2

FORMANT FREQUENCIES OF DUTCH VOWELS IN A TEXT, READ AT NORMAL AND FAST RATE*

Abstract

Speaking rate is thought to affect the spectral features of vowels. Target-undershoot models of vowel production predict more spectral reduction and coarticulation of vowels in fast-rate speech than in normal-rate speech. To test this prediction, a meaningful Dutch text of about 850 words was read twice by an experienced newscaster, once at a normal speaking rate and once as fast as possible. All realizations of seven different vowels and some realizations of the schwa (/ˊ/) were isolated. The first and second formant frequency values of all realizations were measured at five different points, each time by making cross-sections at different points in the vowel realization. The different selections of these points are based on procedures used in literature, such as maximal F_1 or mean formant value. No spectral vowel reduction was found that could be attributed to a faster speaking rate neither was a change in coarticulation found. The only systematic effect was a higher F_1 value in fast-rate speech irrespective of vowel identity. This possibly suggests a generally more open articulation of vowels, speaking louder, or some other general change in speaking style by our speaker when he speaks fast.

*Van Son, R.J.J.H. & Pols, L.C.W. (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", Journal of the Acoustical Society of America 88, 1683-1693.

Introduction

The effects of speaking rate on vowel production have been the objective of many studies (recent examples are e.g., Gopal and Syrdal, 1988; Den Os, 1988; Engstrand, 1988). Speaking rate is thought to affect most, though not solely, coarticulation and spectral reduction (Lindblom, 1963). Both of these are well attested phenomena that play an important role in normal speech (see e.g., the textbooks of O'Shaughnessy, 1987; Clark and Yallop, 1990). The effects of speaking rate on vowels are supposed to be examples of a more general influence of duration on the spectral structure of vowel realizations, an influence described by the target-undershoot model of vowel production, as formulated by Lindblom (1963), Gay (1981), and Lindblom (1983). This model predicts an increase in coarticulation, spectral reduction, or both, in vowels when their realizations shorten.

In its most simple form, the target-undershoot model states that vowels are characterized by their spectrum at a single point in the realization, the vowel target (see also Strange, 1989a). Due to several factors, a vowel realization generally has a target spectrum different from the ideal, or canonical, form. In the target-undershoot model this difference is said to shift the actual target spectrum from the canonical target toward the targets of the neighbouring phonemes (coarticulation) or towards a theoretical neutral vowel (spectral reduction). The articulators are said to miss the ideal target position by undershoot (Lindblom, 1963).

Several factors influencing vowel target spectra are identified and studied, for instance coarticulation (e.g., Pols, 1977; Whalen 1990), speaking style, stress and reduction (e.g., Koopmans-van Beinum, 1980). For the effects of duration on vowel formant frequency targets the results reported are ambiguous. At one hand, several studies support the notion of more target-undershoot with shorter vowel durations (Lindblom, 1963; Broad and Fertig, 1970; Gay et al., 1974; Broad and Clermont, 1987; Lindblom and Moon, 1988). Other studies, however, where unable to detect such an undershoot (Gay, 1978; Nord, 1987; Gopal and Syrdal, 1988; Den Os, 1988; Engstrand, 1988) or found the effect of speaking rate on vowel undershoot to be speaker dependent (Kuehn and Moll, 1976). It can be noted that support for the target-undershoot is mostly found when vowel realizations from only one speaking rate and style are studied, whereas it seems to be difficult to find support when differences between speaking rates or styles are studied. One reason for the ambiguity in the results of these studies might have been the experimental designs used in them. In all studies the speech is uttered under controlled conditions. The level of control often causes the distance between the experimental procedures and natural speech to be large and does not allow the results of these studies to be generalized to more normal modes of speech easily. Most studies used semantically empty words in carrier phrases, and the vowels are often placed in only a limited phoneme context. The experimental procedures used in different studies are often incompatible with one another and comparisons are therefore very difficult.

Three problems especially hamper investigations analyzing the influence of vowel duration on vowel target spectra. First, it is very difficult to elicit

vowel realizations with different durations without altering other factors like context and stress (but note the elegant method used by Lindblom and Moon, 1988), especially if the speech uttered should be close to natural. Second, there seems to be no consensus about how the position of the spectral target in a vowel realization should be determined, different studies use different procedures. For instance, the procedures to determine the point where the target spectrum should be measured of Lindblom (1963), Delattre (1969), Gay (1978), Koopmans-van Beinum (1980), Lisker (1984), Vaissiere (1987), Engstrand (1988), Den Os (1988), and Gopal and Syrdal (1988), all differ largely in definition. Third, there also seems to be a lack of consensus about the representation of the spectral structure of a vowel target. In general, the frequencies of the first two formants are used to characterize a vowel target spectrum. Beside differences in the way these frequencies are measured, there exists a number of ways to represent them (e.g., linear frequencies, logarithmic frequencies, Bark scales) and there is at the moment no reason to prefer one of them over the others for studies testing the target-undershoot model.

To address these problems, an experimental design was selected in which the factors that influence vowel reduction and coarticulation are optimally controlled, and the speech sample was as natural as possible. This was attained by using only a single, experienced, speaker who read a long, meaningful text twice, once at a normal speaking rate and once as fast as possible. A possible drawback of this approach is that stress and vowel context are inherent to that text and are inaccessible to manipulation without losing naturalness. A large collection of vowel realizations was obtained, almost all of which could be used to construct vowel pairs, containing realizations of the same text item at both speaking rates.

The use of only a single speaker could pose a problem if the effects of speaking rate are somehow speaker dependent, as Kuehn and Moll (1976) found. But in this study we are investigating the possibility that a single speaker (this may be any normal speaker) does NOT display any increased articulatory undershoot with an increased speaking rate. If changes in articulatory undershoot are not required in normal speech with an increase in speaking rate, there are profound implications for articulatory theory and research in automatic speech synthesis and recognition.

