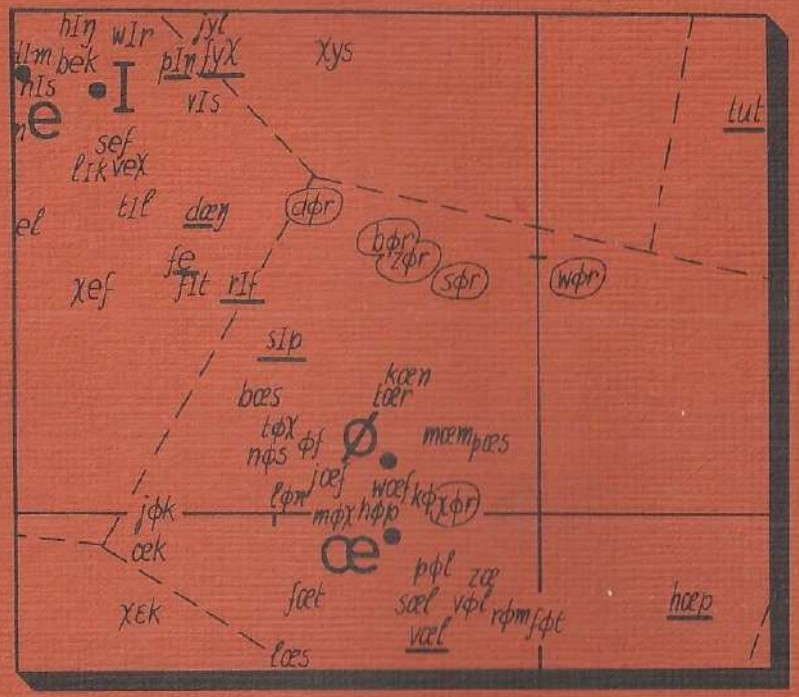
LOUIS C.W. POLS SPECTRAL ANALYSIS AND IDENTIFICATION OF DUTCH VOWELS IN MONOSYLLABIC WORDS

SPECTRAL ANALYSIS AND IDENTIFICATION OF DUTCH VOWELS IN MONOSYLLABIC WORDS

LOUIS C.W. POLS



SPECTRAL ANALYSIS AND IDENTIFICATION OF DUTCH VOWELS IN MONOSYLLABIC WORDS

VRIJE UNIVERSITEIT TE AMSTERDAM

SPECTRAL ANALYSIS AND IDENTIFICATION OF DUTCH VOWELS IN MONOSYLLABIC WORDS

Academisch Proefschrift

ter verkrijging van de graad van
doctor in de wiskunde en natuurwetenschappen
aan de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus mr. J. de Ruiter,
hoogleraar in de faculteit der rechtsgeleerdheid,
in het openbaar te verdedigen
op donderdag 16 juni 1977 te 15.30 uur
in het hoofdgebouw der universiteit,
De Boelelaan 1105

door

LOUIS CORNELIS WILHELMUS POLS

geboren te Tiel

Academische Pers B.V. – Amsterdam
1977

Promotor: Prof. dr. ir. R. Plomp
Coreferent: Dr. S.G. Nooteboom

Dit proefschrift is de voorlopige afronding van een, reeds geruime tijd lopend, stuk spraakonderzoek waaraan velen hun bijdrage hebben geleverd. Mijn promotor Reinier Plomp heeft niet alleen de kiem gelegd voor de hier gepresenteerde benaderingswijze, maar heeft bovenal op nimmer aflatende wijze dit werk begeleid, ondersteund en helpen uitbouwen. Ik ben hem voor zijn vele hulp, ook nu weer bij de feitelijke totstandkoming van dit proefschrift, veel dank verschuldigd. De indringende wijze waarop mijn coreferent Sieb Nooteboom zich in korte tijd heeft ingewerkt in de probleemstelling van dit onderzoek, heeft duidelijk zijn weerslag gevonden in de uiteindelijke vorm van deze studie.

Ik prijs me zeer gelukkig met de werkomgeving waarin ik nu al meer dan tien jaar mag vertoeven. Dit betreft de Rijksverdedigingsorganisatie TNO, in het bijzonder het Instituut voor Zintuigfysiologie, door wie ik in de gelegenheid ben gesteld dit onderzoek te doen en bovendien voldoende tijd heb gekregen voor de afronding in de vorm van een proefschrift.

Stukken vooronderzoek zijn enthousiast uitgevoerd door de diverse gedetacheerden die met mij sinds 1968 uiterst plezierig hebben samengewerkt: Wout Klein, Herman Tromp, Dick van Nierop, Johan van Rooijen en Eef van Heusden. Voor diskussie, bijstand en goede raad kon ik altijd aankloppen bij Guido Smoorenburg. Dit geldt in nog sterkere mate voor mijn kamergenoot sinds vele jaren: Tammo Houtgast die mij steeds weer imponeert door zijn grote veelzijdigheid en scherp inzicht. Herman Steeneken en Ton Mimpen wisten altijd goede raad op mijn vele technische problemen, zoals ik ook veelvuldig van de grote deskundigheid van de medewerkers van de elektronische afdeling heb kunnen profiteren. Leo Spiekman heeft veel bijgedragen tot het "komputeriseren" van de meetmethoden.

De vele diskussies met Bert Schouten hebben mij meer inzicht verschaft in de problematiek van klinkerkoartikulatie. Bovendien ben ik hem zeer dankbaar voor de nauwgezette wijze waarop hij steeds weer bereid was mijn geschreven Engels te korrigeren. De snelle en akkurate eindkontrole van het manuskript door de heer J.B.A. Nijssen van de Vertaaldienst BIP-TNO heb ik erg gewaardeerd. Resterende fouten zijn echter uitsluitend mijzelf te verwijten. Bij de uiterlijke vormgeving van dit proefschrift zijn de heer Huigen, alsook Koos Wolff en Henny Gebbink mij zeer ten dienste geweest. Zij, en alle andere stille werkers en niet genoemde kollega's die op welke wijze dan ook aan de totstandkoming van dit proefschrift hebben bijgedragen, dank ik van harte.

Ik hoop dat dit proefschrift mijn moeder veel voldoening zal geven, zoals het dat zeker mijn te vroeg overleden vader zou hebben gedaan. Tenslotte hoop ik dat na 16 juni het, door mij zo gewaardeerde, gezinsleven meer dan ooit zal opbloeien na een periode waarin Anneliese, Edith en Mirjam onvoldoende aandacht kregen, maar mij wel steeds tot grote steun waren.

CONTENTS

The research reported in this study was carried out in the Institute for Perception TNO, Soesterberg, the Netherlands.

CHAPTER 1

INTRODUCTION, AND REVIEW OF THE LITERATURE

1.1. A FIRST SPECIFICATION OF THE SUBJECT MATTER

In normal speech communication listeners can, in general, easily understand meaningful spoken messages in their own language. Native listeners are able to do this because of (1) their knowledge of the relationship between the acoustic characteristics of the speech signal and the speech elements to be recognized such as phonemes, morphemes, and words; (2) their knowledge of the contextual, grammatical and semantic constraints of the language; (3) their knowledge of the world, and (4) the strategies they have developed in applying these different types of (mainly implicit) knowledge to the decoding of speech signals.

