Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 17 (1993), 33-64.

VOWEL PERCEPTION: A CLOSER LOOK AT THE LITERATURE*

R.J.J.H. van Son

Abstract

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

1 Introduction

1.1 Perceptual-overshoot and dynamic-specification in vowel identification

In natural speech there is a substantial variation in vowel realizations, even when spoken by a single person. Vowels spoken in isolation or in a neutral context, such as /hVd/ in English, are considered to approach the ideal with regard to vowel quality. Such ideal vowel realizations are called canonical realizations. Numerous factors change these canonical realizations to the realizations actually found in natural speech, e.g. speaking style, prosody, context. A popular model to describe this variability for the speech of a single speaker is the target-undershoot model as proposed by Lindblom (1963, 1983). In this model it is assumed that timing constraints prevent the articulators (i.e., jaw, tongue, lips) to reach the canonical target positions for the vowel that is to be pronounced. Both the articulators and the corresponding formant frequencies are stopped short of reaching their targets. The corresponding realization displays "undershoot" with respect to the canonical target.

The variability in vowel realizations could give the impression that vowels are difficult to recognize in normal, connected speech. But, in a normal utterance, vowels are generally identified accurately, whatever the context or speaker characteristics. This raises the question of how listeners accomplish this feat (at the moment,

*This paper is an updated and extended version of chapter 6 of my Ph.D. thesis: *Spectro-temporal features of vowel segments*, in *Studies in Language and Language use* **3**, IFOTT, University of Amsterdam, 1993.

machines cannot). Certain models of vowel perception try to answer this question by looking for acoustic features in vowel realizations that are invariant to coarticulation, reduction, and speaker identity.

In general, models of vowel perception are tied to models of vowel production. The target-undershoot model of vowel production as introduced by Lindblom (1963) inspired the development of a complementary model for vowel perception. In this perceptual model, listeners would compensate for undershoot in production by overshoot in perception. The perceptual-overshoot theory was first proposed and tested by Lindblom and Studdert-Kennedy (1967, see below). In this model, the hypothetical canonical formant target value that was not reached due to target-undershoot could be determined (i.e., calculated) by extrapolating the formant tracks in the Consonant-Vowel (CV) and/or Vowel-Consonant (VC) transition. It is also possible that vowel duration is used together with the articulatory or perceptual "distance" between the vowel realization and its context to factor out the undershoot without a direct recourse to a dynamical perceptual-overshoot (Nearey, 1989). In this latter case, the listener needs to relate the amount of undershoot to the duration of the vowel.

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

1.2 Dynamic-specification versus elaborate target models of vowel perception

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

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

Vowel-inherent spectral changes, like diphthongization, are modelled by assuming a double, compound, target in the vowel nucleus instead of only a single target (Andruski and Nearey, 1992). Still, the transition parts of the vowel realizations (i.e.,

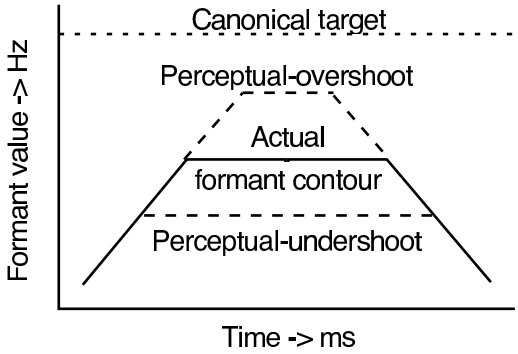

Figure 1. Perceptual over- and undershoot. Hypothetical relation between perceived (dashed lines) versus actual formant contours (solid line). The differences are exaggerated for clarity.

the vocalic parts of CV and VC transitions) do not influence vowel recognition according to these theories. Target-undershoot in production would change the spectral contents of the vowel mid-points depending on vowel duration. This could make it necessary to include duration in the normalization procedure in order for this procedure to compensate for the undershoot in production.

2) On the other side there are theories that acknowledge that dynamical information from parts outside the vowel nucleus is also used to disambiguate the information from the vowel nucleus itself (e.g., Lindblom and Studdert-Kennedy, 1967; Huang, 1991, 1992; Di Benedetto, 1989a, b; Fox, 1989; Strange, 1989a, b). These theories are spectro-temporal and rely on "dynamic-specification" to disambiguate the vowel realizations (also called dynamic-cospecification, Andruski and Nearey, 1992). It is assumed that the shape of a vowel formant track is indicative of the direction and amount of (formant) undershoot. Knowing the amount of undershoot enables a listener to deduce the position of the canonical target of the vowel. A commonly proposed mechanism to achieve this is perceptual-overshoot (see figure 1).

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

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

1.3 Evidence pro and contra dynamic-specification

Evidence for the use of dynamic-specification in vowel recognition comes from several studies. It was noted that coarticulated vowel realizations in a CVC context were identified better, or at least not worse, than vowels spoken in isolation (see discussions in e.g., Strange and Gottfried, 1980; O'Shaughnessy, 1987, p.177; Fox,

1989; Nearey, 1989; Strange, 1989a; Andruski and Nearey, 1992). The same was demonstrated for the medial vowel of /VVV/ sequences (Kuwabara, 1985). Also, vowel realizations from which the kernel was removed (silent-center vowels), leaving only the Consonant-Vowel and Vowel-Consonant transitions up to the borders of the kernels, were often recognized better than the isolated kernel parts alone. Recognition of silent-center vowels was generally only moderately compromised and sometimes recognition was even indistinguishable from that of complete syllables (Strange, 1989b; p.2144). Even when the initial and final transition parts of the silent-center vowels were from speakers of opposite sex, the number of errors remained quite low (Verbrugge and Rakerd, 1986). In all these cases, the vowel mid-point spectrum differed strongly from the canonical case (i.e., vowels pronounced in isolation) or was even absent altogether. This fact did not seem to bother the listeners and as long as the transition parts were present, recognition was hardly compromised. Fox (1989) even found that reducing the transitions in synthetic silent-center realizations to the outermost single pitch periods still allowed quite accurate vowel identification.

In a completely different set of experiments, Di Benedetto (1989b) concluded that F_1 transitions and timing were used to distinguish between high (/i ɪ/) and non-high (/e ɛ/) vowels (her terminology). She discussed perceptual-overshoot as a possible explanation but preferred an explanation in which her subjects had used a weighted average of the F_1 contours. Support for dynamic-specification also came from the fact that information about formant track shape could help to distinguish automatically realizations of different vowels with comparable F_1 mid-point or extreme values (Kuwabara, 1985; Di Benedetto, 1989a; Huang, 1991, 1992; Akagi, 1993).

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

1.4 Distinguishing models of vowel perception

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

In general, dynamic-specification is expected to work in the same direction as perceptual-overshoot. The shape of a formant curve always signifies a vowel with a

target on or beyond the mid-point value actually reached. There are no reports of contexts for which the formant mid-point value of any vowel would systematically overshoot the target it reaches when pronounced and sustained in isolation. For example, an open vowel (like /a/) is generally characterized by a strongly curved, rising-falling F_1 track. The (canonical) F_1 target of this vowel can be found by extrapolating the on- or offglide of this same track. In a first approximation, both the strongly rising-falling curve shape and the target found by extrapolation will indicate an open vowel (i.e., a high F_1-target). Therefore, perceptual-overshoot and dynamic-specification predict the same behaviour of subjects: response targets should overshoot the mid-point values actually present in the tokens. The amount of overshoot should be related to the curvature of the formant tracks and the duration of the tokens.

On the other hand, target-models of vowel perception state that listeners use a cross-section to characterize the complete formant track. In practice, listeners are expected to take the average of some small part of the formant track. This should result either in subject responses that are independent of formant track shape, or alternatively, in some undershoot in strongly curved tracks due to the averaging process (see figure 1). A complicating factor is that listeners could use the wider *context* of the realization, instead of the formant track shape, to compensate for the *expected* undershoot in production. This would result in an apparent "overshoot" in the responses. However, because this apparent overshoot depends *not* on formant track shape (by definition), the overshoot would *only* depend on context and duration. Therefore, it should be easy to discriminate it from perceptual-overshoot and dynamic-specification.

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

In resent experiments we have tested these predictions with a listening experiment (Van Son, 1993). The results of these experiments motivated us look more closely at the experiments discussed in the literature. Since our own experiments are discussed extensively elsewhere (notably, Van Son, 1993; but see also Pols and Van Son, 1993; Van Son and Pols, 1993), I will only summarize them below.

1.5 The influence of formant track shape on the identification of synthetic vowels (Van Son, 1993; Van Son & Pols, 1993; Pols & Van Son, 1993)

In our experiment we compared the responses of Dutch subjects to isolated synthetic vowel tokens with curved formant tracks (F_1 and F_2) with their responses to corresponding stationary tokens with level formant tracks. We also investigated the effects of presenting these vowel tokens in a synthetic context (/nVf/, /fVn/).

Tokens were synthesized with (constant) synthesis parameters: F_0=159, F_3=2490, F_4=3500, and F_5=4500 (Hz). All bandwidths were 50 Hz. The source amplitude was constant over the duration of the tokens. Nine formant "target" pairs (F_1, F_2) were defined using published values for Dutch vowels. These pairs corresponded

approximately to the vowels /i u y ɪ o ɛ ɑ a œ/ and were tuned to give slightly ambiguous percepts. For these nine targets, smooth formant tracks were constructed for F_1 and F_2 that were either level or parabolic curves according to the following equation (see figure 2):

$F_n(t) = Target - \Delta F_n \cdot (4 \cdot (t/Duration)^2 - 4 \cdot t/Duration + 1)$
in which:

$F_n(t)$	- the value of formant n (i.e., F_1 or F_2) at time t.
ΔF_n	- the excursion size: F_n(mid-point) - F_n(on/offset). ΔF_1= 0, +225 or -225, ΔF_2= 0, +375 or -375 (Hz)
Target	- the formant target frequency.
Duration	- the total token duration (0≤t≤Duration).

No tracks were constructed that would cross other formant tracks or F_0. All tracks were synthesized with durations of 25, 50, 100, and 150 ms. Stationary tokens with level formant tracks (i.e., ΔF_1=ΔF_2=0) were also synthesized with durations of 6.3 and 12.5 ms. Of the other tokens (with either ΔF_1= +/- 225 Hz or ΔF_2= +/- 375 Hz), the first and second half of the tracks (i.e., on- and offglide-only) were also synthesized with half the duration of the "parent" token (12.5, 25, 50, and 75 ms). Some other tokens with smaller excursion sizes were used too, these will not be discussed in this paper (but see Van Son, 1993; Pols & Van Son, 1993). All tokens were presented in a pseudo random order to 29 Dutch subjects. They were asked to mark the vowel identity on a sheet in which all 12 Dutch monophthongs were printed (forced choice).