Vowel target formant values were measured using several procedures in parallel to select the target points in the vowel realizations. This way it is possible to determine whether the detection of durational effects on vowel targets depends on the definition of the targets themselves. The problem of the different representations of the formant frequencies is solved by using statistic tests that are insensitive to the representation of the data. These tests are unlike commonly used statistic tests whose results can be invalidated if, for instance, logarithmic values are substituted for linear values. We therefore will use these distribution-free statistic tests (tests based on rank, see Ferguson, 1981).

In this paper we will investigate whether our speaker produces vowels with more articulatory undershoot (spectral reduction or coarticulation) when he speaks at a fast rate than when he speaks at a normal rate.

2.1 Methods

2.1.1 Speech material

In this study, a long text of about 850 words was used. The text was originally used in a radio broadcast and was informative (concerning economics, see appendix C). The text was read by an experienced, over 60 years old, professional speaker who was selected for his good reading and whose voice was known to give good results with LPC analysis. He speaks the standard form of Dutch (Koopmans-van Beinum, 1980, male speaker #1).

The recorded speech is part of a larger body of speech (in total, 2.5 h of speech recordings) recorded in a 1-day session. The text was read twice. The speaker was instructed to read the text first as he would do for an audience, i.e. at a normal speaking rate. For the second reading, he was instructed to read it as fast as possible. The two readings of this text were done with several hours in between. The speaker was unaware of the specific aims of this project.

The speech was recorded on a commercial Sony PCM recorder, low-pass filtered at 4.5 kHz and digitised at 10 kHz, with 12 bit resolution. Subsequent storage, handling and editing were done in digital form only.

Reading this text took 330 s for the normal speaking rate and 220 s for the fast speaking rate. The overall reduction in duration of the fast-rate realization as compared to the normal-rate realization was one-third when pauses longer than 200 ms were included, and one-fourth when these longer pauses were excluded from both readings.

2.1.2 Segmentation

A waveform editing computer program was used to display the waveform and regenerate the sound of the stored vowels. The waveform and the audio signal were used to identify the boundaries of the vowels (see below). The vowel segments thus identified were copied with a leading and trailing edge of 50 ms of speech to ensure correct spectral analysis at the boundaries of the vowels.

The vowels for this study were selected from the original written text based on their orthographic form. Subsequently, the speech material was searched for realizations of the chosen vowels. Any vowel-like sound that could be attributed to the chosen realization was copied. Only a few vowels were completely absent in the recordings. In some instances, complete words were added to the text. These were used as if they had been in the original text. Both phenomena together resulted in four unpaired vowel realizations. No restriction was imposed on the selection of the vowels except that words and names with a non-Dutch orthography were excluded.

The vowel boundaries were chosen at a zero crossing in the speech waveform. Always, a whole number of pitch periods was used. Any pitch period that could be attributed to the target vowel, and not to the neighbouring phonemes, was considered to be part of that vowel. This included vowel periods that were changed severely by coarticulation. In a plosive-vowel-plosive context this would mean that everything, from the first period following the release burst to (and including) the last discernible period within

the closure, was used (note that Dutch plosives are unaspirated). Some vowels could not be separated from the neighbouring phonemes, especially in vowel-vowel contexts. When this occurred, the whole cluster was used, but the use of these vowel realizations was restricted to formant measuring methods (see below) which are insensitive to segmentation errors.

The read text was labelled for sentence-accent by an experienced phonetician. Labelling for actual phoneme realizations was done by one of the authors. Only standard Dutch phoneme labels were used.

2.1.3 Vowels used

For practical reasons, not all Dutch vowels were used in this experiment. Out of the twelve Dutch monophthongs, only seven were used in this study: the vowels /i y u o A a E/. These vowels were selected on their frequency of use and their representativeness in the vowel space. Five of these are short or half-long vowels (/i y u A E/) and two are long vowels (/o a/). All realizations of these vowels were isolated from the text and used in the analysis. Some realizations differed from their inferred pronunciation and these were labelled according to their actual spoken form. Additionally, some realizations of the schwa, which is a legitimate vowel in Dutch, were selected to serve as a neutral "anchor" in the vowel space. The schwa realizations used came from the words "HET" = /´t/ (English: "THE") and "ER" = /´r/ or /d´r/ (English: "THERE"). In Dutch, these two words are occasionally pronounced with an /E/ instead of with a schwa, but this pronunciation never occurred in the readings of this speaker. In Dutch, the /r/ in "ER" can be an alveolar or a velar consonant (our speaker uses the alveolar variant) and strongly colours vowels towards the /´/ (Pols, 1977). This colouring is expected to change the dynamics of the vowel, but since in this study we only use differences between static features of vowels (i.e. point measurements), this will not pose problems. Some other vowels which were reduced to schwa were included in this group of schwa vowels as well. The schwa in Dutch cannot carry stress in normal (i.e. not contrastive) situations. The various numbers of vowels thus obtained are listed in table 2.1. A grand total of 1178 vowel realizations were isolated existing of 587 pairs of realizations of the same text item at different speaking rates and 4 unpaired realizations. These four unpaired realizations originated from

Table 2.1: Number of vowels occurring in the text that has been analysed in this study. The number of incorrectly segmented vowels is given in parenthesis.

vowel	stressed	unstressed	fast	normal	total
E	59	191 (2)	126	124 (2)	250 (2)
A	58 (2)	181 (6)	116 (2)	123 (6)	239 (8)
a	54	157 (3)	106 (1)	105 (2)	211 (3)
i	52 (1)	132 (7)	92 (1)	92 (7)	184 (8)
o	45	132 (4)	88 (1)	89 (3)	177 (4)
´	0	56 (4)	30 (1)	26 (3)	56 (4)
u	13	19	16	16	32
y	11	4	13	12	25
others	···	···	3	1	4
Total	292 (3)	882 (26)	587 (6)	587 (23)	1178 (29)

vowels inserted by the speaker or deleted from one of the two realizations that were read. Within these 1178 realizations, another four vowels had to be labelled as vowels outside the set studied in this paper. Of the 587 pairs, 17 had different vowel realizations in terms of pronunciation for the two speaking rates and these pairs could not be used in pairwise tests. This leaves us with 570 pairs of realizations that can be used in pair-wise comparisons, as is listed in table 2.2. The 17 vowel pairs with differently labelled phonemes did not show any systematic differences between speaking rates and contained the four vowels labelled outside the set studied in this paper.

