5

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

Abstract

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

Introduction

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

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

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

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

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

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

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

5.1 Methods

5.1.1 Isolated vowels

5.1.1.1 Token synthesis

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

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

5.1.1.2 Token construction

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

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

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

$$F_n(t) = \text{Target} - \Delta F_n(4(t/\text{Duration})^2 - 4t/\text{Duration} + 1) \quad [5.1]$$
$$dF_n(t)/dt = -\Delta F_n(8t/\text{Duration} - 4)/\text{Duration}$$

in which:

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

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

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

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

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

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

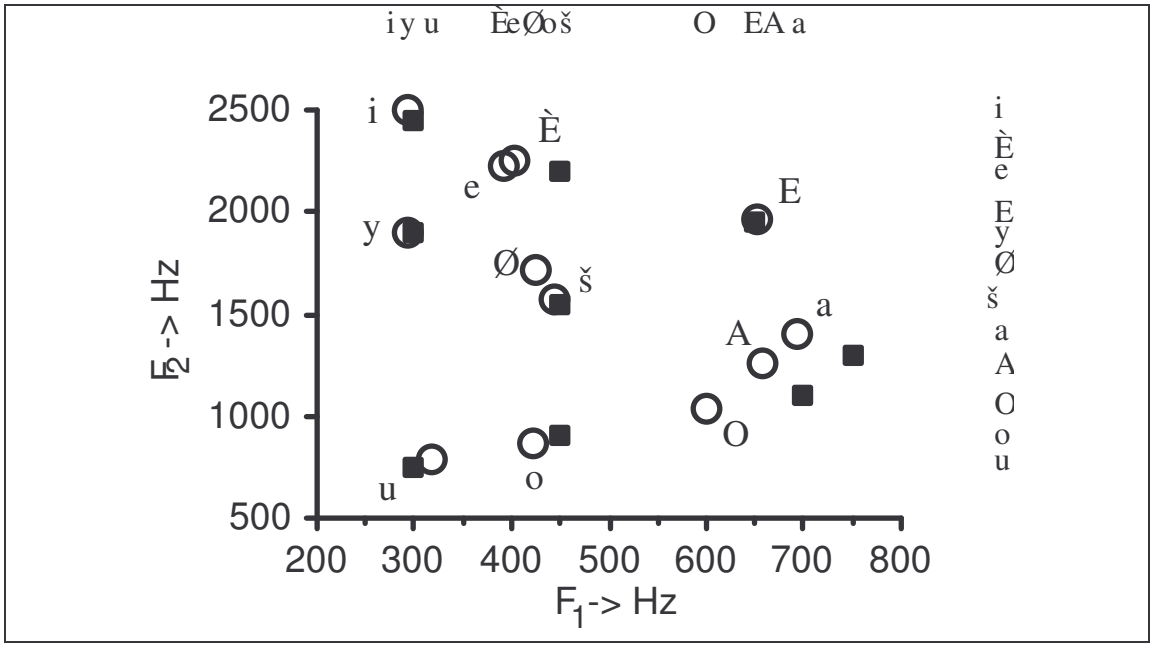

Figure 5.1: Formant frequencies of a single set of Dutch vowels pronounced in isolation by a male speaker (open circles, data taken from Koopmans-van Beinum, 1980). Above and to the right are the vowel labels in the established order along the F_1 and F_2 axis respectively (combined from 8 sets, see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

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

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

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

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

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

5.1.1.3 *Presentation*

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

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

5.1.1.4 Subjects

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

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

5.1.2 Presentation in synthetic CVC syllables

5.1.2.1 Consonants

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

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

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

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

	F_1	F_2	F_3	F_4	NP	NZ	Z_2
/n/	Frequency	180	1600	2600	3800	320	240	-
	Bandwidth	75	300	300	150	100	100	-
/f/	Frequency	300	1600	2600	3500	-	-	250
	Bandwidth	50	500	1200	600	-	-	400

5.1.2.2 Vowel segments and syllable construction

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

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

5.1.2.3 Presentation and subjects

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

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

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

5.2 Results

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

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

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

decide which specific vowel was heard.

