-79/80***, 167-174.

Macchi, M.J. (**1980**). "Identification of vowels spoken in isolation versus vowels spoken in consonantal context", *Journal of the Acoustical Society of America* **68**, 1636-1642.

Mack, M. & Blumstein, S.E. (**1983**). "Further evidence of acoustic invariance in speech production: The stop-glide contrast", *Journal of the Acoustical Society of America* **73**, 1739-1750.

Mann, V. & Soli, S.D. (**1991**). "Perceptual order and the effect of vocalic context on fricative perception", *Perception and Psychophysics* **49**, 399-411.

Miller, J.D. (**1989**). "Auditory-perceptual interpretation of the vowel", *Journal of the Acoustical Society of America* **85**, 2114-2134.

Miller, J.L. (**1981a**). "Effects of speaking rate on segmental distinctions", in *Perspectives on the study of speech*, edited by P.D.Eimas & J.L.Miller (Lawrence Erlbaum, Hillsdale NJ), 39-74.

Miller, J.L. (**1981b**). "Some effects of speaking rate on phonetic perception", *Phonetica* **38**, 159-180.

Miller, J.L. (**1986**). "Limits on later-occuring rate information for phonetic perception", *Language and Speech* **29**, 13-24.

Miller, J.L. & Baer, T. (**1983**). "Some effects of speaking rate on the production of /b/ and /w/", *Journal of the Acoustical Society of America* **73**, 1751-1755.

Moon, S.-J. (**1990**). An acoustic and perceptual study of formant undershoot in clear- and citation-form speech", *Journal of the Acoustical Society of America*, Supplement 1 **88**, **S**129 (A).

Nearey, T.M. (**1989**). "Static, dynamic, and relational properties in vowel perception", *Journal of the Acoustical Society of America* **85**, 2088-2113.

Nord, L. (**1987**). "Vowel reduction in Swedish", in *Papers from the Swedish Phonetics Conference*, edited by O.Engstrand (Department of Linguistics of the University of Uppsala, Uppsala, Sweden), 16-21.

Nossair, Z.B. & Zahorian, S.A. (**1991**). "Dynamic spectral shape features as acoustic correlates for initial stop consonants", *Journal of the Acoustical Society of America* **89**, 2978-2991.

Öhman, S.E.G. (**1966**). "Coarticulation in VCV utterances: Spectrographic measurements", *Journal of the Acoustical Society of America* **39**, 151-168.

O'Shaughnessy, D. (**1987**). *Speech Communication* (Addison-Wesley, Reading, MA).

Peeters, W.J.M. (**1991**). "Diphthong dynamics", Ph.D. Thesis, State University of Utrecht, the Netherlands.

Polka, L. & Strange, W. (**1985**). "Perceptual equivalence of acoustic cues that differentiate /r/ and /l/", *Journal of the Acoustical Society of America* **78**, 1187-1197.

Pols, L.C.W. (**1977**). "Spectral analysis and identification of Dutch vowels in monosyllabic words", Ph.D. Thesis, Free University of Amsterdam, The Netherlands.

Pols, L.C.W. (**1979**). "Coarticulation and the identification of initial and final plosives", in *ASA*50 Speech Communication Papers*, edited by J.J.Wolf & D.H.Klatt (Acoustic Society of America), 459-462.

Pols, L.C.W. & Schouten, M.E.H. (**1978**). "Identification of deleted consonants", *Journal of the Acoustical Society of America* **64**, 1333-1337.

Pols, L.C.W. & Van Son, R.J.J.H. (**1993**). "Acoustics and perception of dynamic vowel segments", accepted for publication in *Speech Communication*.

Repp, B.H. (**1993**). Review of Tohkura et al., 1992, *Language and Speech* **36**, 99-107.

Rietveld, A.C.M. & Koopmans-van Beinum, F.J. (**1987**). "Vowel reduction and stress", *Speech Communication* **6**, 217-229.

Schouten, M.E.H. & Pols, L.C.W. (**1979**). "Vowel segments in consonantal contexts: a spectral study of coarticulation-Part I", *Journal of Phonetics* **7**, 1-23.

Schulman, R. (**1989**). "Articulatory dynamics of loud and normal speech", *Journal of the Acoustical Society of America* **85**, 295-312.

Smits, R. (submitted). "Accuracy of quasi-stationay analysis of highly dynamic speech signals". submitted to *Journal of the Acoustical Society of America* .

Stevens, K.N. & House, A.S. (**1963**). "Pertubation of vowel articulations by consonantal context: an acoustical study". *Journal of Speech and Hearing Research*, **6**, 111-128.

Strange, W. (**1989a**). "Evolving theories of vowel perception", *Journal of the Acoustical Society of America* **85**, 2081-2087.

Strange, W. (**1989b**). "Dynamic specification of coarticulated vowels spoken in sentence context", *Journal of the Acoustical Society of America* **85**, 2135-2153.

Strange, W. & Gottfried, T.L. (**1980**). "Task variables in the study of vowel perception", *Journal of the Acoustical Society of America* **68**, 1622-1625.

Strange, W., Jenkins, J.J. & Johnson, T.L. (**1983**). "Dynamic specification of coarticulated vowels", *Journal of the Acoustical Society of America* **74**, 695-705.

Strange, W., Verbrugge, R.R., Schankweiler, D.P. & Edman, T.R. (**1976**). "Consonant environment specifies vowel identity", *Journal of the Acoustical Society of America* **60**, 213-224.

Tohkura, Y., Vatikiotis-Bateson, E. & Sagisaka Y. (editors) (**1992**). *Speech perception, production and linguistic structure* (Ohmsha, Tokyo; IOS Press, Amsterdam), 463 pp.

Traunmüller, H. (**1988**). "Paralinguistic variation and invariance in the characteristic frequencies of vowels", *Phonetica* **45**, 1-29.

Vaissiere, J. (**1987**). "The use of allophonic variations of /a/ in automatic continuous speech recognition in French", *Research Laboratory of Electronics, Massachusetts Institute of Technology, Speech Communication Group Working Papers* **V**, 15-25.

Van Bergem, D.R. (**1988**). "The first step to a better understanding of vowel reduction", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **12**, 61-75.

Van Bergem, D.R. (**1993**). "Acoustic vowel reduction as a function of sentence accent, word stress, and word class", *Speech Communication* **12**, 1-23.

Van der Kamp, L.J.Th. & Pols, L.C.W. (**1971**). "Perceptual analysis from confusions between vowels", *Acta Psychologica* **35**, 64-77.

Van Son, R.J.J.H. (**1987**). "Automatic slope measurements on formant tracks", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **11**, 67-78.

Van Son, R.J.J.H. & Pols, L.C.W. (**1989**). "Comparing formant movements in fast and normal rate speech", in *Eurospeech 89, the European Conference on Speech Communication and Technology*, Paris, edited by J.P.Tubach & J.J.Mariani (CEP Consultants, Edinburgh, UK), Vol.**2**, 665-668.

Van Son, R.J.J.H. & Pols, L.C.W. (**1990**). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", *Journal of the Acoustical Society of America* **88**, 1683-1693.

Van Son, R.J.J.H. & Pols, L.C.W. (**1991a**). "The influence of speaking rate on vowel formant track shape as modelled by Legendre polynomials". *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **15**, 43-59.

Van Son, R.J.J.H. & Pols, L.C.W. (**1991b**). "The influence of formant track shape on the perception of synthetic vowels". *Eurospeech 91, the European Conference on Speech Communication and Technology*, Genua, Vol. **3**, 1117-1120.

Van Son, R.J.J.H. & Pols, L.C.W. (**1992**). "Formant movements of Dutch vowels in a text, read at normal and fast rate", *Journal of the Acoustical Society of America* **92**, 121-127.

Van Son, R.J.J.H. & Pols, L.C.W. (**1993**). "Eindelijk wat meer aandacht voor medeklinkers in spraak", NWO projekt no. 300-173-029.

Verbrugge, R.R. & Rakerd, B. (**1986**). "Evidence of talker-independent information for vowels", *Language and Speech* **29**, 39-57.

Vogten, L.L.M. (**1986**). "LVS speech processing programs on IPO-VAX 11/780", Manual **67**, Institute for Perception Research, Eindhoven, The Netherlands.

Weismer, G., Kent, R.D., Hodge, M. & Martin, R. (**1988**). "The acoustic signature for intelligibility test words", *Journal of the Acoustical Society of America* **84**, 1281-1291.

Whalen, D.H. (**1990**). "Coarticulation is largely planned", *Journal of Phonetics* **18**, 3-35.

Willems, L.F. (**1986**). "Robust formant analysis", *Annual progress report* **21**, Institute for Perception Research, Eindhoven, The Netherlands, 34-40.

APPENDIX A:

AUTOMATIC SLOPE MEASUREMENT ON FORMANT TRACKS*

This appendix describes the theory behind the peak-picker which was used in chapter 2 to determine the maximal and minimal values of formant tracks. The peak-picker was based on an automatic segmentation algorithm that dissects formant tracks into near-linear segments. Peaks and troughs are the points between segments where the formant slopes switch sign.

---

*Adapted from: Van Son, R.J.J.H. (1987). "Automatic slope measurements on formant tracks", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 11, 67-78.

Introduction

In chapter 2, we used a peak-picker to determine the point of extreme F_1 or F_2 values. This peak-picker is based on an automatic segmentation algorithm to measure formant track slope. This appendix describes that algorithm.

With natural speech it is often very difficult to measure spectral changes in a speech signal. The spectrum of a constant or slowly changing signal can be determined almost to the theoretical limits. The measurements of a rapidly changing signal, however, suffer from a lack of theoretical understanding and comprehensible representation. The central question is which changes are to be measured on a given set of spectra, measured on different points in time (e.g., spectral envelop, band-filters, harmonics). Even the status of a spectrum, measured on a changing signal, is often not clear due to the implicit assumption of stationarity that underlies most spectral representations.

Interest in the spectral changes of speech signals is most often concentrated on the behaviour of spectral peaks. There are several ways to measure and represent spectral peaks. One possibility is to transform the speech waveform into a spectrum, essentially making some type of time representation of bandpass filter outputs or the energy distribution resulting from a Fourier Transforms. The problem is to identify peaks and follow their course in time and frequency. This is no trivial matter because it is difficult to decide what is a peak and what is not and which parts of the spectrum are instances of the same peak at subsequent points in time.

Another possibility is to formulate a model of human speech production and measure the changes in the parameters of this model that affect the spectral contents of the speech signal. This last approach is followed in this paper with the use of Linear Predictive Coding (LPC). This LPC analysis can encode the spectral peaks of the speech signal with a fixed number of variable, second order, bandpass filters. The spectral parameters of interest are the centre frequency and the bandwidth of each peak encoded this way. Here a method will be described for measuring the rate of spectral changes as used to study the course in time of spectral peaks. For this kind of study all spectral peaks have to be defined at all times. In normal LPC analysis, with the Levinson algorithm, sometimes a peak is "lost". To prevent this disturbance, a different algorithm is used here for the LPC analysis, the so called Split-Levinson algorithm. This algorithm was implemented by Willems (1986, pp.34-40).

Choosing an LPC representation has some advantages over approaches that use FFT or banks of fixed band-filters. It is possible to manipulate all parameters of an LPC analysis and still resynthesize recognizable speech. Small changes in the parameters result in small changes in the resynthesized speech. In this way it is possible to test for clues for speech synthesis or speech quality by changing the relevant parameters and, the other way around, to hear whether a change in parameters removes the quality or clue of interest.

The spectral peaks that result from LPC analysis are often called formants. This is because the production model that forms the root of this ap-

proach, can model the effects of resonances in the speech organs quite well, at least in vowels. These resonances are, by definition, formants. The fit between the model and reality is, however, not good enough to ensure a perfect fit between the LPC spectral peaks and the formants. Sometimes there is a discrepancy between the measured peaks and the real formants. Resynthesized speech however, mostly is of acceptable quality. In spite of the imperfect fit, the spectral peaks extracted from an LPC analysis will be called formants hereafter.

A.1 Modeling formant tracks

If the objective of measurement is to determine the spectral change, i.e. the spectral slope, then it is necessary to perform differentiations on the spectral data. Differentiation is an operation that is very sensitive to random measurement errors or noise. It amplifies those errors and noise in such a way that even small, local errors can completely corrupt slope measurements. To deal with this phenomenon it is necessary to remove, at least part of, the noise from the data. To successfully separate the desired signal and the noise, it is necessary to develop a model of the signal and the noise. If there would have been a model for speech production available from which accurate estimations of the course of formants in natural speech could be obtained, the problem could be solved without major problems. But since such a model is not available yet, it is necessary to develop an accurate description of the signal without much reference to production.

It is very often possible to approximate a signal of unknown composition, a posteriori, to any desired accuracy by constructing a sum of model functions. The remaining discrepancy between data and description is treated as noise and removed, only the modeled part is kept. It is important to choose the right class of functions to model the signal. An inappropriate model function will lead to a disturbed signal. Functions that can be made orthogonal are to be preferred.

Choosing functions for modeling is always a guess. The guess made here is that an LPC formant track, f(t), on a given interval $[t_0,t_1>$ can be modeled a posteriori with any desired precision with a polynomial function that has the form:

$$\begin{aligned} f(t) &= a_0 + a_1 \cdot t + a_2 \cdot t^2 + ... \\ &= \sum_{j=0}^{\bullet} a_j \cdot t^j \\ &= H^{\bullet}(t) \end{aligned} \qquad \text{(A.1)}$$

with: $t \quad [t_0,t_1>$

For any given maximal order J of the polynomials the coefficients a_j of

$$H^J(t) = \sum_{j=0}^{J} a_j \cdot t^j \qquad \text{(A.2)}$$

are chosen such that $H^J(t)$ is the best approximation of f(t) for this order of polynomials on this interval. It is possible to rearrange the terms of equation A.2 in such a way that $H^J(t) = b_J h_J(t) + H^{J-1}(t)$, i.e. the best fitting polynomial of order J is the sum of the best fitting polynomial of order J-1 and some order specific polynomial $h_J(t)$. The $h_J(t)$ form a set of orthogonal polynomials. Using a set of orthogonal functions to describe a function f(t) has great methodological and computational advantages, especially if J • 2. A short description of one such set of orthogonal polynomials, the shifted Legendre polynomials, is given in Appendix B.

After the calculations of $H^J(t)$ the original formant track is replaced by

$$f(t) = H^J(t) + \varepsilon(t) \qquad \text{(A.3)}$$

in which $\varepsilon(t)$ is an error term. For high orders of J it will be difficult to determine (a posteriori) the best intervals $[t_i,t_{i+1}>$ of f(t) to fit $H^J(t)$ on. The order of the model function should therefore be as low as possible. For measuring formant slopes (=velocity) an order of 1 will do, for measuring formant acceleration an order of 2 is necessary. In the discussion below an order of 1 will suffice, the order indication of the model functions $H^1(t)$ will be omitted hereafter.

For this first order polynomial model to make a good fit it is important to choose appropriate intervals. The formant track is modeled as a succession of simple straight line segments. If the boundaries between successive line segments are chosen wrongly, the resulting modeled track will hardly have any resemblance to the originally measured formant track. In this model therefore the original formant track f(t) is divided into intervals $T_i = [t_i,t_{i+1}>$ that do not overlap. In every interval T_i the formant is modeled with:

$$f(t) = H_i(t) + \varepsilon(t) = a_i \cdot t + b_i + \varepsilon_i(t) \qquad \text{(A.4)}$$

t $T_i = [t_i,t_{i+1}>$
$H_i(t)$: a straight line on T_i
$\varepsilon_i(t)$: the error term on T_i, defined by $\varepsilon_i(t) = f(t) - H_i(t)$

Next $\varepsilon_i(t)$ can be modeled by a noise term $e_i(t)$ with a Gaussian distribution with expected value $E(e_i(t)) = 0$ and variance $E(e_i(t)^2) = \sigma_i^2$. $H_i(t)$ becomes the straight line that minimizes σ_i^2. In this model the value of the formant at time t T_i is $H_i(t)$ and the slope is a_i.

The assumption that $\varepsilon_i(t)$ can be modeled by a Gaussian distributed noise term is made for convenience. It is possible to use other distributions but calculating the best fit becomes time consuming and for the simple example described here there is no point in using any other distribution. The minimizing criterion for the best fitting function can be altered to emphasize the errors in special parts of the interval, e.g. the centre of the interval, by using a weighting function.

The preceding argument can be summarized as follows:

With LPC analysis it is possible to extract formant frequencies from a speech signal. These formant frequencies form tracks in time. Each of these tracks, represented by the function f(t), can be modeled by dividing the track in non-overlapping intervals T_i and replacing the measured track f(t) with:

$$f(t) \spadesuit H_i(t) + e(t) = a_i \cdot t + b_i + e_i(t) \qquad \text{(A.5)}$$

$t \quad T_i = [t_i, t_{i+1}>$
$H_i(t)$: a straight line on T_i
$e_i(t)$: a Gaussian noise term on T_i, defined by
$E(e_i(t)) = 0$
$E(e_i(t)^2) = \sigma_i^2$ (i.e., independent of t)

In equation A.5 the best guess for $H_i(t)$ is the linear regression line on T_i.

A.2 Segmentation

In the above model, segmenting the tracks in independent intervals is crucial for a good fit of the model on the tracks. Such intervals are called line segments here. A line segment is defined as the largest interval in which the formant track can be modeled by a straight line according to equation A.5. The segmentation can be done in an automatic way if there is a smallest interval length τ for which there is no smaller line segment. If there is such a smallest length of a line segment, then it is possible to find all the boundaries between the segments. This is done by deciding whether a test segment of the track (called Δ_0) with a length smaller than or equal to the smallest interval length (i.e., $|\Delta_0| \leq \tau$) contains a boundary between line segments. If it is concluded that the test segment does contain a boundary between line segments, then the best point to place this boundary can be found. This test segment is shifted over the track until all possible boundaries are found.

The decision whether or not the test segment contains a boundary between line segments is made by trying to find a subdivision of Δ_0 in two sub-segments Δ_1 and Δ_2 that have a lower expected value for the remaining variance of their regression lines (called $E(v_1^2)$ and $E(v_2^2)$) than the undivided test segment (called $E(v_0^2)$). If there is no boundary present in Δ_0, that is, Δ_0 is completely confined in a segment (T_i) of the track with only one straight line segment, then all subdivisions of Δ_0 will have the same expected values for the remaining variance of their regression lines as Δ_0 itself. Or, for all subdivisions Δ_1 and Δ_2 of Δ_0 lying in segment T_i :

$$E(v_0^2) = E(v_1^2) = E(v_2^2) = \sigma_i^2 \qquad (A.6)$$

with:

$E(v_0^2)$, $E(v_1^2)$, $E(v_2^2)$: the expected values of the remaining variance of the regression lines in the segments Δ_0, Δ_1 and Δ_2

v_0^2, v_1^2 and v_2^2: the estimated or calculated values of the remaining variance of the regression lines in the segments Δ_0, Δ_1 and Δ_2

σ_i^2 : the variance of the model noise term in segment T_i (cf. equation A.5)

If, however, the test segment Δ_0 contains a boundary between two segments, T_i and T_{i+1}, with different model lines (not only different noise terms), then there exists at least one subdivision of Δ_0 in two segments Δ_1 and Δ_2 that has a lower expected value of the remaining variance than the test segment itself. Or

$$|\Delta_0| \cdot E(v_0^2) > |\Delta_1| \cdot E(v_1^2) + |\Delta_2| \cdot E(v_2^2) \qquad (A.7)$$

with: $|\Delta_0| = |\Delta_1| + |\Delta_2|$ the lengths of the segments

The subdivision with the lowest remaining variance, $|\Delta_1| \cdot E(v_1^2) + |\Delta_2| \cdot E(v_2^2)$, has expected values of the remaining variance of the regression lines that are equal to the variances of the noise terms in T_i and T_{i+1}. That is:

$$E(v_1^2) = \sigma_i^2 \qquad (A.8)$$
$$E(v_2^2) = \sigma_{i+1}^2$$

and

$$|\Delta_1| \cdot E(v_1^2) + |\Delta_2| \cdot E(v_2^2) = |\Delta_1| \cdot \sigma_i^2 + |\Delta_2| \cdot \sigma_{i+1}^2$$

These equations are valid for a continuous formant track and then for one for which all parameters are known a priori (we used espectation values). If track parameters have to be estimated from a limited number of measuring points, then equation A.7 will become:

$$n_0 \cdot v_0^2 > n_1 \cdot v_1^2 + n_2 \cdot v_2^2 \qquad (A.9)$$

with: $n_0 = n_1 + n_2$ the number of measured points in the segments Δ_0, Δ_1 and Δ_2

If a subdivision is found for which this inequality holds, then there is a segment boundary in Δ_0. The best guess for the position of this boundary is the point that separates the sub-segments Δ_1 and Δ_2 with the lowest value of $n_1 \cdot v_1^2 + n_2 \cdot v_2^2$. If this value is not equal to zero then take this subdivision and rewrite equation A.9 to:

$$\varphi^2 = \{ (n_1+n_2) \cdot v_0^2 - (n_1 \cdot v_1^2 + n_2 \cdot v_2^2) \} / \{n_1 \cdot v_1^2 + n_2 \cdot v_2^2\} > 0 \qquad (A.10)$$

φ^2 Is the largest value possible for the quotient on this test segment (see figure A.1). If both sides of equation A.9 are equal to zero, there is no boundary in the test segment. If only the right hand side of equation A.9 is equal to zero, then there is a boundary in the test segment. Because of definition and the fact that v_0^2 is calculated from the same points as

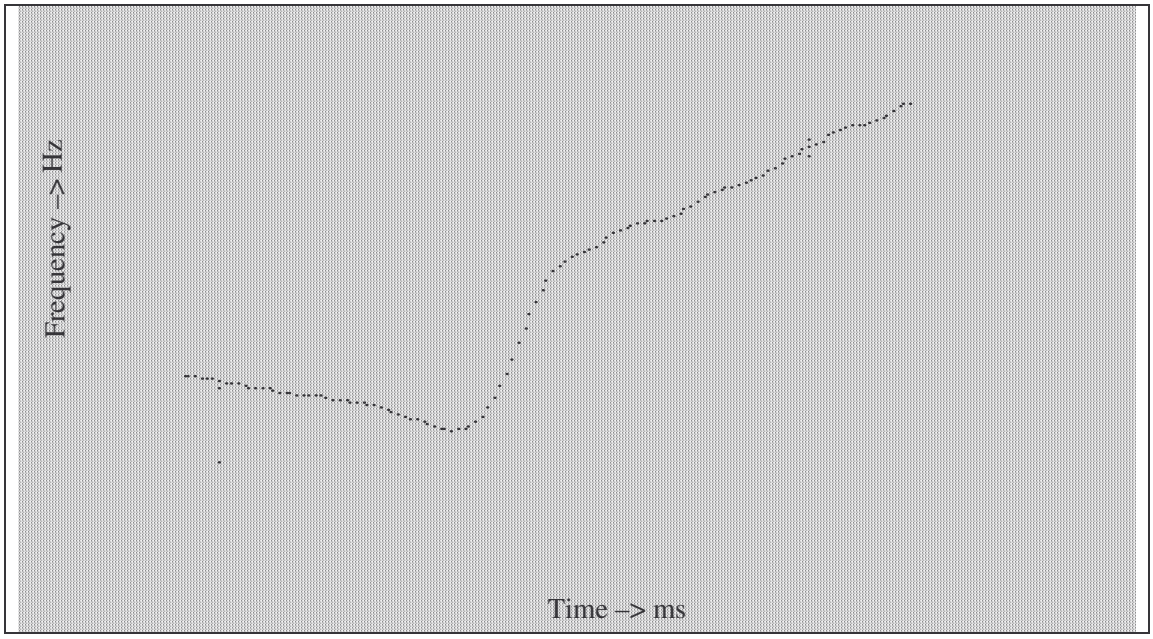

Figure A.1: An example of a formant track f(t) and the calculated values of the parameters of equation A.10 on a test segment Δ_0 (see text for explanation). The division used is indicated by a dashed line and is the one with the highest value of φ^2. Δ_0 is a test segment with $n_1+n_2=100$ points. Δ_1 and Δ_2 are the two neighbouring sub-segments of Δ_0, each containing $n_1=n_2=50$ points. H_0, H_1 and H_2 are the regression lines on these three segments. It can be seen that the test segment, Δ_0, is chosen too large. Three line segments are actually present inside the test segment Δ_0, which results in a total of two boundaries. But inside a test segment only one boundary between line segments can be found with the method described here. As is shown in this figure.