2.1.4 Spectral Analysis

A standard software package for speech research was used for LPC analysis (linear predictive code, Vogten, 1986). The vowel segments were analysed with a 10-pole LPC analysis, using a 25 ms Hamming window. The window was shifted in 1 ms steps. This was the basis for formant extraction. The LPC analysis was based on the Split-Levinson algorithm which gives continuous formant tracks (Willems, 1986).

Five different methods were used in parallel to extract five different "target" values from each formant track of each vowel realization. Using the segment boundaries, the value at the mid-point of the realization is read (method Centre), and the (linear) formant frequency average over the complete vowel realization is calculated (method Average). Both these methods were only used on the subset of vowel realizations for which segmentation could be done reliably.

Using a peak (and trough) picking algorithm (a slope segmentator based on Van Son, 1987, see appendix A; see also André-Obrecht, 1988), the point of maximal energy (method Energy) and maximal or minimal value of the appropriate formant (method Formant) were determined to within 3 ms (using a shifting interval one-eighth of the total length of the realization) and the formant frequencies were read at that point. For method Formant, the appropriate formant maximal or minimal value is chosen for each vowel independently, considering its position in the vowel plane. The realizations of the vowels /a A E/ are measured at the point of maximal F_1, the vowels /u o/ at the point of minimal F_2, the vowel /i/ at maximal F_2, and the vowel /y/

Table 2.2: Number of vowel pairs matched on normal versus fast rate. Both realizations in each pair are from the same text item (see text). The number of pairs with incorrectly segmented vowels is given in parenthesis.

vowel	stressed	unstressed	unequal stress	total
E	23	86 (1)	13 (1)	122 (2)
A	25 (2)	82 (3)	8	115 (5)
a	21	72 (2)	11	104 (2)
i	24 (1)	63 (6)	4	91 (7)
o	17	59 (3)	11	87 (3)
´	0	23 (2)	0	23 (2)
u	4	7	5	16
y	5	6	1	12
total	119 (3)	398 (17)	53 (1)	570 (21)

at minimal F_1. With the Formant method, the schwa /´/ was not measured and the values obtained with method Energy are used instead. Peak picking was not perfect, and in about one out of every five formant and energy tracks the "right" peak had to be selected from the suggested alternatives by visual inspection of the tracks. As a fifth method to determine a suitable target point (method Stationary), an automated method for selecting the most stable part of a vowel realization is used (the section with the least variance in the logarithm of the first three formants, Van Bergem, 1988). The last three methods (Energy, Formant, and Stationary) were used on all vowel realizations.

2.2 Results

To determine whether differences in speaking rate introduce differences in vowel formant target values, the properties of vowels realized at normal and at fast rate are compared. It is possible to detect these differences without relying on a specific representation or statistical distribution of the measured values. To decide on statistical significance we used rank-order statistics which is distribution-free. These distribution-free statistical tests are less sensitive and less efficient (Ferguson, 1981) than tests based on a specific distribution (e.g., Normal, Chi-square, or Student's distributions), but they also lack the methodological problems concerning applicability. The range of different stochastic processes for which a distribution-free test can be used is generally much larger than for other statistical tests.

The test results are recalculated to a normal (Gaussian, *z* scores) or Student's (*t* scores) distribution as appropriate, or probabilities are calculated directly (sign-test for small *n*). All tests are derived from Ferguson (1981). Determination of statistical significance is carried out using tables from Abramowitz and Stegun (1965). To obtain a repeated-test result which still has a probability lower than 5% (single test level, indicated by "+") of one or more spurious results that reach the level of significance, a threshold level of 0.1% (10^{-3}, two-tailed, indicated by "++") was used to determine statistical significance in individual tests. In this way, it still is possible to identify the samples that deviate from the H_0 hypothesis out of a large set (up to 50 samples) with an error probability of less than 5%.

2.2.1 Median values

A general way to compare two sets of values is to test for differences in their median values. The standard target-undershoot model predicts a smaller distance between the median formant values of a specific vowel and the schwa for fast-rate speech than for normal-rate speech. This implies lower median formant values for both F_1 of vowels /E a A/ and F_2 of vowels /i E/ for fast-rate speech than for normal-rate speech and higher median formant values for both F_1 of vowels /i y u/ and F_2 of vowels /u o A/. The other values should be more or less the same under both speaking conditions. An analysis of the data per vowel was made, the results of which are shown in table 2.3. In this table, median formant values and a Mann-Whitney U test

were used to test for differences between the distributions of all normal-rate and all fast-rate realizations of one specific vowel in the set.

First, there is a global shortening of vowel duration detectable in fast-rate speech as compared to normal-rate speech, when all vowels are pooled (total row in table 2.3). However, only long vowels, /a/ and /o/, prove to be shorter in fast-rate speech (0.1% level, ++), the other vowels are ambiguous

Table 2.3: Median values for formant frequencies (Hz) and duration (ms).
Statistical significance is determined with a Mann-Whitney U test. Statistical significance is indicated by "++" (at the 0.1% level); a 5% error level for a result is indicated by "+"; other statistically insignificant results are indicated by "ns". Abbreviations of method names: Form.-Formant, Stat.- Stationary, Ener.-Energy, Cent.-Centre, Aver.- Average. In all columns: normal-rate value left (n), fast-rate value right (f). The total mean values and standard deviation of the duration are: normal rate 99 ± 41 ms, fast rate 84 ± 31 ms (correctly segmented vowels only).