A single realization each of a synthetic /n/ and /f/ sound with durations of 95 ms were obtained from a Dutch speech synthesizer. These two specific realizations were used in mixed pseudo-syllabic stimuli. Vowel tokens with durations of 50 and 100 ms and mid-point formant frequencies corresponding to /ɪ ɛ ɑ o/ were combined with these consonants in both /nVf/ and /fVn/ pseudo-syllables. Furthermore, the corresponding vowel tokens with only the on- or off-glide part of parabolic formant tracks (50 ms durations only) were used in CV and VC structures respectively. For comparison, corresponding *stationary* vowel tokens with 50 ms duration were also used in CV and VC pseudo-syllables. Each vowel token, both in isolation and in these pseudo-syllables, was presented twice to 15 Dutch subjects who were asked to write down what they heard (open response).

The synthetic vowel-like sounds were not identified by every listener in the same way. However, the actual label used was not really important. What was important, was the difference in the identification between one set of stimuli (e.g., with level formant tracks) and another (e.g., with non-level formant tracks). Such a difference indicates how differently the stimuli were perceived.

As an example, we will compare the responses to vowel tokens with an excursion size ΔF_1 of 225 Hz with the corresponding stationary tokens (ΔF_1=ΔF_2=0). There were 696 responses available to tokens with this formant track shape (all four durations pooled). These were compared to the corresponding 696 responses to the stationary tokens of equal duration. Of the 696 labels, 348 (50%) were different from one case to the other. Given the rather consistent fixed rank order of the Dutch vowels along the F_1, viz. /i y u ɪ e ø o œ ɔ ɛ ɑ a/, 304 (44%) of these responses had a lower rank number for the curved F_1-stimulus than for the stationary stimulus. Only 44 (12%) had a higher rank number. The net effect is thus a shift towards a lower rank number in 44-304=-260 (-37%) of the responses. Using a two-tailed sign-test, it is clear that this difference is statistically significant (p≤0.001). The -37% net difference is called the net shift. It is the combined response of our subjects to tokens with an

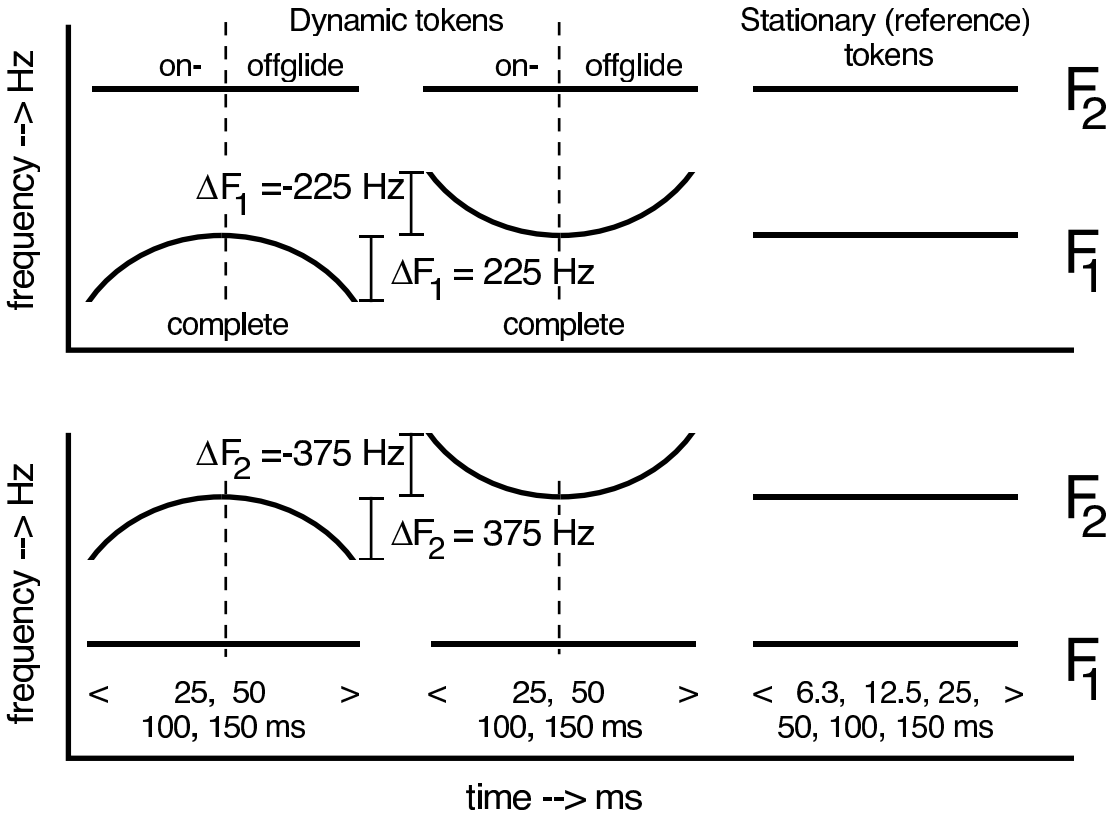

Figure 2. Formant track shapes as used in the experiments discussed in section 1.5. The dynamic tokens were synthesized with durations of 25, 50, 100, and 150 ms. The stationary tokens were synthesized with durations of 6.3, 12.5, 25, 50, 100, and 150 ms. The dynamic tokens were also synthesized as onglide- and offglide-only tokens, i.e., respectively the parts to the left and right of the dashed lines.

excursion size, ΔF_1, of 225 Hz. The negative sign is added to indicate a shift towards a *lower* F_1 rank number. This shift is entered in the left-hand panel of figure 3 as the second (white) column from the left. The same procedure was applied to every combination of token duration, F_1 or F_2 excursion size, and shape (complete, onglide-only, offglide-only). The fixed rank order of the Dutch vowels along F_2 was /u o ɔ ɑ a œ ø y ɛ e ɪ i/.

It showed that a shortening of token duration only induced an exchange of long-vowel labels for the corresponding short-vowel labels. Very short token durations (6.3 and 12.5 ms) induced a rather indiscriminate increase in /ɪ ɔ/ responses. Differences in duration did not result in appreciable changes to the shift in responses. At all durations, as in the above example, the sign of the net shift in the responses was opposite to the sign of the excursion size ΔF_n, i.e. the shift in the responses was towards the on-/off-set points of the formant tracks. This means that the formant track curvature induced perceptual-*under*shoot instead of perceptual-*over*shoot. This is shown in figure 3 for both the results of the experiment with tokens presented in isolation (leftmost columns) as for those presented in a synthetic context (rightmost columns). From the results of figure 3 it is clear that the largest shift was induced by the offglide-only tokens (r), and the smallest by the onglide-only tokens (l). This indicates that listeners put most emphasis on the offset part of the tokens.

The introduction of a non-silent context had no effect on the direction of the shift in the responses. The only effect of the (artificial) context used was a change in the number of long/short-vowel responses. Closed "syllables" (i.e., VC and CVC) received less, open syllables (i.e., CV) more long-vowel responses than isolated

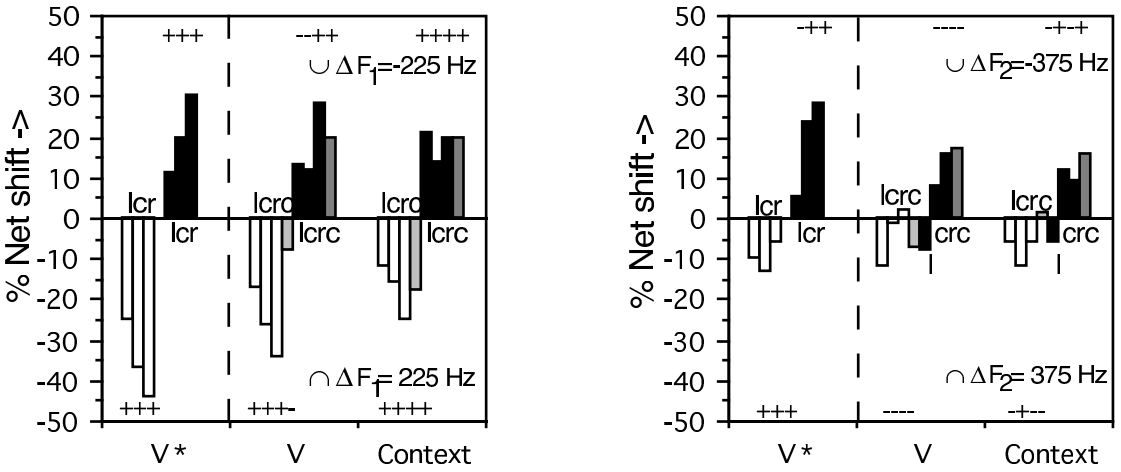

Figure 3. Net shift in responses as a result of curvature of the formants. 'V*' are the results of experiment 1 (all tokens pooled on duration, n≥ 696). 'V' and 'Context' are the results of experiment 2 with 4 mid-point pairs for tokens presented in isolation with n=120 ('V' left), and n=90 ('V' right), or in context (CV, CVC, VC; C one of /n f/) n=240 ('Context' left) and n=180 ('Context' right). Gray bars: 100 ms, white/black bars: 50 ms, l=onglide-only, c=complete, r=offglide-only tokens. +: significant (p≤0.001, sign test), -: not significant. Panel on the left: curved F_1, panel on the right: curved F_2.

vowels (i.e., V). This could be the results of phonotactic constraints in Dutch, open syllables cannot contain short vowels.

We did find a small effect of the task on the size of the net shift. When we compared the responses to identical tokens in the first experiment (forced choice) and second experiment (open response), we saw that the shift in the forced choice experiment was larger than in the open response experiment. This might have been the result of the perception of diphthongs. In the second experiment subjects used a sizeable number of diphthong or triphthong labels. Being unable to use these in a forced choice experiment might have distorted the sizes of the net shifts.

From these results we concluded that the presence of curved formant tracks themselves did not induce perceptual-overshoot in the listeners, nor did it induce any other compensatory mechanism for undershoot in production. On the contrary, it seems that listeners used some kind of average over the formant tracks with most emphasis on the offset part. We did not find any compensatory effect on behalf of the listeners when we added of a non-integrated, synthetic context to our vowel tokens.

1.6 Summary

To summarize the above discussion, there is a disagreement in the literature with regard to the role of Consonant-Vowel (CV) transitions in vowel recognition (see Strange, 1989a; Andruski and Nearey, 1992). Several studies lead to the conclusion that dynamic features, and especially formant transitions, are used to identify vowel realizations. However, no evidence for such a mechanism could be found in other studies. The evidence that was presented in favour of perceptual-overshoot and dynamic-specification could also be interpreted against it (Andruski and Nearey, 1992). We too could find no evidence of dynamic-specification or perceptual-overshoot (see 1.5). On the contrary, we found that non-level formant tracks would lead subjects away from the mid-point values towards perceptual-undershoot. This

means that, instead of alleviating the effects of coarticulation, curved formant tracks would aggravate them. The cause of all these contradictory results remains unknown.