The present study concentrates on the relationship between the acoustic characteristics of speech sounds and one type of speech elements to-be-recognized: the phonemes. One might object that possibly in normal speech perception the recognized units are features, or whole morphemes, or words, rather than phonemes. Particularly phonemes in unstressed and sloppely spoken syllables are often hard, or even impossible, to recognize on the basis of the acoustic speech signal alone. There seems little doubt, however, that listeners are able to recognize phonemes, particularly vowel phonemes, from well pronounced, stressed syllables and this may contribute considerably to their ability to perceive speech. Most theories of speech perception include such a level of phoneme perception, and most automatic speech recognition and understanding systems include phoneme classification in their set of strategies. It thus seems worth while to specify the difficulties that a theory of phoneme perception will encounter owing to the non-unique relation between acoustic signals and phonemes.

Not much research is needed to show that enormous differences can exist between various acoustic realizations of the same phoneme. These differences are,

for instance, visible in spectrograms. This implies that often phonemes cannot be differentiated on the basis of their acoustic realizations alone, which is particularly clear for unstressed vowels in conversational speech. The lack of acoustic constancy of phoneme realizations means a challenge to those who study speech production and perception, and to those who work on speech synthesis and automatic speech recognition.

Although there is a great deal of variation in the acoustic realizations of speech sounds, this does not mean that it is completely unsystematic. A systematic description of acoustic variation in the realizations of phonemes, and a study of the way in which perception is affected by this variation can be of help in obtaining a better understanding of human speech perception and can help to solve the problem of automatic speech recognition.

The aim of this study is to obtain such a systematic description of acoustic variation in the realizations of phonemes, and the perceptual consequences this variation has. In such a project it is necessary to specify certain restrictions in order to limit the number of experimental variables.

These limitations are mentioned in the title of this monograph. They are, with respect to the speech signal: Dutch vowels in monosyllabic words; with respect to the approach: study of the acoustic signal and of speech perception, but not of speech production; with respect to the method: *spectral* analysis and *identification* experiments. A further specification of these points will be given below.

(a) First of all we have limited ourselves to the vowel phonemes of Dutch. This is a group of 15 speech sounds, 12 monophthongs and three diphthongs. We had studied earlier the average spectral characteristics of Dutch vowel sounds in a neutral context (Pols *et al.*, 1973) and wanted to apply the same approach in more complex contexts.

(b) The physical parameters by which any sound can be described are duration, intensity, periodicity, and spectrum. We have restricted ourselves to studying the power or level spectrum, irrespective of phase, as a function of time. Along with vowel duration, spectral variation is the most important variable with respect to vowel realization, and probably also the most important one in vowel perception. Spectral variation is not a variable which can be measured easily, but with the analysis system used, it appears to be able to give a useful description of the speech signal.

(c) There are different ways of measuring and representing spectral differences between vowel sounds. For a long time, the spectrograph has been the most commonly used instrument. The power spectrum, as a function of time, is measured

with a fixed-bandwidth filter (50 or 300 Hz), the midfrequency of which is continuously varied. In a frequency vs. time plot, the filter levels are represented as levels of greyness. Maxima in the amplitude spectrum, the so-called formants, are visible as dark bands. Using the frequencies of the first two formants as co-coordinate values, this results in a frequently applied two-dimensional representation of the vowel spectra. Newly developed, mainly digital, techniques allow faster spectral analysis. Also other parameters, like prediction coefficients, have been introduced to describe the envelope of the amplitude spectrum.

We developed an analysis system based on a parallel set of bandfilters, which allows real-time processing. The filter characteristic resembles the auditory filter in the ear. The bandfilter spectrum is not used to recover the formant parameters (the spectral resolution would be too limited to make that possible), but a statistically defined weighting of all filter levels is used to get an optimal spectral representation of the dynamic spectral vowel characteristics.

(d) In our earlier research projects we had studied the differences between vowels for many speakers, both male (Klein *et al.*, 1970; Pols *et al.*, 1973), and female (van Nierop *et al.*, 1973), in carefully spoken, isolated, stressed words of the type h(vowel)t. In the present study we have made one extension in the direction of variable initial and final consonants in carefully spoken monosyllabic words of the type initial consonant (C_i), vowel (V), final consonant (C_f). Although this condition is still rather artificial, it is one step further towards natural speech. In order to arrive at more general conclusions we used *three* male speakers. Consonant environment is only one of the reasons for spectral variation in vowel realization; there are many others like: individual physiological differences among speakers with respect to their speech production apparatus, going from minor individual differences to the differences between men and children; individual differences in speaking habits; temporary changes in the speaking habits, like "pipe speech"; dialectal differences; overall rate and carefulness of speaking; amount of stress in the pronunciation; different styles of speech, like vowels spoken in isolation, in read-aloud text, or in conversational speech; influence of the direct environment of the vowel phoneme, like preceding and following phoneme, or syllable.

(e) Apart from the spectral variations in the vowel sounds, which have their origin in the way in which those vowel sounds are produced, the acoustic properties of any transmission system between speaker and listener also introduce certain variations. Think for instance of a telephone line, a noisy environment, or a reverberant room. In the present project, this type of distortion was excluded

by using high-quality recordings and an optimal acoustic environment.
(f) Finally we have to specify how the perceptual differences between, and within, vowel phonemes will be studied. There are two different basic concepts underlying the study of vowel perception: one is similarity judgment, the other is the identification paradigm. In this study we use the second method to evaluate the perceptual differences between vowel sounds in different contexts. This method is less discriminative, in so far as two stimuli can be perceptually different but nevertheless be identified in the same way; however, identification results are more relevant with respect to speech recognition. Presented in their original consonant environment, the vowels would be correctly identified with a close to 100% correct score. In order to find out whether this high score is solely due to the physical structure of the vowels themselves, or rather partly due to the presence of prevocalic and postvocalic consonants, the vowel segments were isolated from the original words. Identification results will be compared with the average positions of the vowel segments in the spectral representation as well as with the dynamic characteristics.

The present study will try to give an answer to questions like:
(a) How large is, for a single speaker and with carefully spoken syllables, the spectral variation in the vowel caused by different consonant environments?
(b) When some measure has been obtained of this variation, can this spectral variation be predicted on the basis of the particular consonant environment?
(c) If this variation is not too large, is it then possible to recognize automatically the vowel segments in carefully spoken syllables, ignoring the consonant environment?
(d) Can misidentifications of the isolated vowel segments in the identification experiment be understood from the acoustic properties of the vowel segment, both with respect to the number of confusions per presented vowel, and to the type of confusions made by the subjects?

In the next paragraph (1.2) we will first give a short outline of the different ways of measuring spectral differences between vowel sounds. After that introduction it will be easier to give a review of the literature on spectral vowel differences (paragraph 1.3). In paragraph 1.4 we will give a brief survey of the different ways of measuring perceptual differences between vowel sounds, followed by a review of the literature on that topic (paragraph 1.5). Both surveys are concluded by a short justification of the approach used in this study.

The position of this study with respect to the presented literature will also be given.

1.2. MEASURING SPECTRAL DIFFERENCES BETWEEN VOWELS

Modern technology offers a wide variety of possible analysis techniques. In the spectral domain this variety includes speech spectrograms, zero-crossing distributions, bandfilter analysis, and digital techniques such as digital filtering, discrete Fourier analysis using the fast Fourier transform, and linear prediction. These techniques usually give large amounts of raw data which have to be processed further. Usually, formant parameters are applied to describe this raw spectral information in a reduced way.