Vowel	Form.	Stat.		Ener.		Cent.		Aver.		Duration			
	Rate	n	f	n	f	n	f	n	f	n	f	n	f
E	F_1	554	574	545	565	524	548	544	557	493	520	81	74
		+		+		++		+		++		+	
	F_2	1527	1514	1527	1526	1523	1521	1521	1527	1503	1501		
		ns		ns		ns		ns		ns			
A	F_1	597	618	587	608	581	600	589	609	539	564	81	76
		+		+		+		+		++		+	
	F_2	1151	1153	1112	1133	1128	1133	1119	1131	1133	1129		
		ns		ns		ns		ns		ns			
a	F_1	639	655	631	649	623	637	630	645	579	609	131	97
		+		+		+		+		++		++	
	F_2	1331	1330	1313	1330	1324	1334	1329	1329	1335	1321		
		ns		ns		ns		ns		ns			
i	F_1	312	325	316	332	327	341	313	335	316	333	80	72
		+		+		+		+		+		+	
	F_2	2130	2105	2081	2074	2002	2010	2072	2036	1946	1925		
		ns		ns		ns		ns		ns			
o	F_1	391	413	419	432	412	435	417	439	411	434	121	109
		+		+		+		+		++		++	
	F_2	854	897	930	964	943	972	925	959	995	1029		
		ns		ns		ns		+		ns			
´	F_1	407	440	411	438	407	440	414	434	393	422	52	56
		ns		ns		ns		+		ns		ns	
	F_2	1440	1455	1434	1454	1440	1455	1435	1464	1433	1444		
		ns		ns		ns		ns		ns			
u	F_1	369	368	370	375	376	390	372	373	362	368	83	74
		ns		ns		ns		ns		ns		ns	
	F_2	782	776	800	805	836	821	880	851	947	1012		
		ns		ns		ns		ns		ns			
y	F_1	297	332	313	336	329	364	317	334	316	350	77	76
		ns		ns		+		ns		ns		ns	
	F_2	1452	1416	1576	1442	1624	1566	1590	1476	1582	1504		
		ns		ns		ns		+		+			
Total	F_1	526	553	526	553	498	528	520	535	476	501	89	78
		+		+		+		+		+		++	
	F_2	1339	1351	1341	1347	1343	1361	1334	1357	1345	1360		
		ns		ns		ns		ns		ns			

Figure 2.2. Median values of first and second formant measured with the "Average" method for pairs of realizations of Dutch vowels. Open squares: normal speaking rate values. Filled squares: fast speaking rate values; crosses: /a/, *triangles:* /i/.

in this respect (at most at the 5% level, +). The averaged shortening of vowel duration due to speaking rate is smaller than the overall shortening of the spoken text (only 15% in vowels versus 25% in the total text, see also section 2.1.1, and 2.3.1 below), but the differences are systematic and present in all but one vowel, the schwa.

The number of vowels, for which significant differences ($p \bullet 0.1\%$, ++) between median formant values at different speaking rates are found, is small. Especially for methods for which inter-vowel spectral distances are large (Formant, Stationary, and Centre) none of the vowels shows a significant difference between speaking rates. The number of (not significant) test results with a low probability ($p \bullet 5\%$, +) is sufficiently high to suggest that there is indeed some difference between speaking rates. The probability to obtain at least 5 out of 8 test results at the 5% level is less than 0.1% (++). For only one method, Average, it is possible to identify the vowels which change with some confidence (at the 0.1% level, ++). Using this measuring method, the vowels /E A a o/ show a statistically significant higher first formant value in fast-rate speech as compared to normal-rate speech (see figure 2.1 and table 2.3). No statistically significant differences between second formant frequencies are found (table 2.3).

Comparing columns in table 2.3, the differences between the different measuring methods are small and seem to be limited to a small reduction in overall size of the vowel triangle going from method Formant to method Average. Although the differences between speaking rates are not always statistically significant, the median values all show the same response to an increase in speaking rate. The differences found here between formant values from vowels spoken at different rates are inconclusive in that for only one method, Average, is it possible to identify statistically significant changes in vowel formant values. Apparently, this kind of statistical analysis is not sensitive enough to show the differences between fast- and normal-rate vowels from unrestricted text reliably. Whether or not a test will show a difference between speaking rates depends on the measuring method used.

2.2.2 Consistency

The consistency with which our speaker reproduces the text in each reading and the ability of our measuring methods to capture the within-speaking-rate variation over different readings must be estimated, before comparisons between the members of vowel realization pairs in both readings can be made. This estimation can be performed by checking the similarity between the measurements in the two readings. The similarity of within-speaking-rate rank order of measurements between different speaking rates is an indicator of the desired consistency. It was measured with a Spearman rank correlation test, the results of which are shown in table 2.4. To illustrate graphically the similarity of rank order, a choice has

been made from the data presented in table 2.4. In figure 2.2, the F_2 frequencies of individual vowel pairs spoken at normal and fast rate, measured with the Average method, are plotted against each other for just three vowels: /o a i/. It can be seen that the formant value pairs are ordered along the diagonal of the plot for /o a/ displaying a fairly monotonic relation between normal-rate and fast-rate F_2 values, and thus a high Spearman rank correlation coefficient. The F_2 values of the /i/ are scattered over a large area, indicating that only a minimal relation exists between normal-rate and fast-rate values of the F_2 for this vowel, and thus only a very small correlation coefficient. As a consequence, the F_2 values measured of /o a/ are consistent over speaking rates whereas the F_2 values of /i/ are not.