$n_1 \cdot v_1^2 + n_2 \cdot v_2^2$ is calculated, φ cannot be smaller than zero. It is however easily seen that $\varphi > 0$ is possible with no boundary present.This erroneous boundary detection results from stochastic errors in the estimators v_0^2 , v_1^2 and v_2^2. For this reason equation A.10 should be changed to:

$$\varphi^2 > \delta^2 \qquad \text{(A.11)}$$

for detection of a boundary. δ Is a dimensionless number which gives a threshold for detection in numbers of standard deviations difference between $(n_1+n_2) \cdot v_0^2$ and the smallest possible $n_1 \cdot v_1^2 + n_2 \cdot v_2^2$ value in the test segment.

Because there are different numbers of points involved in calculating the different estimated variances, it is important to use unbiased estimators. Here the following unbiased estimators are used:

$$v_0^2 = \{ \sum_{i=1}^{n1+n2} (f(t_i) - H_0(t_i))^2 \} / \{n_1+n_2-2\} \qquad \text{(A.12)}$$

and

$$v_{12}^2 = \{ \sum_{i=1}^{n1} (f(t_i) - H_1(t_i))^2 + \sum_{j=1}^{n2} (f(t_j) - H_2(t_j))^2 \} / \{ n_1+n_2-4\} \qquad (A.12')$$

with: $v_{12}^2 = \{ n_1 \cdot v_1^2 + n_2 \cdot v_2^2 \} / \{ n_1+n_2 \}$

$t_i \quad \Delta_1$

$t_j \quad \Delta_2$

$H_0(t)$, $H_1(t)$ and $H_2(t)$ the regression lines in the segments Δ_0, Δ_1 and Δ_2

In this notation φ^2 will become:

$$\varphi^2 = \{ v_0^2 - v_{12}^2 \} / v_{12}^2 > \delta^2 \qquad (A.13)$$

for boundary detection.

Two assumptions are critical for the fit of the model track to the formant track. First there is no more than one segment boundary in any part of the track with a length $\leq \tau$, with τ being defined as some minimal length greater than or equal to the length of the test segment. Second the formant tracks consist of straight line segments with additive Gaussian noise. If the first assumption does not hold and a test segment contains two or more segment boundaries, then the behaviour of φ^2 will become dependent on where the boundaries are inside the test segment. The detection and assignment of boundaries between line segments becomes very erratic. If the second assumption does not hold and the formant tracks are curved, then boundaries will be placed in such a way that the regression lines will fit the curve with more or less constant variance.

In an actual implementation of the described boundary detector one shifts the test segment one point at a time and accepts only subdivisions with lowest v_{12}^2 which divide the test segment in two parts of equal length. This secures the use of the most accurate estimation of v_{12}^2 for boundary detection. Every boundary is shifted in the centre of the test segment only once and so can be detected only once.

To calculate two regression lines in a test segment, this segment must contain at least 6 points (three points for each regression line). This constraint determines the minimal time resolution needed for the formant measurements.

A.3 Segmentation of several tracks simultaneously

If more than one formant track is used simultaneously to detect synchronous segment boundaries, a total v_0^2 and a total v_{12}^2 are calculated by summing the individual v_0^2 values for all tracks and by summing the individual v_{12}^2 values for every subdivision of the test segment for all tracks. Equation A.13 for boundary detection will not change, but total values will be used for the estimated variances instead of individual values. This is the equivalent of treating the frequency values of different tracks as independent dimensions and stating that each segment contains a multidimensional straight line.

A.4 Other parameters for detecting boundaries

The method to detect boundaries in formant tracks described above is purely statistical. It is possible to use other clues to find those segment boundaries. For example, a change in the voicing of speech (voiced to unvoiced or the reverse) signifies an important change in speech that is likely to have an important effect on formant tracks. It is also possible to use threshold values for the energy of the speech signal or other threshold values to find important changes in the signal. Of all possible parameters that could be used to detect segment boundaries, only the voicing transition is currently implemented, complementary to the formant tracks themselves, of course.

A.5 Comparing straight lines

After the segmentation, the formant track is divided into a large number of segments. The regression lines of many of these segments will not differ markedly from that of their neighbours. It is possible to remove a considerable number of those segment boundaries and merge segments by comparing the regression lines of neighbouring segments.

Comparing straight lines is done by calculating a distance between lines in a shared interval. The distance of the straight lines in two neighbouring segments T_i and T_{i+1} is defined here as the Root Mean Square difference between the two lines in the total interval ($T_i \approx T_{i+1}$). The difference between the lines is measured perpendicular to some standard line. This standard line can be the time axis, a regression line through the combined interval, the bisector line that divides the arc between the lines evenly in two, or it can be some other line. Using the bisector line as the standard line for distance measurement results in the shortest distance between lines and is currently implemented.

The distance between two lines is calculated as follows. Define two straight lines (see figure A.2):

$$g(t) = a \cdot t + b$$
$$h(t) = c \cdot t + d$$

The distance between these two lines is defined here in the interval [0,T >. Any other interval can be transformed to this interval easily. The distance is defined perpendicular to the bisector line. The bisector line between g(t) and h(t), i.e. the line that divides the angle γ into two equal halves, is calculated as follows.
Define the bisector line as:

$$b(t) = e \cdot t + f$$

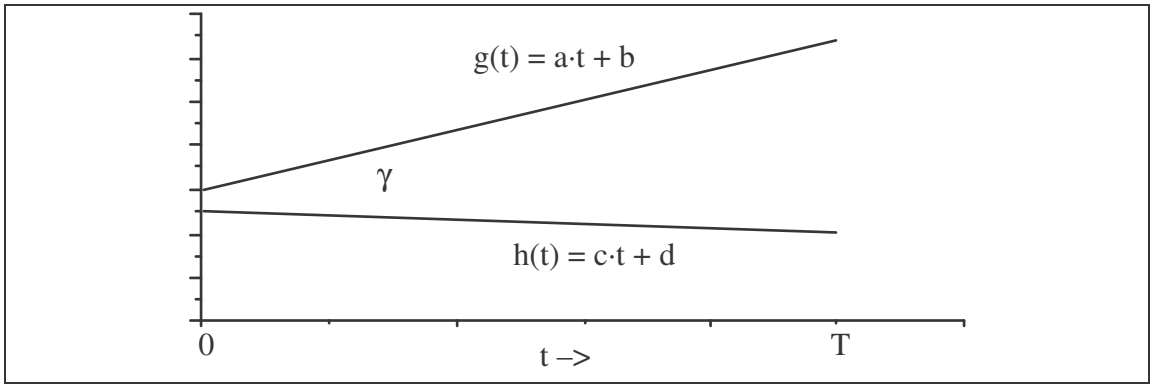

Figure A.2: Two straight lines, g(t) and h(t), with an angle of γ in between.

The angle between g(t) and h(t) is called γ and is:

$$\gamma = \text{arctangent} ((a-c) / (1+a \cdot c))$$

define:

$$\Gamma = \text{tangent} (\gamma / 2)$$

then the parameters of the bisector line become:

$$e = (c+\Gamma) / (1-c \cdot \Gamma)$$
$$f = \{ (1+c \cdot e) \cdot b + (1+a \cdot e) \cdot d \} / \{ 2 + (c+a) \cdot e \}$$

To calculate the distance perpendicular to b(t) all lines are rotated and translated such that b(t) lies on the horizontal axis. In this reference frame the new lines g'(t) and h'(t) are:

$$g'(t) = a' \cdot t + b'$$
$$h'(t) = c' \cdot t + d'$$

and
$$a' = \{ a-e \} / \{ 1+a \cdot e \}$$
$$b' = \{ b-f \} \cdot \bullet [\{ a'^2+1 \} / \{ a^2+1 \}]$$

$$c' = \{ c-e \} / \{ 1+c \cdot e \}$$
$$d' = \{ d-f \} \cdot \bullet [\{ c'^2+1 \} / \{ c^2+1 \}]$$

The distance D is defined in this reference frame as:

$$D^2 = [\overset{T}{\underset{0}{\bullet}} \{ g'(t)-h'(t) \}^2 \, dt] / T$$

This can be simplified to:

$$D^2 = (a'-c')^2 \cdot T^2 / 3 + (a'-c') \cdot (b'-d') \cdot T + (b'-d')^2$$

The mean distance is D.

The mean line distance, defined as above, depends on the total interval length and tends to infinitely large values if the interval length becomes infinite. So this distance is not a quality of the two lines but of the two lines in an interval and depends on the interval. Long intervals must resemble each other more than short intervals in order to be merged into one interval. This distance can be calculated over several formants simultaneously by treating each formant as an independent dimension and the lines as multidimensional straight lines. The total squared distance is calculated by summing the individual squared distances.

Using the line distance to remove unwanted segment boundaries gives the opportunity to segment with high sensitivity and to remove excess boundaries afterwards. This is important because while the segmentation stage has only a narrow, local, scope, the comparing stage has a scope that can be infinite in principle. A local scope is noise sensitive and error prone.

A.6 Conclusions

An implementation of the theory described above was made on a μVAX II mini-computer. Some minor changes were introduced. First, the condition that there should be no more than ONE segment boundary in the test segment was relaxed. Instead of this strict condition, only a minimal segment length was required. This proved to work well. Second, it appeared that the condition of dividing the test segment into two equal sized sub-segments to signal a segment boundary sufficed to select only few excess boundaries. There was no need for an additional threshold for boundary detection (δ^2 in equation A.11). When a minimal RMS line distance is used to decide whether a boundary separates distinct parts of the formant track, then it is possible to eliminate these excess segment boundaries as well as some others that do not separate distinct parts of the formant track. The above theory was used to implement a peak- and trough-picker. This peak- and trough-picker was used in chapter 2 to determine the optimum point for taking cross-sections through vowel realizations (method formant).

APPENDIX B:

CALCULATING LEGENDRE POLYNOMIAL COEFFICIENTS

Legendre polynomial functions are used in chapter 4 to quantify formant track shape. The definition of these functions and the way the numerical calculation of the Legendre polynomial coefficients was performed, is described in this appendix.

Adapted from: M.Abramowitz, I.A.Stegun, Handbook of mathematical functions, Dover publications 1965[9], National Bureau of Standards 1964[10]. The sections on orthogonal polynomials (pp.774,780,798) and numerical integration (pp.886-887)

B.1 Shifted Legendre polynomials

A Legendre polynomial of order J is a function defined for t [-1,1] or t [0,1] of the form:

$$L_J(t)=\sum_{j=0}^{J} a_{Jj}\cdot t^j$$

The functions defined on t [0,1] are called shifted Legendre polynomials. Shifted Legendre polynomials are orthogonal polynomials. That is, they obey the relation:

$$\int_0^1 L_I(t)\cdot L_J(t)\,dt = \begin{cases} 0 & \text{if } I\neq J \\ h_J \neq 0 & \text{if } I=J \end{cases}$$

and for the Shifted Legendre polynomials: $h_J = 1/\{2\cdot J+1\}$

The first five polynomial functions are (see Abramowitz and Stegun, 1965):

$$L_0(t) = 1$$
$$L_1(t) = 2\cdot t - 1$$
$$L_2(t) = 6\cdot t^2 - 6\cdot t + 1$$
$$L_3(t) = 20\cdot t^3 - 30\cdot t^2 + 12\cdot t - 1$$
$$L_4(t) = 70\cdot t^4 - 140\cdot t^3 + 90\cdot t^2 - 20\cdot t + 1$$

If the interval is t [0,k], i.e. the length of the intervals is not zero, then the first five functions change into:

$$L_0(t) = 1$$
$$L_1(t) = 2\cdot t / k - 1$$
$$L_2(t) = 6\cdot t^2 / k^2 - 6\cdot t / k + 1$$
$$L_3(t) = 20\cdot t^3 / k^3 - 30\cdot t^2 / k^2 + 12\cdot t / k - 1$$
$$L_4(t) = 70\cdot t^4 / k^4 - 140\cdot t^3 / k^3 + 90\cdot t^2 / k^2 - 20\cdot t / k + 1$$

and $h_J = k / \{2\cdot J+1\}$

These functions can be translated to another interval, t' $[k_1,k_2]$, by substituting $t=t'-k_1$ and $k=k_2-k_1$.

Any continuous function, f(t), that is defined and is finite in every point of [0,k] can be approximated by a sum of these polynomials

$$f(t) = \sum_{j=0}^{\bullet} A_j\cdot L_j(t)$$

Because of orthogonallity it is possible to calculate the factors A_j <u>independent</u> of one another with the following relation:

$$A_j = \left[\int_0^k f(t) \cdot L_j(t)\, dt \right] / h_j$$

With this relation it is possible to calculate the factors A_j in a very efficient way.

It is straightforward to calculate the Legendre coefficients from a given polynomial representation and vice versa. For instance, any straight line on the interval [0,k] can be written as:

$$\begin{aligned} f(t) &= a \cdot t + b \\ &= A_0 + A_1 \cdot L_1(t) \end{aligned}$$

$t \in [0,k]$

and: $A_0 = b + a \cdot k / 2 \qquad b = A_0 - A_1$

$A_1 = a \cdot k / 2 \qquad a = 2 \cdot A_1 / k$

B.2 Numerical integration using Newton-Cotes formulas

To calculate the Legendre polynomial coefficients, it suffices to integrate the product of the chosen track with the Legendre polynomial of the appropriate order. In practice, the track is generally only available in a sampled form. Integration then becomes a summation. For convenience we assume that the duration of the interval has been normalized to 1.

$$A_j = \left[\int_0^1 f(t) \cdot L_j(t)\, dt \right] / h_j$$

$$\approx \left[\Re \sum_{n=1}^{N} \{ f(n) \cdot L_j((n-1)/N) \} \right] / (N \cdot h_j)$$

The summation will only approximate the integration when N is large compared to the order of the Legendre polynomial j. When N is relatively small, the summation does not represent the integration properly. For small N, a good approximation of the integration can be obtained by using special formulas for numerical integration that, in a way, first interpolate $f(n) \cdot L_j((n-1)/N)$ and then perform the summation. We choose to use the closed Newton-Cotes formulas. These Newton-Cotes formulas are given in table B.1. One should be aware that calculations using the values from table B.1 are very sensitive to rounding errors, so one should always use the highest precision available. For calculating the Legendre polynomial coefficients we used a POP11 system that could handle quotients without converting them to binary fractions (e.g., 1/2·2/3 was evaluated to 1/3 instead of to 0.333…).

We will clarify the procedure used by an example. To calculate the n^{th} order Legendre coefficient, A_n, of a function f(t) whose value is only known at 16 equidistant points between t=0 and t=1, we assume that g(t) = $f(t)\cdot L_n(t)$ (i.e., $g_i = f_i \cdot L_n((i-1)/N)$). Then, the value of the Legendre coefficient (i.e., A_n) is the sum of an eighth and seventh order numerical integration of the product (i.e., g(t)), i.e.,

$$A_n = 1/h_n \cdot \int_0^1 g(t)\,dt \bullet 1/h_n \cdot (C_8/15 \cdot \sum_{i=1}^{9} \Re a^8_i \cdot g_i + C_7/15 \cdot \sum_{i=9}^{16} \Re a^7_i \cdot g_i) =$$

$$(4/14175 \cdot (989g_1 + 5888g_2 - 928g_3 + 10496g_4 - 4540g_5 + 10496g_6 - 928g_7 + 5888g_8 + 989g_9) + 7/17280 \cdot (751g_9 + 3577g_{10} + 1323g_{11} + 2989g_{12} + 2989g_{13} + 1323g_{14} + 3577g_{15} + 751g_{16})) / (h_n \cdot 15).$$

Note that g_9 is used twice, it is the last point of the 8^{th} order sum and the first of the 7^{th} order sum.

Table B.1: Newton-Cotesformulas for numerical integration (closed form).
The unknown value of the integral is approximated by calculating a weighted sum of function values at equidistant points:

$$\int_0^T g(t)\,dt \bullet C \cdot T/N \cdot \sum_{i=1}^{N+1} \Re a^N_i \cdot g(T \cdot (i-1)/N) = C \cdot T/N \cdot \sum_{i=1}^{N+1} \Re a^N_i \cdot g_i$$

In which a^N_i is the i^{th} coefficient of the N^{th} order. Note that the i^{th} point g_i = g(T·(i-1)/N), e.g. g_1 = g(0) and g_{n+1} = g(T). This way the value of the function g(t) *is evaluated BETWEEN the first and the last point (adapted from Abramowitz and Stegun, 1964[10] page 886-887).*

Ord	C	$\int g(t)\,dt \approx C/N \cdot \sum \Re a_i \cdot g_i$
0	1	g_1
1	1/2	g_1+g_2
2	1/3	$g_1+4g_2+g_3$
3	3/8	$g_1+3g_2+3g_3+g_4$
4	2/45	$7g_1+32g_2+12g_3+32g_4+7g_5$
5	5/288	$19g_1+75g_2+50g_3+50g_4+75g_5+19g_6$
6	1/140	$41g_1+216g_2+27g_3+272g_4+27g_5+216g_6+41g_7$
7	7/17280	$751g_1+3577g_2+1323g_3+2989g_4+2989g_5+1323g_6+3577g_7+751g_8$
8	4/14175	$989g_1+5888g_2-928g_3+10496g_4-4540g_5+10496g_6-928g_7+5888g_8+989g_9$
9	9/89600	$2857g_1+15741g_2+1080g_3+19344g_4+5778g_5+5778g_6+19344g_7+1080g_8+15741g_9+2857g_{10}$
10	5/299376	$16067g_1+106300g_2-48525g_3+272400g_4-260550g_5+427368g_6-260550g_7+272400g_8-48525g_9+106300g_{10}+16067g_{11}$

APPENDIX C:

ANNOTATED TEXTS WITH ACCENT TRANSCRIPTION

These are the texts as our speaker read them. We changed the original text somewhat to reflect the exact words our speaker used when reading it. Therefore, the texts presented below differ somewhat from the original text as presented to the speaker. Stressed syllables of words bearing sentence accent are written in uppercase characters. In some of these words, other syllables carried such heavy stress that they too were considered to be accented.
Each vowel realization in the text has a letter code and a number written on the lines below it. The letter codes were NOT intended to represent the correct or expected pronunciation, they were for coding convenience only. The codes correspond to those used in appendix D.

? Vowel from substituted word, not always present in both readings.
-x- Word(s) deleted from reading, corresponding vowels were absent.
+x+ Word(s) inserted in reading, corresponding vowels were added.

Normal rate

De ondernemende samenleving

als er een soort van ONomstREden, bijna HEIlige LEERstelling is in het DENKen
a e oo a aa e s e
1 1 1 2 1 2 1 3

rondom het LEIden van beDRIJven, dan is het de geDACHte dat er TWEE SOORten
s a a s a a e oo
2 3 4 3 5 6 4 2

TOPmensen bestaan.
e aa
5 2

de Ene zijn de MAnagers. dat is een VAK dat je kunt LEren, daar bestaan
a a a aa aa
7 8 9 3 4

HEEL goeie SCHOlen voor en wie zo'n school met sucCES heeft doorLOpen, en
oe oo oo e ie oo oo e uu e oo oo e
1 3 4 6 1 5 6 7 1 8 7 8 9

dan nog wat EXtra-intelliGENtie heeft, en wat amBItie, en een beetje geLUK,
a a e aa ie e ie e a a ie ie e
10 11 10 5 2 11 3 12 12 13 4 5 13

die kan BEST een goeie MAnager worden, EN een steeds HOgere manager. maar
ie a e oe e oo aa
6 14 14 2 15 9 6

DIT soort MAnagers valt EIgenlijk in de categoRIE van wat je BIJna zou kunnen
oo a a oo ie a a aa
10 15 16 11 7 17 18 7

noemen "ergens tussen SUper-administraTEUR en SUper-persoNEELSchef en
oe e uu a ie ie aa e uu e oo e e
3 16 2 19 8 9 8 17 3 18 12 19 20

uitSTEkende manager" IN. dat is de Ene categoRIE, die ALgeMEEN in de
a a oo ie ie a
20 21 13 10 11 22

LEERstelling wordt erKEND.
e e e
21 22 23

de ANdere categorie, en die is HEEL-veel KLEIner, bestaat uit de
a a oo ie e ie e? aa
23 24 14 12 24 13 25 9

onderNEmers. mensen met oorSPRONKelijke geDACHten, OPzetters van NIEuwe
e e oo a e a ie
26 27 15 25 28 26 14

DINGen, mensen die ALtijd op zoek zijn naar iets NIEUWS, en die daarbij
e ie a oe aa ie ie e ie aa
29 15 27 4 10 16 17 30 18 11

INgebouwde onZEkerheden bePAALD niet SCHUwen. DAT zijn de onderNEmers, en
aa ie uu a e
12 19 4 28 31