In addition to a formant representation there are other ways of describing the spectral differences between vowel sounds. For example, Makhoul and Wolf (1972) used a number of prediction coefficients to describe the spectral envelope. As far as I know there has not yet been any systematic study of the spectral differences between vowels in terms of these prediction coefficients. This technique is frequently applied directly to speech synthesis (Markel and Gray, 1974; Atal and Hanauer, 1971) and automatic speech recognition (Itakura, 1975). Linear prediction spectra are frequently used to obtain the formant frequencies and levels (Markel, 1973; McCandless, 1974a).

Some researchers prefer to use the zero-crossing analysis technique to determine the formant frequencies (Scarr, 1968; Koopmans, 1973). In automatic speech recognition too, the zero crossings of the filtered or unfiltered signal are sometimes used as a parameter (Reddy *et al.*, 1973; De Mori, 1971; Niederjohn, 1975).

Another type of parameter used to describe the differences between speech sounds, both in the time domain and in the frequency domain, is the wave function. Speech, filtered in five frequency bands, is approximated by Gaussian modulated cosine waves, defined by so-called ASCON parameters (Culler, 1970; Pfeifer, 1972; Retz, 1973). Also, the Walsh-Hadamard Transforms are suggested for speech wave analysis (Tanaka, 1972; Ying *et al.*, 1973). Beninghof and Ross (1970) approximated speech spectra by means of a linear combination of orthogonal Gram polynomials. A 20-term polynomial fit turned out to be better than a 20-term trigonometric fit. The original spectra were described with 100 points over 4000 Hz. These last few approaches are mainly mathematically interesting but are also used successfully for speech segmentation and automatic speech recognition.

Most of the spectral-analysis methods described so far include a detailed analysis in the frequency domain from which then in some way the formant parameters are derived. Advanced computer programs are necessary for more or less successful automatic formant extraction. Continuity constraints over the utterance have to be taken into account and interactive processing is sometimes unavoidable. Real-time processing is impossible. These difficulties can partly be explained by the fact that formant positions are parameters directly related to the speech production process, whilst they are extracted from spectral data where the formant positions themselves are only available in an indirect way.

Another approach is to do a less detailed frequency analysis, more in accordance with the frequency-analyzing properties of the human ear, for instance by using a set of parallel bandpass filters roughly 1/3-octave wide. These bandfilter spectra are hardly detailed enough to extract formant parameters. The original filter levels can, however, be used directly as descriptive parameters. A further data reduction is possible, by means of, for instance, a principal-components analysis resulting in an m-dimensional spectral representation where each coordinate value in a lower-than-m-dimensional subspace is a linear combination of the original m filter levels. This possibility was first described by Kramer and Mathews (1956) for low-bit-rate speech transmission and later used by various research groups for speech analysis, synthesis, and automatic recognition.

In order to describe spectral vowel differences, this dimensional approach was first used in our institute for Dutch vowels (Plomp *et al.*, 1967; Klein *et al.*, 1970; Pols *et al.*, 1973), soon followed elsewhere by Boehm and Wright (1968), Li *et al.* (1968), Hughes *et al.* (1969), and Wright (1972) for American English vowels, and by Seidel and Paulus (1971), and Seidel (1974) for German vowels.

1.3. REVIEW OF THE LITERATURE ON SPECTRAL VOWEL DIFFERENCES

As was indicated in the Introduction, we will limit ourselves in this study to the spectral differences within and between vowel phonemes in different consonant environments, spoken by different speakers. However, in order to place this problem in its proper framework we will start this review of the literature at a somewhat more elementary level and discuss first the basic spectral differences between the vowel phonemes in a null context (section 1.3.1). All extra variation can then be related to these average spectral differences between vowel phonemes. Next, we will go one step further and briefly review the interindividual vowel differences introduced by different speakers, with the vowels still in a neutral context (section 1.3.2). Only then will we come to a review of the literature most directly related to the present research topic, namely the spectral vowel differences as a function of consonant environment (section 1.3.3).

Some concluding remarks (1.3.4) complete this paragraph. At that point we will also briefly indicate the position of this study with respect to the literature.

1.3.1. Average spectral differences between different vowel phonemes in a neutral context

In order to limit the context effect, isolated vowels are used, or vowels spoken in a neutral context like h(vowel)t. Stevens and House (1963) call this the "null" context for the vowels. In order to limit the speaker effect, most investigators average the spectral information of the vowels over a number of speakers having a pronunciation representative of the language concerned.

In most studies, the spectral vowel information for different languages is expressed at least in terms of the frequency of the first few formants. The frequency of the i-th formant, being the i-th maximum in the amplitude spectrum, is represented by F_i and its level by L_i.

Peterson and Barney (1952) did such measurements for 10 American-English vowels in a context of h(vowel)d with 33 male speakers. In that study there were some dialectal differences between the speakers (Nordström and Lindblom, 1975). Fant (1959) analyzed 14 sustained Swedish vowels spoken by 7 males. Here again the choice of the speakers, in terms of dialect control, could have been stricter (Nordström, 1975). In a later study (Fant *et al.*, 1969) the number of male speakers was increased to 24. Pols, Tromp, and Plomp (1973) analyzed 12 Dutch vowels spoken in a context of h(vowel)t by 50 males. In a recent study, Fant (1975) mentions some other less well-known data sets for languages like Danish (Frøkjær-Jensen, 1967), Estonian (Liiv and Remmel, 1970), Serbo-Croatian (Lehiste and Ivič, 1963), Dutch (Koopmans, 1973), Japanese (Suzuki *et al.*, 1967), and Italian (Ferrero, 1968).

Lobanov (1971) gives F_1-F_2 data of five Russian vowels and their soft variants for two male speakers. The vowels were embedded in 90 different nonsense syllables. Lobanov combines his data with data from Fant (1960) and Öhman (1964) on two other Russian speakers. Hess (1972) gives formant regions in the F_1-F_2 plane for the German vowels from 12 male and four female speakers. German vowel formant data have also been published by Jörgensen (1969) for six males. Majewski and Hollien (1967) give the first two formant frequencies of six Polish vowels,

both sustained and in a context of b(vowel)t, spoken by seven men. Jassem (1968), too, gives formant frequencies for Polish vowels. F_1 and F_2 from nine Hungarian vowels, repeatedly spoken by four men in 32 two-syllabic words, were measured by Tarnóczy and Radnai (1971). Fujisaki and Kawashima (1968) measured the formant frequencies of the five Japanese vowels spoken in isolation by 14 male speakers. In so far as numerical values for F_1 and F_2 were available from these publica-

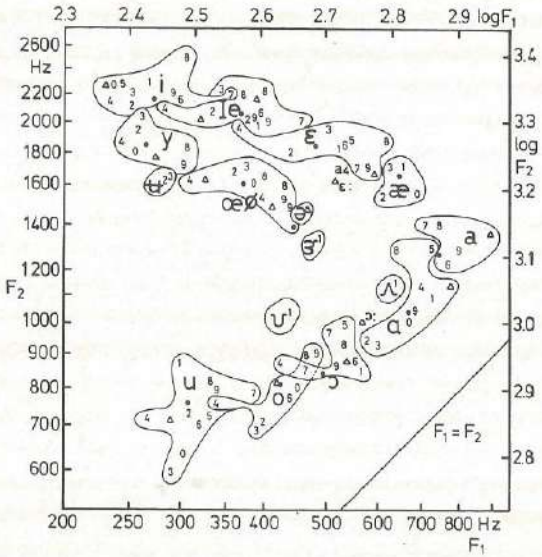


Fig. 1.3.1. Average formant frequencies F_2 versus F_1 for a number of vowels in different languages for male speakers. The vowels for which the same phonetic symbol was used in the original publications, are grouped together.

