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 /ˈXrywˊl/, 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).

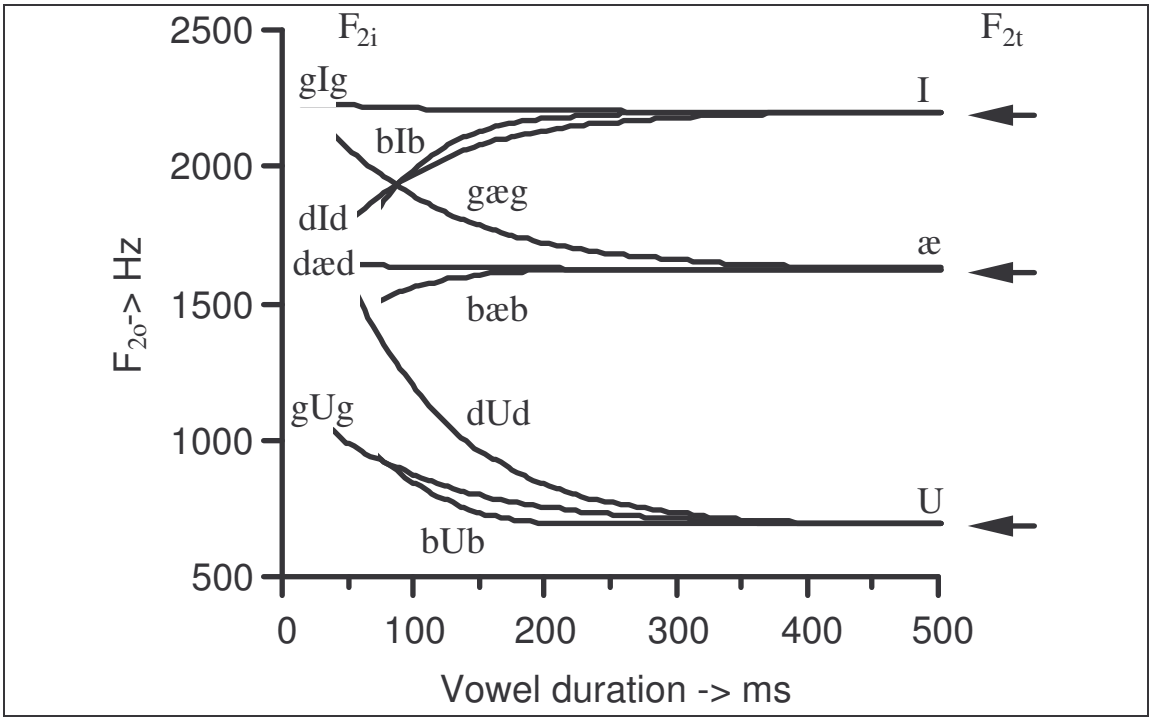

Figure 1.1. The effect of vowel duration on F_2 target-undershoot.
The relation between vowel mid-point value (F_{2o}) and vowel duration, as described by equation 1.1, is illustrated in this example taken from Lindblom (1963). The vowel formant target values (F_{2t}) are indicated by the arrows on the 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.

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

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.