DAT (zegt de LEERstelling) is een VAK dat je NIET op enig SCHOOL of
a e e a a ie oo
29 32 33 30 31 20 16

universiTEIT kunt LEren: daar wordt je mee geBOren, dat "heb-je-in-je-VINGers"
uu ie e ie aa oo a e
5 21 34 22 13 17 32 35

of NIET. en betrekkelijk WEInig mensen HEBben het in hun VINGers.
ie e e e e s
23 36 37 38 39 4

```
DAT is de LEERstelling, en VEEL mensen TWIJfelen er NIET aan dat de stelling
a            e         e       e               e  ie   aa  a      e
33           40        41      42              43 24   14  34     44

JUIST is. de VRAAG is overigens WEL: IS de stelling juist? het feit dat
            aa       oo        e          e              s        a
            15       18        45         46             5        35

VEEL mensen er in "geLOven" zegt op zichzelf NIETS.
     e      e        oo     e           e    ie
     47     48       19     49          50   25


MAAR er is tenminste EEN MAN, EEN van de meest gerenomMEERde DENKers en
aa   e                   a        a               oo        e       e
16   51                  36       37              20        52      53

SCHRIJvers in de WEreld over organiSAtie-problemen, de amerikaan PEter DRUcker,
                        oo    aa ie aa ie oo           aa ie aa
                        21    17 26 18 27 22           19 28 20

die het LEF heeft om de STELling DRAStisch en beredeNEERD onderUIT te halen.
ie  s   e               e        a   ie    e                          aa
29  6   54              55       38  30    56                         21

DRUcker, die nu bijna TACHtig JAAR is, was bij MIJN WEten de EERste MAN
         ie  uu    aa a       aa       a                             a
         31  6     22 39      23       40                            41

die systeMAtisch over MAnagement en onderNEmen is gaan DENKen, en de laatste
ie       aa ie   oo              e               aa   e       e     aa
32       24 33   23              57              25   58      59    26

ACHTenveertig JAAR zijn er VOORTdurend NIEewe BOEken met NIEuwe geDACHten
a             aa        e  oo   uu     ie     oe     e   ie       a
42            27        60 24    7     34      5     61  35       43

over dat ONderwerp van HEM UITgekomen. hij is een erKENde autoriTEIT in
oo   a        e    a   e        oo               e e       oo ie
25   44       62   45  63       26              64 65      27 36

de wereld, en dat WEET ie. Hij is zo langzamerhand GEEStelijk-intellectuEEL
           e  a        ie         oo a   aa   a                    e  uu
           66 46       37         28 47  28   48                   67  8

een FORmiDAbele IJdeltuit geworden, maar SOMmige mensen hebben MEER recht
       ie aa                        aa          e      e          e
       38 29                        30          68     69         70

op die status dan ANderen vind ik, en dat geldt OOK voor DRUcker die tegen
   ie  aa     a   a                e  a   e     oo  oo           ie
   39  31     49  50               71 51  72    29  30           40

zijn TACHtigste jaar nog BOEken produceert waarin hij, zorgVULdig beredeNEERD,
     a          aa       oe     oo uu      aa
     52         32        6     31  9      33

DOGma's onderUIT haalt, en met NIEuwe iDEEen komt.
   aa            aa     e  e   ie     ie
   34            35     73 74  41     42


ik heb net een VRIJ NIEUW BOEK van hem gelezen, -en- dat heet inNOvatie en
   e   e            ie    oe   a   e             e   a          oo aa ie e
   75  76           43     7   53  77            78  54         32 36 44 79

onderNEmerschap. en in dit BOEK betoogt DRUcker niet alleen dat de wereld
          a      e         oe     oo            ie   a      a
          55     80         8     33            45   56     57

voor het oplossen van zijn ecoNOmische en sociAle problemen ontzettend DRINGend
oo   s            a         oo oo ie   e  oo  aa  oo           e
34   7            58        35 36 46   81 37  37  38           82

beHOEFte heeft aan zoVEEL mogelijk onderNEmers, maar OOK dat onderNEmen
  oe           aa  oo     oo                   aa   oo  a
   9           38  39     40                   39   41  59
```

```
(verNIEuwen, verANderen, aan NIEuwe combinaties van DINGen denken, beREID
    ie          a        aa  ie        ie aa ie a          e
    47          60       40  48        49 41 50 61         83

zijn om ZORGvuldig beREkende onZEkerheden te aanVAARden, om maar een PAAR
                                          aa aa          aa       aa
                                          42 43          44       45

voorbeelden te noemen) +dat onderNEmen+ GEENSzins is VOORbehouden aan een soort
oo             oe       a                              oo           aa      oo
42             10       61.5                           43           46      44

inspiRErende bedrijfsKUNdige geNIEen met speciAle taLENten, maar dat het
  ie                          ie    e        aa  a e       aa   a   s
  51                          52    84       47 62 85      48   63  8

gewoon een VAK is dat je kunt LEren en waar je dan HARD en systeMAtisch aan
  oo       a      a                e  aa      a   a    e       aa ie   aa
  45       64     65               86 49      66  67   87      50 53   51

moet WERken om het in de prakTIJK te brengen. Hij gaat dan DOOR met
oe   e         s         a           e           aa   a   oo   e
11   88        9         68          89          52   69  46   90

DRIEhonderd pagina's lang te verTELlen WAAR dat vak uit beSTAAT.
ie          aa ie aa a          e      aa   a   a         aa
54          53 55 54 70         91     55   71  72        56


daar kan IK in vijf miNUten uiteraard niet zo erg veel over verTELlen, maar
aa   a             ie uu        aa   ie   oo e        oo      e       aa
57   73            56 10        58   57   47 92       48      93      59

het is WEL nuttig om hier bijvoorbeeld Even MELding te maken van het FEIT
s      e             ie      oo             e          aa    a   s
10     94            58      49             95         60    74  11

dat naar DRUckers MEning de EERste PAgina van een MAANdelijks rapPORT over
a   aa                           aa ie aa a      aa          a       oo
75  61                           62 56 63 76     64          77      50

de gang van ZAken in een beDRIJF NIET alleen hoort te beSTAAN uit een OPsomming
   a    a   aa                   ie   a      oo         aa
   78   79  65                   60   80     51         66

van de proBLEmen die er zijn geweest, maar vooral OOK over wat ONverwacht
a      oo        ie  e               aa   oo  a  oo  oo   a        a
81     52        61  96              67   53  82 54  55   83       84

GOED is gegaan, en WELke KANsen die BOven verWACHting verlopen GOEde gang van
oe        aa    e  e     a      ie  oo       a           oo    oe    a    a
12        68    97 98    85     62  56       86          57    13    87   88

ZAken zou kunnen bieden voor nog VERdere verBEteringen, inNOvaties, nieuwe
aa              ie     oo       e                        oo aa ie  ie
69              63     58       99                       59 70 64  65

INzichten en beNAderingen. en de VAK-onderNEmer praat regelMAtig in speciAle
          e    aa          e     a               aa         aa            aa
         100   71         101    89              72         73            74

bijEENkomsten met de JONGe mensen in zijn beDRIJF om van HEN te HOren WAT
              e           e                          a   e      oo    a
             102         103                         90 104     60    91

er BEter kan. de ondernemer is NIET vaak met GISteren bezig, maar met MORgen
e        a                    ie   aa   e                    aa   e
105      92                   66   75   106                  76   107

en Overmorgen. en hij neemt OOK niet, zoals VEEL wordt gezegd, RIsico's.
e  oo          e            oo  ie    oo a               e     ie ie oo
108 61        109           62  67    63 93             110    68 69 64

hij HOUDT niet van RIsico's, en hij vindt het voorAL risKANT om ALSmaar
          ie   a   ie ie oo  e           s   oo  a     a       a  aa
          70   94  71 72 65 111          12  66  95    96      97 77
```

```
achterOM te kijken naar GISteren. hij NEEMT geen RIsico's, maar hij ZOEKT
a                 aa                          ie ie oo  aa       oe
98                78                          73 74 67  79       14

die zorgVULdig beredeneerde onZEkerheden OP. en dat ALles zegt drucker,
ie                                          e  a   a     e
75                                        112 99 100   113

vraagt GEEN speciAle taLENten, het vraagt WERK en een geTRAINde,
aa              aa  a  e      s   aa     e    e
80              81 101 114    13  82    115  116

gedisciplinEERde manier van DENKen. daarbij moet natuurlijk OOK worden
  ie ie ie       aa ie  a   e       aa      oe   aa uu      oo
  76 77 78       83 79 102 117      84      15   85 11      68

"geMAnaged", geadminiSTREERD, geORganiseerd en al die ANdere dingen, MAAR
              a ie ie            aa ie     e  a  ie  a              aa
             103 80 81           86 82   118 104 83 105             87

het onderNEmen staat voorOP.
s             aa    oo
14            88    69


DRUcker is er van overTUIGD dat de WESterse SAmenleving op WEG is naar een
          e  a   oo       a      e        aa             e      aa
         119 106  70      107    120      89            121     90

'onderNEmende samenleving' (zoals ie VEle jaren lang te WEInig is geweest),
             aa           oo a  ie      aa    a
             91           71 108 84     92   109

en dan heeft hij het NIET (zoals SOMmige NEderlanders dan misschien meteen
e   a        ie? s   ie    oo a              a       a       ie
122 110     84.5 15  85    72 111            112     113     86

denken) over een KEIHARde SAmenleving van "RIJKaards-en-de-ANderen", maar
e       oo          a     aa          a        aa    e     a        aa
123     73         114    93         115       94   124   116       95

over een SAmenleving die door OPtimale onderNEmingslust, en door wat MINder
oo       aa          ie  oo     ie aa                    e  oo   a
74       96          87  75     88 97                  125 76  117

te bouwen op de EINdeloze WIJSheid van de Overheid en haar voorZIENingen,
                    oo           a      oo       e  aa   oo  ie
                    77          118     78      126 98   79  89

OPlossingen vindt voor (ook in ZIJN Ogen) ONaanvaardbare ZAken als de HOge
                  oo    oo          oo      aa aa   aa   aa    a      oo
                  80    81          82      99 100 101  102   119     83

werkLOOSheid in veel WESterse landen van DIT ogenBLIK. er zijn VOORTDUrend
e   oo              e        a      a       oo        e       oo   uu
127 84             128      120    121      85       129      86   12

nieuwe BAnen te creeren, zegt DRUcker, maar daar heb je GEEN
ie     aa               e            aa   aa   e
90    103              130          104  105  131

"op-de-WINKel-passers" voor nodig, en NIET de Overheid, maar mensen die
             a        oo   oo     e  ie      oo        aa   e      ie
            122       87   88    132 91      89       106  133     92

het VAK "onderNEmen" beGRIJPen en uitOEFenen.
s   a                        e     oe
16 123                      134    16
```

Fast rate

```
      De ondernemende samenleving

als er een soort van ONomstreden, bijna HEIlige LEERstelling is in het DENKen
a   e      oo    a                  aa             e              s   e
1   1      1     2                  1              2              1   3

rondom het LEIden van beDRIJven, dan is het de gedachte dat er TWEE soorten
       s          a             a      s       a      a   e       oo
       2          3             4      3       5      6   4       2

TOPmensen bestaan.
   e        aa
   5        2


de Ene zijn de MAnagers. dat is een vak dat je kunt LEren, daar bestaan
                         a          a   a                  aa     aa
                         7          8   9                  3      4

HEEL goeie SCHOlen voor en wie zo'n school met sucCES heeft doorLOpen, en
     oe    oo      oo   e  ie  oo   oo     e   uu e         oo  oo     e
     1     3       4    6  1   5    6      7   1  8         7   8      9

dan nog wat EXtra-intelliGENtie heeft, en wat amBItie, EN een beetje geLUK,
a       a   e aa      ie e  ie          e  a   a ie ie  e
10      11 10 5        2 11 3           12 12 13  4  5  13

die kan BEST een goeie MAnager worden, en een steeds HOgere MAnager. maar
ie  a   e        oe                    e              oo               aa
6   14  14        2                    15              9                6

DIT soort managers valt eigenlijk in de categorie van wat je bijna zou kunnen
    oo             a                    a   oo ie a   a         aa
    10             15                   16  11  7 17  18         7

noemen "ERgens tussen SUper-administraTEUR en SUper-persoNEELSchef en
oe      e            uu    a ie ie aa     e  uu    e  oo     e    e
 3      16            2   19  8  9  8     17  3    18 12     19   20

uitSTEkende MAnager" IN. dat is de Ene categoRIE, die ALgemeen in de
                         a              a   oo ie  ie  a
                         20             21  13 10  11  22

LEERstelling wordt ERkend.
    e              e e
    21            22 23


de ANdere categoRIE, en die is VEEL-en-veel KLEIner, bestaat uit de
   a      a   oo ie  e  ie         e?               aa
   23     24  14 12  24 13         25                9

onderNEmers. mensen met oorSPRONKelijke geDACHten, OPzetters van NIEuwe
             e      e   oo                a          e       a   ie
             26     27  15                25         28      26  14

DINGen, mensen die ALtijd op zoek zijn naar iets NIEUWS, en die daarbij
        e      ie  a         oe        aa   ie   ie       e  ie  aa
        29     15  27         4        10   16   17       30 18  11

INgebouwde onZEkerheden bePAALD niet SCHUwen. DAT zijn de onderNEmers, en
                          aa    ie   uu       a                        e
                          12    19    4       28                       31

DAT (zegt de LEERstelling) is een vak dat je NIET op enig school of
a    e           e                a   a      ie           oo
29   32          33               30  31     20           16

universiTEIT kunt LEren: daar wordt je mee geBOren, dat "heb-je-in-je-
uu ie e ie               aa                  oo     a    e
5 21 34 22               13                  17     32   35

VINGers" of NIET. en betrekkelijk WEInig mensen HEBben het in hun VINGers.
          ie    e    e                 e      e      s
          23    36   37                38     39     4
```

```
DAT is de LEERstelling, en veel mensen TWIJfelen er NIET aan dat de STELling
a            e         e       e               e  ie   aa  a      e
33           40        41      42              43 24   14  34     44

JUIST is. de vraag is overigens WEL: IS de stelling juist? het feit dat
            aa       oo        e          e              s        a
            15       18        45         46             5        35

VEEL mensen er in "geLOven" zegt op zichzelf NIETS.
     e      e        oo     e           e    ie
     47     48       19     49          50   25


MAAR er is tenminste EEN MAN, een van de meest gerenomMEERde DENKers en
aa   e                   a        a               oo        e       e
16   51                  36       37              20        52      53

SCHRIJvers in de wereld over organiSAtie-proBLEmen, de ameriKAAN PEter DRUcker,
                        oo    aa ie aa ie oo           aa ie aa
                        21    17 26 18 27 22           19 28 20

die het LEF heeft om de stelling DRAStisch en beredeNEERD onderUIT te halen.
ie  s   e              e        a   ie    e                          aa
29  6   54             55       38  30    56                         21

DRUcker, die nu bijna TACHtig JAAR is, was bij MIJN weten de EERste man
         ie  uu    aa a       aa       a                            a
         31  6     22 39      23       40                           41

die systeMAtisch over MAnagement en onderNEmen is gaan DENKen, en de laatste
ie       aa ie   oo             e               aa   e       e     aa
32       24 33   23             57              25   58      59    26

ACHTenveertig jaar zijn er VOORTDUrend NIEuwe boeken met NIEuwe gedachten
a            aa        e  oo   uu     ie     oe     e   ie       a
42           27        60 24    7     34      5     61  35       43

over dat onderwerp van hem UITgekomen. hij is een erKENde autoriteit in
oo   a        e    a   e        oo               e e       oo ie
25   44       62   45  63       26              64 65     27 36

de WEreld, en dat WEET ie. hij is zo langzamerhand GEEStelijk-intellectuEEL
           e  a        ie         oo a   aa   a                    e  uu
           66 46       37         28 47  28   48                   67  8

een formiDAbele IJdeltuit geworden, maar sommige mensen hebben MEER recht
       ie aa                       aa          e      e          e
       38 29                       30          68     69         70

op die STAtus dan ANderen vind ik, en dat geldt OOK voor DRUcker die tegen
   ie  aa     a   a               e  a   e     oo  oo           ie
   39  31     49  50              71 51  72    29  30           40

zijn TACHtigste jaar nog BOEken produceert waarin hij, zorgVULdig beredeNEERD,
     a          aa       oe     oo uu      aa
     52         32        6     31  9      33

DOGma's onderUIT haalt, en met NIEuwe iDEEen komt.
   aa            aa     e  e   ie     ie
   34            35     73 74  41     42


ik heb net een VRIJ NIEUW BOEK van hem gelezen, en dat heet INnovatie en
   e   e            ie    oe   a   e            e  a          oo aa ie e
   75  76           43     7   53  77           78 54         32 36 44 79

onderNEmerschap. en in dit boek betoogt DRUcker niet alleen dat de wereld
          a      e         oe     oo            ie   a      a
          55     80         8     33            45   56     57

voor het OPlossen van zijn ecoNOmische en sociAle proBLEmen ontzettend DRINGend
oo   s            a         oo oo ie   e  oo  aa  oo           e
34   7            58        35 36 46   81 37  37  38           82

beHOEFte heeft aan zoveel mogelijk onderNEmers, maar ook dat onderNEmen
  oe           aa  oo     oo                   aa   oo  a
   9           38  39     40                   39   41  59
```

```
(verNIEuwen, verANderen, aan NIEuwe combinaties van DINGen denken, beREID
    ie          a        aa  ie        ie aa ie a          e
    47          60       40  48        49 41 50 61         83

zijn om zorgVULdig berekende onZEkerheden te aanVAARden, om maar een paar
                                          aa aa          aa       aa
                                          42 43          44       45

voorbeelden te noemen) GEENSzins is voorbehouden aan een soort
oo             oe       a           oo           aa      oo
42             10       61.5        43           46      44

inspiRErende bedrijfsKUNdige geNIEen met speciAle taLENten, maar dat het
  ie                          ie    e        aa  a e       aa   a   s
  51                          52    84       47 62 85      48   63  8

gewoon een VAK is dat je kunt LEren en waar je dan HARD en systeMAtisch aan
  oo       a      a                e  aa      a   a    e       aa ie   aa
  45       64     65               86 49      66  67   87      50 53   51

moet WERken om het in de prakTIJK te brengen. Hij gaat dan door met
oe   e         s         a           e           aa   a   oo   e
11   88        9         68          89          52   69  46   90

DRIEhonderd pagina's lang te verTELlen waar dat vak uit beSTAAT.
ie          aa ie aa a          e      aa   a   a         aa
54          53 55 54 70         91     55   71  72        56


daar kan ik in VIJF miNUten uiteraard niet zo erg veel over verTELlen, maar
aa   a             ie uu        aa   ie   oo e        oo      e       aa
57   73            56 10        58   57   47 92       48      93      59

het is wel NUTtig om hier bijvoorbeeld even MELding te maken van het feit
s      e             ie      oo             e          aa    a   s
10     94            58      49             95         60    74  11

dat naar DRUckers mening de EERste pagina van een MAANdelijks rapport over
a   aa                            aa ie aa a      aa          a       oo
75  61                            62 59 63 76     64          77      50

de gang van ZAken in een beDRIJF NIET alleen hoort te bestaan uit een opsomming
   a    a   aa                   ie   a      oo         aa
   78   79  65                   60   80     51         66

van de proBLEmen die er zijn geWEEST, maar VOORal ook over wat onverwacht
a      oo        ie  e               aa   oo  a  oo  oo   a        a
81     52        61  96              67   53  82 54  55   83       84

GOED is gegaan, en WELke kansen die BOven verWACHting verlopen goede gang van
oe        aa    e  e     a      ie  oo       a           oo    oe    a    a
12        68    97 98    85     62  56       86          57    13    87   88

ZAken zou kunnen bieden voor NOG verdere verBEteringen,+voor+innoVAties, nieuwe
aa              ie     oo       e                      oo     oo aa ie  ie
69              63     58       99                     58.5   59 70 64  65

INzichten en beNAderingen. en de VAK-ondernemer praat regelMAtig in speciAle
          e    aa          e     a               aa         aa            aa
         100   71         101    89              72         73            74

bijEENkomsten met de JONGe mensen in zijn beDRIJF om van HEN te horen WAT
             e           e                        a   e      oo    a
            102         103                       90 104     60    91

er BEter kan. de ondernemer is NIET vaak met GISteren bezig, maar met  MORgen
e        a                   ie   aa   e                  aa   e
105      92                  66   75   106                76   107

en Overmorgen. en hij neemt OOK NIET, zoals VEEL wordt geZEGD, RIsico's.
e  oo          e           oo  ie    oo a              e     ie ie oo
108 61        109          62  67    63 93            110    68 69 64

hij HOUDT niet van RIsico's, en hij vindt het voorAL riskant om ALSmaar
          ie   a   ie ie oo  e           s   oo  a     a       a  aa
          70   94  71 72 65 111          12  66  95    96      97 77
```

```
achterOM te kijken naar GISteren. hij NEEMT GEEN RIsico's, maar hij zoekt
a                 aa                          ie ie oo  aa       oe
98                78                          73 74 67  79       14

die zorgVULdig beredeneerde onZEkerheden op. en dat ALles zegt drucker,
ie                                        e  a   a     e
75                                      112 99 100   113

vraagt GEEN speciale taLENten, het vraagt WERK en een geTRAINde,
aa              aa  a  e      s   aa     e    e
80              81 101 114    13  82    115  116

gedisciplINEERde manier van DENKen. daarbij moet natuurlijk OOK worden
  ie ie ie       aa ie  a   e       aa      oe   aa uu      oo
  76 77 78       83 79 102 117      84      15   85 11      68

"geMAnaged", geadminiSTREERD, geORganiseerd en al die ANdere dingen, maar
               a ie ie           aa ie     e  a  ie  a              aa
             103 80 81           86 82   118 104 83 105            87

het onderNEmen staat voorOP.
s              aa    oo
14             88    69


DRUcker is er van overTUIGD dat de WESterse SAmenleving op weg is naar een
          e  a   oo       a      e       aa            e      aa
        119 106  70       107    120     89           121     90

'onderNEmende samenleving' (zoals ie VEle jaren lang te WEInig is geweest),
             aa           oo a  ie      aa    a
             91           71 108 84     92   109

en dan heeft hij het NIET (zoals sommige NEderlanders dan misschien meteen
e   a        ie?  s   ie    oo a              a       a       ie
122 110     84.5  15  85    72 111           112     113      86

denken) over een KEIharde SAmenleving van "RIJKaards-en-de-ANderen", maar
e       oo          a     aa          a       aa    e     a        aa
123     73         114    93         115      94   124   116       95

over een SAmenleving die door OPtimale onderNEmingslust, en door wat MINder
oo       aa          ie  oo     ie aa                   e  oo   a
74       96          87  75     88 97                 125 76  117

te bouwen op de EINdeloze WIJSheid van de Overheid en haar voorzieningen,
                    oo           a      oo       e  aa   oo  ie
                    77          118     78      126 98   79  89

OPlossingen vindt voor (ook in ZIJN ogen) ONaanvaardbare zaken als de  HOge
                  oo    oo          oo      aa aa   aa   aa    a      oo
                  80    81          82      99 100 101  102   119     83

werkLOOSheid in veel WESterse landen van DIT ogenBLIK. er zijn VOORTDUrend
e   oo              e       a      a       oo        e       oo   uu
127 84             128      120    121      85       129      86   12

nieuwe BAnen te creeren, zegt DRUcker, maar daar heb je GEEN
ie     aa               e            aa   aa   e
90    103              130           104  105  131

"op-de-WINKel-passers" voor nodig, en NIET de Overheid, maar MENsen die
             a       oo   oo     e  ie      oo        aa   e      ie
           122       87   88    132 91      89       106  133     92

het VAK "onderNEmen" beGRIJPen en UIToefenen.
s   a                        e     oe
16 123                      134    16
```

Speaking rate: Fast

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
s	y	p	+	51	204	297	62	1215	24	2461	−68	44	uu	3
s	y	p	+	77	227	337	98	1140	−198	2362	−72	47	uu	2
d	y	r	+	133	189	369	18	1449	245	2373	−353	46	uu	7
d	y	r	+	135	204	329	−54	1712	264	2304	−412	49	uu	12
n	y	t	+	94	208	344	45	1559	−115	2394	−176	52	uu	10
X	y	w	+	76	154	298	−12	1235	96	2419	14	43	uu	4
n	y	b	−	89	175	351	98	1397	297	2382	77	45	uu	6
s	y	k	−	47	156	296	−17	1403	−35	2412	−43	45	uu	1
f	y	n	−	56	167	317	−34	1428	−29	2397	28	45	uu	5
m	y	n	−	59	175	367	62	1416	−21	2354	156	47	ie	56
t	y	r	−	76	156	430	−40	1346	−75	2322	53	47	uu	11
d	y	s	−	60	175	314	−1	1609	−116	2368	−140	45	uu	9
t	y	w	−	83	143	332	−3	1429	−131	2373	8	43	uu	8

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
X	u	d	+	77	213	390	60	699	−257	2516	16	47	oe	12
&	u	f	+	174	185	350	6	736	−684	2609	390	48	oe	9
b	u	k	+	64	233	338	1	698	−228	2364	−18	45	oe	6
b	u	k	+	98	213	400	−64	769	−373	2448	285	50	oe	7
t	u	f	−	105	152	335	−28	919	−585	2452	−97	44	oe	16
X	u	j	−	49	204	369	36	803	−200	2486	5	46	oe	13
X	u	j	−	64	227	367	34	699	−222	2456	189	49	oe	1
X	u	j	−	67	182	370	4	850	−318	2410	39	51	oe	2
b	u	k	−	67	185	339	−1	736	−135	2513	134	49	oe	5
b	u	k	−	94	233	355	24	666	−299	2330	73	51	oe	8
z	u	k	−	77	213	347	79	783	−407	2478	−82	46	oe	4
z	u	k	−	90	222	350	6	806	−358	2424	−190	48	oe	14
n	u	m	−	71	143	392	33	842	−67	2508	82	42	oe	10
n	u	m	−	72	169	390	23	798	−80	2477	−35	44	oe	3
m	u	t	−	73	137	369	21	750	−319	2471	61	41	oe	11
m	u	t	−	75	167	428	−39	892	−381	2464	296	45	oe	15