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

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

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

5.2.1 Isolated vowel presentation

5.2.1.1 Effects of duration on tokens with level formant tracks

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

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

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

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

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

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

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

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

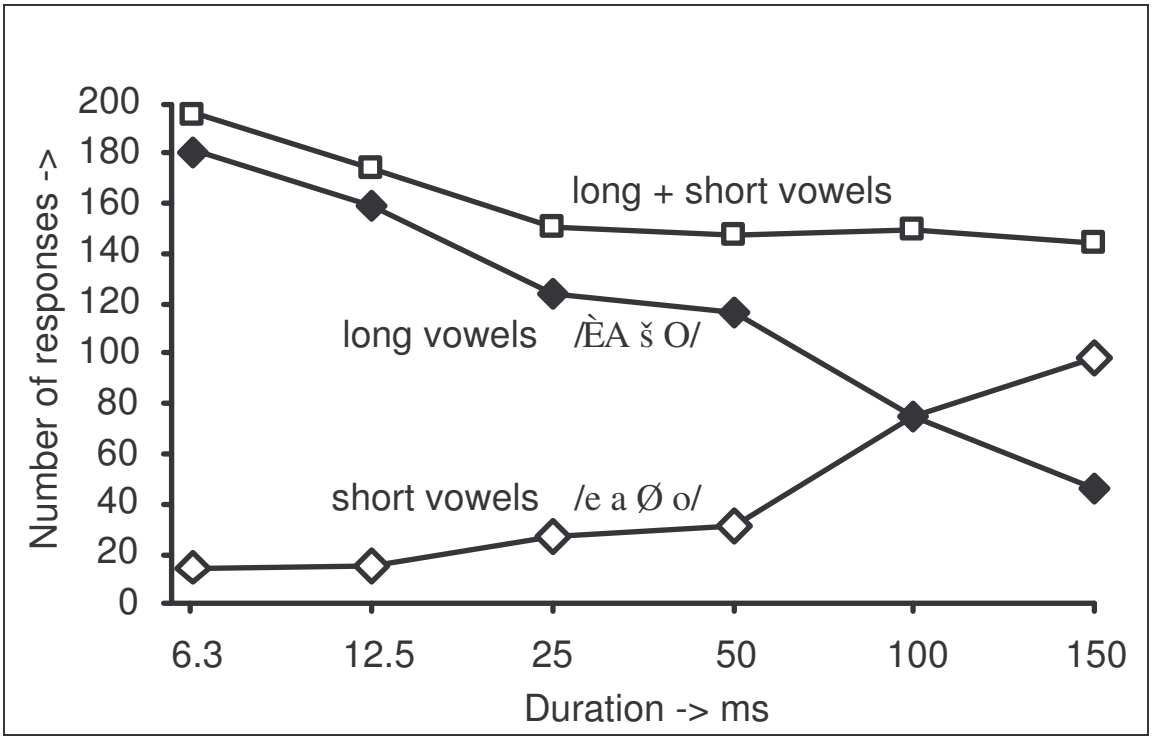

Figure 5.3.a: Absolute number of long- and short-vowel responses (/e a Ø o/ *and* /È A π O/, *respectively) for different durations. The total number of responses per duration was 261.*