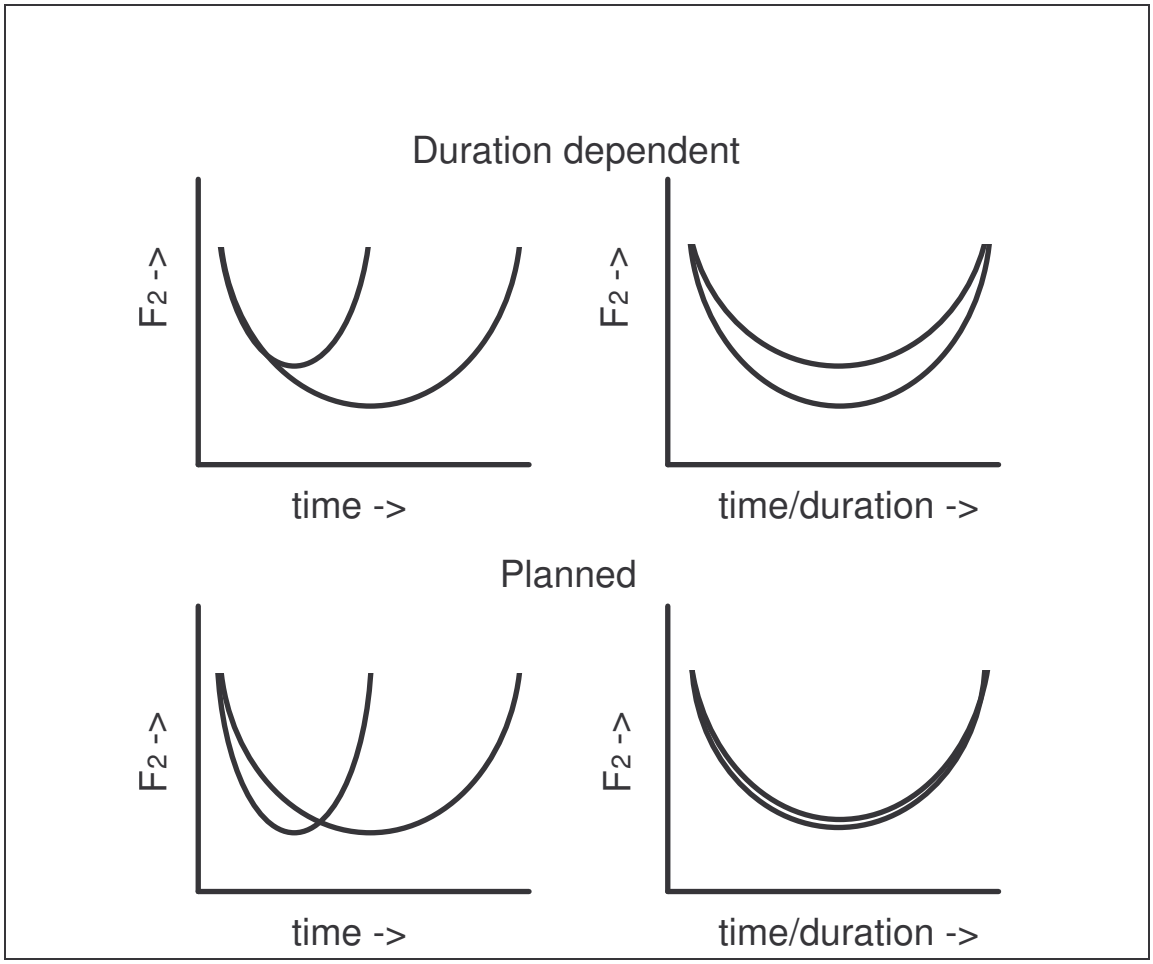

Figure 1.2. The influence of vowel duration on formant track shape. Tracks from two vowels with different durations are depicted for an input-driven model, i.e. duration-dependent undershoot (with excess undershoot, top row), and for an output-driven model (with no excess undershoot, bottom row). The 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.

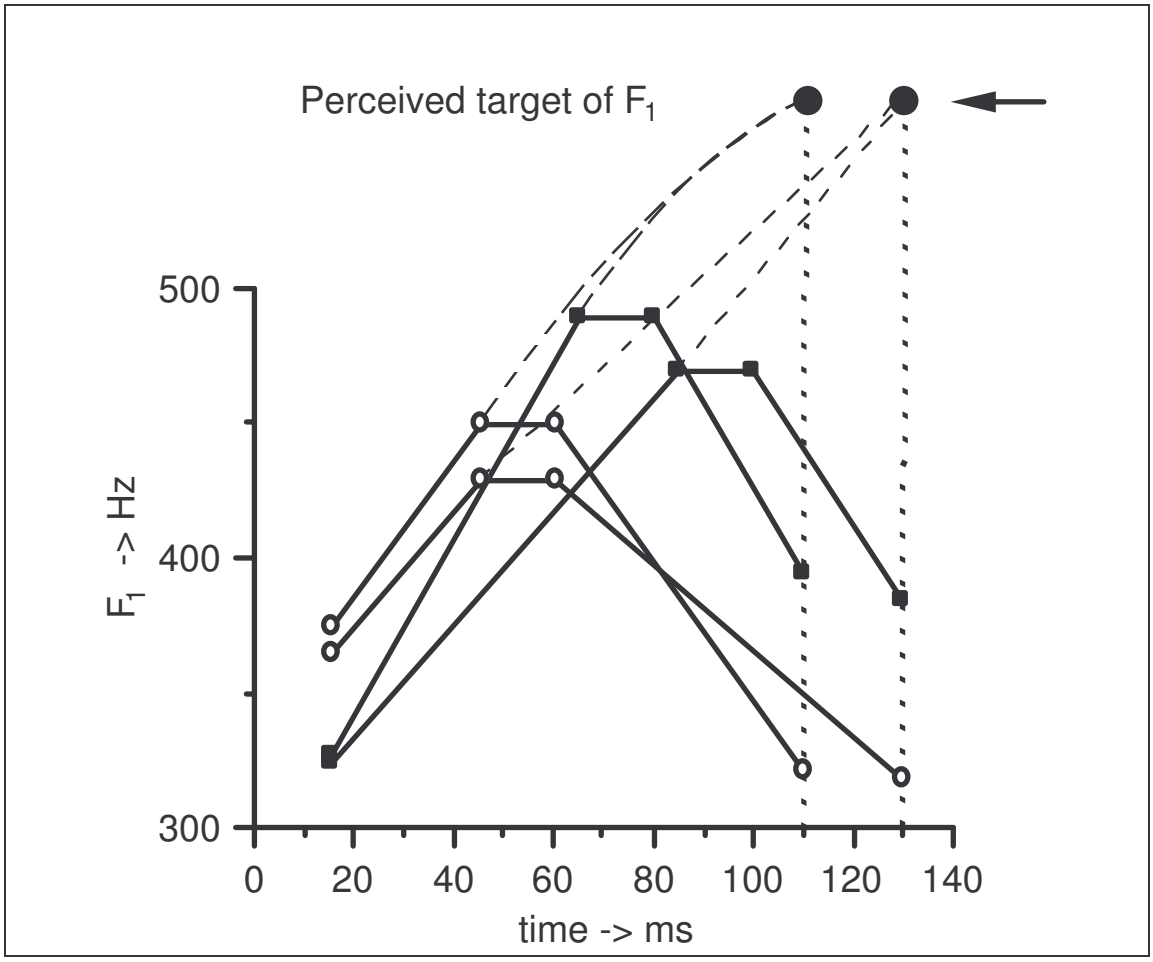

Figure 1.3. Perceptual-overshoot.
The F_1 tracks of four tokens are drawn in a frequency versus time plot. All four tokens lead to the same F_1 "target percept". This target was interpreted to be positioned beyond the maximal values reached in the tokens (indicated by the thin lines). Reproduced from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