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
n	&	m	−	35	145	464	55	1284	129	2473	119	44	e	77
n	&	n	−	73	167	582	90	1500	50	2653	21	47	e	43
v	&	n	−	42	182	437	−28	1410	−36	2523	−27	48	a	37
d	&	r	−	43	164	394	−15	1454	55	2845	77	46	e	48
d	&	r	−	56	182	383	3	1488	−70	2780	63	48	e	51
s	&	r	−	60	192	442	−27	1309	39	2717	−88	54	e	1
t	&	r	−	41	200	517	56	1389	44	2765	−71	47	e	105
t	&	r	−	63	141	420	46	1121	−173	2893	182	43	e	4
#	&	t	−	20	85	392	−4	1430	17	2645	4	39	schwa	14
#	&	t	−	58	167	430	63	1495	3	2520	−72	49	schwa	5
#	&	t	−	62	132	392	117	1456	70	2605	108	41	schwa	13
d	&	t	−	48	159	461	41	1454	−10	2664	−5	48	a	57
d	&	t	−	65	167	457	60	1610	134	2478	15	49	a	9
d	&	t	−	65	169	461	82	1691	128	2420	−121	47	a	31
d	&	t	−	73	167	448	57	1629	−29	2537	−233	48	a	65
i	&	t	−	45	68	342	15	1760	18	2631	17	39	schwa	15
i	&	t	−	65	139	417	39	1856	−276	2606	6	46	schwa	16
i	&	t	−	*103**	156	327	*24*	2071	*279*	2585	*−67*	48	schwa	6
m	&	t	−	18	78	380	−7	1386	19	2653	23	42	schwa	9
m	&	t	−	66	139	439	52	1307	−124	2578	174	43	schwa	2
n	&	t	−	51	149	440	36	1554	104	2656	1	45	schwa	4
n	&	t	−	52	149	547	68	1256	25	2753	56	46	schwa	11
n	&	t	−	55	147	498	135	1590	59	2596	11	43	schwa	1
n	&	t	−	65	172	406	82	1492	−174	2731	30	49	schwa	12
r	&	t	−	41	156	431	19	994	−15	2724	−4	44	schwa	7
r	&	t	−	46	175	500	16	1427	52	2657	−13	51	schwa	10
s	&	t	−	68	179	380	44	1529	22	2632	−11	47	schwa	3
t	&	t	−	48	152	441	32	1435	−28	2496	31	47	aa	85
t	&	t	−	68	185	459	85	1294	−14	2822	152	48	schwa	8
n	&	v	−	68	122	528	142	1486	68	2410	−25	38	e	60

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
h	o	X	+	133	213	433	−24	848	−13	2542	17	57	oo	83
h	o	X	+	140	222	432	53	761	52	2660	−15	62	oo	9
k	o	k	+	76	213	480	57	885	6	2365	−47	55	oo	68
t	o	k	+	89	222	410	40	823	−12	2389	14	55	oo	29
t	o	k	+	131	233	430	29	715	−171	2649	278	59	oo	62

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
X	o	l	+	129	192	424	51	787	-110	2486	33	53	oo	3
n	o	m	+	94	192	567	96	984	-13	2701	559	45	oo	36
l	o	p	+	138	154	390	136	754	-279	2619	277	54	oo	8
b	o	r	+	173	200	379	-16	918	-59	2388	-99	58	oo	17
v	o	r	+	79	204	474	-62	906	-216	2500	29	55	oo	53
v	o	r	+	134	208	388	34	819	-178	2089	-397	51	oo	24
v	o	r	+	171	244	427	38	887	-195	2261	-381	52	oo	86
l	o	s	+	113	196	419	44	925	-338	2523	178	51	oo	84
&	o	v	+	128	213	467	-42	839	-201	2684	373	55	oo	78
&	o	v	+	142	175	386	27	740	-351	2723	326	51	oo	89
b	o	v	+	143	213	426	38	723	-249	2716	376	61	oo	56
l	o	v	+	122	200	404	74	799	-146	2502	234	52	oo	19
n	o	v	+	129	185	384	-58	752	-316	2759	609	55	oo	61
z	o	A	-	41	192	545	-49	978	-60	2536	-133	53	oo	63
z	o	A	-	*122**	161	439	*104*	951	*-305*	2827	*-28*	52	oo	71
z	o	E	-	41	164	421	-21	1201	-123	2336	8	46	oo	47
z	o	O	-	46	192	445	-1	1001	-100	2513	-274	52	oo	72
s	o	S	-	73	164	415	42	1080	-437	2265	-235	44	oo	37
m	o	X	-	102	161	448	8	772	25	2841	72	49	oo	40
n	o	X	-	160	185	374	-75	744	-119	2733	21	52	oo	82
t	o	X	-	109	167	413	10	782	-245	2756	-80	50	oo	85
t	o	X	-	131	213	439	35	921	-355	2622	260	55	oo	33
r	o	b	-	86	182	387	-16	745	-233	2507	560	49	oo	38
r	o	b	-	88	152	405	7	764	7	2552	143	46	oo	22
r	o	b	-	89	139	378	38	709	-117	2528	365	48	oo	52
n	o	d	-	138	161	473	86	911	-234	2576	556	51	oo	88
r	o	d	-	79	200	380	75	1226	-71	1971	114	50	oo	31
l	o	k	-	139	196	358	73	721	30	2304	428	57	oo	54
r	o	k	-	109	167	380	73	733	18	2581	-29	51	oo	81
r	o	k	-	149	233	403	-65	808	55	2517	101	59	oo	41
X	o	l	-	87	185	461	-2	897	-147	2353	240	51	oo	16
z	o	l	-	55	164	392	-9	989	-190	2107	134	47	oo	28
X	o	m	-	99	204	469	35	837	-77	2699	300	55	oo	6
k	o	m	-	115	123	433	187	840	14	2556	214	40	oo	26
k	o	n	-	58	152	437	28	905	-130	2671	179	46	oo	35
s	o	n	-	36	137	383	21	1073	-68	2508	68	38	oo	12
w	o	n	-	115	208	507	61	897	-99	2560	174	54	oo	45
z	o	n	-	125	182	512	-21	933	-575	2469	-98	51	oo	5
l	o	p	-	105	167	420	68	789	-113	2655	510	53	oo	57
X	o	r	-	78	156	369	6	1082	-107	2599	16	38	oo	11
X	o	r	-	111	137	370	33	946	-376	2623	33	42	oo	13
X	o	r	-	126	143	378	42	912	-156	2620	34	48	oo	14
d	o	r	-	99	154	454	75	801	-176	2683	-85	50	oo	75
d	o	r	-	126	179	386	7	1046	-288	2399	8	53	oo	7
d	o	r	-	141	204	416	-15	1231	-23	2459	-24	57	oo	46
h	o	r	-	168	196	361	-19	919	9	2441	-103	57	oo	60
n	o	r	-	120	179	432	49	965	-563	2543	-128	50	oo	51
s	o	r	-	93	172	384	19	1211	-167	2278	-230	46	oo	44
s	o	r	-	109	213	440	66	1164	-172	2385	-128	57	oo	10
s	o	r	-	120	222	429	37	1129	-284	2236	-199	56	oo	1
s	o	r	-	123	182	373	22	1138	-282	2341	-167	53	oo	2
t	o	r	-	37	164	352	-10	1251	-77	2668	-36	43	oo	27
t	o	r	-	121	167	398	-117	947	-447	2503	-55	46	oo	15
v	o	r	-	56	152	422	32	913	-201	2578	71	46	oo	59
v	o	r	-	63	208	401	3	944	-179	2369	26	53	oo	30
v	o	r	-	70	167	368	-49	897	-194	2448	-200	46	oo	34
v	o	r	-	81	156	358	-26	895	-318	2634	202	45	oo	58
v	o	r	-	89	161	454	-42	908	-127	2558	-8	50	oo	66
v	o	r	-	93	182	380	-60	996	-99	2512	153	53	oo	49
v	o	r	-	94	154	385	-40	933	-297	2704	276	44	oo	87
v	o	r	-	107	159	392	23	828	-69	2510	-18	46	oo	69
v	o	r	-	119	182	377	15	896	-180	2387	-58	53	oo	80
v	o	r	-	123	182	383	-12	909	-10	2328	195	49	oo	42
v	o	r	-	132	204	387	23	1004	245	2379	145	53	oo	43
v	o	r	-	178	133	359	38	808	-360	2391	-77	44	oo	4
k	o	s	-	113	156	465	66	971	-297	2587	201	49	oo	67
k	o	s	-	117	132	450	123	889	-383	2554	-79	43	oo	65
k	o	s	-	130	123	467	150	965	-361	2513	-116	44	oo	64
d	o	v	-	133	143	412	20	812	-218	2581	68	48	oo	21
k	o	v	-	129	172	402	46	766	-93	2664	299	52	oo	55
l	o	v	-	86	156	430	42	882	-289	2684	531	50	oo	48
n	o	v	-	78	147	403	9	872	-175	2655	443	46	oo	73
n	o	v	-	82	161	418	21	827	-31	2571	421	46	oo	25
n	o	v	-	91	182	409	13	892	-386	2703	363	53	oo	59
n	o	v	-	125	169	434	-18	792	-559	2642	261	51	oo	32
r	o	v	-	87	182	479	22	881	-40	2691	25	52	oo	74
s	o	v	-	82	161	430	9	872	-174	2541	-16	44	oo	23

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
s	o	v	-	125	182	416	82	859	-269	2557	-40	50	oo 18
t	o	v	-	64	161	409	26	875	-47	2546	59	46	oo 50
z	o	v	-	82	147	392	-18	962	-323	2411	-136	44	oo 39
d	o	w	-	62	167	390	-22	906	-149	2503	74	49	oo 76
l	o	z	-	90	167	415	-1	981	-467	2447	-113	48	oo 77
v	o	z	-	67	169	364	-68	876	-714	2489	29	40	oo 79

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
r	i	E	+	148	149	308	-65	2278	721	2565	-41	45	ie 12
r	i	O	+	102	227	347	-77	2115	584	2534	-46	47	ie 54
&	i	d	+	79	161	305	-9	2267	105	2620	-40	45	ie 42
r	i	d	+	68	152	314	-15	2157	255	2578	-7	45	ie 10
n	i	j	+	42	185	344	21	1951	251	2605	6	41	ie 52
r	i	s	+	88	182	281	-3	2267	217	2584	-68	49	ie 68
r	i	s	+	106	196	331	36	2159	46	2817	-249	49	ie 73
r	i	s	+	107	141	283	-13	2102	308	2585	-149	43	ie 71
b	i	t	+	84	222	338	61	1893	166	2456	-2	46	ie 4
n	i	t	+	72	204	374	-8	1754	120	2642	-50	44	ie 24
n	i	t	+	72	233	353	29	1894	-336	2635	-106	48	ie 20
n	i	t	+	74	233	281	10	2018	292	2533	-47	50	ie 60
n	i	t	+	82	227	261	-21	2177	271	2653	-1	51	ie 91
n	i	t	+	87	156	340	-11	2129	679	2628	169	40	ie 25
n	i	t	+	95	213	398	73	1909	-251	2577	32	49	ie 66
n	i	t	+	97	172	312	-34	2280	845	2645	273	44	ie 23
n	i	t	+	111	208	373	127	2041	133	2520	-258	50	ie 67
n	i	t	+	112	222	341	91	1955	476	2562	-115	51	ie 85
n	i	w	+	60	227	358	39	2031	-50	2643	-5	49	ie 14
n	i	w	+	66	147	363	33	2020	244	2600	33	42	ie 17
n	i	w	+	67	213	371	4	1962	14	2611	38	46	ie 48
n	i	w	+	86	204	314	36	1801	-325	2454	74	44	ie 41
n	i	w	+	101	196	325	25	1822	-188	2570	128	43	ie 35
n	i	w	+	102	227	333	11	2045	70	2688	31	46	ie 34
n	i	w	+	115	238	365	17	1830	130	2489	306	48	ie 43
n	i	w	+	132	200	380	26	2090	438	2646	148	50	ie 47
s	i	#	-	31	167	330	-12	2101	237	2610	-13	42	ie 27
s	i	#	-	73	182	327	38	2084	60	2431	-124	42	ie 5
t	i	#	-	68	118	274	38	2202	192	2663	52	33	ie 37
d	i	&	-	72	137	279	-45	2275	285	2637	32	39	ie 92
d	i	&	-	*103**	156	283	*24*	2211	*279*	2733	*-67*	48	ie 29
t	i	&	-	47	169	305	-6	2137	246	2619	35	51	ie 85
d	i	A	-	62	141	292	-63	2291	334	2560	10	40	ie 11
d	i	A	-	72	192	333	-69	2271	393	2633	85	48	ie 15
d	i	A	-	74	182	350	-95	2054	488	2642	25	49	ie 83
d	i	E	-	77	161	320	-44	2204	294	2601	61	44	ie 61
s	i	E	-	53	189	355	-9	2077	124	2613	-111	49	ie 44
d	i	I	-	92	156	309	-8	2238	282	2520	-1	45	ie 13
l	i	X	-	39	135	368	23	1982	181	2634	70	40	ie 2
d	i	b	-	78	161	288	-19	2357	720	2667	113	44	ie 62
b	i	d	-	51	179	347	22	1972	378	2460	102	44	ie 63
d	i	d	-	56	145	286	-37	2169	258	2579	5	39	ie 18
d	i	d	-	68	167	318	2	2300	295	2623	19	45	ie 87
m	i	d	-	88	169	319	7	2329	976	2570	230	45	ie 38
s	i	e	-	59	169	385	-1	1931	52	2453	-37	45	ie 3
d	i	k	-	78	147	297	44	2252	401	2525	76	47	ie 6
r	i	k	-	35	152	327	-8	2429	83	2706	18	42	ie 28
s	i	k	-	59	128	300	56	2017	117	2560	-39	38	ie 72
s	i	k	-	63	145	298	24	2173	41	2621	-7	42	ie 74
s	i	k	-	72	141	318	49	1984	130	2521	-2	44	ie 69
t	i	m	-	53	222	297	14	2002	374	2651	86	46	ie 88
X	i	n	-	47	169	340	-9	2202	381	2654	61	45	ie 86
X	i	n	-	54	196	367	10	2023	140	2606	23	44	ie 55
X	i	n	-	67	156	368	49	2016	415	2545	57	44	ie 59
b	i	n	-	56	164	338	5	1947	289	2451	17	45	ie 49
d	i	n	-	73	182	321	11	2242	426	2579	33	48	ie 31
l	i	n	-	65	143	298	-28	2148	384	2603	71	41	ie 78
m	i	n	-	12	175	335	55	1215	224	2528	13	42	ie 8
m	i	n	-	33	156	328	-8	1352	-43	2432	37	38	ie 80
z	i	n	-	80	167	330	-15	2145	367	2629	-55	46	ie 89
s	i	p	-	74	156	306	30	2148	-196	2559	-181	42	ie 77
h	i	r	-	78	192	342	37	2206	628	2569	65	56	ie 58
n	i	r	-	97	179	324	-40	2162	469	2597	31	43	ie 79
p	i	r	-	103	143	307	-2	2029	362	2547	12	38	ie 51
d	i	s	-	53	167	292	-16	2163	6	2581	-155	39	ie 39
d	i	s	-	61	169	328	-10	2071	95	2622	4	48	ie 76
d	i	s	-	68	152	303	-33	2157	251	2595	-103	44	ie 32
m	i	s	-	47	182	308	4	2033	147	2953	-305	42	ie 46

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
n	i	s	-	39	152	373	9	1892	148	2549	-50	39	ie	9
n	i	s	-	54	156	312	-6	2121	216	2602	-97	39	ie	81
n	i	s	-	73	169	300	-19	2073	141	2598	-156	41	ie	26
n	i	s	-	92	145	357	-27	2182	190	2587	24	42	ie	82
s	i	s	-	51	182	287	8	2073	16	2791	-80	38	ie	50
s	i	s	-	67	192	305	45	2060	-87	2756	-275	42	ie	30
s	i	s	-	94	137	261	2	2249	210	2553	-160	39	ie	64
t	i	s	-	62	208	357	49	1936	-122	2985	-403	43	ie	53
t	i	s	-	64	204	335	68	1936	-15	2870	-355	39	ie	33
d	i	t	-	57	145	301	-15	2159	265	2588	1	41	ie	40
n	i	t	-	69	196	368	67	2040	484	2619	0	45	ie	45
n	i	t	-	72	164	363	19	2091	236	2550	-28	44	ie	57
n	i	t	-	73	167	305	19	1945	51	2571	-147	42	ie	19
n	i	t	-	85	169	329	49	2197	277	2655	10	47	ie	70
r	i	t	-	82	175	324	-10	2250	518	2658	20	45	ie	16
r	i	t	-	83	167	313	0	2280	352	2592	-67	48	ie	36
s	i	t	-	52	139	308	-5	2110	120	2593	-210	37	ie	22
n	i	v	-	62	149	349	23	2108	629	2505	-17	40	ie	21
r	i	v	-	96	169	339	13	2217	584	2624	37	46	ie	7
s	i	v	-	95	156	317	-8	2097	487	2585	-349	46	ie	84
n	i	w	-	76	196	339	-30	2313	437	2694	70	49	ie	65
n	i	w	-	87	213	396	45	2135	445	2644	43	50	ie	90
d	i	z	-	83	167	300	-33	2148	149	2517	-110	45	ie	75
w	i	z	-	56	189	319	18	2091	99	2695	-23	49	ie	1

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
d	a	b	+	144	200	665	214	1330	-95	2537	78	54	aa	29
n	a	d	+	132	196	635	113	1351	15	2643	-121	53	aa	71
z	a	k	+	123	196	731	198	1315	-23	2245	-50	53	aa	65
z	a	k	+	129	175	737	280	1343	35	2219	-323	51	aa	69
S	a	l	+	121	196	731	238	1309	-164	2262	-131	48	aa	47
j	a	l	+	85	213	606	80	1137	-37	2222	-493	48	aa	74
j	a	l	+	86	204	632	91	1180	-120	2302	-129	48	aa	37
p	a	l	+	137	172	726	201	1233	160	2408	-164	54	aa	12
s	a	m	+	103	196	687	146	1271	148	2350	-260	48	aa	89
s	a	m	+	117	172	692	151	1238	62	2223	-232	49	aa	93
s	a	m	+	122	204	725	152	1360	132	2543	103	55	aa	96
b	a	n	+	119	200	654	81	1218	24	2613	138	58	aa	103
k	a	n	+	93	182	707	167	1304	-94	2685	121	51	aa	20
m	a	n	+	105	204	657	77	1190	95	2417	-105	54	aa	64
j	a	r	+	140	182	662	200	1505	-87	2366	-82	49	aa	23
m	a	r	+	160	200	648	80	1471	128	2512	20	54	aa	16
v	a	r	+	159	208	675	263	1436	309	2254	-258	52	aa	43
m	a	t	+	110	204	650	100	1204	55	2437	91	52	aa	73
m	a	t	+	118	227	687	317	1364	244	2590	238	50	aa	50
m	a	t	+	127	204	643	272	1372	156	2530	-8	49	aa	24
s	a	t	+	126	196	679	308	1293	-505	2591	-241	51	aa	18
t	a	t	+	134	169	699	379	1260	-28	2173	-279	51	aa	31
t	a	t	+	140	143	637	285	1283	-61	2424	-143	46	aa	56
v	a	t	+	160	204	672	203	1330	170	2573	112	58	aa	70
r	a	I	-	54	169	597	48	1514	-80	2545	-2	47	aa	5
p	a	X	-	119	213	662	103	1369	67	2411	-3	56	aa	53
p	a	X	-	131	204	715	143	1302	45	2393	-12	56	aa	62
r	a	X	-	109	167	714	104	1326	82	2518	-11	56	aa	80
r	a	X	-	113	161	674	121	1327	92	2324	-250	54	aa	82
r	a	X	-	138	204	668	90	1358	59	2582	-65	54	aa	15
n	a	h	-	97	156	655	70	1513	-22	2451	24	53	aa	1
d	a	k	-	73	167	664	215	1414	7	2337	47	57	aa	57
m	a	k	-	89	175	719	223	1234	48	2487	75	49	aa	60
v	a	k	-	121	200	792	257	1261	-3	2477	299	56	aa	75
z	a	k	-	142	161	677	203	1321	-108	2493	15	52	aa	102
h	a	l	-	118	149	677	150	1127	-49	2626	169	49	aa	35
h	a	l	-	133	123	626	36	1206	49	2234	-60	42	aa	21
j	a	l	-	95	208	685	118	1254	-79	2294	107	55	aa	81
m	a	l	-	116	204	634	81	1223	114	2605	236	56	aa	97
t	a	l	-	53	182	588	68	1235	9	2280	-82	52	a	101
t	a	l	-	79	189	666	177	1256	-57	2247	-250	50	a	62
&	a	m	-	*98**	147	691	*250*	1448	*41*	2525	*49*	47	aa	19
m	a	m	-	66	154	637	56	1275	204	2567	177	51	aa	76
s	a	m	-	145	137	647	274	1224	17	2541	106	47	aa	91
z	a	m	-	53	179	578	110	1233	15	2449	195	47	aa	28
&	a	n	-	80	149	735	75	1379	-96	2582	-74	46	aa	40
&	a	n	-	96	169	726	95	1329	21	2300	-64	47	aa	42
X	a	n	-	40	149	551	39	1419	-10	2466	52	40	aa	17
X	a	n	-	55	149	620	185	1272	18	2592	114	44	aa	25
X	a	n	-	59	179	628	162	1415	16	2450	-9	49	aa	86

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
X	a	n	-	126	145	649	164	1167	-26	2434	112	48	aa	68
m	a	n	-	47	172	649	90	1275	51	2575	45	48	aa	83
n	a	n	-	67	167	651	74	1413	109	2392	-104	46	aa	46
n	a	n	-	118	196	704	284	1323	292	2765	444	55	aa	99
s	a	n	-	72	167	682	178	1419	53	2610	48	47	aa	51
t	a	n	-	68	159	631	89	1263	11	2653	169	49	aa	4
t	a	n	-	77	185	693	174	1314	36	2355	28	50	aa	14
t	a	n	-	95	139	624	145	1246	-161	2272	-569	43	aa	38
t	a	n	-	131	145	622	188	1290	16	2502	106	43	aa	2
t	a	n	-	131	172	702	201	1226	-28	2143	-277	52	aa	66
b	a	r	-	73	196	600	103	1426	-13	2354	58	55	aa	101
d	a	r	-	57	164	592	95	1449	-35	2546	-49	51	aa	105
d	a	r	-	61	152	602	77	1471	143	2382	22	50	aa	3
d	a	r	-	62	145	564	115	1473	78	2375	24	46	aa	13
d	a	r	-	80	154	646	176	1464	59	2515	82	54	aa	84
d	a	r	-	92	143	669	235	1433	-44	2374	8	51	aa	11
j	a	r	-	105	172	681	144	1577	-66	2229	-214	55	aa	92
j	a	r	-	107	179	698	129	1491	-35	2220	-242	48	aa	32
j	a	r	-	190	152	634	310	1567	-249	2415	-315	48	aa	27
k	a	r	-	169	164	673	256	1508	-76	2305	-190	54	aa	94
m	a	r	-	47	154	624	20	1229	45	2447	9	48	aa	79
m	a	r	-	54	154	633	84	1399	93	2500	14	49	aa	48
m	a	r	-	64	175	603	146	1331	59	2481	-83	52	aa	6
m	a	r	-	65	208	625	40	1118	59	2645	4	52	aa	39
m	a	r	-	67	133	591	94	1320	137	2348	103	44	aa	106
m	a	r	-	67	149	546	91	992	-61	2611	42	47	aa	95
m	a	r	-	69	164	635	201	1397	102	2485	72	51	aa	59
m	a	r	-	80	179	596	168	1158	87	2341	27	54	aa	67
m	a	r	-	83	175	729	176	1289	107	2419	20	50	aa	77
m	a	r	-	139	196	717	244	1335	481	2449	106	49	aa	44
m	a	r	-	162	167	671	141	1378	91	2434	-28	56	aa	87
m	a	r	-	179	169	688	82	1522	212	2498	-43	57	aa	104
n	a	r	-	61	152	615	41	1373	74	2634	36	46	aa	61
n	a	r	-	64	156	634	80	1411	12	2557	11	50	aa	78
n	a	r	-	65	167	607	145	1452	41	2656	4	45	aa	90
n	a	r	-	87	182	600	79	1500	-47	2493	-44	49	aa	10
n	a	r	-	95	182	672	146	1406	276	2437	148	48	aa	98
p	a	r	-	117	175	694	149	1479	433	2241	-52	49	aa	45
r	a	r	-	89	147	633	115	1363	-103	2276	-234	53	aa	58
v	a	r	-	145	196	670	112	1457	165	2358	-133	57	aa	100
w	a	r	-	40	149	415	13	1407	-17	2622	33	43	aa	33
w	a	r	-	59	154	627	105	1386	-86	2144	-79	51	aa	49
m	a	s	-	56	161	590	7	1348	-43	2375	-120	46	aa	30
m	a	s	-	95	208	655	133	1257	64	2402	-203	55	aa	34
n	a	s	-	83	167	609	90	1470	26	2571	-90	48	aa	54
v	a	s	-	164	196	642	269	1300	22	2485	-18	52	aa	36
X	a	t	-	83	204	668	78	1302	-68	2540	-109	55	aa	52
l	a	t	-	129	227	690	205	1355	-101	2562	444	58	aa	26
n	a	t	-	91	152	616	235	1550	197	2572	15	49	aa	22
n	a	t	-	121	185	677	275	1353	-164	2605	64	48	aa	41
r	a	t	-	58	145	579	155	1308	11	2589	-74	47	aa	8
r	a	t	-	123	200	673	139	1240	135	2621	-64	53	aa	72
s	a	t	-	124	172	712	290	1236	4	2340	-93	50	aa	9
t	a	t	-	123	159	653	254	1302	106	2508	-25	52	aa	88
n	a	v	-	97	147	601	239	1390	201	2546	94	48	aa	63
n	a	z	-	60	175	543	59	1480	-56	2285	-213	45	aa	7