1. Peterson and Barney (1952) , 10 American-English vowels, hVd, 33 males;
2. Fant (1959) , 14 Swedish vowels, sustained, 7 males;
3. Fant *et al.* (1969) , 14 Swedish vowels, isolated long, 24 males;
4. Frøkjær-Jensen (1967) , 11 Danish vowels, sustained, 10 males;
5. Majewski and Hollien (1967), 6 Polish vowels, sustained and bVt, 7 males;
6. Ferrero (1972) , 7 Italian vowels, isolated, 25 males;
7. Lehiste and Ivić (1963) , 6 Serbocroatian vowels, stressed syllables in words, 5 males;
8. Koopmans (1973) , 12 Dutch vowels, monosyllables, 10 males;
9. Pols *et al.* (1973) , 12 Dutch vowels, hVt, 50 males;
0. Liiv and Remmel (1970) , 9 Estonian vowels, in isolation, 6 males;
Δ. Govaerts (1974) , 15 Dutch vowels, sustained, most representative from 10 males, twice;
•. Fant (1975) , 16 vowels, averaged over 6 languages.

	1		2		3		4		5		6		7		8		9		0		Δ		•	
	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
ɑ	640	1190																					640	1190
ɑ:	730	1090	580	940	600	925	698	1144	729	1281					639	1292	679	1051					670	1045
a							541	1708			770	1220	700	1400	738	1409	795	1301	666	1001	890	1360	735	1270
æ	660	1720	610	1550	625	1720													684	1575			650	1670
ɛ	530	1840	440	1795	505	1935	367	1973	543	1872	540	1870	555 "short/e/"	1715	618	1877	583	1725			590	1695	480	1840
ɛ:																					520	1640		
ɪ	390	1990							380	1993					374	2196	388	2003			320	2000	} 360	2045
e			335	2050	345	2250	283	2091			390	2010	460	1995	403	2218	407	2017	381	2074	390	2170		
ɤ	270	2290	255	2065	255	2190	235	2119	245	2286	300	2140	355	2192	306	2494	294	2208	239	2279	225	2280	275	2190
ɔ	570	840					426	876	534	999	550	870	505 "short/o/"	970	538	928	523	866			540	875	500	840
ɔ:																					570	1000		
o			400	710	390	690	340	755			440	770	470	860	464	999	487	911	454	779	420	815	425	815
ʊ	440	1020																					440	1020
u	300	870	310	730	290	595	266	722	332	725	320	700	390	790	334	813	339	810	304	666	290	720	310	760
ʏ	490	1350																					490	1350
ɜ																					465	1475		
œ							380	1546							436	1596	438	1498			325	1640	450	1390
ø			365	1690	380	1730	308	1633	308	1633					406	1650	443	1497	386	1613	413	1500	375	1610
ʉ			285	1638	285	1640																	283	1633
y			260	1930	260	2060	242	1862							308	1819	305	1730	255	1813	275	1775	265	1835

Table 1.3.1. Average formant frequencies of vowels in various languages. All data concern male speakers. A graphical representation of these data is given in Fig. 1.3.1. The legend accompanying that figure presents the specifications of the studies concerned.

tions, these are combined in Table 1.3.1 and Fig. 1.3.1. Added are data from Govaerts (1974) representing one set of 15 Dutch vowels which were selected from sustained vowels spoken twice by 10 different speakers, and judged by 130 listeners as being perceptually most representative of the language area concerned. These formant positions are somewhat more extreme than the Dutch vowel formant positions as averaged over 50 male speakers (Pols *et al.*, 1973). Apparently, if subjects are free to select an "ideal" vowel they tend to choose somewhat extreme vowel positions. This effect is also clearly seen in a study on tolerance and intolerance in vowel perception by Cohen *et al.* (1967).

The general distribution of the different vowel sounds in the formant plane is nicely demonstrated in Fig. 1.3.1.

Also included in Table 1.3.1 and Fig. 1.3.1 are the formant frequencies averaged over six different languages as gathered by Fant (1975). However, it is very questionable if this averaging over more or less similar vowels in different languages makes sense. Fig. 1.3.1 illustrates that the use of the same IPA (1967) phonetic symbols does not al all imply equal spectra. This is shown in a similar way by Nordström and Lindblom (1975), who give a picture of formant positions of male vowels all written as /i, e, ɛ, a/ from 38 studies of 22 langua-

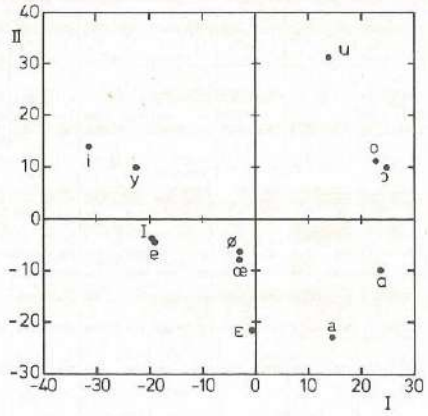


Fig. 1.3.2. Two-dimensional principal-components representation of the one-third octave spectra of vowels, spoken in a context of h(vowel)t, and averaged over 50 male Dutch speakers (Klein, Plomp, and Pols, 1970).

ges or dialects. Considerable variation and overlap were shown to exist.

If, instead of the formant parameters, a principal-components representation of average bandfilter spectra is used, one can equally well illustrate the average spectral differences between different vowel phonemes in a null context. Fig. 1.3.2 illustrates this for Dutch vowels, spoken in a context of h(vowel)t, by 50 male speakers (Klein *et al.*, 1970). The analogy with the formant representation is striking. For more details see Chapter 2.

1.3.2. Spectral differences between vowels pronounced by different speakers

In the previous section we considered vowel data averaged over many speakers. We will next consider differences between the individual data in a neutral context.

When a number of vowels is repeatedly spoken by the same speaker under identical conditions, then the corresponding vowel points in the two-formant plane form strictly bounded areas with no overlap at all between the different vowels (Potter and Steinberg, 1950; Fant, 1959). However, different speakers introduce considerable variation. For example, in Fig. 1.3.3 all individual vowel points in the log F_1-log F_2 plane for 50 male Dutch speakers are represented (Pols *et al.*, 1973). This diagram is provided with logarithmic rather than linear frequency scales because this is more in line with the properties of the hearing system.