In table 2.4, the Spearman rank correlation coefficients of formant values and duration are presented for all methods and vowels used. Except for F_1 of /u/ and F_2 of /i/, all correlation coefficients are above 0.5 for at least some of the methods used. Except for /´/, all durational correlation coefficients are above 0.5. For most vowels, the F_2 formant values correlate with coefficients around 0.7 or well above. These correlation coefficients are comparable in size to those found by Kuehn and Moll (1976) when they compared articulatory velocities from vowel-consonant transitions spoken at different rates. The correlation coefficients show peculiar differences between vowels that are not easily explained without a detailed analysis of the distribution of context features over the different vowels, an analysis that is outside the scope of this paper. The very low correlation of F_2 from /i/ can probably be attributed to problems with the LPC formant analysis of this vowel formant. The F_2 and F_3 values of the /i/ might be too close for the

Table 2.4: Coefficients of a Spearman Rank Correlation test on formant frequency values and durations between the realizations within pairs (normal-rate versus fast-rate) of vowels. For indication of statistical significance and abbreviations see table 2.3.

Vowel	Form.	Stat.	Ener.	Cent.	Aver.	Duration	
E	F_1	0.65 ++	0.56 ++	0.60 ++	0.61 ++	0.61 ++	0.68++
	F_2	0.70 ++	0.64 ++	0.58 ++	0.67 ++	0.71 ++	
A	F_1	0.81 ++	0.74 ++	0.75 ++	0.79 ++	0.79 ++	0.65++
	F_2	0.85 ++	0.86 ++	0.87 ++	0.88 ++	0.89 ++	
a	F_1	0.61 ++	0.57 ++	0.55 ++	0.59 ++	0.65 ++	0.77++
	F_2	0.72 ++	0.73 ++	0.75 ++	0.76 ++	0.84 ++	
i	F_1	0.58 ++	0.53 ++	0.44 ++	0.53 ++	0.58 ++	0.66++
	F_2	0.16 ns	0.13 ns	0.24 +	0.10 ns	0.24 +	
o	F_1	0.78 ++	0.73 ++	0.80 ++	0.86 ++	0.87 ++	0.81++
	F_2	0.79 ++	0.63 ++	0.70 ++	0.69 ++	0.86 ++	
´	F_1	0.70 ++	0.62 +	0.70 ++	0.52 +	0.44 +	-0.06ns
	F_2	0.92 ++	0.89 ++	0.92 ++	0.83 ++	0.91 ++	
u	F_1	0.39 ns	0.31 ns	0.12 ns	0.16 ns	0.27 ns	0.57 +
	F_2	0.63 +	0.62 +	0.53 ns	0.73 +	0.59 ns	
y	F_1	0.01 ns	0.45 ns	0.73 +	0.70 +	0.76 +	0.60 +
	F_2	0.32 ns	0.58 ns	0.69 +	0.54 ns	0.65 +	
Total	F_1	0.94 ++	0.93 ++	0.93 ++	0.94 ++	0.94 ++	0.77++
	F_2	0.96 ++	0.92 ++	0.94 ++	0.93 ++	0.96 ++	

analysis method to resolve the differences between these two formants, resulting in aberrant F_2 values. The total absence of a correlation for the duration of /´/ is to be expected because all pairs of this vowel were taken from only two different, unstressed, high frequency words (/´t/ and /(d)´r/), giving only a very small variation in context.

As before (section 2.2.1), all measuring methods seem to capture the same kind of features with only a difference in sensitivity, and no method behaves at variance with the others. The strong correlations found between values measured for vowels uttered at different speaking rates indicates that whatever systematic differences exist between these vowel realizations, it is conserved by the measurements. This means that a pairwise comparison should indeed be able to discover systematic differences in formant values between speaking rates.

2.2.3 Pairwise changes in formant frequencies and duration

The measured formant and duration values of the vowel pairs were divided into two sets. One set contained all value pairs for which the fast-rate value was higher than the normal-rate value. The other set contained all value pairs for which the fast-rate value was lower than the normal-rate value. Pairs in which both values are equal were omitted. This was done for each of the parameters, F_1, F_2 and duration, and for each method. In table 2.5, the fractions of pairs with a higher fast-rate formant frequency or a lower fast-rate duration are presented as percentages of total. Statistical signifi-

Table 2.5: Percentage of pairs for which the fast-rate realization has a higher formant value than its normal-rate counterpart. Last column (Duration): Percentage of pairs for which the fast-rate realization is shorter than its normal-rate counterpart. Significance is given for a Sign test, ties (fast-rate value = normal-rate value) are omitted. For indication of statistical significance and abbreviations see table 2.3.

Vowel	Form.	Stat.	Ener.	Cent.	Aver.	Duration	
E	F_1	70 ++	73 ++	71 ++	71 ++	80 ++	74++
	F_2	47 ns	48 ns	47 ns	49 ns	44 ns	
A	F_1	70 ++	70 ++	72 ++	70 ++	76 ++	71++
	F_2	64 +	60 +	62 +	64 +	63 +	
a	F_1	62 +	66 +	63 +	68 +	77 ++	92++
	F_2	58 ns	52 ns	51 ns	51 ns	46 ns	
i	F_1	62 +	67 +	64 +	73 ++	71 ++	72++
	F_2	48 ns	55 ns	47 ns	38 +	42 ns	
o	F_1	81 ++	74 ++	81 ++	84 ++	88 ++	81++
	F_2	68 +	64 +	68 +	68 +	71 ++	
´	F_1	61 ns	70 ns	61 ns	76 +	76 +	43ns
	F_2	61 ns	52 ns	61 ns	62 ns	67 ns	
u	F_1	73 ns	55 ns	64 ns	45 ns	55 ns	46ns
	F_2	73 ns	70 ns	64 ns	64 ns	73 ns	
y	F_1	82 ns	73 ns	100 +	82 ns	91 +	73ns
	F_2	36 ns	9 +	17 +	18 ns	9 +	
Total	F_1	69 ++	70 ++	70 ++	72 ++	78 ++	75++
	F_2	57 +	55 +	54 ns	53 ns	53 ns	

cance was determined with a sign-test. Based on the duration figures, most vowels can be said to be shorter when spoken at a fast rate (75%), thus confirming the overall shortening of the vowels in fast-rate speech (section 2.2.1).