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
&	A	X	+	81	213	816	44	1365	6	2467	-48	51	a	42
d	A	X	+	92	154	678	165	1201	-48	2320	-112	54	a	25
t	A	X	+	72	222	692	105	1272	50	2267	-274	57	a	52
t	A	X	+	77	213	667	41	1275	-52	2226	-116	56	a	39
w	A	X	+	80	196	695	94	1161	103	2346	25	57	a	86
v	A	k	+	72	222	717	104	1166	45	2369	234	60	a	123
v	A	k	+	95	222	689	118	1234	-104	2183	75	60	a	64
v	A	k	+	101	238	816	194	1199	-120	2372	158	60	a	89
i	A	l	+	104	233	686	178	1164	-239	2590	-116	54	a	27
i	A	l	+	166	204	686	212	1079	-462	2553	-90	54	a	22
m	A	l	+	123	222	676	60	1022	-203	2622	-11	53	a	97
r	A	l	+	139	244	647	155	935	-7	2325	-256	57	a	95
t	A	l	+	101	213	683	94	1118	43	2329	-287	56	a	100
&	A	n	+	82	244	706	112	1061	-79	2581	55	53	a	23
&	A	n	+	89	213	721	72	1140	-39	2659	178	52	a	116
i	A	n	+	73	192	617	83	1248	-164	2578	46	51	a	105
m	A	n	+	130	161	650	210	1051	55	2904	529	55	a	36
n	A	n	+	97	204	670	135	1093	-12	2588	498	48	a	50

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
r	A	n	+	89	185	717	135	1055	-69	2311	186	54	a	60
h	A	r	+	66	213	683	155	1487	60	2201	-201	51	a	67
r	A	s	+	103	213	641	105	1155	-144	2371	-362	56	a	38
d	A	t	+	95	185	640	247	1313	-19	2572	29	56	a	33
d	A	t	+	104	192	623	183	1264	-144	2546	-10	57	a	28
d	A	t	+	120	185	689	267	1251	-129	2532	3	58	a	29
w	A	t	+	94	204	654	215	1129	55	2651	155	61	a	91
X	A	N	-	84	175	654	218	1134	222	2392	-151	47	a	87
X	A	N	-	91	172	655	153	1115	134	2262	-182	47	a	78
l	A	N	-	47	172	661	106	1230	60	2531	129	50	a	109
l	A	N	-	70	196	618	127	1163	77	2447	-28	55	a	70
l	A	N	-	71	182	675	148	1111	-7	2364	-25	53	a	47
d	A	X	-	82	172	698	134	1227	-4	2267	-409	52	a	43
d	A	X	-	87	189	635	146	1176	-32	2477	-128	57	a	5
r	A	X	-	65	192	816	-24	1264	-25	2579	-62	54	a	98
w	A	X	-	100	196	695	149	1152	90	2423	-102	59	a	84
&	A	d	-	81	182	739	205	1278	97	2499	-118	49	a	19
&	A	d	-	*102**	154	691	*112*	1134	*-33*	2455	*-27*	51	a	103
w	A	d	-	72	172	652	120	1234	22	2333	135	53	aa	55
r	A	k	-	60	179	640	159	1122	35	2287	-101	50	a	68
v	A	k	-	73	169	658	196	1139	54	2163	-54	54	a	72
v	A	k	-	77	182	694	205	1241	35	2278	109	58	a	8
v	A	k	-	92	179	696	281	1219	-18	2122	-83	56	a	30
#	A	l	-	109	196	644	73	1038	-144	2597	-35	59	a	1
n	A	l	-	92	161	642	180	1178	-40	2557	82	51	a	104
n	A	l	-	105	141	588	70	1031	-333	2443	-280	44	a	119
o	A	l	-	76	182	619	113	894	-108	2728	-20	53	a	93
o	A	l	-	*122**	161	512	*104*	1026	*-305*	2687	*-28*	52	a	108
r	A	l	-	141	227	656	189	936	-51	2687	381	58	a	82
t	A	l	-	62	200	595	80	1354	-121	2481	-54	58	a	80
t	A	l	-	91	172	630	230	1342	-107	2317	-395	50	a	56
v	A	l	-	122	182	615	150	943	20	2406	-183	53	a	15
t	A	m	-	61	149	617	86	1081	35	2768	270	45	a	13
d	A	n	-	31	161	583	56	1408	30	2426	50	47	a	4
d	A	n	-	50	167	610	143	1427	72	2562	65	50	a	110
d	A	n	-	54	167	573	98	1128	24	2483	69	49	a	113
d	A	n	-	56	182	633	149	1203	19	2713	129	47	a	49
d	A	n	-	65	143	614	125	1329	22	2390	-91	47	a	66
d	A	n	-	77	182	597	146	1057	-103	2746	196	51	a	69
d	A	n	-	84	143	601	97	1103	-1	2673	-10	45	a	10
k	A	n	-	49	189	574	52	1352	-12	2500	57	55	a	73
k	A	n	-	87	141	583	228	1225	-52	2681	321	46	a	14
k	A	n	-	93	200	623	171	1057	-103	2737	325	55	a	85
k	A	n	-	98	204	564	132	1131	-163	2421	-136	51	a	96
k	A	n	-	103	127	606	228	1120	-22	2383	-18	42	a	92
l	A	n	-	52	172	639	67	973	-29	2661	161	46	a	112
l	A	n	-	84	149	620	138	931	-25	2582	-41	51	a	120
m	A	n	-	88	164	635	121	1066	-2	2786	234	50	a	41
r	A	n	-	75	167	645	129	1029	1	2755	99	52	a	48
v	A	n	-	42	143	594	51	1122	69	2571	-10	42	a	45
v	A	n	-	46	156	607	121	1005	16	2743	207	47	a	121
v	A	n	-	47	154	600	61	1123	35	2570	83	47	a	76
v	A	n	-	48	152	562	101	1114	27	2786	111	44	a	81
v	A	n	-	49	167	576	102	1058	29	2760	266	46	a	74
v	A	n	-	50	167	557	43	1063	-30	2719	221	45	a	102
v	A	n	-	50	185	484	43	1086	24	2467	26	53	a	2
v	A	n	-	54	149	588	156	1027	28	2612	121	46	a	118
v	A	n	-	56	161	522	26	981	-132	2525	274	44	a	17
v	A	n	-	57	204	590	30	1086	-49	2824	244	51	a	106
v	A	n	-	58	164	597	129	1157	-31	2528	141	48	a	53
v	A	n	-	58	182	592	136	1016	-175	2579	295	46	a	61
v	A	n	-	65	141	552	126	1089	35	2598	9	46	a	88
v	A	n	-	66	175	583	133	1158	-80	2655	114	50	a	26
v	A	n	-	69	149	574	133	1004	-51	2755	606	47	a	3
v	A	n	-	69	167	639	113	1094	-31	2652	120	53	a	90
v	A	n	-	72	161	530	86	1053	-23	2684	51	42	a	58
v	A	n	-	79	159	589	167	1190	-22	2518	97	48	a	94
v	A	n	-	83	137	523	114	1058	-105	2371	-250	43	a	79
X	A	p	-	99	128	553	224	1037	-39	2486	196	38	a	55
r	A	p	-	57	167	641	125	1154	37	2447	-47	52	a	77
h	A	r	-	97	179	697	153	1331	27	2178	-255	50	a	114
v	A	r	-	131	147	627	58	1111	-30	2345	254	50	a	115
p	A	s	-	87	159	618	138	1042	10	2406	-82	52	a	122
w	A	s	-	79	169	613	146	1097	-48	2493	-178	52	a	40
d	A	t	-	48	149	567	69	1194	-3	2511	-69	45	a	44
d	A	t	-	55	152	462	74	1346	57	2783	36	47	a	107
d	A	t	-	55	167	576	91	1226	-40	2615	-90	50	a	51

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
d	A	t	-	56	145	517	82	1390	14	2596	-10	46	a 7
d	A	t	-	57	154	580	85	1343	-11	2571	-49	48	a 75
d	A	t	-	61	161	584	115	1179	26	2697	157	50	a 71
d	A	t	-	62	161	529	93	1178	-78	2598	-1	49	a 59
d	A	t	-	65	169	569	190	1329	-36	2593	-9	51	a 20
d	A	t	-	66	159	554	100	1282	-36	2677	-52	49	a 46
d	A	t	-	71	179	541	132	1244	-42	2654	44	49	a 63
d	A	t	-	72	161	457	47	1279	11	2803	-31	50	a 6
d	A	t	-	77	152	591	150	1263	-41	2667	77	47	a 54
d	A	t	-	81	159	608	231	1287	-8	2587	114	52	a 32
d	A	t	-	82	147	590	182	1218	-148	2763	20	49	a 99
d	A	t	-	83	169	573	171	1254	-68	2558	-78	52	a 35
k	A	t	-	62	164	626	168	1230	-20	2518	-50	49	a 16
k	A	t	-	81	149	575	204	1317	-118	2475	-8	48	a 21
k	A	t	-	97	169	636	256	1243	-39	2461	50	53	a 24
n	A	t	-	72	175	531	78	1293	56	2471	7	47	a 34
w	A	t	-	51	161	603	109	1212	-78	2360	-142	48	a 18
w	A	t	-	62	133	580	113	1074	24	2500	-37	46	a 11
w	A	t	-	71	132	590	183	1008	-28	2564	6	44	a 12
w	A	t	-	72	143	583	163	1110	92	2434	-135	49	a 117
w	A	t	-	80	182	622	210	1026	94	2474	-48	55	a 83

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
d	E	N	+	83	185	496	100	1837	394	2528	-59	52	e 3
d	E	N	+	83	189	554	138	1642	428	2501	-49	48	e 58
d	E	N	+	86	149	561	123	1725	393	2564	-44	48	e 117
d	E	N	+	95	192	579	219	1799	352	2597	-1	51	e 52
z	E	X	+	92	169	478	23	1697	153	2505	-231	49	e 110
h	E	b	+	64	217	618	177	1568	133	2520	-17	54	e 39
l	E	f	+	103	227	642	209	1595	94	2450	7	58	e 54
&	E	k	+	66	213	672	86	1727	124	2471	48	52	e 10
m	E	l	+	99	196	585	114	1296	81	2776	46	54	e 95
t	E	l	+	70	189	587	77	1287	76	2172	156	53	e 91
t	E	l	+	74	167	602	112	1393	33	2458	50	51	e 93
t	E	l	+	76	182	629	217	1415	95	2446	18	51	e 44
w	E	l	+	121	200	598	198	1161	41	2697	78	60	e 98
w	E	l	+	170	152	628	171	1315	-78	2506	92	52	e 45
#	E	n	+	77	217	673	95	1463	183	2370	-13	57	e 13
X	E	n	+	55	200	588	120	1425	140	2120	13	47	e 11
h	E	n	+	69	222	586	113	1705	260	2580	149	54	e 104
k	E	n	+	77	233	630	177	1410	-115	2627	184	55	e 65
l	E	n	+	69	152	509	111	1500	77	2554	28	50	e 114
l	E	n	+	78	159	515	95	1554	81	2576	26	47	e 85
m	E	n	+	75	189	624	115	1617	176	2563	93	54	e 133
#	E	r	+	146	182	664	1	1630	-53	2404	-79	55	e 16
t	E	r	+	55	143	560	34	1635	-121	2517	97	40	e 22
w	E	r	+	77	204	566	65	1646	181	2609	-46	59	e 115
w	E	r	+	107	217	609	183	1543	140	2478	-5	50	e 88
b	E	s	+	101	204	584	114	1583	54	2432	-50	57	e 14
s	E	s	+	95	227	566	110	1442	-109	2251	-379	51	e 8
w	E	s	+	84	227	590	162	1387	11	2495	-266	50	e 120
w	E	s	+	90	196	590	95	1522	93	2503	-119	52	e 128
d	E	N	-	85	156	512	97	1717	35	2499	27	48	e 123
d	E	N	-	88	137	507	106	1603	110	2591	68	43	e 83
r	E	N	-	112	128	540	135	1439	-149	2480	54	40	e 89
r	E	X	-	92	196	614	90	1655	56	2467	-101	52	e 70
w	E	X	-	89	185	504	65	1531	135	2355	-78	48	e 121
z	E	X	-	66	192	528	38	1661	89	2635	-45	51	e 113
z	E	X	-	70	196	558	49	1674	74	2577	-81	54	e 32
z	E	X	-	72	179	546	34	1716	97	2542	-98	49	e 49
z	E	X	-	86	149	512	115	1578	77	2598	-102	46	e 130
h	E	b	-	75	179	604	162	1699	146	2496	-46	56	e 35
k	E	b	-	61	189	573	109	1593	96	2456	90	52	e 75
n	E	b	-	38	164	558	82	1537	120	2592	114	46	e 100
n	E	b	-	55	185	549	115	1535	121	2601	145	48	e 69
r	E	b	-	76	164	657	107	1528	3	2508	51	50	e 131
S	E	f	-	85	172	535	142	1606	87	2392	-104	48	e 19
l	E	k	-	71	145	524	100	1515	9	2348	-24	45	e 67
r	E	k	-	88	175	587	162	1469	184	2485	-134	51	e 37
X	E	l	-	97	169	646	193	1160	-138	2670	-1	52	e 72
s	E	l	-	50	161	578	98	1375	25	2422	-2	50	e 21
s	E	l	-	84	141	563	144	1441	14	2368	-18	42	e 40
t	E	l	-	68	179	612	149	1397	-27	2323	-104	50	e 55
t	E	l	-	69	182	594	73	1315	-4	2417	-27	54	e 2
t	E	l	-	81	169	613	104	1406	18	2324	-45	50	e 33
t	E	l	-	92	139	592	215	1444	51	2377	-67	46	e 46

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
w	E	l	-	124	222	607	159	979	-190	2747	180	59	e	94
z	E	l	-	146	179	628	233	1219	14	2557	-203	50	e	50
n	E	m	-	62	143	590	50	1499	163	2502	104	44	e	63
#	E	n	-	22	139	366	71	1443	30	2562	-6	40	e	71
#	E	n	-	27	139	451	95	1552	47	2288	51	40	e	97
#	E	n	-	28	141	564	78	1451	23	2489	-18	44	e	122
#	E	n	-	30	133	300	68	1471	8	2489	28	42	e	59
#	E	n	-	33	145	554	96	1481	5	2527	134	43	e	80
#	E	n	-	35	145	505	69	1612	-106	2457	35	43	e	101
#	E	n	-	40	147	597	113	1514	-36	2476	193	46	e	109
#	E	n	-	45	132	583	91	1485	-16	2566	-66	39	e	112
#	E	n	-	49	141	599	131	1616	109	2406	67	45	e	36
#	E	n	-	51	135	563	24	1567	120	2493	97	44	e	86
#	E	n	-	62	152	590	100	1537	30	2508	-115	47	e	111
#	E	n	-	65	161	581	72	1595	87	2550	116	48	e	41
#	E	n	-	71	149	597	114	1553	106	2534	139	47	e	31
#	E	n	-	72	182	665	38	1597	41	2619	42	49	e	134
#	E	n	-	75	137	589	76	1585	144	2498	11	43	e	116
#	E	n	-	79	152	654	61	1622	82	2503	-3	50	e	15
X	E	n	-	78	133	647	94	1596	16	2541	-13	43	e	132
d	E	n	-	36	145	527	71	1366	42	2564	-30	41	e	118
d	E	n	-	43	189	575	9	1367	-26	2402	-7	47	e	126
d	E	n	-	50	159	486	58	1453	18	2707	-15	49	e	66
f	E	n	-	86	152	646	125	1516	97	2605	86	40	e	20
i	E	n	-	28	182	548	92	1568	-15	2426	13	41	e	79
i	E	n	-	61	154	514	71	1620	-154	2415	-142	47	e	24
k	E	n	-	60	141	530	111	1572	44	2499	4	45	e	23
l	E	n	-	71	161	539	94	1669	56	2494	0	51	e	25
m	E	n	-	59	167	609	113	1500	122	2454	189	47	e	103
m	E	n	-	67	217	619	159	1502	177	2404	210	51	e	68
m	E	n	-	68	196	574	97	1560	83	2480	134	53	e	29
m	E	n	-	77	189	582	107	1507	66	2507	109	53	e	26
m	E	n	-	77	208	600	115	1461	46	2636	231	54	e	42
m	E	n	-	81	204	578	107	1493	90	2449	56	50	e	47
m	E	n	-	83	182	564	137	1512	165	2546	214	49	e	38
n	E	n	-	60	154	543	131	1510	63	2651	60	42	e	78
n	E	n	-	71	156	586	123	1507	135	2702	29	47	e	108
n	E	n	-	72	141	556	104	1394	60	2570	5	43	e	9
p	E	n	-	77	143	513	124	1416	37	2487	56	46	e	5
r	E	n	-	70	169	570	63	1275	-33	2634	-216	44	e	17
s	E	n	-	63	167	534	52	1439	30	2524	45	47	e	56
s	E	n	-	70	167	564	178	1632	99	2578	-59	42	e	30
s	E	n	-	79	145	601	228	1600	124	2546	110	44	e	124
s	E	n	-	88	164	625	137	1501	104	2660	-49	47	e	125
s	E	n	-	102	156	572	163	1535	95	2646	-83	45	e	53
t	E	n	-	55	152	487	57	1505	51	2627	-2	45	e	57
t	E	n	-	60	149	541	97	1511	164	2565	122	46	e	73
t	E	n	-	78	141	611	167	1536	216	2597	144	44	e	12
#	E	r	-	92	149	567	45	1468	-183	2401	-126	46	e	129
i	E	r	-	72	145	454	35	1566	-227	2595	-110	47	e	96
n	E	r	-	79	179	666	181	1570	50	2620	39	53	e	64
o	E	r	-	89	156	592	92	1404	143	2243	-60	47	e	92
p	E	r	-	79	167	510	66	1506	138	2482	43	47	e	18
s	E	r	-	62	204	447	41	1414	104	2757	112	52	e	119
v	E	r	-	75	152	461	93	1519	64	2580	-22	47	e	34
v	E	r	-	87	182	543	93	1419	107	2519	-29	50	e	99
w	E	r	-	85	161	583	74	1557	221	2556	70	51	e	127
w	E	r	-	120	145	578	166	1444	508	2519	127	45	e	62
d	E	s	-	54	196	578	103	1566	71	2666	-35	46	e	87
s	E	s	-	90	172	545	117	1590	-34	2513	-289	42	e	81
m	E	t	-	35	152	513	63	1405	87	2491	76	43	e	102
m	E	t	-	41	152	560	114	1407	49	2608	55	48	e	107
m	E	t	-	48	137	556	62	1438	94	2359	39	45	e	61
m	E	t	-	49	130	470	70	1514	60	2504	51	42	e	74
m	E	t	-	53	161	550	39	1514	100	2382	-7	46	e	84
m	E	t	-	55	169	513	84	1286	82	2392	-95	51	e	7
m	E	t	-	62	192	558	206	1456	232	2554	100	49	e	106
m	E	t	-	80	182	567	125	1469	259	2473	133	53	e	27
m	E	t	-	87	179	568	154	1478	267	2500	304	52	e	90
n	E	t	-	77	233	624	248	1573	105	2634	12	54	e	76
z	E	t	-	70	213	533	111	1572	65	2073	-444	49	e	82
z	E	t	-	101	204	593	206	1491	86	2507	89	56	e	28
#	E	w	-	55	149	582	34	1577	79	2386	-21	48	e	6

APPENDIX D:

FORMANT VALUES AND EXCURSION SIZES

This appendix contains durations (ms), formant values measured with method Formant (Hz, see chapter 2), and excursion sizes (Hz) calculated as $\Delta F = -3/2 \cdot P_2 - 5/8 \cdot P_4$*, in which* P_n *is the Legendre polynomial coefficient of order n (see chapters 4 and 7). Next to these values that were used in our study we included* F_0*,* F_3*, and the RMS energy in dB. All vowel realizations were sorted on accented versus not-accented, post-vocalic consonant (*C_2*), pre-vocalic consonant (*C_1*), and duration (dur), in that order. The contents of the last two columns, i.e. base and number, correspond to the codes used in appendix C.*
An asterisk is attached to the durations of vowel realizations that could not be segmented reliably. The corresponding values that are dubious as a result of the problems with the segmentation are written in italic.

Vowel symbols: y-y, u-u, &-´, o-o, i-i, a-a, A-A, E-E, O-O, I-È
Consonant symbols: N-N

These data are available in ASCII format (MS Dos, Macintosh, VAX/VMS) upon request to the author.