We notice in Fig. 1.3.3 considerable individual variation around the average

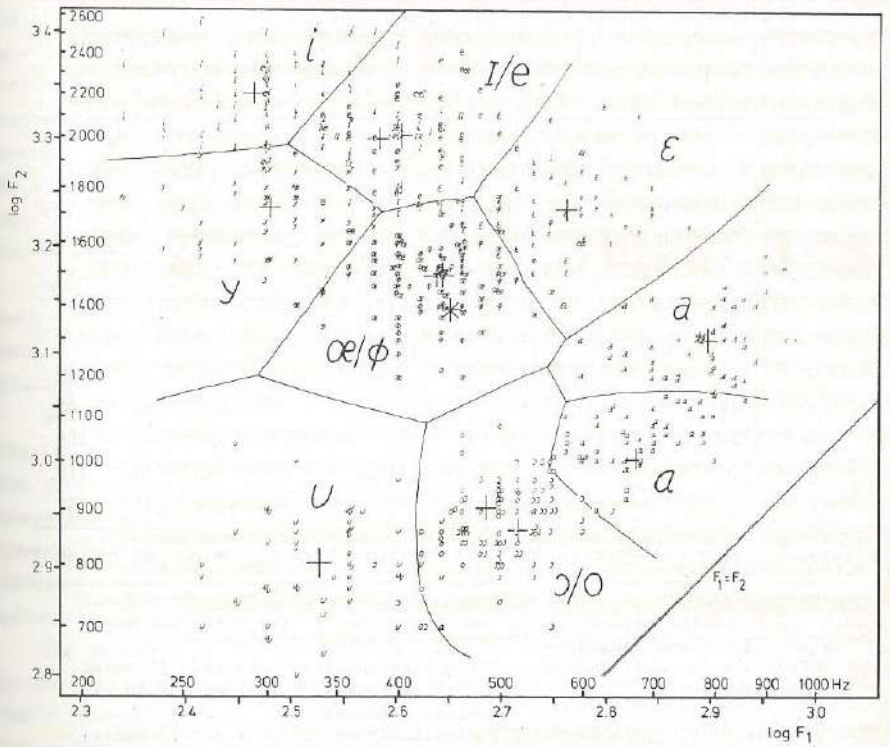


Fig. 1.3.3. Individual vowel positions in the log F_1-log F_2 plane for 50 male Dutch speakers. The crosses represent the average vowel positions, and the lines represent the boundaries of the maximum-likelihood regions (Pols *et al.*, 1973).

vowel positions. As a measure for this variation the spread σ, being the square root from the variance, can be used. Table 1.3.2 gives σ for F_1 and F_2 for the Dutch vowel data as well as for data from other publications. The largest spread is in our own data; this is probably caused by the large number (50) of phonetically naïve speakers, and the use of an h(vowel)t context instead of sustained vowels. For F_1 the "average" spread ($\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2}$) is about 45 Hz, and for F_2 about 110 Hz. In general the spread is higher for higher frequencies of the formants. Another component of the overall variance in the data is the variance between the average vowel points. For the data of Fig. 1.3.3 the square root of

	σ_{F_1}					σ_{F_2}				
	4	5	6	8	9	4	5	6	8	9
ɑ	79	53		31	80	57	96		82	89
a	73		52	37	95	177		84	73	113
ɛ	27	38	40	37	67	124	143	86	93	164
I		38		21	53		102		109	180
e	29		32	35	52	147		90	124	161
i	23	48	28	24	38	171	93	110	120	169
ɔ	30	40	39	64	49	60	96	77	89	72
o	27		41	42	42	35		59	81	90
u	39	46	28	24	46	94	97	55	68	85
œ	31			28	48	77			110	159
ø	26			52	46	81			78	115
y	23			29	42	79			131	152
average	41.5	44.2	38.0	37.3	57.3	110.0	105.9	82.0	98.7	134.3

Table 1.3.2. The spread σ per vowel for F_1 and F_2 for different data sets in different languages:
4. Frøkjær-Jensen (1967) , 11 Danish vowels, sustained, 10 males;
5. Majewski and Hollien (1967), 6 Polish vowels, sustained and in bVt, twice, 7 males;
6. Ferrero (1972) , 7 Italian vowels, isolated, 25 males;
8. Koopmans (1971) , 12 Dutch vowels, monosyllables, 10 males;
9. Pols *et al.* (1973) , 12 Dutch vowels, hVt, 50 males.

this variance, on a linear scale, is 145 Hz for F_1, and 464 Hz for F_2.

Part of the individual variation between the vowel points can be eliminated by some form of speaker normalization. In mathematical terms such a normalization can have the form of a translation of the individual vowel points, or a linear compression or expansion, or a rotation, or combinations of these (Boehm, and Wright, 1971; Lobanov, 1971; Fant, 1966; Gerstman, 1968). The Peterson and Barney (1952) data have frequently been used to test different types of speaker normalization (Gerstman, 1968), axis transformations (Foulkes, 1961; Beninghof and Ross, 1970), and recognition algorithms (Welch and Wimpress, 1961; Broad and Shoup, 1975).

In Klein *et al.* (1970), Pols *et al.* (1973) and van Nierop *et al.* (1973), different speaker normalization procedures were evaluated in terms of percentage of correctly recognized vowels, using maximum likelihood regions both for formant

data as well as for principal-components data. A simple and relatively efficient way of speaker normalization turned out to be a translation of the individual centres of gravity into a fixed point, for instance the origin of the coordinate system. The data corrected in this way were called the *centred* data. For the formant data this is similar to the parallel shift, as used by Gerstman (1968); he also included a linear compression or expansion. For the principal-components data a translation is interpretable as a subtraction of the average spectrum per speaker from the individual vowel bandfilter spectra.

Recently some speaker-normalization procedures have been suggested which are more directly based on the fact that differences between individual vowel parameters are partly due to differences in vocal-tract length. With some form of scaling these size-dependent factors can be eliminated.

Wakita (1973, 1975) uses linear predictive coding to calculate, along with the LPC coefficients, also the cross-sectional area as a function of place, and the total length of the vocal tract. By normalizing any speaker's vocal tract shape and length to some standard value, a major step forward can be achieved towards making automatic speech-recognition systems speaker-independent (White, 1976). Different speaker-normalization procedures have been used extensively in automatic vowel and word recognition systems.

The uniform scale factor suggested by Nordström and Lindblom (1975) is equal to the ratio of the actual vocal-tract length to the average length for the "male" vowels. In order to determine the actual vocal-tract length, a fixed relation between this length and the F_3 of open vowels is assumed. This scale factor was successfully applied to average formant data from males, females and children. A non-uniform extension of this procedure defines a correction factor that is a function of both formant number and vowel category. Fant (1975) has applied this to average vowel data of different languages, which resulted in small residual female-male differences. The final goal is to bring out real dialectal differences between speakers more clearly.

This brings us to a topic closely related to the differences between individual data, namely that of the differences between average data for males, females and children.

A number of studies give formant data for male and female speakers, and sometimes also for children, for various languages: American-English (Peterson and Barney, 1952; Eguchi and Hirsh, 1969; Strange *et al.*, 1976), Swedish (Fant, 1959), Danish (Frøkjær-Jensen, 1967), Dutch (Pols *et al.*, 1973; van Nierop *et al.*, 1973 and Koopmans, 1973), Estonian (Liiv and Remmel, 1970), Serbo-Croatian (Lehiste and Ivič, 1963), Japanese (Suzuki *et al.*, 1967), Italian (Ferrero,

1968), German (Hess, 1972), Polish (Majewski and Hollien, 1967) and Hungarian (Tarnóczy and Radnai, 1971).