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

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

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

2 An evaluation of the relevant literature

The results of our experiments seemed to disagree with at least some that were reported in the literature (see section 1.5). In this section we will interpret our results in the light of results reported in the literature. We will first discuss two questions that are related to the question of whether dynamic information is used to identify vowels. First, is there dynamic information in the spectro-temporal structure of vowel segments that could be used to identify vowel realizations (section 2.1). Second, is the ambiguity found in the responses to synthetic stimuli also found in natural speech or are natural vowels always recognized well (section 2.2). The remainder of section 2 will be dedicated to findings that are directly related to the question of whether listeners use dynamic information from consonant-vowel (or vowel-vowel) transitions to identify vowel realizations. I divided the experiments reported in the literature into two groups:

1. Experiments using synthetic speech (section 2.3)
2. Experiments using natural speech (section 2.4)

2.1 Information present in formant dynamics

Several studies have tried to determine whether vowel realizations contain dynamic information that could be used to identify them. Earlier studies about the relation between vowel formant track shape and formant target frequencies indicated that vowel formants started and ended, on average, from a closed (low-F_1) and non-high/non-low (mid-F_2) position (see e.g., Pols and Van Son, 1993). Stressing the fact that, in these studies, these starting and ending points are averages, this seems not to be unreasonable from an articulatory point of view. Furthermore, the strong correspondence between formant spaces constructed from "excursion size" and "mid-point" values (Pols and Van Son, 1993) indicates that the link found between formant excursion size and vowel identity is unlikely to be a statistical artefact.

Examining natural speech, Di Benedetto (1989a) found that she could use the time at which the maximum in the F_1 was reached to distinguish realizations of the vowels /ɪ ɛ/. Huang (1991, 1992) reported that characterizing a vowel formant track with three points (at 25%, 50%, and 75% of duration) instead of only at a single point, could increase the recognition score of a Gaussian classifier. This shows that information on formant track shape could help classification. Kuwabara (1985) and Akagi (1990, 1993) also concluded that information from spectral dynamics could be used to improve automatic vowel classification in natural speech. Both Huang and Akagi suggested that a mid-point "overshoot" mechanism that compensates for coarticulatory undershoot could do the job.

These studies show that the spectral dynamics of vowel realizations can be used to help classify vowel realizations automatically. This was found using several different methods to measure these dynamic features. The systematic nature of the relation between formant track shape and vowel identity suggested the possibility that human listeners would use this information too. However, our own study has shown that the matter is not that simple (section 1.5). It is clear that some conditions must be met before listeners will actually use the dynamic information present in vowel realizations.

2.2 Natural versus synthetic speech

In our experiments (section 1.5), we used synthetic stimuli with simplified formant contours. The formant trajectories in our vowel tokens were in a sense quite unnatural, moving mostly along one formant at a time. It could be that, for each *natural* vowel realization, the combined trajectory of the formants in formant space (i.e., F_1/F_2 space) would spend most of its time within the boundaries of the perceptual area of that vowel. This way it would not matter on which part of a natural vowel realization its identity was determined. In most experiments using synthetic speech, it is tried to make the trajectories in formant space similar to those in natural speech (c.f. Lindblom and Studdert-Kennedy, 1968; Fox, 1989). However, it is known that reduced vowels and vowels excised from their context are identified less well than vowels spoken in isolation (Koopmans-van Beinum, 1980; Van Bergem, 1993). From this we can conclude that in natural speech too, formant trajectories seem to leave the perceptual area of the vowel, just as in our experiments. Therefore, some other mechanism seems to ensure correct identification.

Formant excursion sizes in natural speech are generally smaller than the extreme excursion sizes used in our listening experiments (section 1.5). It is to be expected that vowel realizations from natural speech, with “natural” mid-point formant frequencies and moderate formant excursion sizes, will generally be identified

correctly. This might in part explain the generally high recognition scores for natural vowel realizations uttered in context (see discussions in Strange, 1989a; Nearey, 1989; Andruski and Nearey, 1992). However, this fact cannot explain everything, because of the above mentioned fact that vowel realizations from natural speech are identified much worse when presented out of context.

2.3 Experiments using synthetic speech

The strongest claims for the existence of perceptual-overshoot were based on experiments using synthetic vowel tokens with well defined formant tracks. The oldest and most cited paper that reported perceptual-overshoot is the study of Lindblom and Studdert-Kennedy (1967). This study contrasts with our own study in which we did find the opposite results: clear perceptual-undershoot (section 1.5). Their stimuli were similar to ours and it certainly requires some explanation why the results of both studies disagreed. I will therefore discuss their experiments extensively. I will also discuss several other papers.

A preliminary remark must be made about an important difference between the experiments discussed below and that of our own (section 1.5). Almost all experiments discussed in this section, 2.3, used a forced choice paradigm for the responses. Listeners were always asked to respond with only one of a limited set of possibilities, often only two labels were available, irrespective of what they actually heard. In our experiments we either asked our listeners to respond with any of the Dutch monophthongs (forced choice) or they were asked to respond whatever they heard (open response). In section 1.5 we saw that restricting the response categories to all Dutch monophthongs, therefore excluding diphthongs and triphthongs, already increased the size of the perceptual-undershoot found. Restricting the response categories still further to only two labels (e.g., /U ɪ/ or /ɪ ɛ/) will result in even more dramatic changes in the outcome of the experiments. Essentially, in the experiments discussed below, the *listeners* were forced to place their responses on a single continuum. In our experiments, *we* constructed these continua ourselves by rank-ordering the response labels along the F_1 and F_2 directions. It is certain that these two different procedures for ordering responses along a continuum will give different results. However, it is very *un*likely that this methodological difference will change perceptual-overshoot in the responses into perceptual-undershoot and therefore I will not elaborate on this difference. The number and quality of response categories might, however, have a very strong effect on the sizes of the over- or undershoot found. Therefore, between-paper comparison of results can only be done in a qualitative way, not in a quantitative way.

In the following sections, 2.3.1-2.3.6, I will discuss several papers in detail. Readers who are not interested in technical discussions can skip these sections and go straight to section 2.3.7 which summarises the conclusions obtained in these earlier sections.

2.3.1 The paper of Lindblom and Studdert-Kennedy (1967)

Lindblom and Studdert-Kennedy (1967) used vowel tokens in a well defined and integrated context. Vowel token mid-point values spanned a continuum in the range between /U ɪ/ (F_1 = 350 Hz, F_2 = 1-2 kHz, F_3 = 2.3-2.8 kHz). Vowel tokens were presented to subjects in isolation with level formant tracks and in /wVw/ and /jVj/ syllables with parabolically shaped formant tracks (see figure 4). The vowel on- and

offset frequencies were F_1 = 250 Hz, F_2 = 800 Hz, F_3 = 2200 Hz in /wVw/ context and F_1 = 250 Hz, F_2 = 2200 Hz, F_3 = 2900 Hz in /jVj/ context. The consonants were synthesized as two stationary 20 ms sounds with formant frequencies that were identical to the vowel formant on- and offset frequencies. The responses of the subjects were limited to only two categories: /U/ and /ɪ/. Stimuli of different durations and with or without context were presented in a blocked fashion. Ten native speakers of American English participated in the experiments. Four were tested in Sweden (KTH, Stockholm) and six in the USA (Haskins Laboratories, New York). Pseudo-random sequences of tokens of each duration in context and in isolation were presented in four blocks, /wVw/ and /jVj/ combined versus #V# for each duration (i.e., 200 ms and 100 ms). This means that tokens of different durations were never mixed, nor were vowel tokens with and without context.

Next to the similarities in stimuli, several important differences with our experiments are apparent (cf. section 1.5). Spectral changes from consonants to vowels and vice versa were continuous in the experiment of Lindblom and Studdert-Kennedy (1967). The formant tracks of the vowel parts always started and ended at the values used for the consonants. Furthermore, their consonants were synthesized as "vowel-like" sounds. The consonants and vowels in the Consonant-Vowel-Consonant (CVC) syllables were therefore well integrated. Next, the F_2 excursion sizes were often larger than those used in our experiments, up to 1200 Hz (compared to a maximum of 375 Hz in section 1.5). With our relatively small excursion sizes we already induced a sizeable amount of diphthong responses. It is to be expected that the stimuli of Lindblom and Studdert-Kennedy induced an even stronger perception of diphthongs than our own. This might have influenced the responses of the subjects in ways unaccounted for in their experiments.

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

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

The median difference between the F_2 mid-point values in /wVw/ context and in /jVj/ context for which /U/ changed into /ɪ/ responses, i.e. the 50% cross-over point in the responses, was 180 Hz for 200 ms vowel tokens and 274 Hz for 100 ms tokens. The cross-over point for /jVj/ syllables had a higher F_2 value than that for /wVw/ syllables, showing clear perceptual-overshoot. However, three out of the eight subjects showed consistent perceptual-undershoot instead of overshoot (all three tested in Sweden). If only the responses of the five subjects showing consistent overshoot were used, the median differences in F_2 mid-point value between /wVw/