Speaking rate: Normal

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
s	y	p	+	74	204	374	68	1395	78	2251	-137	46	uu 3
s	y	p	+	78	233	314	66	1380	53	2120	-215	45	uu 2
d	y	r	+	168	185	367	21	1159	218	2543	-280	47	uu 12
n	y	t	+	94	182	323	30	1664	-62	2405	-96	47	uu 10
X	y	w	+	107	133	261	-80	1203	64	2381	-59	40	uu 4
n	y	b	-	92	143	256	39	1170	294	2292	33	46	uu 6
s	y	k	-	54	143	289	-117	1486	22	2206	-255	46	uu 1
f	y	n	-	69	149	291	-12	1420	-47	2268	-116	44	uu 5
d	y	r	-	193	175	372	20	1484	167	2386	-135	45	uu 7
t	y	r	-	76	143	303	-47	1623	-35	2298	-95	47	uu 11
d	y	s	-	67	141	284	-33	1702	88	2422	-10	41	uu 9
t	y	w	-	37	130	287	-11	1675	3	2411	-4	40	uu 8

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
X	u	d	+	77	213	383	26	695	-242	2559	310	48	oe 12
&	u	f	+	208	172	344	-105	721	-657	2645	131	47	oe 9
t	u	f	+	96	132	353	-28	953	-575	2432	-206	43	oe 16
X	u	j	+	78	204	380	35	783	-447	2477	88	45	oe 13
b	u	k	+	71	213	367	-56	647	-257	2436	155	44	oe 6
b	u	k	+	87	169	357	-36	744	-225	2509	-1045	52	oe 7
b	u	k	+	97	185	388	14	795	-189	2447	-136	50	oe 5
b	u	k	+	104	161	348	-92	702	-204	2503	-102	50	oe 8
z	u	k	+	120	196	419	-55	811	-405	2293	-315	54	oe 14
X	u	j	-	48	189	374	-4	776	-166	2460	15	47	oe 2
X	u	j	-	53	169	339	4	819	-155	2368	59	50	oe 1
z	u	k	-	62	208	332	-1	769	-232	2408	-41	43	oe 4
n	u	m	-	67	147	374	21	937	-123	2408	67	46	oe 3
n	u	m	-	88	128	338	45	798	-184	2523	214	41	oe 10
m	u	t	-	79	185	370	26	781	-241	2514	194	45	oe 11
m	u	t	-	88	159	374	-13	865	-326	2477	71	46	oe 15

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
r	&	I	-	35	145	530	25	1428	-72	2430	-123	45	aa 5
d	&	r	-	60	169	376	13	1403	-93	2649	57	47	e 51
i	&	r	-	*181**	169	362	*-16*	1862	*465*	2477	*-41*	50	e 96
n	&	r	-	57	141	471	66	1416	38	2577	29	43	e 48
n	&	r	-	74	128	492	138	1487	159	2504	164	38	e 60
s	&	r	-	39	172	495	-158	1242	-3	2587	41	51	e 1
s	&	r	-	68	169	423	-38	1258	-74	2561	-102	48	e 119
t	&	r	-	32	122	391	30	1028	-214	2866	142	38	e 4
t	&	r	-	114	139	454	44	1496	246	2638	25	42	e 105
#	&	t	-	46	179	445	55	1523	37	2477	-169	48	schwa 5
#	&	t	-	51	130	376	19	1432	-102	2624	128	39	schwa 14
#	&	t	-	57	127	411	77	1415	-3	2443	94	40	schwa 13
d	&	t	-	56	167	403	32	1629	-64	2296	-247	49	a 9
i	&	t	-	59	130	390	72	1728	226	2506	99	45	schwa 16
i	&	t	-	*88**	175	366	*15*	2000	*127*	2557	*-26*	54	schwa 15
i	&	t	-	*105**	122	370	*106*	1936	*217*	2554	*-120*	43	schwa 6
m	&	t	-	29	139	395	127	1214	108	2462	99	42	schwa 9
m	&	t	-	40	127	398	26	1261	70	2505	72	44	schwa 2
n	&	t	-	50	123	398	70	1553	-1	2269	-273	42	schwa 1
n	&	t	-	50	127	444	77	1178	-40	2629	113	42	schwa 11
n	&	t	-	52	164	385	-6	1447	-55	2671	46	45	schwa 12
n	&	t	-	92	141	411	87	1599	146	2564	75	47	schwa 4
r	&	t	-	50	156	429	37	1102	18	2684	-10	48	schwa 7
r	&	t	-	61	172	502	62	1453	82	2586	10	48	schwa 10
s	&	t	-	46	204	403	12	1521	4	2532	-74	43	schwa 3
t	&	t	-	55	164	431	15	1264	-173	2802	192	48	schwa 8

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
h	o	X	+	160	417	441	-94	767	-249	2498	-35	61	oo 9
h	o	X	+	169	400	431	13	810	65	2534	67	61	oo 83
n	o	X	+	211	167	377	-85	779	-214	2585	116	52	oo 82
&	o	k	+	137	182	423	-21	894	91	2449	7	56	oo 54
k	o	k	+	104	182	378	69	728	-12	2419	129	52	oo 68
r	o	k	+	179	204	380	-23	715	18	2473	131	56	oo 41
t	o	k	+	92	189	386	19	874	-42	2512	-30	49	oo 29
t	o	k	+	175	217	430	-90	709	-386	2571	158	57	oo 62
X	o	l	+	138	172	430	21	803	-217	2409	254	51	oo 16
X	o	l	+	140	213	425	40	863	-73	2453	188	58	oo 3

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
n	o	m	+	113	172	400	167	868	-28	2351	451	46	oo	36
l	o	p	+	128	159	380	70	750	-14	2509	305	52	oo	8
b	o	r	+	213	179	361	-57	859	-266	2335	-85	59	oo	17
d	o	r	+	248	385	362	8	1092	-206	2397	-207	56	oo	46
h	o	r	+	227	182	349	-18	916	79	2349	-104	56	oo	60
s	o	r	+	145	161	336	-19	1006	-450	2217	-177	49	oo	2
v	o	r	+	145	192	384	26	1006	144	2284	196	52	oo	43
v	o	r	+	167	208	414	46	903	-21	2301	-280	48	oo	24
v	o	r	+	169	238	400	1	831	-349	2267	-150	48	oo	86
l	o	s	+	143	196	406	49	844	-471	2560	346	55	oo	84
&	o	v	+	174	179	385	-41	705	-355	2560	405	53	oo	89
&	o	v	+	188	385	401	-71	771	-169	2639	318	54	oo	78
b	o	v	+	209	400	436	25	696	-427	2574	307	59	oo	56
l	o	v	+	147	154	390	83	798	-266	2500	246	51	oo	19
n	o	v	+	129	152	375	3	831	-428	2535	92	50	oo	32
n	o	v	+	131	139	369	0	761	-546	2585	192	49	oo	59
n	o	v	+	166	370	370	-108	738	-582	2600	156	60	oo	61
z	o	A	-	*59**	152	522	*14*	918	*-17*	2610	*-20*	50	oo	71
z	o	A	-	*147**	164	489	*121*	904	*-289*	2862	*-100*	50	oo	63
z	o	A	-	*153**	169	510	*86*	854	*-169*	2663	*24*	52	oo	72
z	o	E	-	82	143	417	-13	945	-374	2297	-207	48	oo	47
s	o	S	-	81	137	408	-51	1045	-538	2225	-357	45	oo	37
m	o	X	-	125	152	471	8	815	-60	2799	49	50	oo	40
t	o	X	-	119	143	403	27	787	-214	2621	0	49	oo	85
t	o	X	-	148	200	405	10	814	-308	2771	-71	57	oo	33
r	o	b	-	81	143	370	71	777	-207	2651	237	44	oo	38
r	o	b	-	89	127	377	30	746	-101	2519	158	42	oo	22
r	o	b	-	104	139	375	-16	732	-213	2612	290	49	oo	52
n	o	d	-	161	141	418	70	875	-251	2579	468	48	oo	88
r	o	d	-	105	167	424	36	863	-208	2582	239	48	oo	31
X	o	i	-	100	120	331	-13	949	-212	2409	-8	45	oo	14
#	o	k	-	93	143	390	-16	734	-15	2423	-41	52	oo	81
X	o	l	-	115	204	423	18	823	-83	2626	122	55	oo	6
z	o	l	-	73	179	386	-14	905	-380	2161	47	49	oo	28
k	o	m	-	95	119	372	95	839	-30	2603	227	39	oo	26
n	o	m	-	84	143	472	122	924	-126	2687	210	48	oo	20
k	o	n	-	65	137	394	65	813	-108	2546	142	45	oo	35
s	o	n	-	38	130	334	27	894	-16	2285	136	41	oo	12
w	o	n	-	148	167	408	50	938	-48	2422	153	54	oo	45
z	o	n	-	128	167	501	58	960	-483	2237	-249	49	oo	5
l	o	p	-	126	167	387	72	832	-38	2507	343	51	oo	57
X	o	r	-	86	141	329	-22	978	-115	2587	57	42	oo	11
X	o	r	-	136	122	344	6	864	-531	2466	105	42	oo	13
d	o	r	-	99	167	366	4	979	-69	2422	-3	55	oo	76
d	o	r	-	115	130	378	25	947	-326	2237	21	46	oo	7
d	o	r	-	162	149	449	5	801	-370	2627	-28	51	oo	75
h	o	r	-	113	179	430	48	961	-179	2334	-55	53	oo	51
s	o	r	-	89	213	442	29	1248	-131	2257	-41	57	oo	10
s	o	r	-	130	208	431	29	1109	-192	2221	-180	53	oo	1
s	o	r	-	136	159	355	34	1160	-51	2292	-188	48	oo	44
t	o	r	-	70	161	340	-28	1094	-227	2586	-36	45	oo	27
t	o	r	-	112	137	398	-80	916	-288	2348	-10	47	oo	15
v	o	r	-	75	175	379	-35	953	-106	2482	227	47	oo	30
v	o	r	-	77	161	359	-56	997	-48	2466	-147	48	oo	34
v	o	r	-	93	139	355	39	862	-119	2404	317	43	oo	87
v	o	r	-	99	147	352	-99	833	-792	2469	-46	38	oo	79
v	o	r	-	99	149	424	-54	905	-175	2461	-67	49	oo	66
v	o	r	-	104	172	406	-133	846	-375	2462	-15	52	oo	53
v	o	r	-	105	154	361	-45	905	-343	2347	120	51	oo	58
v	o	r	-	113	167	372	-51	1125	154	2258	-74	50	oo	49
v	o	r	-	119	182	393	-10	851	-73	2490	-6	54	oo	69
v	o	r	-	138	152	372	45	963	-29	2316	188	45	oo	42
v	o	r	-	210	120	341	-47	846	-310	2231	-201	44	oo	4
v	o	r	-	236	156	333	-115	957	-202	2334	-83	50	oo	80
k	o	s	-	152	119	391	162	917	-260	2567	152	42	oo	64
k	o	s	-	158	118	455	119	927	-529	2373	-162	44	oo	65
k	o	s	-	176	119	443	96	902	-369	2532	107	41	oo	67
#	o	v	-	90	149	390	-56	768	-125	2505	98	52	oo	73
#	o	v	-	105	145	398	19	735	-455	2566	167	48	oo	23
#	o	v	-	112	161	392	-20	787	-406	2596	79	50	oo	21
#	o	v	-	113	169	361	-23	747	-69	2521	183	53	oo	55
l	o	v	-	116	149	423	18	839	-298	2445	407	51	oo	48
n	o	v	-	109	149	406	-6	838	-346	2515	-18	50	oo	70
n	o	v	-	113	145	396	-10	787	-305	2587	468	46	oo	25
r	o	v	-	83	204	410	-43	839	-91	2584	237	57	oo	74
s	o	v	-	130	161	396	50	808	-207	2389	-76	47	oo	18
t	o	v	-	85	141	372	14	791	-103	2400	182	47	oo	50

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
z	o	v	-	124	139	360	44	849	-619	2498	-327	44	oo 39
l	o	z	-	127	159	375	-27	949	-537	2393	-55	51	oo 77

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
r	i	#	+	81	111	263	14	2130	298	2504	25	29	ie 10
r	i	#	+	171	167	318	57	2290	612	2703	56	46	ie 7
n	i	&	+	181	179	330	-46	2313	357	2748	-11	46	ie 52
&	i	d	+	*162**	161	312	*21*	2343	*513*	2675	*252*	45	ie 42
r	i	h	+	104	244	313	-48	2187	491	2611	177	47	ie 54
z	i	n	+	94	204	366	27	2088	303	2522	-109	45	ie 89
r	i	s	+	110	156	308	20	2137	147	2569	-237	46	ie 73
r	i	s	+	113	200	357	-72	2083	106	2484	-116	52	ie 68
r	i	s	+	125	139	283	-36	2049	354	2587	-278	45	ie 71
b	i	t	+	92	169	305	27	2207	434	2553	91	47	ie 4
n	i	t	+	80	233	382	23	2056	337	2624	-34	50	ie 20
n	i	t	+	81	233	322	46	2091	396	2636	48	50	ie 60
n	i	t	+	84	182	346	71	2233	568	2657	132	45	ie 24
n	i	t	+	89	147	299	13	2175	557	2724	214	42	ie 25
n	i	t	+	89	222	360	37	2200	593	2686	80	49	ie 66
n	i	t	+	99	204	378	70	2262	861	2638	92	51	ie 85
n	i	t	+	123	222	269	-10	2249	440	2668	213	52	ie 91
n	i	t	+	128	137	321	46	1930	746	2488	102	45	ie 23
n	i	w	+	22	185	339	-19	1932	113	2536	67	47	ie 14
n	i	w	+	85	227	399	-2	2102	294	2665	-29	48	ie 48
n	i	w	+	89	172	295	27	2277	446	2754	291	48	ie 41
n	i	w	+	104	213	359	5	2044	651	2649	62	46	ie 35
n	i	w	+	105	222	352	-25	2149	306	2724	8	49	ie 34
n	i	w	+	155	204	360	27	2200	567	2569	48	51	ie 47
n	i	w	+	161	128	295	-49	1780	-454	2500	-175	44	ie 17
n	i	w	+	182	200	370	5	2222	-85	2625	179	49	ie 43
t	i	#	-	55	108	246	19	2227	113	2705	65	33	ie 37
d	i	&	-	77	130	262	-61	2181	230	2534	1	43	ie 92
d	i	&	-	*105**	122	268	*106*	2122	*217*	2601	*-120*	43	ie 29
d	i	&	-	*181**	169	322	*-16*	2172	*465*	2608	*-41*	50	ie 61
t	i	&	-	*88**	175	349	*15*	2223	*127*	2676	*-26*	54	ie 85
d	i	A	-	*72**	128	266	-77	2179	*-51*	2572	*-111*	41	ie 11
d	i	A	-	92	169	330	-77	2155	648	2602	4	55	ie 83
d	i	A	-	93	94	336	-51	2192	396	2549	10	38	ie 15
o	i	E	-	150	137	292	-95	2159	452	2549	31	44	ie 12
s	i	E	-	50	182	350	-7	2194	111	2666	16	48	ie 44
s	i	E	-	*50**	132	294	*-15*	2028	*-48*	2524	*-22*	43	ie 5
d	i	I	-	59	137	283	17	2056	50	2467	-6	44	ie 13
l	i	X	-	43	167	332	-22	2038	223	2626	56	44	ie 2
d	i	b	-	85	133	261	11	2125	524	2625	195	42	ie 62
b	i	d	-	69	164	323	2	2176	299	2518	29	46	ie 63
d	i	d	-	68	145	284	-7	2098	112	2478	-42	45	ie 87
d	i	d	-	77	145	293	35	2169	-96	2529	-37	43	ie 18
m	i	d	-	91	169	311	-9	2306	788	2566	163	48	ie 38
s	i	e	-	75	175	346	2	2005	125	2515	-26	44	ie 3
d	i	k	-	64	128	295	26	2076	38	2638	-542	43	ie 6
r	i	k	-	*88**	141	292	*27*	2370	*-46*	2695	*84*	43	ie 28
s	i	k	-	53	122	270	17	1930	38	2383	-189	38	ie 72
s	i	k	-	66	141	286	25	2145	118	2512	-27	45	ie 69
s	i	k	-	75	132	265	4	2026	57	2459	-214	43	ie 74
t	i	m	-	58	182	321	5	2276	376	2694	231	44	ie 88
X	i	n	-	50	154	397	35	1834	109	2445	56	41	ie 55
X	i	n	-	52	149	326	-20	2021	178	2533	-4	41	ie 86
X	i	n	-	55	137	292	54	1140	-331	2446	125	42	ie 59
b	i	n	-	52	161	313	-3	2243	448	2566	55	45	ie 49
d	i	n	-	60	130	266	8	2168	425	2621	21	39	ie 31
l	i	n	-	66	137	277	-27	2194	470	2578	79	43	ie 78
m	i	n	-	13	152	352	19	1765	148	2407	15	42	ie 80
m	i	n	-	33	149	362	13	1909	314	2459	26	43	ie 8
m	i	n	-	83	133	282	-1	1841	335	2357	10	46	ie 56
s	i	p	-	50	120	299	62	2112	165	2639	-50	35	ie 27
s	i	p	-	73	149	303	33	2034	142	2472	-100	46	ie 77
h	i	r	-	120	385	372	49	2183	1004	2569	189	54	ie 58
n	i	r	-	154	182	340	-16	2247	626	2637	79	50	ie 79
p	i	r	-	135	132	284	17	2115	240	2524	58	38	ie 51
d	i	s	-	74	135	272	16	2265	55	2641	-144	40	ie 32
d	i	s	-	78	167	301	-16	2359	381	2670	2	40	ie 39
m	i	s	-	65	164	331	-47	2003	452	2503	-95	41	ie 46
n	i	s	-	54	133	290	-28	1983	84	2496	-216	41	ie 9
n	i	s	-	62	156	311	-60	2164	205	2564	-89	44	ie 81
n	i	s	-	85	147	287	-44	2132	170	2546	-263	42	ie 26
n	i	s	-	97	132	316	-30	2057	161	2579	60	40	ie 82

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
s	i	s	-	51	169	333	5	2079	-37	2655	-117	35	ie 50
s	i	s	-	119	156	293	-13	2130	136	2475	-116	41	ie 64
t	i	s	-	73	204	390	51	2091	82	2587	-169	41	ie 53
t	i	s	-	117	175	291	1	2151	88	2541	-286	40	ie 30
t	i	s	-	118	192	347	25	2214	345	2582	-144	43	ie 33
d	i	t	-	49	123	266	61	1975	84	2534	-7	36	ie 40
n	i	t	-	58	145	305	15	2016	406	2464	-41	43	ie 57
n	i	t	-	70	175	319	18	2120	374	2610	78	44	ie 45
n	i	t	-	74	204	378	63	2094	522	2611	33	47	ie 70
n	i	t	-	77	145	308	10	2121	433	2619	56	42	ie 19
n	i	t	-	113	175	326	12	2135	794	2547	43	46	ie 67
r	i	t	-	57	161	327	-3	1489	112	2443	33	42	ie 16
r	i	t	-	78	161	295	3	2248	253	2588	-25	47	ie 36
s	i	t	-	77	127	272	28	2054	135	2489	-172	38	ie 22
n	i	v	-	67	137	305	12	1897	295	2399	5	44	ie 21
s	i	v	-	60	141	291	-19	2050	60	2455	-61	40	ie 84
n	i	w	-	88	189	335	21	2194	416	2596	248	47	ie 90
n	i	w	-	100	189	350	35	2239	572	2718	260	52	ie 65
d	i	z	-	109	156	302	-43	2122	152	2480	-278	45	ie 75
w	i	z	-	80	196	387	-4	1976	121	2413	-135	47	ie 1

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
p	a	X	+	148	182	680	200	1292	79	2552	84	53	aa 62
r	a	X	+	135	213	710	110	1368	72	2548	38	55	aa 15
d	a	b	+	194	175	662	267	1336	-97	2743	252	51	aa 29
n	a	d	+	141	161	637	204	1324	-45	2766	83	48	aa 71
z	a	k	+	136	159	677	220	1341	-76	2385	42	50	aa 69
z	a	k	+	148	200	663	192	1304	-126	2521	-7	51	aa 65
z	a	k	+	181	147	606	235	1332	-102	2538	151	50	aa 102
S	a	l	+	113	204	628	161	1473	-130	2442	22	49	aa 74
S	a	l	+	125	169	675	235	1221	-183	2503	-5	46	aa 37
S	a	l	+	133	152	700	374	1213	-240	2548	167	48	aa 47
S	a	l	+	156	182	648	221	1252	-234	2582	62	50	aa 81
p	a	l	+	157	156	666	197	1170	-4	2642	348	54	aa 12
s	a	m	+	125	182	681	187	1257	115	2667	268	50	aa 89
s	a	m	+	126	204	638	159	1289	118	2647	104	51	aa 96
s	a	m	+	132	169	674	192	1258	5	2653	157	51	aa 93
b	a	n	+	145	169	672	212	1206	-3	2667	100	54	aa 103
m	a	n	+	121	196	675	86	1215	58	2659	53	54	aa 64
t	a	n	+	152	164	637	171	1281	62	2734	47	49	aa 66
j	a	r	+	206	139	662	296	1541	-161	2465	-10	47	aa 23
j	a	r	+	225	139	604	292	1511	-297	2369	-128	46	aa 27
m	a	r	+	185	179	655	76	1394	230	2494	-28	54	aa 16
m	a	r	+	229	143	722	182	1405	221	2444	-92	51	aa 87
p	a	r	+	156	161	656	205	1515	287	2280	-67	50	aa 45
v	a	r	+	204	182	712	266	1480	468	2436	-112	50	aa 43
w	a	r	+	178	323	685	227	1488	288	2395	-177	56	aa 55
m	a	t	+	142	204	687	228	1282	200	2548	212	50	aa 50
m	a	t	+	161	196	662	216	1307	113	2706	472	51	aa 24
m	a	t	+	162	196	596	157	1252	184	2606	2	52	aa 73
s	a	t	+	146	147	639	303	1275	-295	2616	18	49	aa 18
t	a	t	+	159	147	667	344	1240	-159	2536	21	48	aa 56
d	a	&	-	57	137	571	227	1485	153	2432	64	43	aa 3
p	a	X	-	124	182	726	194	1337	97	2462	99	52	aa 53
r	a	X	-	137	149	666	145	1308	45	2401	-107	51	aa 80
r	a	X	-	147	154	672	192	1293	174	2463	-153	54	aa 82
n	a	h	-	116	143	647	102	1504	-53	2416	-23	51	aa 1
m	a	k	-	121	147	609	109	1267	24	2531	122	49	aa 60
v	a	k	-	134	204	637	144	1297	-94	2542	342	54	aa 75
h	a	l	-	118	112	620	96	1171	45	2461	-102	40	aa 21
h	a	l	-	171	127	603	218	1135	61	2498	-51	43	aa 35
m	a	l	-	150	189	592	60	1148	70	2633	167	53	aa 97
t	a	l	-	80	159	620	163	1318	-16	2350	-145	53	a 101
&	a	m	-	*114**	122	690	*376*	1407	*79*	2444	*67*	44	aa 19
s	a	m	-	132	127	624	180	1224	73	2568	135	46	aa 91
z	a	m	-	53	172	578	92	1205	50	2439	25	44	aa 28
#	a	n	-	69	135	644	76	1335	44	2385	-33	45	aa 46
#	a	n	-	84	133	709	241	1341	79	2483	35	46	aa 40
#	a	n	-	105	132	658	188	1305	7	2493	39	46	aa 38
&	a	n	-	81	167	711	191	1331	105	2585	185	46	aa 51
&	a	n	-	*187**	132	681	*334*	1318	*4*	2351	*27*	44	aa 42
X	a	n	-	58	159	547	104	1537	110	2295	29	45	aa 86
X	a	n	-	62	149	545	113	1269	40	2276	-46	41	aa 17
X	a	n	-	65	156	637	169	1212	-37	2615	66	46	aa 25
X	a	n	-	145	122	575	162	1089	-7	2428	13	42	aa 68
k	a	n	-	162	169	665	349	1371	-214	2699	366	51	aa 20

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
m	a	n	-	70	145	592	151	1238	42	2531	54	51	aa	83
n	a	n	-	151	179	648	160	1230	144	2810	172	54	aa	99
s	a	n	-	81	149	584	251	1159	-39	2299	-170	48	aa	4
t	a	n	-	151	137	584	162	1158	21	1970	-56	42	aa	2
t	a	n	-	152	156	624	259	1421	24	2543	171	47	aa	14
b	a	r	-	100	185	595	146	1420	82	2388	31	55	aa	101
d	a	r	-	77	132	608	156	1436	-22	2461	-99	42	aa	105
d	a	r	-	89	147	628	177	1461	13	2420	54	53	aa	84
d	a	r	-	102	149	622	205	1497	-150	2384	52	56	aa	57
d	a	r	-	130	141	671	226	1546	-131	2383	-66	52	aa	11
h	a	r	-	131	137	584	156	1475	331	2340	-63	45	aa	98
j	a	r	-	113	161	632	133	1597	-107	2293	-96	53	aa	92
j	a	r	-	212	154	650	295	1534	-156	2332	-148	47	aa	32
k	a	r	-	222	152	635	225	1469	-24	2330	-210	49	aa	94
m	a	r	-	64	159	638	51	1343	170	2483	-16	48	aa	44
m	a	r	-	67	149	627	103	1289	17	2404	-52	50	aa	48
m	a	r	-	78	156	595	114	1386	44	2367	-57	47	aa	30
m	a	r	-	83	143	525	84	1200	121	2161	-10	47	aa	67
m	a	r	-	84	132	619	206	1310	132	2479	154	45	aa	106
m	a	r	-	87	164	647	106	1471	86	2370	9	53	aa	59
m	a	r	-	89	154	613	195	1222	78	2393	-56	51	aa	95
m	a	r	-	95	139	565	37	1395	133	2342	16	48	aa	79
m	a	r	-	98	172	670	132	1351	194	2468	39	51	aa	39
m	a	r	-	106	196	682	269	1411	36	2414	-51	56	aa	6
m	a	r	-	131	182	627	69	1367	249	2434	29	51	aa	77
m	a	r	-	207	141	663	210	1511	152	2307	-40	50	aa	104
m	a	r	-	212	161	652	147	1463	224	2383	24	54	aa	76
n	a	r	-	77	156	620	62	1426	1	2405	20	48	aa	78
n	a	r	-	82	147	592	96	1487	44	2486	-41	51	aa	90
n	a	r	-	101	139	611	150	1392	156	2526	77	46	aa	61
n	a	r	-	101	169	633	128	1428	-137	2484	6	47	aa	10
r	a	r	-	137	135	652	239	1406	113	2280	-216	48	aa	58
v	a	r	-	182	179	672	186	1298	128	2425	-168	55	aa	100
w	a	r	-	65	141	406	38	1361	-37	2495	21	45	aa	33
w	a	r	-	141	139	660	143	1366	238	2146	-123	47	aa	49
m	a	s	-	156	192	646	134	1301	161	2593	28	53	aa	34
n	a	s	-	105	137	501	90	1269	-603	2114	-460	45	aa	54
X	a	t	-	93	169	724	146	1273	-32	2344	44	56	aa	52
l	a	t	-	116	208	720	177	1339	44	2617	92	54	aa	26
n	a	t	-	94	147	630	269	1483	16	2550	-46	50	aa	22
n	a	t	-	144	182	677	293	1339	-249	2661	106	47	aa	41
r	a	t	-	82	128	569	258	1194	-37	2480	-76	44	aa	8
r	a	t	-	161	196	624	171	1351	138	2571	28	54	aa	72
t	a	t	-	115	167	633	234	1285	14	2752	152	47	aa	88
t	a	t	-	152	154	654	270	1288	18	2592	46	49	aa	31
t	a	t	-	155	169	664	382	1327	70	2577	64	51	aa	9
v	a	t	-	164	196	684	313	1289	142	2578	157	50	aa	36
v	a	t	-	180	167	701	254	1318	190	2597	208	51	aa	70
n	a	v	-	143	128	568	314	1316	28	2327	-114	44	aa	63
d	a	w	-	74	132	572	148	1431	72	2328	-94	45	aa	13
n	a	z	-	70	185	570	105	1588	33	2497	-34	49	aa	7