With only one exception (i.e., /u/ analysed using the Centre method) the majority (> 50%) of pairs of all vowels with all measuring methods show a fast-rate F_1 value which is higher than the normal-rate formant value. This higher fast-rate F_1 value is found, independent of the identity of the vowel. This means that the first-formant values generally rise with speaking rate, which conforms with the results of the tests using median values (section 2.2.1). This time, however, the differences found are statistically significant (level 0.1%, ++) with all methods used for /E A o/ and vowels pooled (total), and not just for method Average, as was the case when analysing median values (section 2.2.1, see table 2.3). Method Average gives statistical significant differences (level 0.1%, ++) for 5 out of the 8 vowels used (/E A a i o/).

When it comes to vowel formant differences between speaking rates, no clear picture emerges for the second formant. No statistical significant changes can be found except for F_2 of /o/ with the Average method. This averaging method seems to be the most sensitive method for analysis of differences between formant values of vowel realizations, both for F_1 and F_2.

2.2.4 Correlation between formant frequency and duration

The target-undershoot model presupposes a relation between spectral vowel reduction and vowel duration. If vowel formant values move to the schwa value (i.e. show spectral reduction) with shorter vowel durations, there should be a (strong) correlation between vowel duration and vowel formant values. The strength of this correlation, in relation to the correlation between different speaking rates (section 2.2.2), is an indication of the importance of vowel duration in determining the vowel formant value, relative to the other important factors (e.g., stress, context).

The rank correlation between vowel formant values and duration shows very small, but often statistically significant ($p \bullet 0.1\%$, ++), correlation coefficients (table 2.6) which implies that only a very small part of the variation in formant values between vowel realizations can be explained by the differences in duration. This was found for realizations of both speaking rates pooled (table 2.6.a) and for the fast rate realizations (table 2.6.b) and normal rate realization individually (data not shown, they are comparable to those of table 2.6.b). The correlations seem to be stronger when realizations from both speaking rates are used independently instead of pooled together (compare table 2.6.b with table 2.6.a). Of all correlations, only the coefficients of the F_1 values of the vowels /E A a/ are statistically significant.

In contrast, the correlation between formant values of realizations that differ in speaking rate only (table 2.4) is high and statistical significant for both formants and almost all vowels and can thus explain a great part of the variation in formant values. Based on these correlations, it must be concluded that vowel duration has only a marginal power in explaining the vowel formant targets. This small explanatory power holds just as much be-

tween as within speaking rates. The correlation coefficients are so extremely small compared with the pairwise correlations (table 2.4) that it is even possible for these correlations to be the result of a residual correlation stemming from the correlation between both formant target frequency and duration and the stress and context of the vowel.

2.2.5 Influence of phoneme context

Analysis of how the influences of speaking rate depend upon the phonetic context in which the vowels occur (coarticulation) is hampered by the large number of different contextual phonemes per vowel which is inherent to unrestricted (near-natural) text. Consequently, there are so few realizations of any specific vowel-context combination, that a statistical analysis is almost impossible with the amount of text and the statistical methods used in this paper.

As a first attempt, vowels and consonants were pooled on articulatory features. Of all the consonants, the alveolar consonants were most common. In Dutch, the alveolar consonants encompass /n t d s z r l/. Alveolar consonants are articulated very close to the /i/, they can be described as high, closed and fronted phonemes. The vowels were divided into several overlapping sets. A set of closed vowels, /i y u/, versus a set of open vowels, /a A E/, and a set of fronted vowels, /i E/, versus a set of back vowels, /o u/. The vowel realizations in alveolar context were pooled on these groups and the pairwise differences between speaking rates were tested (like in section 2.2.3). Three arrangements are possible: CV*, *VC, and CVC, in which the C is an alveolar consonant and * can be any context. It showed that in, all three arrangements, the same pattern emerged. Because the trailing consonant has the greatest importance in determining stationary vowel spectra (Pols, 1977), and the vowel realizations in this context were most numerous, we only show the *VC results (table 2.7).

It appears that all vowels, grouped on different features, behave identical. The trend of higher F_1 values in fast-rate speech, already found for the individual vowels, without regarding context, emerges again. Also, the lack of significant differences between F_2 values measured at different speaking rates is found again. Despite the fact that open vowels are "distant" in an articulatory sense from the (closed) alveolars, these vowels do not behave different from the more "nearby" closed vowels. The same is found for the distant back vowels and the nearby front vowels. The higher F_1 value in fast-rate speech implies, in these articulatory terms, a more open articulation where a more closed articulation (i.e. lower F_1 values) is expected if a higher speaking rate should result in more coarticulation.

2.2.6 Influence of stress

Thus far, vowels were considered to be comparable when different speaking rates were used. However, the effects of speaking rate could very well be different for stressed and unstressed vowels. This was investigated by comparing the changes between pairs of vowels for the two speaking rates just as in table 2.5, but now for stressed and unstressed vowels separately. Because of the small number of stressed vowel pairs, all vowels were pooled and only these total figures per formant value were used (table 2.8). These total scores indicate a small difference in percentage of pairs changing in one direction for stressed and unstressed vowels. The differences between speaking rates are somewhat less pronounced for the formant values of stressed vowels than for unstressed vowels. The reverse is true for differ-

Table 2.6.a: Similar to table 2.4 but this time the coefficients indicate the Spearman Rank Correlation coefficients between formant values and duration for each vowel realization. Only correctly segmented vowels are used. Normal-rate and fast-rate realizations pooled.