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

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

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

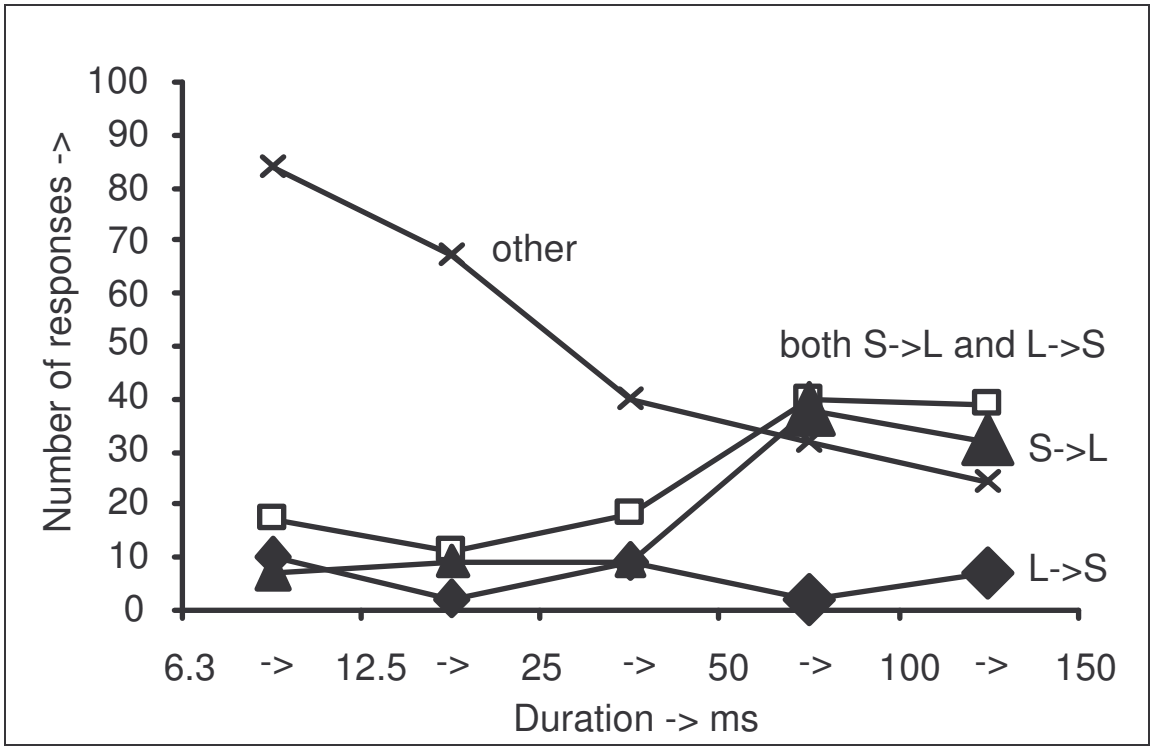

Figure 5.3.b: The number of short-to-long vowel label exchanges when tokens become longer (max. 261). Large symbols indicate cases where S->L and L->S differed significantly (p≤0.1%, two tailed sign-test). S->L: Short-vowel labels that are exchanged for the corresponding long-vowel labels (/A/->/a/,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

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

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

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

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

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

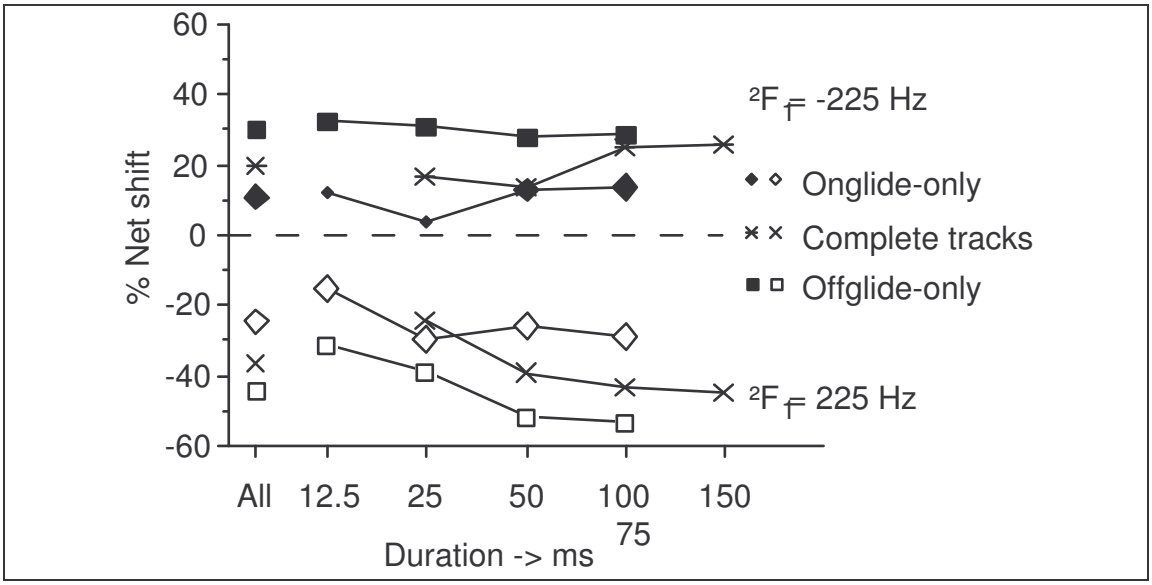

Figure 5.4.a: Net shifts in responses to isolated vowel tokens as a result of formant track curvature of the F_1.
All values in percent of total number of responses (see text). Positive shifts are towards higher formant frequencies, negative shifts towards lower frequencies. Large symbols indicate statistically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