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
&	A	X	+	93	204	927	50	1356	-12	2536	-87	51	a	42
d	A	X	+	63	159	695	88	1195	-5	2476	-133	55	a	25
d	A	X	+	78	196	626	101	1204	-10	2503	-33	57	a	5
d	A	X	+	97	169	686	142	1211	16	2410	-100	50	a	43
t	A	X	+	88	217	678	84	1315	15	2469	-41	55	a	52
t	A	X	+	94	189	736	207	1303	-71	2560	-36	54	a	39
w	A	X	+	107	182	704	136	1156	67	2509	9	54	a	86
v	A	k	+	89	169	685	177	1283	79	2115	88	57	a	8
v	A	k	+	91	189	730	156	1195	-55	2122	-53	57	a	64
v	A	k	+	101	222	664	273	1170	-1	2370	20	56	a	123
v	A	k	+	107	233	784	167	1158	-90	2320	46	60	a	89
v	A	k	+	114	161	657	192	1258	14	2278	74	54	a	30
i	A	l	+	108	233	690	50	1069	-120	2782	269	52	a	27
i	A	l	+	*163**	172	700	*-79*	1242	*-375*	2600	*288*	55	a	22
m	A	l	+	165	204	756	126	1096	-199	2719	104	55	a	97
r	A	l	+	179	233	629	162	936	-95	2719	148	55	a	95
t	A	l	+	115	222	719	120	1167	33	2502	1	54	a	100
#	A	n	+	79	200	412	64	1116	-76	2581	193	54	a	60
&	A	n	+	*82**	233	697	*145*	1173	*-111*	2699	*381*	55	a	23
&	A	n	+	106	169	712	135	1129	-31	2599	268	51	a	116
A	A	n	+	66	156	658	86	1039	-37	2807	213	47	a	50
i	A	n	+	84	182	651	141	1108	-190	2859	497	53	a	105
k	A	n	+	107	172	614	196	1160	-144	2668	334	50	a	96

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
k	A	n	+	109	189	609	123	1070	-162	2796	279	50	a	85
m	A	n	+	112	141	565	124	1069	-16	2728	239	49	a	41
m	A	n	+	129	141	600	176	1052	19	2781	392	52	a	36
h	A	r	+	80	204	684	39	1377	-45	2138	-200	50	a	67
h	A	r	+	135	182	654	116	1308	28	2292	-25	51	a	114
r	A	s	+	104	196	632	66	1134	-104	2555	-35	52	a	38
d	A	t	+	85	169	613	260	1289	-145	2559	100	53	a	33
d	A	t	+	98	182	642	180	1214	-92	2598	-9	55	a	28
d	A	t	+	120	185	662	225	1176	-152	2572	-105	55	a	29
w	A	t	+	87	200	635	155	1103	-36	2570	-38	60	a	91
d	A	A	-	84	147	576	147	1190	-68	2659	19	46	a	49
X	A	N	-	88	147	612	174	1118	61	2289	-20	49	a	78
X	A	N	-	112	167	646	169	1151	138	2344	11	50	a	87
l	A	N	-	74	189	658	104	1174	-41	2472	145	53	a	47
l	A	N	-	79	152	619	163	1183	53	2478	31	53	a	109
l	A	N	-	96	164	651	251	1173	108	2642	350	51	a	70
r	A	X	-	68	169	785	-45	1245	-6	2438	-72	54	a	98
w	A	X	-	104	167	654	124	1151	76	2521	5	53	a	84
#	A	d	-	86	156	732	86	1216	8	2522	-128	50	a	19
&	A	d	-	*109**	130	658	*271*	1161	*-30*	2508	*-163*	44	a	103
r	A	k	-	75	152	626	83	1176	14	2320	-78	46	a	68
v	A	k	-	111	139	648	228	1079	-34	2376	142	49	a	72
#	A	l	-	95	128	584	-4	1021	-220	2545	-133	42	a	119
#	A	l	-	128	182	637	65	1029	-207	2572	62	57	a	1
E	A	l	-	124	169	682	204	1160	-204	2817	289	52	a	104
o	A	l	-	*59**	152	549	*14*	973	*-17*	2590	*-20*	50	a	108
o	A	l	-	*147**	164	604	*121*	984	*-289*	2720	*-100*	50	a	93
o	A	l	-	*153**	169	595	*86*	945	*-169*	2689	*24*	52	a	111
r	A	l	-	118	192	614	104	981	6	2642	138	56	a	82
t	A	l	-	76	169	655	193	1327	35	2437	-93	49	a	62
t	A	l	-	92	200	635	155	1273	-188	2413	-53	57	a	80
t	A	l	-	111	149	595	160	1292	-238	2374	-96	49	a	56
v	A	l	-	136	175	652	223	1020	-26	2596	65	56	a	15
t	A	m	-	88	185	632	205	984	-44	2763	262	49	a	13
d	A	n	-	46	169	563	96	1347	-5	2429	-34	50	a	4
d	A	n	-	58	156	602	104	1274	-16	2583	32	51	a	110
d	A	n	-	75	145	581	66	1131	-100	2482	-116	50	a	10
d	A	n	-	79	130	618	208	1196	-28	2493	-93	43	a	66
d	A	n	-	87	135	583	188	1006	-168	2695	170	48	a	69
h	A	n	-	110	172	631	149	996	-50	2749	146	47	a	48
k	A	n	-	66	115	619	121	1079	-58	2497	120	42	a	14
k	A	n	-	76	143	612	152	1284	-81	2474	94	50	a	73
k	A	n	-	108	112	593	209	1089	-206	2489	177	40	a	92
l	A	n	-	80	147	597	167	1018	-34	2855	267	50	a	120
l	A	n	-	88	149	575	127	1048	0	2749	277	47	a	112
s	A	n	-	86	139	549	182	1069	-54	2603	99	46	a	113
v	A	n	-	38	135	462	79	1051	35	2380	-62	42	a	74
v	A	n	-	50	141	539	65	1153	0	2414	71	46	a	53
v	A	n	-	54	152	552	204	1066	-30	2592	240	45	a	37
v	A	n	-	55	159	479	14	980	-22	2332	-36	45	a	17
v	A	n	-	59	133	544	152	1037	-4	2478	167	47	a	26
v	A	n	-	59	135	503	113	982	-39	2372	25	43	a	76
v	A	n	-	62	133	471	102	1003	-20	2504	70	45	a	81
v	A	n	-	64	161	556	25	1057	-58	2731	23	48	a	106
v	A	n	-	66	135	545	156	1024	-27	2545	165	44	a	118
v	A	n	-	67	169	569	104	1098	-216	2778	72	44	a	61
v	A	n	-	69	130	510	153	980	-55	2684	177	44	a	121
v	A	n	-	69	182	592	82	1014	-20	2609	12	52	a	2
v	A	n	-	71	159	558	80	1105	-51	2733	177	48	a	102
v	A	n	-	73	137	530	101	1021	-15	2511	-6	43	a	45
v	A	n	-	76	147	532	86	998	-10	2560	70	46	a	3
v	A	n	-	87	130	574	123	1001	-119	2458	-25	46	a	90
v	A	n	-	102	152	553	117	1143	-76	2509	166	48	a	94
v	A	n	-	103	130	504	130	1015	-90	2254	-338	43	a	88
v	A	n	-	129	128	540	198	1061	-29	2439	-273	42	a	79
v	A	n	-	151	128	612	151	1113	-50	2409	224	45	a	115
X	A	p	-	81	114	538	242	1051	-51	2371	444	37	a	55
r	A	p	-	74	128	627	296	1105	-44	2544	-5	44	a	77
p	A	s	-	110	152	627	127	1017	-341	2473	-174	54	a	122
w	A	s	-	103	169	574	123	1051	-103	2527	-108	51	a	40
d	A	t	-	45	130	496	79	1147	17	2746	247	43	a	6
d	A	t	-	53	128	502	90	1383	-8	2347	-245	42	a	32
d	A	t	-	54	130	565	234	1219	-54	2747	117	42	a	46
d	A	t	-	56	143	470	92	1236	-43	2476	74	46	a	7
d	A	t	-	57	161	566	145	1319	-29	2689	5	51	a	63
d	A	t	-	58	135	489	99	1430	-10	2482	-86	40	a	65
d	A	t	-	65	172	553	151	1357	-91	2496	-66	52	a	20

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
d	A	t	-	69	133	485	59	1274	-51	2509	-92	45	a	34
d	A	t	-	69	137	558	115	1256	-27	2527	-32	47	a	54
d	A	t	-	69	139	497	158	1217	-110	2460	-121	46	a	107
d	A	t	-	69	141	561	108	1295	-7	2462	-139	45	a	75
d	A	t	-	70	132	523	108	1235	-46	2616	-46	44	a	51
d	A	t	-	70	141	490	149	1427	-20	2461	-27	46	aa	85
d	A	t	-	72	143	548	127	1271	-70	2560	-111	50	a	57
d	A	t	-	73	135	505	72	1441	-111	2377	-235	46	a	31
d	A	t	-	76	147	550	123	1119	-111	2540	-188	48	a	59
d	A	t	-	77	127	609	129	1129	-14	2532	-107	45	a	44
d	A	t	-	77	164	548	126	1261	-51	2582	-34	52	a	35
d	A	t	-	80	147	584	156	1222	-82	2547	-122	50	a	99
d	A	t	-	83	169	576	236	1241	41	2795	119	50	a	71
d	A	t	-	90	149	612	203	1163	-105	2586	-58	51	a	62
k	A	t	-	74	119	559	145	1230	-6	2369	27	43	a	21
k	A	t	-	74	145	590	160	1243	-176	2346	-39	49	a	16
k	A	t	-	79	141	584	177	1213	-54	2500	31	51	a	24
w	A	t	-	71	169	547	139	1270	-191	2317	-139	51	a	18
w	A	t	-	71	196	586	153	1094	-38	2561	-40	54	a	12
w	A	t	-	82	164	572	136	1140	117	2548	-33	55	a	117
w	A	t	-	91	123	587	205	1034	-50	2461	-7	44	a	11
w	A	t	-	101	156	630	299	1091	140	2557	45	54	a	83
v	A	z	-	57	143	475	46	1040	-111	2571	-134	40	a	58

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E	base	number
d	E	N	+	81	154	582	223	1705	324	2498	-31	52	e	117
d	E	N	+	90	145	553	142	1714	341	2464	-44	48	e	58
d	E	N	+	93	172	517	127	1858	150	2571	42	52	e	3
d	E	N	+	102	179	556	150	1780	382	2602	10	51	e	52
w	E	X	+	96	182	552	87	1672	208	2393	41	52	e	121
h	E	b	+	78	204	601	106	1599	65	2465	-29	59	e	39
l	E	f	+	111	200	599	110	1577	-97	2524	-51	56	e	54
#	E	k	+	73	185	613	-32	1697	27	2506	-112	48	e	10
m	E	l	+	135	175	604	126	1353	22	2750	226	50	e	95
t	E	l	+	91	185	613	139	1412	130	2565	-45	52	e	93
t	E	l	+	95	169	605	123	1371	71	2636	-43	51	e	91
t	E	l	+	107	179	646	280	1417	-16	2553	83	51	e	55
w	E	l	+	146	227	663	234	1082	-182	2775	281	56	e	94
w	E	l	+	157	200	607	197	1140	-34	2666	145	59	e	98
w	E	l	+	220	149	635	193	1355	43	2487	123	51	e	45
h	E	m	+	65	141	554	73	1503	152	2457	98	47	e	63
#	E	n	+	113	154	626	129	1563	44	2440	-34	53	e	15
X	E	n	+	61	208	516	126	1621	106	2545	171	45	e	11
h	E	n	+	109	222	564	185	1631	130	2610	59	53	e	104
k	E	n	+	76	145	532	95	1531	55	2442	22	43	e	23
k	E	n	+	79	213	573	200	1685	78	2585	134	55	e	65
l	E	n	+	74	208	539	125	1431	47	2645	121	48	e	85
l	E	n	+	78	145	478	98	1533	105	2522	35	49	e	114
v	E	r	+	146	204	582	144	1616	251	2508	-21	52	e	99
w	E	r	+	95	182	519	67	1669	182	2527	11	54	e	115
w	E	r	+	107	149	550	86	1604	251	2514	26	50	e	88
b	E	s	+	100	196	592	122	1498	62	2407	22	56	e	14
s	E	s	+	102	213	527	95	1516	14	2455	-61	50	e	8
w	E	s	+	100	164	587	182	1494	89	2464	-105	51	e	128
w	E	s	+	123	217	633	315	1413	-10	2275	-182	55	e	120
#	E	A	-	58	152	525	-16	1497	3	2614	-9	46	e	118
d	E	N	-	115	120	521	175	1676	321	2462	-63	42	e	83
d	E	N	-	121	139	533	171	1714	248	2441	13	47	e	123
r	E	N	-	94	119	564	177	1347	146	2340	-47	39	e	89
r	E	X	-	70	189	586	66	1678	22	2520	-57	49	e	70
z	E	X	-	72	133	510	82	1590	39	2423	-130	43	e	130
z	E	X	-	86	139	489	-64	1674	167	2481	-98	48	e	49
z	E	X	-	87	175	553	-27	1684	154	2479	-87	52	e	32
z	E	X	-	100	179	521	61	1591	97	2553	-96	49	e	113
z	E	X	-	141	145	501	123	1667	87	2461	-381	48	e	110
h	E	b	-	68	137	582	175	1512	143	2431	23	45	e	131
h	E	b	-	83	152	522	151	1707	380	2430	-51	51	e	35
k	E	b	-	75	204	530	80	1575	47	2425	46	50	e	75
n	E	b	-	90	149	574	225	1510	152	2524	69	43	e	69
S	E	f	-	127	167	511	141	1584	68	2458	95	48	e	19
l	E	k	-	72	135	512	112	1601	91	2435	-23	46	e	67
r	E	k	-	106	147	574	165	1469	4	2538	40	50	e	37
X	E	l	-	103	137	540	121	1366	-34	2557	64	45	e	72
s	E	l	-	80	118	531	153	1366	61	2120	-121	39	e	40
s	E	l	-	89	154	598	125	1395	19	2299	-104	50	e	33
t	E	l	-	62	141	554	114	1418	20	2381	-40	49	e	21

C_1	V	C_2	acc	dur	F_0	F_1	ΔF_1	F_2	ΔF_2	F_3	ΔF_3	E base	number
t	E	l	-	82	169	579	135	1402	59	2438	-18	51	e 2
t	E	l	-	83	156	587	219	1410	33	2415	-29	50	e 44
t	E	l	-	98	132	594	187	1426	54	2411	-44	48	e 46
z	E	l	-	166	149	618	232	1290	-241	2546	222	49	e 50
#	E	m	-	48	122	569	101	1534	33	2552	66	41	e 73
n	E	m	-	75	135	483	119	1436	229	2442	50	46	e 77
#	E	n	-	16	120	318	49	1442	19	2412	15	36	e 71
#	E	n	-	36	141	438	58	1390	-29	2248	-90	43	e 9
#	E	n	-	39	130	323	39	1484	27	2416	-36	38	e 108
#	E	n	-	45	149	571	85	1562	36	2490	-26	45	e 80
#	E	n	-	46	132	269	56	1370	54	2464	132	40	e 66
#	E	n	-	47	143	520	72	1521	53	2437	-3	45	e 112
#	E	n	-	47	152	584	51	1553	-59	2458	-88	46	e 101
#	E	n	-	49	141	581	59	1527	49	2500	2	45	e 125
#	E	n	-	52	130	582	24	1590	8	2550	-3	45	e 31
#	E	n	-	52	154	565	65	1649	129	2425	25	46	e 109
#	E	n	-	56	141	551	12	1396	119	2519	-8	42	e 116
#	E	n	-	58	154	551	59	1530	85	2328	-62	46	e 6
#	E	n	-	59	135	594	84	1567	-19	2431	-12	45	e 30
#	E	n	-	63	112	476	224	1494	206	2642	-13	36	e 59
#	E	n	-	63	137	592	85	1539	36	2533	60	45	e 56
#	E	n	-	64	154	597	122	1615	-11	2461	-77	50	e 111
#	E	n	-	65	137	610	109	1524	83	2407	63	46	e 122
#	E	n	-	72	141	614	109	1658	29	2513	56	45	e 124
#	E	n	-	73	137	585	48	1563	31	2460	-41	44	e 97
#	E	n	-	73	161	637	111	1542	18	2513	-22	52	e 134
#	E	n	-	74	137	619	101	1577	199	2458	34	43	e 132
#	E	n	-	77	127	547	73	1441	82	2473	105	40	e 86
#	E	n	-	87	128	572	71	1553	-117	2493	-186	41	e 87
#	E	n	-	90	141	613	188	1575	141	2474	132	50	e 36
#	E	n	-	96	147	623	111	1634	33	2474	-29	49	e 20
#	E	n	-	102	137	599	97	1527	4	2400	-11	48	e 17
#	E	n	-	106	143	549	55	1657	65	2495	-117	49	e 41
S	E	n	-	*91**	152	526	*127*	1594	*-3*	2536	*-105*	41	e 81
d	E	n	-	39	130	584	86	1489	43	2457	-45	43	e 126
i	E	n	-	*41**	128	496	*73*	1656	*6*	2275	*-38*	43	e 13
i	E	n	-	55	167	567	181	1651	162	2556	25	47	e 79
i	E	n	-	69	133	517	113	1739	-32	2523	-13	45	e 24
m	E	n	-	71	130	478	78	1411	75	2519	50	44	e 5
m	E	n	-	75	179	568	117	1567	164	2555	97	51	e 133
m	E	n	-	79	179	547	93	1497	90	2509	-130	52	e 103
m	E	n	-	81	167	579	176	1490	151	2615	221	49	e 68
m	E	n	-	84	169	538	166	1568	105	2618	160	52	e 26
m	E	n	-	85	179	568	147	1593	115	2484	69	54	e 29
m	E	n	-	86	175	557	154	1504	154	2494	164	50	e 47
m	E	n	-	98	123	563	215	1508	200	2489	141	40	e 38
m	E	n	-	99	182	558	105	1477	134	2539	9	52	e 42
n	E	n	-	48	145	522	122	1501	122	2458	87	45	e 100
n	E	n	-	88	127	512	109	1595	321	2498	54	44	e 43
s	E	n	-	101	149	526	134	1513	48	2495	-3	43	e 53
t	E	n	-	50	143	512	92	1419	22	2543	-36	45	e 57
#	E	r	-	89	156	566	-38	1533	-82	2307	-418	49	e 129
#	E	r	-	168	161	623	66	1675	73	2365	-67	55	e 16
d	E	r	-	100	137	547	87	1579	-46	2505	-11	41	e 22
n	E	r	-	83	161	628	108	1528	-88	2567	-8	48	e 64
o	E	r	-	98	143	614	54	1549	253	2258	-51	46	e 92
p	E	r	-	92	156	503	84	1446	161	2281	2	50	e 18
v	E	r	-	95	133	424	64	1482	-14	2366	-113	45	e 34
w	E	r	-	89	164	546	79	1478	110	2473	59	52	e 127
w	E	r	-	171	141	541	156	1434	392	2492	187	44	e 62
m	E	t	-	55	132	436	71	1463	118	2317	-180	44	e 102
m	E	t	-	59	161	509	74	1361	132	2405	51	50	e 7
m	E	t	-	63	119	487	140	1475	88	2510	79	40	e 74
m	E	t	-	64	185	522	148	1482	325	2477	118	48	e 106
m	E	t	-	65	149	525	141	1576	75	2507	-90	44	e 84
m	E	t	-	66	128	499	171	1486	54	2486	23	41	e 61
m	E	t	-	68	149	549	88	1491	86	2455	-18	47	e 27
m	E	t	-	69	133	504	102	1437	57	2307	82	45	e 90
m	E	t	-	82	139	437	96	1593	189	2484	121	45	e 107
n	E	t	-	78	238	600	136	1434	59	2614	-66	51	e 76
z	E	t	-	71	192	522	81	1473	23	2575	-67	47	e 82
z	E	t	-	84	182	557	237	1485	60	2556	30	51	e 28
#	E	w	-	70	196	443	44	1317	22	2391	54	53	e 12

Rob van Son

SPECTRO-TEMPORAL FEATURES OF VOWEL SEGMENTS

Rob van Son (1960) graduated in biology at Nijmegen University in 1984. He subsequently worked at the University of Amsterdam for the Dutch national SPIN program *Analysis and synthesis of speech* and for the ESPRIT project *Polyglot* at Nijmegen University. His current research at the University of Amsterdam, on a grant of the Dutch organisation for Scientific Research, concerns variation in the production of consonants.

Current theories in phonetics about vowels are deceptively simple. Vowel identity is fully determined by the position of the first two frequency peaks in the spectrum and to a lesser extent by vowel duration. However, under different conditions of e.g., context and stress, these values will frequently vary in highly systematic ways. What controls this variation? How do listeners cope with it? Theories that try to answer these questions are critically evaluated and a series of experiments is presented that test them. Finally, a unifying view is presented that explains best the current data.

Summary

In this thesis we have investigated several aspects of the spectro-temporal structure of vowel segments, both concerning vowel production as well as vowel perception. Chapter 1 contains a summary of current models on vowel production and perception. Models of vowel pronunciation try to explain why vowel realizations vary so much in natural speech. It is known that vowel production is influenced in highly systematic ways by context, stress, and speaking style (among others). The classical explanation is that of the target-undershoot model. This model states that vowel articulation is limited by the speed of the articulators (e.g., jaw, tongue, lips). Each vowel has a unique target-position for each of the articulators which will produce the ideal, or canonical, realization of that vowel. When vowel realizations are very long, there is ample time for the different articulators to reach their respective target positions. However, when vowel duration is short and the context forces the articulators to cover relatively large distances, there is not enough time and the articulators are stopped short of their targets. The resulting vowel realizations show "undershoot" in their articulatory movements as well as in the resulting formant frequencies, hence the name of the model: target-undershoot.

The classical quantitative study of Lindblom (1963) on the relation between vowel duration and formant-undershoot is discussed in depth. It showed that formant-undershoot increased exponentially with a decrease in vowel duration. However, subsequent studies gave ambiguous results. Some studies did find clear evidence for articulatory- and formant-undershoot. Others showed that there were numerous cases were no relation between vowel duration and target-undershoot could be found. Especially, changes in stress and speaking style could bring about changes in duration that were not accompanied by changes in target-undershoot. In our opinion, these conflicting results can be explained by assuming that target-undershoot is planned by the speaker. In this view, the undershoot serves a purpose that depends on factors like context, prosody, and speaking style. From this it follows that, irrespective of vowel duration, the undershoot itself should not change if the purpose of the undershoot does not change and vice versa.

Considering the conflicting reports in the literature, it seems that any test of the target-undershoot model should introduce changes in vowel duration without changing stress, speaking style, or other prosodic factors that were known to cause ambiguous results. In this study, we settled for changes in speaking rate. A long, meaningful text, read at a normal and at a fast rate, would induce a speaker to use the same stress assignments and the same "style" of speaking, irrespective of reading speed. At the same time, a difference in speaking rate would change the duration of all the vowels. In this study (chapters 2-4), we used all realizations of seven different vowels and some realizations of the schwa (/ˊ/). If vowel duration could control formant-undershoot all by itself, then an increase in speaking rate should induce an increase in undershoot. However, if formant-undershoot is planned, then a change in speaking rate should not necessarily result in a change in formant-undershoot.