Vowel		Form.	Stat.	Ener.	Cent.	Aver.
E	F_1	0.28 ++	0.19 +	0.21 ++	0.30 ++	0.08 ns
	F_2	0.04 ns	0.03 ns	0.07 ns	0.02 ns	-0.03 ns
A	F_1	0.49 ++	0.42 ++	0.39 ++	0.45 ++	0.31 ++
	F_2	-0.18 +	-0.27 ++	-0.23 ++	-0.26 ++	-0.23 ++
a	F_1	0.41 ++	0.37 ++	0.34 ++	0.33 ++	0.15 +
	F_2	-0.02 ns	-0.02 ns	-0.05 ns	0.02 ns	0.03 ns
i	F_1	0.03 ns	-0.04 ns	0.04 ns	-0.05 ns	-0.02 ns
	F_2	0.33 ++	0.22 +	0.07 ns	0.23 +	0.03 ns
o	F_1	-0.02 ns	0.11 ns	-0.01 ns	0.12 ns	0.12 ns
	F_2	-0.27 ++	-0.12 ns	-0.13 ns	-0.16 +	-0.13 ns
´	F_1	0.17 ns	0.18 ns	0.17 ns	0.10 ns	0.00 ns
	F_2	0.40 +	0.34 +	0.40 +	0.36 +	0.32 +
u	F_1	0.02 ns	0.06 ns	-0.15 ns	-0.09 ns	-0.06 ns
	F_2	0.01 ns	0.17 ns	0.08 ns	-0.05 ns	0.11 ns
y	F_1	0.41 +	0.35 ns	0.26 ns	0.38 ns	0.37 ns
	F_2	-0.11 ns	0.23 ns	0.35 ns	0.24 ns	0.09 ns
Total	F_1	0.23 ++	0.22 ++	0.21 ++	0.23 ++	0.19 ++
	F_2	-0.26 ++	-0.27 ++	-0.27 ++	-0.27 ++	-0.27 ++

ences in duration. This time it matters indeed which method is used to determine the formant frequency. For stressed vowels, methods that are sensitive for the exact shape of the formant track with respect to the vowel boundaries (i.e. Energy, Centre, and Average) indicate more change than do methods that try to catch shape-invariant points of the formants (Formant and Stationary). It is not possible to substantiate this any further with the rather limited set of data used here.

2.3 Discussion

The median formant values found in this study (table 2.3) for normal-rate speech are generally lower than those found by Koopmans-van Beinum (1980, male speaker #1) with speech of the same speaker for stressed and unstressed vowels in read text. Apart from methodological differences in vowel selection and labelling, these differences can be attributed to the differences in spectral analysis (LPC versus spectrographic).

2.3.1 Differences between speaking rates: Duration

Although most fast-rate vowel realizations are shorter than their normal rate counterparts, the differences between these vowel durations are quite small. The global decrease in total duration is about 25%, but the decrease in duration of the vowels studied is less than 15% if the fast-rate reading of the text is compared to the normal-rate reading. The exception is the vowel /a/, which seems to shorten by approximately 25% (*median* values from table 2.4, see also section 3.2.1).

Table 2.6.b: As table 2.6.a. Vowel realizations from fast-rate reading only.

Vowel		Form.	Stat.	Ener.	Cent.	Aver.
E	F_1	0.34 ++	0.27 +	0.32 ++	0.38 ++	0.18 +
	F_2	0.02 ns	-0.02 ns	0.04 ns	0.01 ns	-0.03 ns
A	F_1	0.49 ++	0.38 ++	0.31 ++	0.42 ++	0.27 +
	F_2	-0.16 +	-0.27 +	-0.22 +	-0.26 +	-0.22 +
a	F_1	0.57 ++	0.55 ++	0.50 ++	0.53 ++	0.36 ++
	F_2	-0.04 ns	-0.05 ns	-0.06 ns	-0.04 ns	-0.06 ns
i	F_1	0.00 ns	-0.05 ns	-0.03 ns	-0.10 ns	-0.10 ns
	F_2	0.23 +	0.18 ns	0.04 ns	0.16 ns	-0.04 ns
o	F_1	-0.05 ns	0.19 ns	0.08 ns	0.25 +	0.22 +
	F_2	-0.34 +	-0.14 ns	-0.14 ns	-0.18 ns	-0.19 ns
´	F_1	0.18 ns	0.06 ns	0.18 ns	0.06 ns	-0.10 ns
	F_2	0.38 +	0.37 +	0.38 +	0.37 +	0.40 +
u	F_1	-0.05 ns	0.05 ns	0.14 ns	0.07 ns	0.05 ns
	F_2	0.07 ns	0.00 ns	0.20 ns	0.00 ns	0.19 ns
y	F_1	0.57 +	0.15 ns	0.34 ns	0.36 ns	0.45 ns
	F_2	0.44 ns	0.58 +	0.42 ns	0.51 ns	0.38 ns
Total	F_1	0.23 ++	0.23 ++	0.22 ++	0.24 ++	0.20 ++
	F_2	-0.27 ++	-0.27 ++	-0.28 ++	-0.28 ++	-0.28 ++

Different explanations are possible. At one hand, we may have overestimated the global decrease in duration by including too much of the silent parts (pauses shorter than 200 ms, section 2.1.1). These silent parts could be the elements that absorb the shortening. At the other hand, our segmentation may have been biased toward longer fast-rate vowels by including more pitch periods in fast-rate vowel realizations than in normal-rate realizations. This kind of bias is difficult to detect if the context from which the vowel realizations are obtained is as diverse as in this study.

Apart from these methodological problems, another reason for the small difference in vowel duration between speaking rates may be the fact that the normal rate vowel realizations themselves already are quite short. A normal, and pleasant, speaking rate for reading a long text will be faster than the speaking rate used for isolated sentences in a citation style of speaking. The attainable durational differences between speaking rates for vowel realizations in studies using that kind of speech may be higher than what is found in the present study.