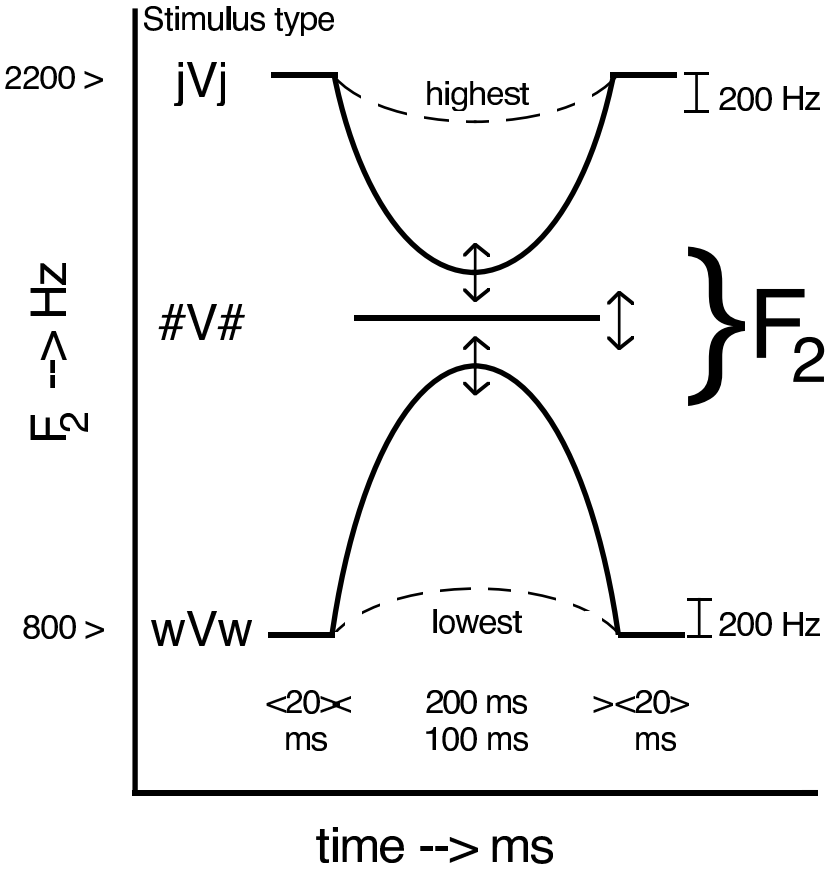

Figure 4. Contrasting parabolically shaped F_2 tracks used in the listening experiments of Lindblom and Studdert-Kennedy (1967). F_1 and F_3 were also parabolically shaped (not shown). For each of the stimulus types (wVw, #V#, and jVj) the midpoint F_2 frequency was varied from lowest (1000 Hz) to highest (2000 Hz) along an /U/-/ɪ/ continuum. Note that two durations were used with fixed 'consonant' durations: a 'vowel' duration of 200 ms and one of 100 ms. This drawing is not to scale

and /jVj/ context, i.e. the combined perceptual-overshoot, became 289 Hz and 363 Hz (200 ms and 100 ms tokens respectively). This is a considerable amount of overshoot, approximately 30% of the combined excursion sizes (by definition: combined excursion sizes + combined overshoot = /jVj/ onset - /wVw/ onset = 1400 Hz for this experiment).

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

excursion size, it was also not related to formant track slope. It might have been related to the /w/ and /j/ context itself (see section 2.3.7).

Lindblom and Studdert-Kennedy also reported that a shorter duration (100 ms) increased the amount of perceptual-overshoot in the /wVw/ syllables for 9 out of 10 subjects (median increase in F_2 overshoot was 68 Hz, all ten subjects completed the answers for the /wVw/ tokens). However, when the significant effect of token duration on the responses to the isolated vowel tokens was taken into account, the increase in perceptual-overshoot in the /wVw/ syllables was found only for 6 out of 10 subjects (median increase in F_2 overshoot was 32 Hz). For the short duration too there was no relation between formant-overshoot and formant excursion size. When I combined their results for 200 and 100 ms tokens there was a strong negative correlation between excursion size and perceptual-overshoot for the /wVw/ tokens ($r \approx -0.93$, $p \leq 0.01$) and no correlation at all for the /jVj/ tokens.

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

Lindblom and Studdert-Kennedy related their results to the overshoot found in diphthong perception. They discussed the fact that in diphthongs, generally only one of the two targets is actually realized. The presence of the other target is only suggested by the movements of the formants. "*Thus, an articulatory movement [ɑe] or [ɑɛ] is heard as [ɑi] by the naive listener*" (quote from Lindblom and Studdert-Kennedy, 1967, p.842). From our results, described in section 1.5, we could infer that the tokens used in their experiments were indeed long enough, and had sufficiently large excursion sizes, to induce diphthong responses. Nearey (1989, p.2103) reported that stimuli with a similar formant track shape produced glide-like percepts. The fact that vowel-like consonants (i.e., /w/ and /j/) were added would only have strengthened this tendency. If their subjects would have interpreted their tokens as diphthongs, this would explain the overshoot in identification found. Subjects would have used the extent of the "glide" part as a co-specification to diphthong or glide identity. The design of the tokens then would cause a negative correlation between formant excursion size and "perceptual-overshoot". Diphthong or glide perception could also make more understandable the large differences between subjects. For some subjects the threshold for glide-perception might be so large that the F_2 track would "overshoot" the #V# cross-over F_2 frequency. In our experiments we also found that the number of diphthong responses varied widely between subjects. But we did not find any variation in the "direction" of the responses (i.e., perceptual over- or undershoot) between subjects when responses to formant curvature in general were examined.

Lindblom and Studdert-Kennedy (1967) concluded that vowel perception in context was influenced by perceptual-overshoot. When we consider the fact that their tokens strongly resembled glides or diphthongs (or even triphthongs), we might conclude instead, that they have only showed perceptual-overshoot for glides and

diphthongs. When their tokens were interpreted as diphthongs, this might also explain the variation in behaviour between the subjects.

2.3.2 The paper of Nearey (1989)

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

The mid-point values of F_1, F_3, and F_4 were fixed at 700, 2400, and 4000 Hz, respectively. The F_2 mid-point value was varied in 20 steps from 900 to 1800 Hz (see figure 5). The vowel tokens were 100 ms long and had an F_0 of 120 Hz. The on/offset values for F_1, F_2, and F_3 were 150, 2000, and 3000 Hz for /dVd/ and 150, 700, and 2100 Hz for /bVb/, respectively. In principle, this would have given F_2 excursion sizes ranging from 200 to 1100 Hz for both /dVd/ and /bVb/ tokens. However, due to the low F_1 on/offset frequencies, the F_2 amplitude was very low at the formant on- and offset points. The real F_2 on- and offset frequencies were measured at the -20 dB point and ranged from about 800 to 1170 Hz for /bVb/ tokens and from about 1510 to 1920 Hz for /dVd/ tokens. This gives F_2 excursion sizes ranging from 100 to 630 Hz and from 120 to 610 Hz for /bVb/ and /dVd/ tokens respectively.

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

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

Nearey compared the perceptual-overshoot he found with the amount necessary to compensate for the target-undershoot predicted by Lindblom (1963) and Broad and Clermont (1987). It was clear that the amount of perceptual-overshoot found in his listening experiments (11 to 125 Hz) was insufficient to compensate for the expected amount of target-undershoot (140 to 260 Hz). Again, there even seemed to be a

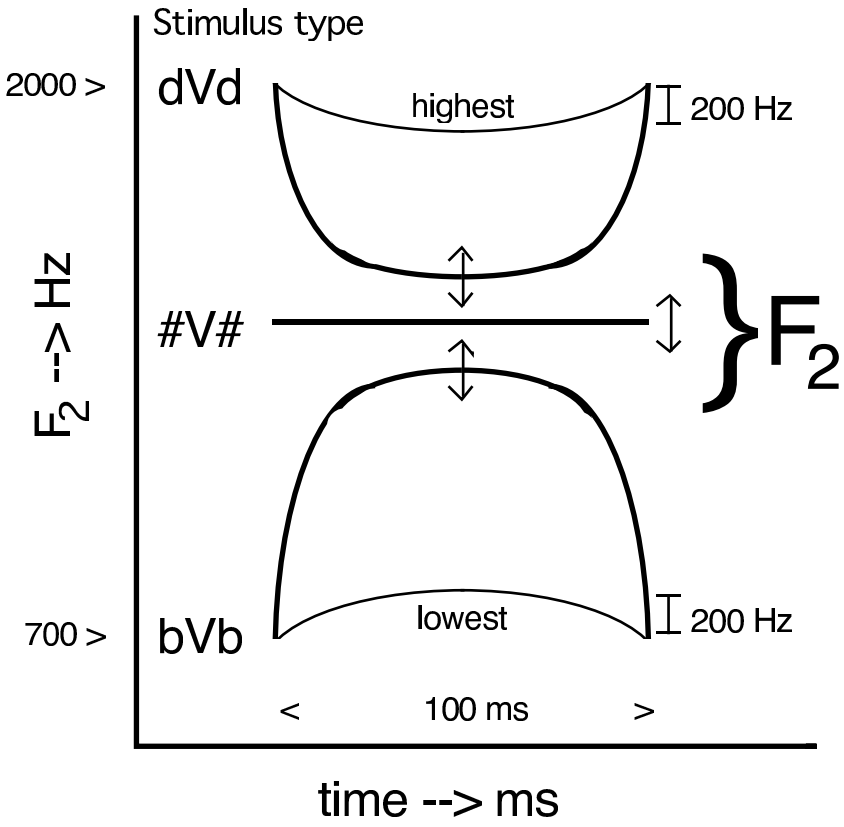

Figure 5. Contrasting F2 track shapes used by Nearey (1989). For all stimulus types (i.e., bVb, #V#, and dVd) the F2 midpoint frequencies were varied between 'lowest' (900 Hz) and 'highest' (1800 Hz) along an /ɒ/-/ʌ/-/ɛ/ continuum. Apart from F_2, also F_2 and F_3 tracks (not shown) were shaped with a sixth order polynomial (see text), $F(t) = F_{target} + (F_{initial} - F_{target}) \cdot (2 \cdot t/Duration - 1)^6$). This drawing is not to scale.

negative correlation between the expected amount of target-undershoot and the amount of perceptual-overshoot actually found, or no relation at all.

Considering the fact that 75% of the formant change was confined to only 20% of the total duration (compared to 50% of duration in Lindblom and Studdert-Kennedy, 1967), it is remarkable that any effect of formant track shape could be detected at all. The fact that these short transitions of the vowel have such a large effect on vowel identity suggests that the "perceptual-overshoot" found in this experiment is not caused by formant track shape itself but by the perception of the context it caused. This would mean that the context, and not the vowel realization, triggers the compensation for coarticulation. Such a mechanism would induce perceptual-overshoot in any vowel realizations presented in the proper context. This mechanism could be tested by presenting stationary tokens in the same context as "correct" and "incorrect" dynamic tokens. However, it is difficult to elicit good stop consonant percepts without the proper formant movements. This means that experiments using stop consonants as a context could not readily distinguish between vowel inherent effects and context effects on perception.

