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 definition: 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 • -0.93, p•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., $F(t) = F_{target} + (F_{initial} - F_{target})\cdot(2\cdot t/\text{Duration} - 1)^6$). 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•-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 experiments 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•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 u/. 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 experiments, 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.