Whatever the explanation of the rather small size of the differences in vowel duration between speaking rates, these differences are highly systematic. Therefore, the fast-rate vowel realizations should nevertheless show the differences in target values associated with speaking rate differences, but actually did not.

2.3.2 Differences between speaking rates: Formant frequencies

Considering the material and methods used here, it is not possible to uncover the cause of the higher F_1 values found in all vowels with a higher speaking rate. An explanation for this higher formant value might be that, given the fact that F_1 is related to the openness of vowels, our experienced speaker lowers his jaw somewhat more in fast-rate speech than in normal-rate speech. This could be the result of overcompensation or overshoot when the speaker accommodates for the high speaking rate. An alternative explanation might be that our speaker reads the fast-rate realization with a louder voice than the normal-rate realization. It is known that differences in speech effort can change the articulation (Schulman, 1989) and the formant values of vowels (Traunmüller, 1988). A louder voice might also be partly responsible for the relatively long vowel durations in fast rate speech

*Table 2.7: Similar to table 2.5 but this time only vowels uttered in *VC context are used, for which the C is an alveolar consonant (one of* /n t d s z l r/*) and * can be any context. The vowel pairs are pooled on the features [+Closed] (*/i y u/*, n=60), [+Open] (*/E A a/*, n=255), [+Front] (*/i E/*, n=141), and [+Back] (*/u o/*, n=46).*

Vowel	Form.	Stat.	Ener.	Cent.	Aver.	Dur.	
Closed	F_1	53 ns	62 ns	64 +	63 ns	67 +	73++
	F_2	45 ns	51 ns	38 ns	38 ns	43 ns	
Open	F_1	66 ++	69 ++	68 ++	70 ++	79 ++	79++
	F_2	54 ns	53 ns	52 ns	52 ns	52 ns	
Front	F_1	62 +	70 ++	67 ++	68 ++	75 ++	75++
	F_2	45 ns	48 ns	44 ns	43 ns	45 ns	
Back	F_1	80 ++	73 +	82 ++	83 ++	87 ++	85++
	F_2	64 ns	61 ns	67 ns	67 +	70 +	

(Schulman, 1989; c.f., section 2.3.1). Because we did not calibrate our recordings for loudness, we are not able to check this. The difference between the F_1 values at different speaking rates is, however, very small and its perceptual relevance is questionable.

These results show that a different style of speaking, fast-rate versus normal-rate reading of a text, can change the duration of the vowels without changing the vowel formant values or can change the vowel formant target values in unexpected ways. Even when using vowels in identical context, a simple correlation between vowel formant target values and vowel duration cannot be extended over different speaking styles. Indications for speaking-style specific correlations between F_1 and duration were also found by Lindblom and Moon (1988) when they compared clear and citation form speech. Also the explanatory power of duration when predicting vowel target values must be judged marginal if compared to other (contextual) factors.

It is known that articulatory adaptation to a fast speaking rate can be speaker dependent (Kuehn and Moll, 1976) and it is to be expected that the ability to read aloud at a fast rate, and still pronounce correctly, depends on experience and training. The speaker used in this experiment has had a very long career as a professional speaker and newscaster, so his capabilities are not likely to be shared by naive, untrained, subjects. The results are nevertheless important for general theories on articulation and the design of systems for automatic speech recognition and synthesis. The experience of the speakers used should also be considered seriously when designing an experiment regarding the effects of speaking rate on speech sounds.

2.3.3 Differences between measuring methods

In this paper different methods to measure vowel formant values in a given formant track were used. Averaging the formant values over the complete vowel is the method most sensitive to speaking rate changes; at the same time this method also produces formant frequencies that deviate most from the values reported in literature (e.g., Pols, 1977; Koopmans-van Beinum, 1980). However, the differences between the various methods used are in most respects marginal and all methods used essentially give the same outcome. When studying vowel targets, the method that is most convenient can be used.

Probably all points in a vowel segment change in concert when speaking rate changes, so it may not be crucially important which cross-section in

Table 2.8: Similar to table 2.5 but this time with all vowel pairs pooled on stress, first row: unstressed; second row: stressed; last row: stressed and unstressed combined. Only pairs with equal stress realization on both readings are used.

VowelForm.		Stat.	Ener.	Cent.	Aver.	Dur.	
no stress	F_1	72 ++	72 ++	71 ++	74 ++	78 ++	73++
	F_2	58 +	56 +	55 ns	54 ns	54 ns	
stress	F_1	57 ns	63 +	69 ++	66 ++	76 ++	84++
	F_2	53 ns	52 ns	50 ns	50 ns	51 ns	
Total	F_1	69 ++	70 ++	70 ++	72 ++	78 ++	75++
	F_2	57 +	55 +	54 ns	53 ns	53 ns	

the realization is actually used to measure the difference. Such a model of vowel dynamics can only be checked with a detailed analysis of the total dynamic shape of vowel formant tracks, not by using point measurements as was done here. This dynamic description of formant tracks is the subject of the next two chapters (see also, Van Son and Pols, 1989, 1991a, 1992).

2.4 Conclusions

With the restriction that speech of only one speaker was used and that the speech was constrained to two readings of one text, our analysis reveals that neither excess vowel reduction (in terms of vowel targets) nor excess coarticulation accompanies a higher speaking rate. The only change in vowel formant frequency that could be detected was a higher value of the first formant frequency in fast-rate speech as compared to normal-rate speech, irrespective of the vowel identity. This shift in formant frequency may be linked to a more open articulation of the vowels or an increase in loudness of the speech. No difference due to stress or consonantal context was found that could explain this behaviour, neither was there an effect of the method with which the target points within the vowel realizations were determined.