Nearey concludes that his experiments have shown the existence of perceptual compensation effects for formant-undershoot in production. The amount of compensation found is quite small and seems to be unrelated to the formant excursion size or the formant track on- and offglide slopes. There also seems to be no relation with the amount of expected formant-undershoot in production. Therefore, the

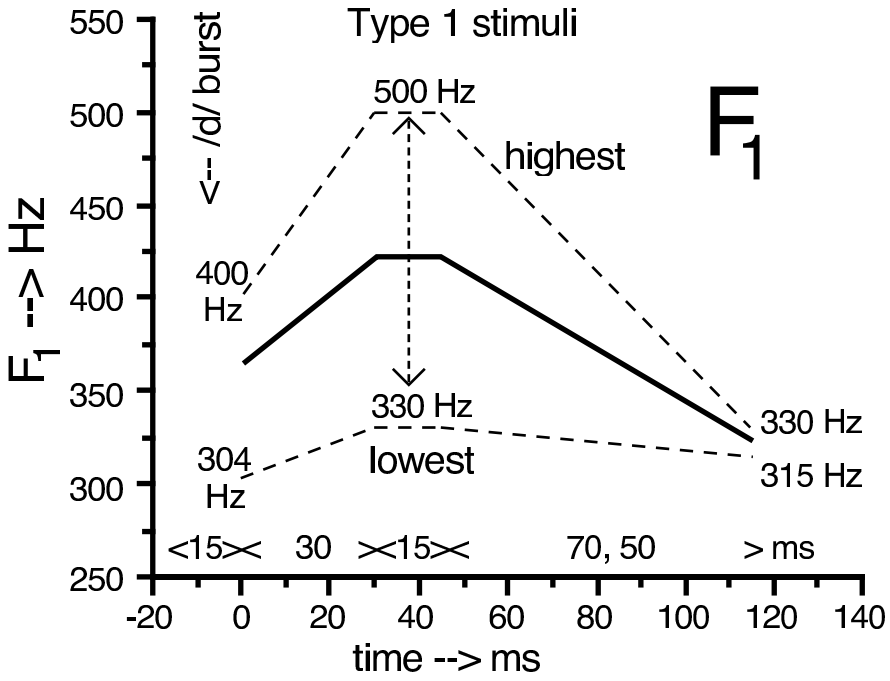

Figure 6. F_1 track shape used by Di Benedetto (1989b). F_1 tracks were varied between lowest (330 Hz top) and highest (500 Hz top) along an /i ɪ/-/e ɛ/ continuum. Higher formants had symmetrically shaped tracks. The contrasting type 2 stimuli were time-reversed (mirror images) of the type 1 stimuli shown. Responses to these type 1 stimuli were compared with the responses to the corresponding (time-reversed) type 2 stimuli. The vowel parts were preceded by a synthetic aperiodic /d/ type burst sound with a length of 15 ms. Note that two token durations were used: tokens with 70 ms offglides and tokens with 50 ms offglides (onglides for type 2 stimuli).

"overshoot" found could have been the result of some high level compensation for coarticulation instead of a low level "perceptual" overshoot.

2.3.3 The paper of Di Benedetto (1989b)

Di Benedetto (1989b) also found evidence that the shape of the F_1 formant tracks did influence vowel identification. She presented vowel tokens in a /dVd/ syllable (i.e., starting with a 15 ms synthetic /d/-burst) with linear on- and offglides and a plateau of 15 ms in F_1 (see figure 6). The F_1 maximum varied between 330-500 Hz in 10 steps, the F_1 excursion size varied between 26-170 Hz (1.4-7.2 semitones). The F_2 changed symmetrically from 2593 to 2800 Hz and back. Her seven subjects had different language backgrounds, i.e. American English (4), Italian (2), and Japanese (1). Subjects were asked to label the tokens as /ɪ i/ (high, closed) or /e ɛ/ (non-high, open) depending on native language (using her terminology).

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

kens with variable F_1 track shapes. From the results of this experiment she concluded that her subjects used the complete formant tracks to identify vowels.

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

However, when I compare her results with those discussed in section 1.5 I am inclined to conclude that the use of a weighted formant time average by the subjects is the more likely explanation. A conclusion that was also favoured by Di Benedetto herself. With her data I made a (very) crude estimate of the relative weights attached to the first and second half of each of her tokens. The relative weights of the first and second half showed to be around 8:1 in favour of the first half (both durations, all subjects). This contrasts sharply with our own results that showed that the final half was most important for identification (section 1.5). This might mean that there was an effect of formant track slope after all. It is possible that the perception of the initial /d/ interfered with the weighting of the formant tracks. We might speculate that the curious effect of formant onset slope on cross-over frequencies mentioned above might be linked to a shift in the perception of the pre-vocalic consonant, which again might have induced a stronger perceptual compensation in the form of overshoot. This could be tested by presenting the tokens from Di Benedetto's experiment in isolation as well as in context.

2.3.4 The paper of Fox (1989)

Fox (1989) performed silent-center experiments with synthetic stimuli using a 7-step /bɪb/-/bɛb/ continuum. Next to the mid-point values, his tokens also modelled the "natural" movements of F_1-F_3 with linear line segments. The total duration of the tokens was 300 ms. The duration of the vowel parts of the tokens was 255 ms, they consisted of symmetrical linear on- and offglides of 30 ms each and a stationary medial part of 195 ms (see figure 7). The vowel parts were preceded by a 5 ms synthetic /b/ type burst and a 40 ms segment of synthetic /b/ murmur. Listeners were asked to identify these tokens as either /bɪb/ or /bɛb/, or to discriminate pairs of tokens to be the same or different. He presented listeners with the full tokens, silent-center tokens, and with medial vowel tokens. The silent-center tokens consisted of only the first and last 4 pitch periods of each vowel token (35 ms and 38 ms respectively) with a silent gap in between. The stationary tokens only contained the stationary medial vowel part (185 ms). The on-/offset to mid-point excursion sizes in the 7 tokens were in the range (maximal-minimal formant frequency), F_1: 30-95 Hz (1.3-3.5 semitones), F_2: 306-265 Hz (3.2-3.0 semitones), and F_3: 177-128 Hz (1.2-0.9 semitones). The formant track excursions in this continuum were such that a higher F_1 excursion size and a lower F_2 or F_3 excursion size indicated a more /ɛ/-like vowel (see figure 7). It would therefore be difficult to distinguish perceptual over- and undershoot of formant mid-point values. Evidence for perceptual-overshoot from one formant would point to perceptual-undershoot for another formant!

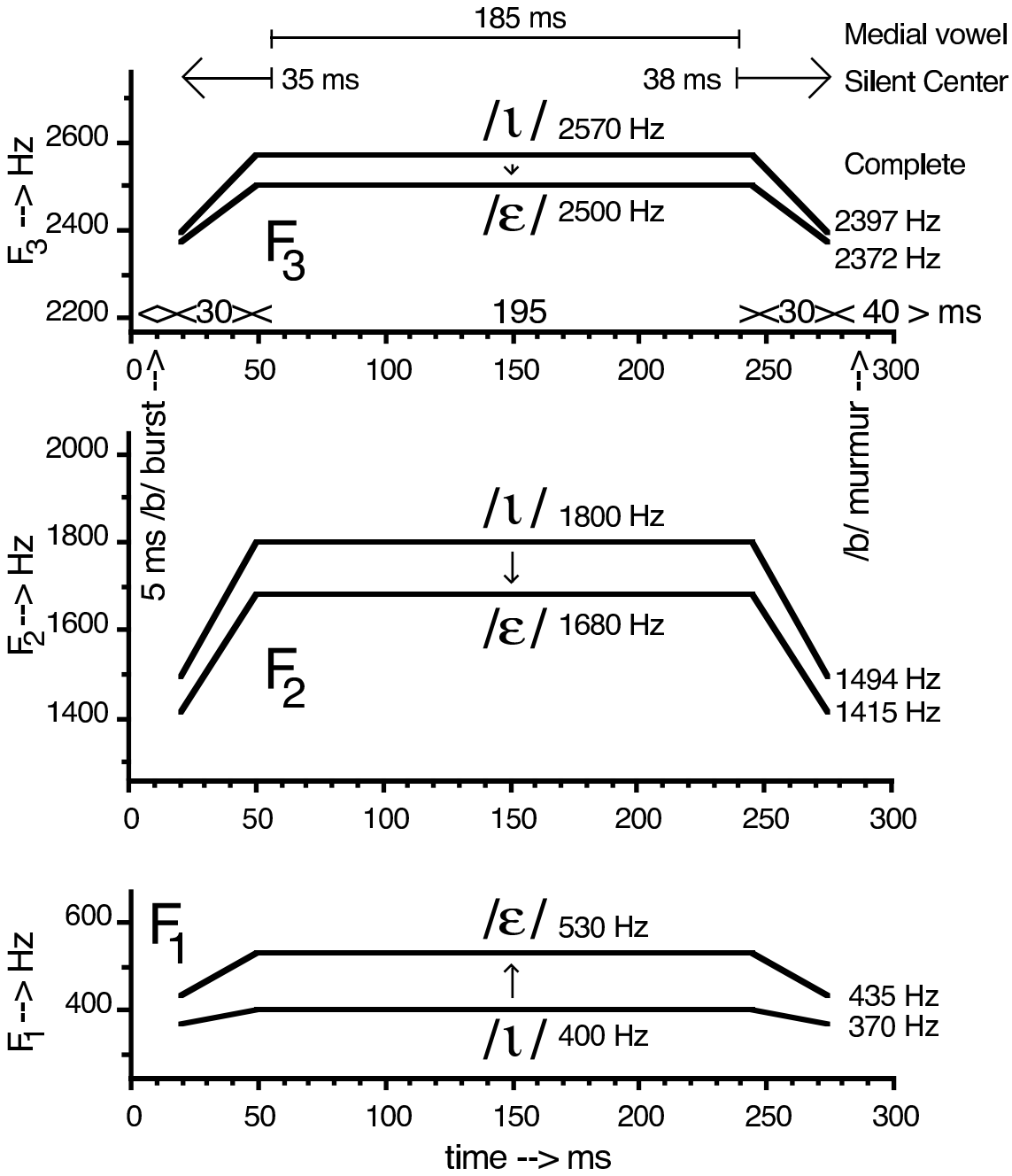

Figure 7. Formant track shapes used by Fox (1989). Vowel formant values varied along a /bɪb/-/bɛb/ continuum. Vowel tokens were preceded by a 5 ms synthetic /b/ type burst and followed by 40 ms of synthetic /b/ murmur. Tokens were presented as complete 'syllables', as Silent Center tokens (excluding the medial 185 ms) and as medial vowels only (including only the central 185 ms of the stationary part).

In a set of discrimination experiments Fox was able to show that the silent-center tokens were perceived differently from the stationary medial vowel tokens. In separate experiments he presented the silent-center tokens also with only the outer 1, 2, 3, or the full 4 pitch periods of the on- and offset transitions, i.e. removing respectively 3, 2, 1, or no pitch periods from the inside of the original silent-center tokens. It appeared that the number of pitch periods present in the tokens influenced the identification scores. In general, the more pitch periods were present in a token,

the more /ɛ/-responses it got. This result could be explained by assuming that subjects identified the tokens on the transition end-point formant frequencies.