All these studies indicate that there are systematic differences between the average formant parameters of male and female speakers. Acoustic theory predicts this on the basis of the different over-all vocal-tract lengths (Fant, 1960). One can define a scale factor $k_i = (\frac{F_{i\ female}}{F_{i\ male}} - 1) \times 100\%$ which transforms the i-th female formant frequency into the i-th male formant frequency. A constant scale factor k_i for all vowels would indicate a translation of the female data on a log F_i scale. For actual data, one scale factor for all vowels is too general. The grand average female-male scale factor for a number of vowels, averaged over different languages, is 17%; for Dutch it is 10%, according to Fant (1975). The average male and female formant data, together with the scale factors per vowel, which might more accurately be called correction factors, are given in Table 1.3.3 for Dutch. Pols *et al.* (1973) and van Nierop *et al.* (1973)

	F_1			F_2			F_3		
	male	female	correction factor	male	female	correction factor	male	female	correction factor
ɑ	679	762	12.2	1051	1117	6.3	2619	2752	5.1
a	795	986	24.0	1301	1443	10.9	2565	2778	8.3
ɛ	583	669	14.8	1725	1905	10.4	2471	2788	12.8
I	388	465	19.8	2003	2262	12.9	2571	2840	10.5
e	407	471	15.7	2017	2352	16.6	2553	2895	13.4
i	294	276	- 6.1	2208	2510	13.7	2766	3046	10.1
ɔ	523	578	10.5	866	933	7.7	2692	2852	5.9
o	487	505	3.7	911	961	5.5	2481	2608	5.1
u	339	320	- 5.6	810	842	4.0	2323	2746	18.2
œ	438	490	11.9	1498	1688	12.7	2354	2568	9.1
ø	443	476	7.4	1497	1690	12.9	2260	2512	11.2
y	305	288	- 5.6	1730	1832	5.9	2208	2520	14.1
average	473	523	→ 10.6	1468	1628	→ 10.9	2489	2742	→ 10.2

Table 1.3.3. Average male and female formant frequencies in Hz, together with the correction factors, for Dutch vowels in an h(vowel)t context. (Adapted from Pols *et al.*, 1973; and van Nierop *et al.*, 1973.)

also studied the differences between male and female vowel spectra in terms of a principal-components representation of the bandfilter spectra. The first two eigenvectors, which define the optimal two-dimensional subspace as a weighting of the original filter levels, are comparable with regard to the male and the female data. The highest weightings for the female data occur at about 1000 Hz and 2000 Hz, in both cases about 1/3-octave band higher than for the male data. This coincides with the average shift in formant frequencies of about 10% for F_1, F_2, and F_3. If one common eigenvector base is used for both male and female data, the averaged centred male and female vowel points are close to each other. They can be brought to almost complete overlap by a small rotation of one of the two configurations. Such a rotation corresponds with a frequency shift in the original spectrum.

In this section we have learned that there is considerable individual spectral variation within the different vowel phonemes, even when these are all pronounced in a neutral context. However, we have also seen that this individual variation is not randomly distributed over all vowels, but that speaker-normalization procedures can effectively be used to reduce considerably the individual differences. This holds just as well for the formant representation as for the principal-components representation of the bandfilter spectra.

So, on the one hand, it is advisable to use a number of different speakers, since this happens to be a source of variation which is of the same order as differences between neighbouring average vowel positions in the spectral vowel space. On the other hand, normalization procedures are available for neutralizing a great deal of this variation, in order to be able to specify better the general trends in the data.

1.3.3. Spectral differences between vowel sounds as a function of consonant environment

In the preceding two sections we considered the spectral differences between vowel phonemes in a neutral context, as well as the individual variation within these vowels introduced by different speakers. Now, as another source of variation, we have to see which spectral differences are introduced by different consonantal environments. These differences will vary from speaker to speaker, but the goal is to find the systematic effects (Stevens *et al.*, 1966).

The term "coarticulation" is often used in this respect, implying the influence of one speech segment upon another (Daniloff and Hammarberg, 1973). In this study I am mainly interested in the extent to which the spectral properties of the *vowel* segment are influenced by their surroundings. The ideal or target val-

ues will normally not be reached. When the sound spectrum of a vowel changes in the direction of a more neutral vowel, we speak about "vowel reduction". This term, however, is mostly restricted to the difference between differently stressed syllables (Delattre, 1969). Some authors use the term vowel reduction also with respect to coarticulation (Lindblom, 1963; Ohde and Sharf, 1975).

Various parameters of the vowel are influenced by coarticulation and have been objects of study, such as duration, fundamental frequency, and relative power (House and Fairbanks, 1953). The spectral effects got most of the attention, although in vowel reduction and coarticulation the duration of the vowel also plays an important role (Lindblom, 1963; Stevens *et al.*, 1966). Only as an illustration, Fig. 1.3.4 gives the effect of context condition on vowel duration. Data are given for different types of speech ranging from well articulated isolated vowels to unstressed vowels in a normal conversation.

Most effects are not restricted to neighbouring-phoneme interactions but may cover more than one phoneme. For vowel duration this is especially clear (Nooteboom, 1972), but spectral effects have also been reported over several phonemes

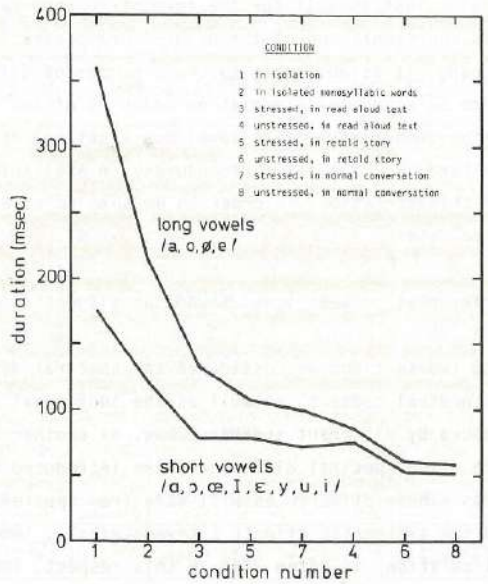


Fig. 1.3.4. Duration of Dutch vowels in the eight different contexts as indicated. The data are for one trained speaker and split up for long and short vowels. (Adapted from Koopmans, 1975.)

(Öhman, 1966; Kido *et al.*, 1968; Derkach *et al.*, 1970; Reitsma, 1974). More or less to his own surprise Gay (1975) recently found, however, no carry-over effects of the first vowel on the EMG signals of the second vowel in VCV syllables, suggesting a neutral tongue body position (relative to the vowel) during consonant production. In general, the coarticulation studies restrict themselves to neighbouring phonemes in monosyllabic stressed words spoken in isolation. The words are usually of the type C_1VC_2 or V_1CV_2 where C_1 and C_2, or V_1 and V_2 may be the same, giving a symmetrical environment, or different phonemes. Sometimes the CVC syllables are preceded by /hə/ or a similar neutral carrier (Stevens and House, 1963; Li *et al.*, 1972; Ohde and Sharf, 1975).

Mainly in those studies dealing with vowel reduction, other styles of speech than just isolated words, have been used (Tiffany, 1959; Li *et al.*, 1972; Stålhammar *et al.*, 1973; Ladefoged *et al.*, 1976; Earle and Pfeifer, 1975; Kanamori, 1975). Vowel reduction, in terms of spectral neutralization and vowel shortening, has been measured for a number of different languages. The experimental results are in good agreement: the vowel formant pattern in the F_1-F_2 plane shifts more and more in the direction of the neutral vowel, going from isolated vowels, via stressed vowels in isolated words, and stressed vowels in connected speech, to unstressed vowels in connected speech. Consonant environment is in general disregarded in those studies, see for instance Stålhammar *et al.* (1973).

