7

GENERAL DISCUSSION

Abstract

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

Introduction

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

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

7.1 Target-undershoot in production

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

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

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

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

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

Below we will discuss them all.

7.1.1 Quasi-stationary formant analysis might give inaccurate values

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

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

There is one area where the window-size does cause problems. At the vowel on- and offset boundaries, half of the analysis window will sample the context of the vowel realization instead of the vowel itself. As formant frequencies tend to be ill-defined in consonants or in rapidly changing consonant/vowel boundaries, formant frequencies measured here might be atypical (Smits, submitted). This is not to say that the (possibly incorrect) formant frequencies at the on- and offset boundaries behave in an irregular way. Correlations between speaking rates for formant frequencies at the boundaries were as high as for those in the middle part (figure 3.2, chapter 3).

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

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

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

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

We used the model and data of Lindblom (1963) and the mean vowel durations from chapter 3 of the present study to estimate the size of the expected excess undershoot in F_2 due to speaking rate in our own data (see figure 1.1, chapter 1). This was done for the three different contexts that Lindblom had used (i.e., b_b, d_d, g_g) and the vowels that were closest to ours (/È œ Ø a O U/ in his study). Of the three values of formant-undershoot predicted for each of our vowels (one for each /b/, /d/, or /g/ context), we used only the median value. Using the median value is more realistic than using the extreme values because of the diverse context in our samples which would tend to average out the excess undershoot. For our realizations of the vowels /A a o u i/, the expected amount of excess undershoot due to a higher speaking rate was in the range of 30-40 Hz. This value is larger than the threshold of detection determined earlier.

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

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

7.1.3 A ceiling (floor) in undershoot was already reached

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

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

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

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

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

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

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

shown up as more centralization.

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

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

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

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

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

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

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

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

7.1.5 Coarticulation was not strong enough to require extra undershoot

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

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

We did test whether a larger "articulatory distance" between vowels and context would have changed our results. To do this, we selected vowel realizations from an alveolar context (i.e., one of the consonants /n s z t d r l/). These consonants would restrict the tongue to a high and fronted position, i.e. close to the position it takes for the vowel /i/. The articulatory distance between the consonants of the context and the high, fronted vowels (i.e., /i E y/ from our sample) would be relatively small. The distance with the low, back vowels (i.e., /u o A a/ from our sample) would be comparatively large and should therefore induce a sizeable amount of excess formant-undershoot with an increase in speaking rate (c.f. Gopal and Syrdal, 1988). But even this subset of realizations with a large articulatory distance between vowels and context did not show any excess formant-undershoot at a fast speaking rate.

7.1.6 Alternative articulating strategies

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

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

7.1.7 Does duration control vowel target-undershoot?

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

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

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

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

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

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

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

7.2.1 Recapitulation of our vowel identification results

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

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

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

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

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

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

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

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

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

7.2.2 Results from the literature

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

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

7.3 Target-undershoot and vowel perception

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

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

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

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

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

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

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

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

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

7.4 Conclusions

We can summarize the preceding discussion by saying that:

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

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

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

7.5 Suggestions for future research

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

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

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