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

Fox interpreted his results as evidence for dynamic-specification without discussing the direction of the perceptual difference between stationary and transition-only stimuli. However, from his figures 8 and 9, it followed that his results could be explained by assuming perceptual-undershoot of the F_2 or perceptual-overshoot of the F_1. For low F_1 values, there is little difference between token responses. At higher F_1 frequencies there is a steady excess of /ɪ/ responses for the stationary tokens. This finding is consistent with both perceptual-undershoot of the F_2 and perceptual-overshoot of the F_1 in the silent-center tokens. However, the excess /ɪ/ responses in the experiments of Fox do remind us of the same excess /ɪ/ responses we found in our own experiments (see section 1.5). In our experiments the increase in the number of /ɪ/ responses at short token durations was indiscriminate and could not be traced to any kind of under- or overshoot. This raises the possibility that the increase of /ɪ/ responses in both experiments might have been caused by some factor unrelated to formant track shape. I will not pursue this matter further because at the moment this possibility cannot be substantiated.

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

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

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

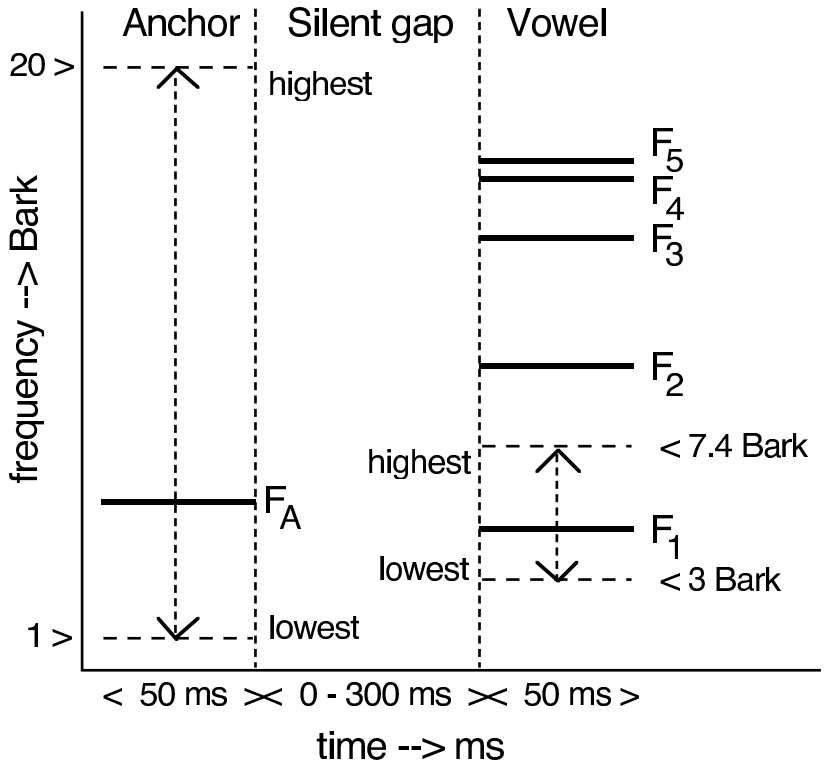

Figure 8. Token structure as used by Akagi (1993). Synthetic stationary vowel tokens were preceded by a variable silent gap and a single formant anchor sound (all sounds had a duration of 50 ms). The F1 frequency varied along an /u/-/a/ continuum The formant value used for the anchor sound varied between 1 and 20 bark.

2.3.5 The paper of Akagi (1993)

As part of a larger effort to model coarticulation, Akagi (1993) studied vowel formant boundary shifts in perception (see also Akagi, 1992; and the review of this work by Repp, 1993). In his experiment, two Japanese subjects were asked to identify synthetic vowels as either /u/ or /a/. The stimuli in this experiment were stationary, five formant, vowel tokens with a duration of 50 ms. They were preceded by a stationary single formant anchor of 50 ms that was separated from the vowel token by a variable silent gap (see figure 8). The F_1 of the vowel tokens varied in such a way as to form a continuum from /u/ to /a/. The formant frequency of the anchor token preceding the vowel varied from below the lowest F_1 frequency to over the F_5 frequency. The duration of the silent gap, separating the anchor from the vowel token, varied from 0-300 ms in 25 ms steps.

The results of his experiments showed that the F_1 values for which /u/ responses changed into /a/ responses depended on both the formant frequency of the anchor and the duration of the silent gap. Akagi concluded that there was an assimilation effect (i.e., perceptual-undershoot) when the duration of the silent gap was below 70 ms (or should that be 75 ms because of the 25 ms step-size?) and a contrast effect when the duration of the silent gap was longer (i.e., perceptual-overshoot). This means that the presence of perceptual under- or overshoot was determined by the duration of the silent gap. Therefore, it seems that it was the temporal structure of the context that influenced the perception of the vowel more than the spectral difference between anchor and vowel token. This points towards an important role for context in the

process of vowel identification. It also shows that perceptual-overshoot is not limited to "natural" stimuli.

2.3.6 The papers of Kuwabara (1983, 1993)

Kuwabara (1983) extended the experiments of Lindblom and Studdert-Kennedy (1967) with synthetic trisyllabic vowel sequences of the type /uVu/ (which replaced the wVw stimulus of Lindblom and Studdert-Kennedy, 1967). He used stationary 100 ms 'context' /u/ vowels and a 200 ms medial vowel with parabolically shaped F_1, F_2 or both (cf. the wVw formant track shape in figure 4). In the first type of stimuli, the F_2 frequency was fixed on the /u/ value of 1300 Hz and the F_1 midpoint frequency varied from 300 Hz (/u/) to 700 Hz (/a/) in 16 steps. In the second type of stimuli, the F_1 frequency was fixed on the /u/ value of 300 Hz and the F_2 midpoint frequency varied from 1300 Hz (/u/) to 2100 Hz (/i/). The third kind of stimuli combined the first two in that the F_1 midpoint frequency varied from 300 Hz to 700 Hz and the F_2 midpoint frequency from 1300 Hz to 2100 Hz (/u/ to /e/). Next to these, ranges of stationary vowel tokens (200 ms) were constructed for comparison.

Both the dynamical and the stationary vowels were each presented in pseudo-random order to four Japanese subjects. The subjects were asked to identify them as one of the five Japanese vowels (open response). One of the subjects could not distinguish the tokens on the /u i/ continuum, and the responses of this listener were not used. For the stationary vowels the boundary frequencies between /u/ and /a/ (type 1), /u/ and /i/ (type 2) and /u/ and /e/ (type 3) were found to lie around stimulus number 7, that is halfway in the continuum.

However, for the trisyllabic uVu stimuli the boundaries lay around the third stimulus step. This means that a frequency difference between the medial V and the u_u context of only 85 Hz in the F_1 or 130 Hz for the F_2 was enough to induce a different vowel response (respectively type 1 and type 2 stimuli, this became 50 Hz and 100 Hz for the type 3 stimuli). These values are close to the difference limens for time varying symmetric vowel tokens (Mermelstein, 1978). This means that the subjects of Kuwabara would respond a non-/u/ vowel whenever they could hear a difference between the u_u context and the medial vowel. It is interesting to note that the subjects responded with /o/ labels to type 1 stimuli (in the /u/-/a/ continuum) when the F1 value was not high enough to warrant an /a/ label. Apparently, no /o/ labels were responded to the stationary tokens.

In a second experiment, which was presented in an extended form in 1993, Kuwabara (1983) constructed uVu and eVe tokens from linear segments (see figure 9). A 200 ms medial vowel was surrounded by two 75 ms stationary vowel segments (either /u/ or /e/-like). The F_1 frequency of the stationary central part of the medial vowels was varied in 25 steps between 300 Hz (/u/) and 550 Hz (/e/). Simultaneously, the F_2 frequency was varied between 1300 Hz (/u/) and 1800 Hz (/e/). The stationary central 100 ms of the medial vowel was connected to the stationary u_u or e_e context with two 50 ms linear segments (see figure 9). Identification results for these tokens showed, again, that a difference of 50 Hz in F_1 and 100 Hz in F2 between medial vowel and u_u or e_e context was enough to induce a change in response.

In an ABX experiment, (unspecified) subjects were asked which stationary vowel token (B or X) was most like a given uVu token (A). Kuwabara (1993) found that, as long as the tokens were identified as the same vowel, stationary vowel tokens had to be 7 steps further along the stimulus continuum to be heard as alike, i.e. the F2 of an uVu stimulus was perceived as 140 Hz higher, F_1 as 70 Hz higher.

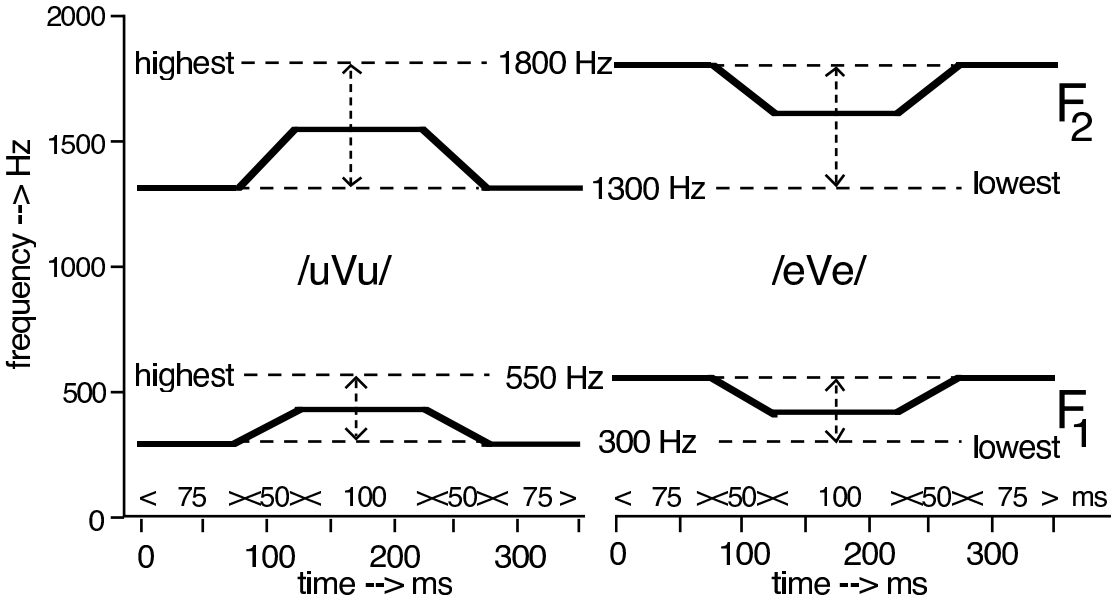

Figure 9. Formant track shapes used by Kuwabara (1993, 1985). Formant values of the medial vowel in trisyllabic uVu and eVe stimuli were varied along an /u/-/e/ continuum (from lowest to highest). These tokens were used in an identification experiment, an ABX comparison experiment and in a dichotic listening experiment.

In a dichotic listening experiment, Kuwabara (1983) presented uVu and eVe tokens simultaneously to both ears (each token to one ear only). Two types of combinations were used:

1: Both medial vowels had equal or close formant frequencies, but had been labelled as different vowels in the identification experiment.