Fig. 1.3.5 presents a clear illustration of vowel reduction in terms of spectral neutralization. These data are adapted from Koopmans (1975) for one trained speaker.

In procedures for the automatic recognition of continuous speech, some investigators prefer to start with the stressed syllables since these are believed to be the most readily decoded portions of continuous speech (Lea *et al.*, 1973; Li *et al.*, 1973b). In speech-understanding research, a first start has been made in defining phonological rules which handle vowel reduction and deletion (Oshika *et al.*, 1975; Cohen and Mercer, 1975).

Several degrees of extensiveness can be introduced in the spectral measurements to study coarticulation. This can vary from a single formant measurement somewhere in the middle, or in the most stationary part of the vowel (Stevens and House, 1963; Delattre, 1969; Stålhammar *et al.*, 1973; Ohde and Sharf, 1975; Earle and Pfeifer, 1975), to a number of formant measurements at discrete points in the vowel (Lehiste and Peterson, 1961; Lindblom, 1963), up to a complete description of the whole formant trajectory (Stevens *et al*, 1966; Broad and Fertig, 1969). Because of the difficulty of measuring formants, human intervention is often necessary, which makes all these measurements very time-consuming.

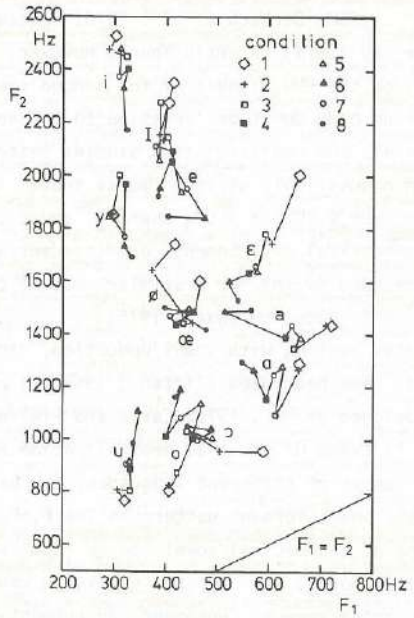


Fig. 1.3.5. Formant positions of Dutch vowels in eight different contexts, varying from vowels spoken in isolation to unstressed vowels in normal conversation. See the preceding figure for a full description of these context conditions. In this figure the conditions per vowel are connected in the following order 1-2-3-5-7-4-6-8. All data are from one trained speaker. (Adapted from Koopmans, 1975.)

Consequently, in some way or another, these studies are all rather restricted in their design. This means that only one or a few speakers are used, that only one style of speech is used, that the number of vowels and/or consonants and their possible combinations is drastically curtailed, or that the number of measurements in time over the utterance is limited. For instance, in a recent study Ohde and Sharf (1975) intended to study specifically the coarticulatory influence on the vowel of preceding consonant (= L(eft) to R(ight) coarticulation) and following consonant (= R to L coarticulation). However, they used only the vowels /i/ and /u/, and the consonants /b, d, g, h/ in the contexts hVCV (for R to L coarticulatory influence of consonant C on the first vowel V), CVhV (for L to R), CVCV (for symmetrical effects), and #V# (for vowel targets). Vowel coarticulation was determined by comparing the target formant values from the isolated vowels with formant measurements made only in the steady-state positions of the vowels in context. They found that L to R effects were considerably greater than

R to L effects, and also greater for /u/ than for /i/. Nevertheless, all these studies combined give a valuable insight into the spectral effects related to coarticulation. On the basis of the assumption that the average values in a neutral or null context define the ideal or target formant frequencies, coarticulation causes a deviation or undershoot relative to these target values. The formant frequencies can thus be lower or higher, depending on the consonantal context. Fig. 1.3.6 gives an example of these effects for the second formant frequency at initial, middle and final positions in the vowel segment for different vowels in hə'C_1VC_1 words, reported by Stevens *et al.* (1966). In automatic speech recognition, formant movements are used to recognize certain sounds (McCandless, 1974b; Weinstein *et al.*, 1974; Fujisaki *et al.*, 1974). With our system for real-time bandfilter analysis and statistical data processing, it takes less time to study coarticulation and, therefore, it will be possible to include more varia-

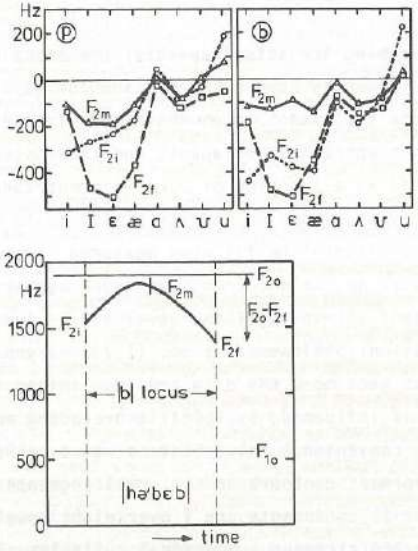


Fig. 1.3.6. In the lower part of this figure, a stylized F_2 transition is drawn for the /ɛ/ segment in the indicated word. The dashed lines represent the vowel boundaries. From these lines the initial (F_{2i}) and final (F_{2f}) second-formant frequency can be read. F_{2m} is midway between these two positions in time. F_{1o} and F_{2o} indicate the target values of the first and second formant of this vowel /ɛ/ determined in a null context. F_{2o}-F_{2i}, F_{2o}-F_{2m}, and F_{2o}-F_{2f} are represented in the upper part of this figure for 8 different vowels, and two different consonant contexts. The data are averages over three talkers and are adopted from Stevens *et al.* (1966).

bles. The research described in Chapter 3 is a first attempt in this direction. In fact the principal-components spectral representation has already been used in this respect by a few researchers (Li *et al.*, 1972; Newman, 1971).

Several authors have tried to explain the acoustical data on coarticulation and vowel reduction in terms of active commands to the speech production mechanism (Liberman *et al.*, 1967), including deliberately programmed forward coarticulation (Kozhevnikov and Chistovich, 1965; Eek, 1970), and in terms of the inertia of the articulatory mechanism (Stevens and House, 1963; Lindblom and Studdert-Kennedy, 1967; Ohde and Sharf, 1975). Delattre (1969) actually measured vowel reduction at the articulatory level by means of motion-picture X-rays which make visible the position of the tongue, lips, and other articulators, as well as the shape and volume of the resonance cavities of the vocal tract. Öhman (1967) did the same and developed a model based on the superposition of articulatory gestures. MacNeilage and DeClerk (1969) used high speed cinefluorograms and electromyography.

Models, mainly describing the actual spectral phenomena of vowel reduction and coarticulation, are given by some authors. Lindblom (1963) uses an exponential function to describe the extent to which formant frequencies in the vowels of C_1VC_1 syllables (eight short Swedish vowels and the consonants /b, d, g/) reach their target values as a function of vowel-segment duration.