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

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

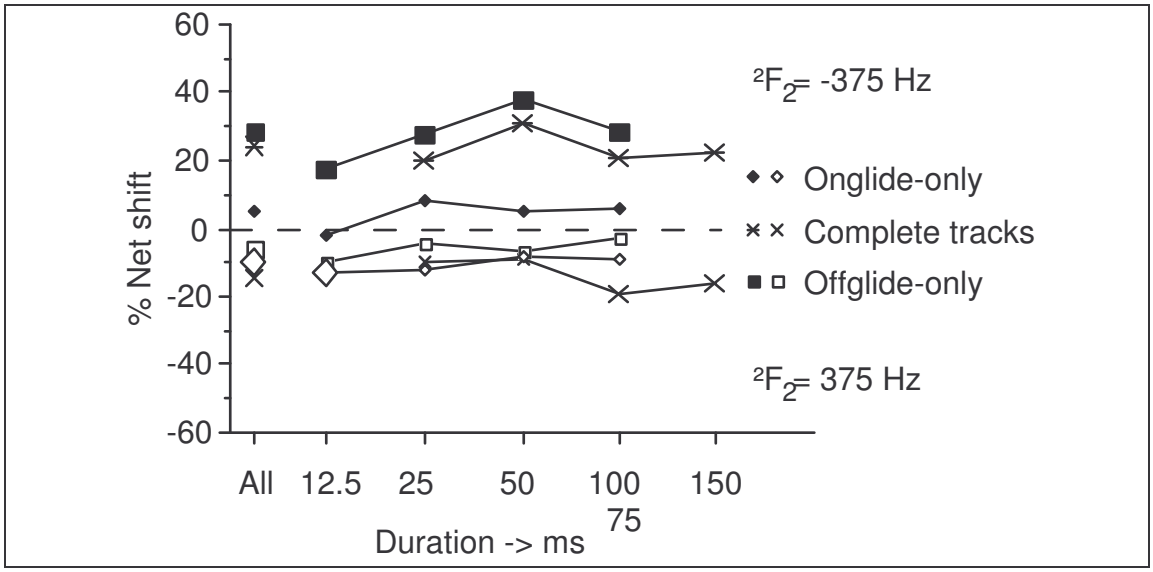

Figure 5.4.b: As figure 5.4.a. Net shifts in responses to isolated vowel tokens as a result of formant track curvature of the F_2. Open symbols: ΔF_2 = 375 Hz (n=928), filled symbols: ΔF_2 = -375 Hz (n=812). The first formant is level (i.e., ΔF_1 = 0) for both.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2.2 Presentation of vowels in context