2: Both medial vowels were labelled as the same vowel in the identification experiment, but had very dissimilar formant frequencies.

Five Japanese subjects were asked whether or not the two medial vowels would 'fuse', i.e. would be heard as a single sound instead of being heard as two separate sounds. Kuwabara reasoned that if the compensation for vowel formant undershoot in production would be a peripheral auditory function, then 'fusion' of the two different medial vowels would depend on whether the 'compensated', overshot, values were close. If, however, the compensation depends on a process in the central brain, 'fusion' would only depend on the actual spectral distance of the medial vowels. His results clearly supported the latter option. Fusion of the medial vowels in simultaneously presented uVu and eVe tokens only depended on the spectral closeness of the medial vowels, and not on the labels responded in identification experiments.

Kuwabara (1983, 1993) concluded that listeners use a compensatory mechanism that reverses coarticulation when they evaluate vowels. This mechanism or process operates *after* sounds from both ears are integrated. Kuwabara interprets his results in terms of boundary shifts. In his identification experiments he reports boundary shifts that were almost twice the excursion size, i.e., a shift of 100/280 Hz with an excursion size of 85/130 Hz in the first experiment (F_1/F_2). In his ABX experiment (Kuwabara, 1993) he reports a 140/70 Hz shift with respect to vowel quality (F_1/F_2). This seems too large a shift with respect to the excursion size to call it perceptual overshoot or dynamical specification (terms that were *not* used by Kuwabara). From his results it looks as if his subjects exaggerated the differences between the medial vowel and the context. A medial vowel would be labelled as different as soon as the difference was audible (Mermelstein, 1978). Dichotic fusion was *not* dependent on the difference

between medial vowel and context. Together these results suggest that the compensatory function, or reverse articulation, is a central brain function that is at least partly dependent of the context, and not only on the vowel segment itself.

2.3.7 What factor could induce perceptual-overshoot?

In addition to experiments with speech-like signals, there are several psycho-physical studies with formant sweeps that indicated that the presence of formant transitions themselves is not enough to induce perceptual-overshoot in listeners (e.g., Brady et al., 1961; Pols et al., 1984; Van Wieringen, 1991). Akagi's (1993) study indicated that the structure of the vowel context might be crucial to the existence of perceptual-overshoot, or dynamic-cospecification in general. When we compare the results of our own study to that of Lindblom and Studdert-Kennedy (1967), we see that it is exactly there that the major differences are located (leaving aside the differences in response categories). They supplied a convincing and contrasting context to their vowel tokens, we did not. Nearey (1989) also ensured that the formant track slopes at the consonant-vowel transitions were as acute as those found in natural speech. He described the percepts of the plosives as convincing. Both Di Benedetto (1989b) and Fox (1989) used linear line segments to model plosive-vowel transitions. The quite long and gradual vowel formant on- and offset transitions used by Di Benedetto and Fox cannot be expected to have added much to the perception of the plosive context (compare these with the very acute on- and offsets of Nearey, 1989). What is more important, in these latter studies all vowel tokens were presented in the same context so any effect of context would have gone unnoticed. It seems therefore, that the presence of perceptual-overshoot depends more on the perception of the context than on the actual formant track shape, i.e. formant excursion size, inside the vowel token itself (see also Tohkura et al., 1992; Repp, 1993; for related studies on context effects). This is supported by the fact that in none of the experiments the size of perceptual-overshoot of formant mid-point values was positively correlated with formant excursion sizes or formant track slopes. Without the perception of a proper context, subjects seemed to have reverted to the use of a weighted formant average to identify the vowel tokens.

2.4 Experiments using natural speech

The above experiments all used *synthetic* speech. The quality of synthetic signals as a model of *natural* speech depends on our current knowledge of natural speech. Because we do not know the cues listeners use to identify vowels, such cues might be improperly represented in synthetic signals. Of course, all the important features can be found in natural speech itself. Therefore, a lot of experiments have been done using natural speech to seek evidence for or against the use of dynamic information in vowel recognition. Of these experiments, only a handful can be discussed here.

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

2.4.1 The influence of context on vowel intelligibility

Vowels spoken in consonantal context have mid-point spectra that differ from spectra taken from canonical realizations, i.e. vowels spoken in isolation (e.g., Stevens and House, 1963; Lindblom, 1963). It is therefore logical to suspect that vowels spoken in context are less well understood than those uttered in isolation. Initial experiments comparing vowel recognition in context with recognition of isolated vowels claimed that vowels in context were actually recognized *better* than those spoken in isolation (10% versus 30% errors, e.g., Strange et al., 1976; Gottfried and Strange, 1980; Strange and Gottfried, 1980). However, by taking more care on various methodological aspects such as dialect background and response procedure, Macchi (1980) found no difference between the intelligibility of isolated vowels and vowels in context (errors around 2%, see also the extensive reviews of Strange, 1989a and Nearey, 1989). Koopmans-van Beinum (1980) found that vowels excised from one-syllable words uttered in isolation were recognized worse than vowels spoken in isolation (16% versus 10% errors, $p \leq 0.0001$, her tables 7.2 and 7.4). Most of the errors in the responses to her isolated vowels were caused by the problems of identifying the realizations of the short vowels /ɔ ɑ ɪ œ/ spoken in isolation because of their relatively long durations. Removing responses to these four tokens made the differences even more dramatic (13% versus 3% errors respectively). This shows that the difficulties with the duration cannot explain the differences in identification scores. Unstressed vowels from free conversation, which were severely reduced, performed extremely poorly (77% *errors*). As these unstressed and reduced realizations were very short, the errors were now concentrated in the responses to the long vowels (/a e/ received only 4.2% *correct* responses). However, even the four short vowels mentioned before were identified incorrectly in more than half of the responses (54% *errors*).

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

Huang presented consonant-vowel-consonant syllables to subjects as well as the excised vowels from these syllables (i.e., without the consonants). The recognition rate for the full syllables was more than 8% higher than that for the excised vowels alone (79% versus 71%, $p \leq 0.001$, Huang, 1991; calculated from her tables 4.4-4.11). Kuwabara found an even more dramatic effect of context. He used Japanese three-vowel sequences, taken from sentences. The medial vowel of each sequence was presented both in context and separately in isolation (i.e., without the two flanking vowels). Recognition of the medial vowel in isolation was much worse than in context (recognition rates of 80% and 96% respectively). However, it was not clear how much of the Vowel-Vowel transitions was included with the medial vowels when they were presented in isolation. It is therefore difficult to assess the significance of his results.

Next to the presence of the context, the nature of the context might also influence vowel recognition (as was also found by Gottfried and Strange, 1980). The results of Koopmans-van Beinum, Huang and Kuwabara show that the conclusion that vowels in context are recognized as well as vowels spoken in isolation (Strange, 1989a;

Nearey, 1989) does not hold for vowel realizations presented without their proper context.

Using the above error rates for excised vowels and vowels in context, it is possible to estimate how much of the information needed to identify individual vowels was extracted from the context in these experiments. I will assume unbiased responses and random errors. The information needed for perfect identification is $^2\log$(number of vowel categories) bits per vowel. The information *lost* by presenting the vowels out of their original context can be calculated from the stimulus-response confusion matrix. For an error rate ε the amount of information lost is of the order $-(\varepsilon\,^2\log(\varepsilon) + (1-\varepsilon)\,^2\log(1-\varepsilon))$. The relative amount of information present in the removed context is then the quotient of these two values. For example, using the numbers of Huang (1991), the amount of information necessary to identify five vowel categories is 2.32 bit. In syllabic context the error rate was 21%, so the amount of information present in vowels in syllabic context = $2.32 - 0.21 \cdot {}^2\log(0.21) - 0.79 \cdot {}^2\log(0.79) = 1.58$ bit (0.74 bit is lost). Presented in isolation, the error rate increases to 29%, which corresponds to a loss of information of 0.87 bit. This latter loss is 38% of the information necessary for perfect recognition. With respect to the syllabic context the loss is 0.87-0.74 = 0.13 bit, which is 8% of the information present in the syllabic context (i.e., 1.58 bit). Using the recognition rates presented above, we see that, in general, 10% or less of the information present is lost by removing the immediate context. However, severely reduced unstressed vowels from free conversation can contain *less* than half of the information necessary to identify them (46% recognition rate for four vowels, Koopmans-van Beinum, 1980).

2.4.2 The importance of the transition for vowel recognition

Experiments that try to determine the importance of consonant-vowel transitions in vowel recognition, generally use the silent-center paradigm (as in section 2.3.4, but now with *natural* speech). Simple syllables, mostly of the stop-vowel-stop type (e.g., /bVb/) are spoken in carrier sentences. The vocalic part of the target syllables is divided into three parts: an initial part which contains all of the consonant-vowel transition (e.g., /bV/), a final part, which contains all of the vowel-consonant transition (e.g., /Vb/), and a medial part which contains the more or less stationary vowel kernel. Generally, care is taken to include only the transitions in the initial and final parts and to exclude parts of the vowel kernel. Then two new kinds of syllables are constructed, one containing only the medial part and one containing only the initial and final transition parts with silence substituted for the medial part. The original as well as the new syllables are then presented to listeners and the number of recognition errors is noted.

Several variations of the basic design of silent-center experiments are in use. The length of the syllables, either the medial vowel kernels or the silent centers, can be manipulated to exclude the original durational information from the tokens. The initial and final parts of the vowels used to create the silent-center syllables can be taken from different realizations or even from different speakers (with opposite sex). Finally, the initial and final parts can also be presented separately in isolation. Sometimes, vowels spoken in isolation are also added for comparison.

Several studies using the silent-center paradigm are reported in the literature (e.g., Strange et al., 1983; Verbrugge and Rakerd, 1986; Strange, 1989b; Andruski and Nearey, 1992). Verbrugge and Rakerd asked listeners to identify /bVb/ syllables. The vowel could be one of /ɪ i ɛ æ ʌ ɑ U u/. They heard the original syllables, silent-center syllables (with the medial 60% removed), hybrid silent-center syllables whose

initial and final part were from different speakers (of opposite sex), and the initial and final parts separately. The pattern of recognition errors was typical for experiments with silent-center syllables. The error rate of the labelling was: whole syllables 8%, silent-centers 20%, hybrid silent-centers 26%, initial parts 48%, and final parts 66% errors. All differences were significant, except for the differences between the two types of silent-center syllables. The error rate was much lower when short-long vowel errors were removed (i.e., 4%, 12%, 18%, 26%, and 40% respectively for four vowel categories). Others found that the centers-only were recognized as well as the silent-center syllables (Strange et al, 1983; Strange, 1989b). From these latter studies it could also be deduced that removing durational information almost doubled the error rate: from 6-8% to 13-33% errors for 10 vowels and 4 long/short pairs. This indicates that durational information constitutes around 10% of the information used to identify vowels (see section 2.4.1). This is less than would have been expected from the number of long/short vowel pairs (if duration was the only cue, 4 long/short pairs out of 10 vowels would induce an error rate of 40%, or somewhat less than 30% of the information). Of course, it is clear that spectral cues are also used.