In chapter 2, we measured formant frequencies in the *vowel kernel*. Vowel realizations uttered at the normal speaking rate were compared to the corresponding realizations uttered at the fast speaking rate. No spectral vowel reduction was found that could be attributed to a faster speaking rate. There was also no change in the amount of coarticulation or stress-induced reduction as a result of speaking rate. The only systematic effect was a higher F_1 value in fast-rate speech irrespective of vowel identity. This possibly suggests a generally more open articulation of vowels, speaking louder, or some other general change in speaking style by our speaker when he speaks fast.

In chapter 3 we looked at the effects of speaking rate on *vowel formant track shape*, using the same material as in chapter 2. The formant track shape was assessed on a point-by-point basis, using 16 samples at the same relative positions in the vowels. Differences in speaking rate only resulted in the same uniform change in F_1 frequency already found in chapter 2. Within each speaking rate, there was only evidence for a weak leveling off of the F_1 tracks of the open vowels /A a/ with shorter durations. When considering sentence stress or vowel realizations from a more uniform, alveolar-vowel-alveolar context, these same conclusions were reached.

In chapter 4 we again looked at the effects of speaking rate on formant track shape. This time we used a more elaborate method for assessing formant track shape. Legendre polynomial functions were used to model and quantify the shape of time normalized formant tracks. No differences in these normalized formant track shapes were found either that could be attributed to differences in speaking rate. A uniform higher F_1 frequency in fast-rate speech relative to normal-rate speech was found. Within each speaking rate, there was only evidence of a weak leveling off of the F1 tracks of the open vowels /E A a/ with shorter durations. Again, as in chapter 3, separately inspecting vowel realizations from a more uniform, alveolar-vowel-alveolar context, did not alter our conclusions.

The target-undershoot model of vowel production inspired a complementary model of vowel perception (Lindblom and Studdert-Kennedy, 1967). As vowel formant tracks will systematically undershoot the canonical target values in natural speech, it was suggested that listeners would compensate for this undershoot automatically by systematically overshooting the formant frequencies actually reached in perception, i.e. perceptual-overshoot. Early studies with synthetic speech did indeed find this kind of perceptual-overshoot. However, it showed to be rather difficult to prove the existence of an automatic mechanism for perceptual-overshoot in natural speech.

At the moment, there are two classes of models on vowel perception. The first class are models with dynamic-specification. In these models it is assumed that listeners use dynamical information from the Consonant-Vowel and/or Vowel-Consonant transitions to improve the recognition of the, stationary, vowel nucleus. Perceptual-overshoot is just one of such models. The second class of models is based on the assumption that a single, spectral, cross-section of the kernel of a vowel realization contains all information necessary to recognize it. In these models the vowel on- and offset transitions are of minor importance in vowel recognition.

The difference between these two types of models is the position of the Consonant-Vowel transition (in the vowel on- and offset). Is it used in vowel recognition, as is stated by models using dynamic-specification, or is it not, as stated by target models? There is evidence for perceptual-overshoot in synthetic speech. It is also known that presenting syllables without a vowel kernel, i.e. with only the vowel on- and offset transitions, hardly impairs vowel recognition. Still, there is no undisputable proof that the recognition of isolated, monphthongal, vowel segments is improved by adding dynamical information to the formant tracks. Exactly such an improvement is expected when listeners use dynamic-specification of vowels.

In natural speech, the amount of variation in durations, vowel formant frequencies and track shapes is limited. These various types of variation are furthermore strongly correlated. It is therefore better to use synthetic speech, for which it is possible to control all features. With synthetic speech, it is also possible to detach formant track shape from formant frequency. This way, the effects of formant track shape can be studied independently of vowel identity and vowel duration. We therefore choose to use synthetic speech to study how vowel duration and formant track shape influence vowel identity. Especially we looked for any evidence for perceptual-overshoot. The result of this study is presented in chapter 5 (see below). In chapter 6 we took a closer look at the existing literature in order to try to find an explanation for the disagreement between our results and those presented in several earlier papers.

In chapter 5 we used synthetic vowels to investigate whether listeners use vowel duration and formant track shape to determine vowel identity. The synthetic vowels had level or parabolically-shaped formant tracks and variable durations. They were presented in isolation as well as in synthetic CVC syllables. There was no evidence of perceptual compensation for expected target-undershoot due to token duration or context. The only asserted effects of duration and context were in the number of long- and short-vowel responses. There was also no evidence that the listeners used the formant track shape or slopes independently to identify the synthetic vowel tokens. Tokens with curved formant tracks were generally identified near their formant offset frequencies.

The results of chapter 5 contradicted claims made in the literature about the way listeners use dynamical information to identify vowel realizations. The literature on vowel perception itself also contains contradictory claims regarding the use of information from CV-transitions in vowel recognition. Our own experiments showed that the information in formant track shape was not always used to compensate for formant-undershoot. In chapter 6 a re-evaluation of the literature is attempted. A closer study of the most relevant papers shows that evidence for compensatory processes, i.e. perceptual-overshoot and dynamic-specification, was only found when vowel realizations from different, and appropriate, context were contrasted. Some studies show that vowel recognition deteriorated when vowel segments were presented out of context. Together, these facts suggest that the presence of an appropriate context is essential for any perceptual compensation of coarticulatory changes. This speculation might be used as a starting hypothesis for further research on vowel perception.

Finally, in chapter 7 we summarize and discuss our findings. We recapitulate the methods used in chapters 2 to 4 to study the effects of speaking rate on formant-undershoot. We argue that, under the circumstances used, any excess undershoot due to an increase in speaking rate should have been detectable, but did not show up. We therefore conclude that, for our speaker, speaking rate did not influence the amount of vowel formant-undershoot or the formant track shape. Therefore, we can conclude that changes in vowel duration alone do not change the amount of target-undershoot and that the undershoot that does occur is probably planned.

The listening experiments presented in chapter 5 showed that our listeners did not use a perceptual-overshoot mechanism or dynamic-specification to help them identifying the synthetic vowel tokens. In general, they seemed to use the offset part of each vowel realization to identify it. We therefore conclude that listeners do not automatically and unconditionally compensate for the formant-undershoot that can be predicted from the formant track shape.

Samenvatting

Beschrijving en identificatie van klinkers lijkt een simpel probleem te zijn. Wanneer klinkers echter door machines herkend moeten worden, of omgekeerd, wanneer machines klinkers moeten produceren, dan wordt de complexiteit van dit probleem al snel duidelijk. Klinkers zoals *aa*, *ie* of *oe* kunnen articulatorisch beschreven worden met slechts drie parameters: 1) de mate waarin de mond open is; 2) de positie van de tong, voor/boven of achter/beneden; 3) de mate waarin de lippen getuit zijn. Bij de klinker *aa*, zoals in *vaas*, is de mond zo ver mogelijk open, ligt de tong "middenin" de mond en zijn de lippen gespreid. Bij de klinker *ie*, zoals in *fiets*, is de mond (bijna) gesloten, ligt de tong vóór in de mond en (bijna) tegen het verhemelte en zijn de lippen gespreid. Bij de klinker *oe*, zoals in *voet*, is de mond ook gesloten, maar ligt de tong achterin de mond en zo ver mogelijk van het verhemelte en zijn de lippen getuit. Deze drie klinkers zijn het meest extreem wat betreft de positie van onderkaak, tong en lippen (de articulatoren). De andere Nederlandse klinkers liggen ertussenin.

Wanneer het geluid van klinkers onderzocht wordt lijkt de zaak in eerste instantie zelfs nog simpeler te worden. Klinkers worden onderscheiden op hun klankkleur (naar analogie van het timbre van muziekinstrumenten). De klankkleur van een klinker kan grotendeels beschreven worden met slechts twee frequenties, die van de eerste twee resonanties van de mond-keelholte. Deze resonanties worden formanten genoemd (F_1 en F_2). De klinkers *ie* en *oe* hebben de laagste waarde voor de F_1 en respectievelijk de hoogste en de laagste waarde voor de F_2. De klinker *aa* heeft de hoogste waarde voor de F_1 en een gemiddelde waarde voor de F_2. Wanneer de frequentie van de tweede formant uitgezet wordt tegen de frequentie van de eerste formant dan vormen de *ie*, *oe* en *aa* de hoekpunten van een driehoek. De waarden voor de formanten van alle andere klinkers (b.v. *uu*, *oo*, *o*, *ee*, *e*, *eu*) liggen binnen deze klinkerdriehoek.

Voor langgerekte klinkers, zoals 'aaaaaaah' of 'ooooooooh' geldt nu een heel simpele regel: bij elke klinker hoort een unieke waarde voor de F_1 en F_2. Als men weet wat voor klinker uitgesproken is, dan weet men ook vrij nauwkeurig wat de waarden van de eerste twee formanten zullen zijn. Omgekeerd, als men de waarden van de eerste twee formanten kent, dan weet men ook wat voor klinker er uitgesproken is.

Helaas is het in werkelijkheid niet allemaal zo eenvoudig. Omdat de formanten resonanties zijn van de mond-keelholte, zijn ze afhankelijk van de grootte van mond en keel. Dit wil zeggen dat de frequenties van deze formanten anders zullen zijn voor mannen, vrouwen en kinderen. En ook binnen deze groepen zijn de individuele verschillen groot. Deze variatie kan berekend worden en men kan ervoor corrigeren. Na correctie, of wanneer men de spraak van één enkele spreker bekijkt, geldt de eenvoudige, één-éénduidige relatie tussen langgerekte klinkers en formantfrequenties weer.

Nu is het verleidelijk om deze eenvoudige relatie tussen formantwaarden en (langgerekte) klinkers door te trekken naar normale spraak. Dit blijkt echter niet zomaar te kunnen. Er zijn verschillende processen die roet in het eten gooien. Allereerst is er een proces dat *coarticulatie* genoemd wordt. Als je goed luistert, dan hoor je dat de *a* uit *kar* niet hetzelfde klinkt als die

uit *tas*. De formantwaarden die men kan meten voor deze twee realisaties van de *a* zijn ook duidelijk verschillend. Het lijkt in deze gevallen of de medeklinkers die om de klinker staan, de formantwaarden ervan in de richting van een zeer specifieke frequentie 'trekken'. Als men alle mogelijke combinaties van klinkers en medeklinkers onderzoekt, dan blijkt dat er een grote spreiding bestaat in de formantwaarden van dezelfde klinkers. Het komt relatief vaak voor dat de ene klinker in de ene contekst dezelfde formantwaarden heeft als een andere klinker in een andere contekst. Zonder de contekst te kennen is het vaak niet meer mogelijk om te voorspellen wat de formantwaarden zullen zijn van een klinker en andersom is het niet meer mogelijk om uit enkel de formantwaarden te bepalen om welke klinker het gaat.

Er is nog een tweede proces dat de klankkleur en daarmee de formantwaarden van klinkers verandert. Dit proces wordt *reductie* genoemd. In dezelfde omgeving van medeklinkers, klinkt de *a* uit *kabouter* toch anders dan die uit *kabbelen*. Er is ook een verschil in formantwaarden. Het verschil tussen *kabouter* en *kabbelen* is te wijten aan woordklemtoon. In *kabouter* zit de *a* in een onbeklemtoonde lettergreep, in *kabbelen* in een beklemtoonde. Naast woordklemtoon en zinsaccent speelt ook de stijl van spreken een rol. Als iemand een tekst voorleest dan praat hij/zij anders dan wanneer hij/zij een ongedwongen gesprek voert. Gemiddeld genomen liggen de formantwaarden van onbeklemtoonde klinkers en klinkers uit ongedwongen conversatie meer in het midden van de klinkerdriehoek dan de beklemtoonde klinkers en de klinkers uit voorgelezen teksten. Het lijkt erop alsof de formantfrequenties gemiddeld naar het centrum van de klinkerdriehoek getrokken worden. Onbeklemtoonde klinkers en klinkers uit ongedwongen conversatie zijn *gereduceerd* ten opzichte van beklemtoonde klinkers en klinkers uit voorgelezen tekst.

Coarticulatie en reductie zijn twee verschijnselen die de klankkleur van klinkers sterk en systematisch veranderen. Als gevolg hiervan is het niet meer mogelijk om op grond van alléén de formantfrequenties de identiteit van de klinker te achterhalen (machinaal of automatisch klinkers herkennen is moeilijk). Toch blijkt dat menselijke luisteraars er weinig moeite mee hebben om klinkers in welke contekst dan ook te herkennen. Met betrekking tot klinkers zijn er nu twee vragen waarop een antwoord gezocht wordt. Ten eerste, hoe verandert de klankkleur van klinkers als gevolg van contekst, klemtoon en spreekstijl? Met andere woorden, welk mechanisme zit er achter coarticulatie en reductie? Ten tweede, hoe zijn luisteraars in staat een klinker te herkennen ondanks het feit dat de klankkleur door coarticulatie en reductie sterk verandert? Over deze twee vragen gaat dit proefschrift.

In hoofdstuk 1 van dit proefschrift wordt een overzicht gegeven van de relevante literatuur. We bespreken o.a. de klassieke studie van Lindblom uit 1963. In deze studie vindt Lindblom dat er een relatie is tussen de duur van een klinker en de formantwaarden. Lindblom formuleerde een model waarbij de duur van een klinker de mate van coarticulatie, en indirect die van reductie, bepaalde. Dit model wordt het 'target-undershoot' model genoemd ('het doel niet bereiken'; er is geen goede Nederlandse vertaling voor deze term). Dit model gaat uit van het feit dat de menselijke articulatoren,

zoals onderkaak, tong en lippen, tijd nodig hebben om de bewegingen te maken die nodig zijn om medeklinker-klinker-medeklinker reeksen uit te spreken. De snelheid waarmee deze organen bewogen kunnen worden is beperkt. Als er te weinig tijd is, kunnen de noodzakelijke bewegingen niet meer afgemaakt worden, het doel wordt gemist, en er ontstaat coarticulatie. Lindblom beweerde nu dat de duren van de klinkers in normale spraak eigenlijk al te kort zijn om nog perfect uitgesproken te kunnen worden. Later werd dit model genuanceerd door te stellen dat de mate waarin de klinkerduren te kort zijn, afhankelijk is van de inspanning die de spreker zich getroost om de klinkers goed uit te spreken. Ook bij deze nuancering blijft echter gelden dat de duur van de klinkers de mate van coarticulatie en reductie bepaalt.

Sinds het target-undershoot model werd geformuleerd zijn er diverse studies uitgevoerd waarvan de resultaten dit model ondersteunden, maar ook studies die het model tegenspraken. Het bleek bijvoorbeeld, dat onbeklemtoonde klinkers best even lang kunnen zijn als beklemtoonde klinkers, terwijl ze toch gereduceerd zijn. Het kwam ook voor dat klinkers wel korter werden, maar zonder dat er meer coarticulatie of reductie optrad. Het lijkt zeer wel mogelijk dat zowel coarticulatie als reductie (ten dele) 'bewust' uitgevoerd worden, en dat de relatie tussen coarticulatie, reductie en klinkerduur ontstaat doordat de duur van een klinker meestal ook korter wordt in omstandigheden die leiden tot coarticulatie en reductie. Dit betekent dat de articulatoren wel degelijk sneller kunnen bewegen dan dat ze normaal doen en dat het mogelijk moet zijn om spraak uit te lokken met veel kortere klinkers maar zonder extra coarticulatie en reductie. Om nu de geldigheid van het target-undershoot model te onderzoeken moet de klinkerduur variëren terwijl alle andere factoren die kunnen leiden tot verschillen in coarticulatie en reductie (zoals klemtoon, contekst, spreekstijl e.d.) hetzelfde blijven. Een van de manieren waarop dit bereikt kan worden is door een spreker te vragen een tekst voor te lezen, eerst in een normaal tempo, daarna zo snel mogelijk. Dit is de methode die wij voor ons onderzoek gekozen hebben.

In de hoofdstukken 2 tot en met 4 hebben wij onderzocht of er inderdaad meer coarticulatie en reductie optreedt wanneer klinkers korter worden in snelle spraak. We gebruikten daarvoor een normale tekst die twee keer werd voorgelezen door een profesionele nieuwslezer, eerst in een normaal tempo, daarna snel. We gebruikten zeven van de twaalf Nederlandse klinkers (de *oe*, *oo*, *a*, *aa*, *e*, *ie* en *uu*) en verder enkele realisaties van de schwa (de *uh* klinker uit *'t* en *d'r*). De klinkers waren zo gekozen dat ze goed verspreid lagen over de 'klinkerdriehoek'. In hoofdstuk 2 onderzoeken we de formantwaarden in het midden van iedere klinkerrealisatie. Het bleek dat er geen noemenswaardig verschil was tussen de formantwaarden in snelle en normale klinkers. Voor alle klinkers was er een lichte stijging in de frequentie van de eerste formant die misschien het gevolg is geweest van een verschil in luidheid. Het kan zijn dat onze spreker niet alleen sneller maar ook harder is gaan praten. Niets wijst er echter op dat er ook maar enig verschil in coarticulatie of reductie is tussen de normaal en de snel gelezen versie van de tekst.

Aangezien de formantwaarden gerelateerd zijn aan de positites van de articulatoren, kunnen verschillen in de bewegingen (meestal) teruggevonden worden door de formantwaarden te volgen in de tijd. Als de vorm van de formantsporen (d.w.z. de sporen van de formantfrequenties in de tijd) verschilt tussen normale en snelle spraak, dan moeten ook de bewegingen van de articulatoren verschillen. Als de vorm van de formantsporen, na normalisatie voor duur, niet verschilt tussen normale en snelle spraak, dan is het onwaarschijnlijk dat de bewegingen van de articulatoren wel verschillen.

In de hoofdstukken 3 en 4 gebruiken we twee verschillende methoden om vormverschillen in formantsporen te onderzoeken, na eerst voor de klinkerduur gecorrigeerd te hebben. In hoofdstuk 3 gebruiken we een rechttoe-rechtaan methode om te onderzoeken of er verschillen zijn tussen begin, midden en eind van iedere klinker. In hoofdstuk 4 gebruiken we een meer geavanceerde methode (hogere orde curve fitting) om te kijken of de formantsporen vlakker worden in snelle spraak. Geen van beide methoden toont enig verschil tussen normale en snelle spraak aan. Hieruit moet geconcludeerd worden dat er wel een verschil is in de *snelheid* van de bewegingen van de articulatoren in normale en snelle spraak van onze professionele spreker, maar geen verschil in het *verloop* van de bewegingen.

In hoofdstuk 7 lichten we nogmaals toe dat het target-undershoot model een meetbare toename van coarticulatie en reductie zou hebben voorspeld in snelle spraak. Wij vinden echter geen verschil. Hieruit moet geconcludeerd worden dat onze spreker in staat is sneller te spreken zonder *extra* coarticulatie en reductie (d.w.z. boven de normale variatie) en dat deze verschijnselen dus niet direct afhangen van de klinkerduur. Hieruit volgt dat het waarschijnlijker is dat coarticulatie en reductie actief geregeld worden door onze spreker en dat zij niet het gevolg zijn van de passieve traagheid van zijn articulatoren.

In het perceptieve deel van dit proefschrift (hoofdstukken 5 en 6) onderzoeken we hoe luisteraars klinkers identificeren. Uit het voorgaande moge gebleken zijn dat klinkers nogal variëren wat betreft klankkleur. De vraag is nu hoe luisteraars toch in staat zijn deze variabele klanken te identificeren. In hoofdstuk 1 zijn de twee belangrijkste theoriën op dit punt besproken. De eerste theorie is op het eerste gezicht de eenvoudigste. Deze stelt dat de formanten in het midden van de klinker genoeg informatie bevatten om de klinker te identificeren. Men kan er echter niet meer mee volstaan om de eerste twee formanten te gebruiken (F_1 en F_2), maar men moet ook de derde formant (F_3) en de grondtoon, d.w.z. de toonhoogte (F_0) gebruiken. Tevens is het noodzakelijk om formanten aan elkaar te relateren (bijvoorbeeld, F_3-F_2 i.p.v. de afzonderlijke formanten).

De tweede theorie stelt dat de problemen ontstaan door de gevolgen van coarticulatie. Als bekend is hoe de coarticulatie 'gericht' is, dan kunnen de gevolgen ongedaan worden gemaakt. De richting en mate van coarticulatie kunnen bepaald worden door de formantsporen aan het begin en einde van een klinker te bekijken, daar waar klinker en medeklinker elkaar raken. Door de hellingen van de formantsporen aan het begin en eind te extrapoleren kan men een goede schatting maken van de 'doelwaarde', d.w.z. de

waarde zonder coarticulatie. Het hypothetische proces in het menselijke gehoor dat daarvoor moet zorgen wordt 'perceptual-overshoot' genoemd ('waarnemingscorrectie door extrapolatie', deze term is ook al niet goed te vertalen). Het gehoor trekt, als het ware, de bewegingen van de formanten, en daarmee die van de articulatoren, door. Het principe dat de vorm van de formantsporen in het overgangsgebied tussen naburige medeklinkers en klinkers een rol speelt bij de identificatie wordt 'dynamische specificatie' genoemd. Er zijn veel artikelen geschreven over de voors en tegens van deze twee theoriën over klinkeridentificatie, maar tot nog toe spreken de resultaten elkaar tegen.

In hoofdstuk 5 onderzoeken wij de effecten van de vorm van formantsporen op de identificatie van klinkers. Dit is gedaan door synthetische klinkers aan luisteraars aan te bieden. In deze synthetische klinkers werden de duur van de klinker en de vorm van de formantsporen (hele en halve parabolen) gevarieerd. Wij kunnen duidelijk aantonen dat onze luisteraars de formantsporen in onze synthetische klinkers niet extrapoleren naar een hypothetische waarde zonder coarticulatie. Integendeel, in plaats van de bewegingen in de klinkeraanzet te extrapoleren wordt over het laatste deel van de klinker gemiddeld.

In hoofdstuk 6 gaan we dieper in op de tegenstrijdige resultaten in de literatuur, alsook in die van ons eigen onderzoek. In experimenten van anderen waarin aanwijzingen gezocht worden voor het bestaan van perceptual-overshoot en dynamische specificatie, worden de reacties van luisteraars onderzocht op klinkers met verschillende typen formantsporen. Het blijkt echter dat de vorm van de formantsporen niet de enige factor is die gevariëerd werd in die experimenten. Telkens wanneer aanwijzingen gevonden worden voor perceptual-overshoot en dynamische specificatie blijkt ook de contekst gevarieerd te zijn. Die experimenten zijn op een dusdanige manier uitgevoerd dat compensatie voor de coarticulatie als gevolg van de context en extrapolatie van formantsporen precies hetzelfde resultaat zouden hebben gehad. Het blijkt ook dat het weglaten van de contekst de verstaanbaarheid van klinkers zeer nadelig beïnvloed. In hoofdstuk 7 concluderen we uit de resultaten van ons eigen onderzoek (hoofdstuk 5) dat perceptual-overshoot en dynamische specificatie niet zonder meer volgen uit de vorm van de formantsporen. Samen met de resultaten van anderen, zoals gedetailleerd beschreven in hoofdstuk 6, suggereert dit dat een 'passende' contekst noodzakelijk is voor compensatie van de effecten van coarticulatie. Het kan zelfs zijn dat de contekst op zichzelf al voldoende is om een luisteraar aan te zetten tot het compenseren van eventuele effecten van coarticulatie.

Acknowledgements

I would like to thank dr. A.C.M. Rietveld of the Catholic University of Nijmegen, the Netherlands for performing the sentence-accent labelling of the speech material used in this study and D.R. van Bergem of the University of Amsterdam for providing the method for measuring vowel formant values at near stationary positions of the realizations. The text used was selected by dr. W. Eefting of the State University of Utrecht, and the speech was recorded by her and dr. J. Terken of the Institute of Perception Research, Eindhoven. This research project was part of the Dutch national program *Analysis and synthesis of speech* funded by the Dutch program for the Advancement of Information Technology (SPIN).

Name Index

Subject Index