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

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

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

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

5.2.2.1 *Consistency in responses to synthetic vowels*

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

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

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

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

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

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

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

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

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

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

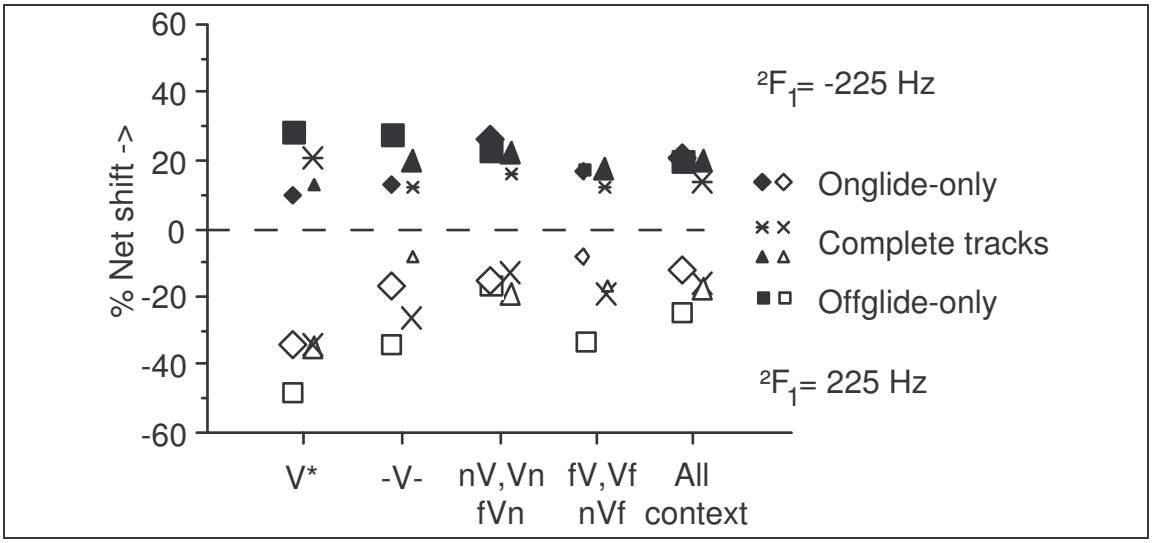

Figure 5.5.a: Net shifts in responses as a result of curvature of the F_1 both for isolated vowel tokens and vowel tokens in context. All values in percent of total number of responses (see text). Large symbols indicate statistically significant shifts (p≤0.1%, two-tailed sign-test). V: results from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

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

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

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

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

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

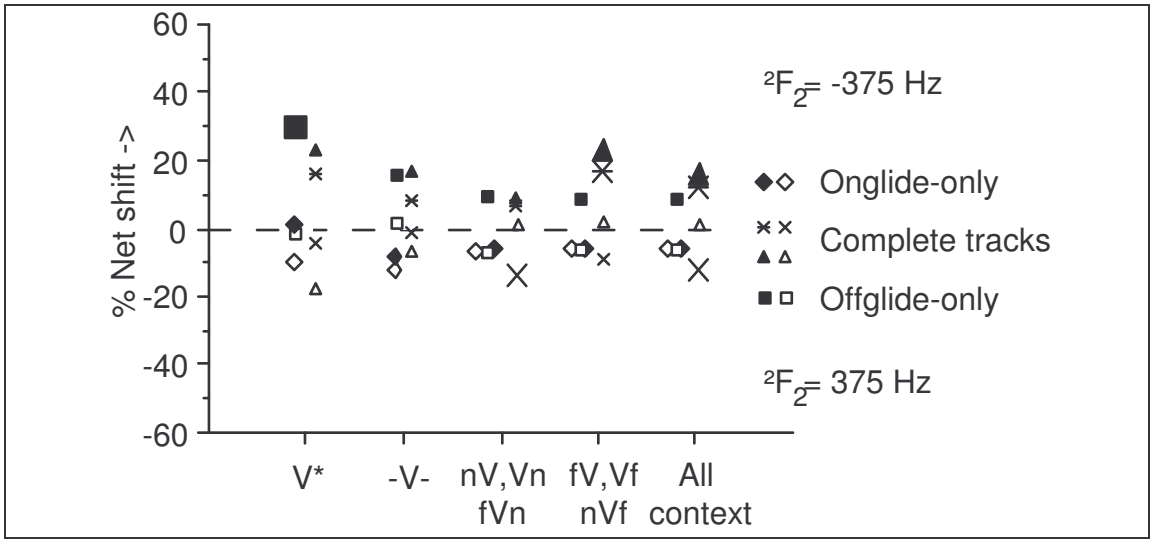

Figure 5.5.b: As figure 5.5.a. Net shifts in responses as a result of curvature of the F_2 for isolated vowel tokens and vowel tokens in context. Open symbols: ΔF_2 = 375 Hz (n=120), filled symbols: ΔF_2 = -375 Hz (n=90). The first formant is level (i.e., ΔF_1 = 0) for both.

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

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

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

5.3 Discussion

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

5.3.1 The effects of duration

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

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

5.3.2 The effects of formant excursion size

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

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

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

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

5.3.3. The effects of context

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

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

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

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

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

5.3.4 Relevance for natural speech

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

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

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

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

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

5.4 Conclusions

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

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

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

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