Verbrugge and Rakerd tried to devise a way to predict the silent-center recognition scores from the individual recognition scores of the initial and final parts. In general, combining the recognition scores of the initial and final parts severely overestimated the recognition errors for the silent-center syllables, even when short-long errors were not counted. This was even so under the unlikely assumption that the recognition would be incorrect only when both parts were not recognized correctly. The same difference between recognition of individual parts and complete silent-center syllables was found in the other studies (Strange et al., 1983; Strange, 1989b). Both Verbrugge and Rakerd (1986) and Strange (1989b) found that the initial parts were recognized significantly better than the final parts. Strange also found that there was no difference in the error rate between the centers and the initial parts when durational information was removed from the centers. This result is similar to our own results. In section 1.5 (figure 3) we found that the difference in responses between onglide-only tokens and stationary tokens was small. Both differed markedly from the offglide-only tokens. The apparent difference in "error rate" in silent-center experiments and our own experiments (section 1.5) can be attributed to methodological differences (type of speech, language). Furthermore, it is difficult to define an error rate for our synthetic stimuli ("net shift" is not synonymous with error rate) as we do not know what the "correct" response should be.

What is striking in most of these studies is the small difference in recognition rate between the original syllables and the silent-center syllables. The 12% difference found by Verbrugge and Rakerd (8% versus 20% errors) was the largest of the studies discussed here. Strange et al. (1983) and Strange (1989b) found no significant difference at all between these two types of syllables. Verbrugge and Rakerd found that combining the initial part of a man's vowel realization with the final part of a female's, and vice versa, did not significantly affect the recognition of these hybrid silent-center syllables. The results of the latter study indicate that the recognition of the vowel "target" frequency could not have been the result of a simple extrapolation of the formant tracks into the silent center. It strongly suggests that both parts were processed separately and that the resulting vowel "targets" were abstracted in such a way that they could be combined into a single, more dependable target.

In general, the results from these silent-center studies support our own results. We saw that the responses to the offglide transition of a vowel were generally shifted (i.e., caused more "errors") from those to the onglide and stationary medial parts. We also saw that there is at most only a small difference between responses to the onglide transition part and to the stationary medial part (Strange et al., 1983; Strange, 1989b).

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

3 Integration of the available results

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

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

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

In our own experiments (section 1.5) we wanted to compare identical vowel realizations in different context (including presentation in isolation). We wanted to test the effects of the presence of a context *an sich* on the identification of vowel tokens. To achieve this, we deliberately did not change the formant track shape to match the context in which the vowel token was presented. Therefore, the vowels in the /nVf/ and /fVn/ pseudo-syllables we used might have been perceived as still being "pronounced" in isolation and not in well integrated syllables. Furthermore, we do not know whether /n/ and /f/ are capable of inducing a detectable amount of compensation even in natural speech. In neither case, any compensation would have been found in our experiments.

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

If the compensation for coarticulation is performed only after the context is "reconstructed" by the listener, this would also explain the good results for hybrid silent-center syllables. Both parts in a hybrid silent-center syllable give the same (hypothetical) "proto-targets" for the vowel and context. These would then have to be combined and the compensation would have to be determined from the combination of these elements. What information is actually used to determine the compensation is not clear at this moment. The results of the experiments using synthetic speech do point towards dynamic information, specifying formant movements. But in these experiments, the dynamic information strongly correlated with the "locus" values of the consonants in the context. This still leaves the possibility that, in these experiments too, the listeners used the *identity* of the perceived consonants to help identify the vowel and not the formant track shape itself. It is therefore not really possible to distinguish between these two possibilities at the moment. They could be distinguished by comparing the identification of vowel kernels (with no or small formant movements) and complete vowel segments (with all or most of the formant movements) presented in isolation with the identification of the complete vowel segments when presented in their natural context. Experiments to do this are currently performed at our institute.

From the experiments using natural speech, we can estimate the magnitude of the amount of information that is extracted from the context. The fraction of the information lost when vowels were presented in isolation rather than in syllables was generally less than 10% of the total. Another 10% may be lost when durational cues are removed. A fraction of 80% or more of the information used to distinguish vowel realizations was apparently extracted from the vowel segments themselves.

4 Conclusions

We can summarize the evidence discussed in this paper as follows. The shape of formant tracks carries information that could be used to compensate for coarticulatory formant-undershoot in production and so could help to improve vowel identification. Experiments with synthetic speech indicated that, when tokens were presented in an

appropriate (contrasting) context, subjects did use “dynamic” information in a way that might compensate for the effects of coarticulation in that context. Without such a context, this dynamic information was not used by subjects and was even detrimental to "identifying" any canonical target, assumed to correspond to the given formant track shape. Experiments with natural speech indicated that (parts of) vowel realizations were identified better in their original context than when excised from it and presented in isolation. In their original context, vowel realizations were equally intelligible as vowels spoken in isolation.

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

Ballpark estimates of the amount of information lost when vowel realizations were excised from syllables and presented in isolation showed that generally less than 10% of the total amount present is extracted from the syllabic context. This amount is comparable to the amount of information lost when durational cues were removed.

Acknowledgements

I would like to thank Louis Pols for his numerous and valuable suggestions and comments that helped me to compile this paper.

References

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

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

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

Andruski, J.E. & Nearey, T.M. (**1992**): 'On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables', *Journal of the Acoustical Society of America* **91**, 390-410.

Brady, P.T., House, A.S., and Stevens, K.N. (**1961**): 'Perception of sounds characterized by a rapidly changing resonant frequency', *Journal of the Acoustical Society of America* **33**, 1357-1362.

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

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

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

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

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

Gottfried, T.L. & Strange, W. (**1980**): 'Identification of coarticulated vowels', *Journal of the Acoustical Society of America* **68**, 1626-1635.

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

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

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

Kuwabara, H. (**1983**): 'Vowel identification and dichotic fusion of time-varying synthetic speech sounds', *Acoustica* **53**, 143-151.

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

Kuwabara, H. (**1993**): 'Temporal effect on the perception of continuous speech and a possible mechanism in the human auditory system', *Proceedings of Eurospeech '93*, Berlin, Germany, 713-716.

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

Lindblom, B. (**1983**): 'Economy of speech gestures' in *The production of speech*, edited by P.F. MacNeilage (Springer-Verlag, New York, NY), 217-246.

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

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

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

Mermelstein, P. (1978): 'Difference limens for formant frequencies of steady-state and consonant-bound vowels', *Journal of the Acoustical Society of America* **63**, 572-580.

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

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

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

Miller, J.L. & Baer, T. (**1983**): 'Some effects of speaking rate on the production of /b/ and /w/', *Journal of the Acoustical Society of America* **73**, 1751-1755.
Nearey, T.M. (**1989**): 'Static, dynamic, and relational properties in vowel perception', *Journal of the Acoustical Society of America* **85**, 2088-2113.
Nossair, Z.B. & Zahorian, S.A. (**1991**): 'Dynamic spectral shape features as acoustic correlates for initial stop consonants', *Journal of the Acoustical Society of America* **89**, 2978-2991.
O'Shaughnessy, D. (**1987**): *Speech Communication* (Addison-Wesley, Reading, MA).
Peeters, W.J.M. (**1991**): 'Diphthong dynamics', Ph.D. Thesis, State University of Utrecht, The Netherlands.
Polka, L. & Strange, W. (**1985**): 'Perceptual equivalence of acoustic cues that differentiate /r/ and /l/', *Journal of the Acoustical Society of America* **78**, 1187-1197.
Pols, L.C.W., Boxelaar, G.W., and Koopmans-van Beinum, F.J. (**1984**): 'Study on the role of formant transitions in vowel recognition using the matching paradigm', *Proceedings of The Institute of Acoustics* **6** (4), 371-379.
Pols, L.C.W. & Van Son, R.J.J.H. (**1993**): 'Acoustics and perception of dynamic vowel segments', *Speech Communication*. **13**, 135-147.
Repp, B.H. (**1993**): Review of Tohkura et al., 1992, *Language and Speech* **36**, 99-107.
Stevens, K.N. & House, A.S. (**1963**): 'Pertubation of vowel articulations by consonantal context: an acoustical study'. *Journal of Speech and Hearing Research* **6**, 111-128.
Strange, W. (**1989a**): 'Evolving theories of vowel perception', *Journal of the Acoustical Society of America* **85**, 2081-2087.
Strange, W. (**1989b**): 'Dynamic specification of coarticulated vowels spoken in sentence context', *Journal of the Acoustical Society of America* **85**, 2135-2153.
Strange, W. & Gottfried, T.L. (**1980**): 'Task variables in the study of vowel perception', *Journal of the Acoustical Society of America* **68**, 1622-1625.
Strange, W., Jenkins, J.J. & Johnson, T.L. (**1983**): 'Dynamic specification of coarticulated vowels', *Journal of the Acoustical Society of America* **74**, 695-705.
Strange, W., Verbrugge, R.R., Schankweiler, D.P. & Edman, T.R. (**1976**): 'Consonant environment specifies vowel identity', *Journal of the Acoustical Society of America* **60**, 213-224.
Tohkura, Y., Vatikiotis-Bateson, E. & Sagisaka Y. (editors) (**1992**): *Speech perception, production and linguistic structure* (Ohmsha, Tokyo; IOS Press, Amsterdam), 463 pp.
Van Bergem, D.R. (**1993**): 'Acoustic vowel reduction as a function of sentence accent, word stress, and word class', *Speech Communication* **12**, 1-23.
Van der Kamp, L.J.Th. & Pols, L.C.W. (**1971**): 'Perceptual analysis from confusions between vowels', *Acta Psychologica* **35**, 64-77.
Verbrugge, R.R. & Rakerd, B. (**1986**): 'Evidence of talker-independent information for vowels', *Language and Speech* **29**, 39-57.
Van Son, R.J.J.H. (**1993**): *Spectro-temporal features of vowel segments*, in *Studies in Language and Language use **3***. Ph.D. Thesis, University of Amsterdam, pp. 195.
Van Son, R.J.J.H. and Pols, L.C.W. (1993): 'Vowel identification as influenced by vowel duration and formant track shape', *Proceedings of Eurospeech '93*, Berlin, Germany, 285-288.
Van Wieringen, A. and Pols, L.C.W. (**1991**): 'Transition rate as a cue in the perception of one-formant speech-like synthetic stimuli', *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, 446-449.