This target is an asymptotic value and is considered to be the frequency that produces the best straight-line fit when measured formant values are plotted semilogarithmically as deviations from the target against vowel duration. The target is an invariant attribute of the vowel and is independent of consonantal context and duration. Stålhammar *et al.* (1973) extended Lindblom's model for vowel reduction, but once more the data are too limited to specify to what extent vowel reduction is influenced by specific preceding and following consonants. For mathematical convenience only, Stevens *et al.* (1966) used parabolic curves to describe the formant contours in the vowel segments of symmetric ha'CVC syllables. C varied over 15 consonants and V over eight vowels, and three talkers were used, resulting in 360 stressed symmetrical syllables. The results demonstrate that the formant frequencies at a point centrally located within the vowel (F_{2m} in Fig. 1.3.6), have shifted away from certain hypothetical target values F_{2o} (derived from vowels spoken in the null environment #V#, or hVd) under the influence of the consonantal environment. Also, the formant frequencies at the initial and final vowel boundaries (F_{2i} and F_{2f}) have shifted away from certain hypothetical consonant locus values, because of the vowel environment. Öhman (1966) concluded from his spectrographic measurements on VCV words, that

these utterances cannot be regarded as a linear sequence of three successive gestures. The stop consonant (C = /b, d, g/) gestures are actually superimposed on a context-dependent vowel substrate, describable as a diphthongal gesture which is present during all of the consonantal gesture.

Broad and Fertig (1970) developed a superposition model for the description of vowel-formant frequency trajectories in CVC syllables, based on statistical analysis of formant frequency measurements at 11 points in the vowel segment. Twenty-four different consonants were used in all combinations of both initial and final consonant, resulting in 576 CVC syllables, which were repeated three times by one male speakers. The vowel was always /I/. This large set of data made it possible to extract separate initial-consonant and final-consonant transition functions. The proposed linear model gives the mean formant trajectory for a given consonantal context as the sum of a target frequency, an initial-consonant transition function, and a final-consonant transition function. Nonlinear interaction between initial and final consonants appeared to be small.

This short survey shows that spectral effects of vowel coarticulation caused by different consonant environments have been studied only fragmentarily. Much more detailed and exhaustive measurements are necessary. Some descriptive and articulatory models exist but are, for instance, not good enough to be used in automatic speech recognition.

1.3.4. Concluding remarks

As was outlined in the preceding section, almost all spectral data on vowel coarticulation are given in terms of formant parameters. These parameters are difficult to measure, especially if one is interested in dynamic variations. Bearing in mind the good results achieved with the principal-components representation of bandfilter spectra of vowel sounds in a neutral context, we should like to introduce the same approach for the study of spectral vowel coarticulation.

The choice of a specific analysis method and the subsequent data processing is mostly based on theoretical as well as on practical grounds. We preferred to work with a fast, reproducible and unambiguous analysis system related to the processing of auditory stimuli in the peripheral hearing organ, rather than to the way the speech sounds are produced. For that reason we preferred not to work with formant parameters but to apply a bandfilter analysis. Further arguments for using this analysis procedure will be discussed in paragraph 2.1. The bandfilter spectra can be represented as points in a multidimensional space, which makes a further data reduction in terms of a principal-components representation possible. The analysis can be executed in real time, which is important if many

data have to be analyzed, and the subsequent data processing is straightforward and statistically well defined. There appeared to be a good correlation with the more traditional formant representation (Pols *et al.*, 1973). In paragraph 2.3 the details of this data representation procedure will be discussed.

For the moment it may suffice to say that this analysis system makes it possible to do a detailed dynamic spectral analysis of natural utterances in real time. The subsequent data processing and data representation give detailed information of the dynamic spectral behaviour of the isolated vowel segments represented in terms of numbers and trace displays.

1.4. MEASURING PERCEPTUAL DIFFERENCES BETWEEN VOWEL SOUNDS

Subjects in general judge a stimulus as an entity. They are able to say: I identify this stimulus as an /a/, or, this stimulus is more similar to sound A than to sound B, but it is much more difficult for them to say in which particular properties the similarity or dissimilarity resides. Consequently, the results of perceptual experiments are confusion matrices, or similarity matrices, or scores on semantic scales. In contrast, physical analyses can be very detailed, with much time-dependent information. However, these data are generally difficult to transform into overall measures, comparable with the perceptual results.

With some processing a certain amount of variation in the natural stimuli can be removed. An adjustment of the overall level is possible by re-recording the utterances using a time-variable amplification, or by using an intermediate digital storage and a variable multiplication factor. A question to be decided is whether perceptual loudness or physical root-mean-square (rms) values should be equalized. Stimuli with a fixed duration can be obtained by gating. Digital tape splicing makes it also possible to repeat or delete single periods until the desired duration is achieved. Minor differences in the fundamental frequency can be corrected by changing the sampling frequency of the digitally stored waveform; however, this also influences the formant structure. If independent control of fundamental frequency and amplitude spectrum is wanted, one has to use synthetic speech produced with, for instance, a formant synthesizer, or by the linear prediction technique.

Even if duration, overall intensity, and pitch of the natural vowel stimuli are equalized, there are still dynamic characteristics which are not completely under control; for example, onset, offset, and dynamic spectral variations. Of course one may argue that these are typical characteristics of the sound, which

have to be taken into account. Even then a straightforward correlation between physical and perceptual data is difficult: Is the average spectral information most representative of a vowel categorization, or is it the initial or final part, or perhaps the transition?

The utmost reduction is achieved by using stationary synthetic vowel stimuli (Cohen *et al.*, 1967; Govaerts, 1974; Bond, 1976; Pols *et al.*, 1969; van der Kamp and Pols, 1971). An additional advantage of synthetic stimuli, very useful for studying the effect of one specific parameter on the perception of vowel sounds, is that single parameters can be systematically varied (Fujisaki and Kawashima, 1968; Slawson, 1968; Ainsworth, 1971).

Simple, constant stimuli are easy to describe but are not very speech-like. More natural stimuli have so many variables that one is never sure which one, or which group, has caused a certain perceptual result. One may find this dilemma everywhere in the literature.

Specific aspects of speech perception can also be studied by the introduction of specific distortions into the naturally spoken sounds. A well-known example is the study on consonant confusion with bandpass filtering and masking noise by Miller and Nicely (1955). Comparable confusion experiments with low- and high-pass filtering, or noise masking, were performed with vowels (Miller, 1956; Pickett, 1957; Carterette, 1964; Castle, 1964). Duration is another parameter that can be controlled.

Confusion experiments need some kind of distortion of the stimuli because the percentage of confusions appears to be too low when only the natural vowel sounds themselves are used. As far as I know, Govaerts (1974) is the only investigator who nevertheless did such an experiment without extra distortions. In order to get enough off-diagonal data in the confusion matrix, many presentations were necessary which made the experiment very time-consuming. Govaerts used 15 isolated Dutch vowels spoken twice by 10 speakers; the vowels were identified by 130 listeners, resulting in 2600 observations per vowel. At first glance, one is tempted to say that this seems to be the optimal set of data. However, once again one cannot exclude the effect of dynamic spectral variations and differences in duration, and speaker-specific effects may also have had some influence. Even with this variation, vowel sounds spoken in isolation cannot be regarded as natural stimuli.

In many identification experiments some sort of distortion has been introduced quite arbitrarily to get enough confusions. Apart from adding noise and filtering, shortening of the signal or resynthesizing speech can also be used as forms of distortion. All these approaches have the drawback that they also

change the stimulus itself.

An alternative is to introduce certain types of distortions in the identification task, like a memory task, a secondary task, shadowing, utilization of choice reaction times, cross-linguistic settings (Singh, 1975), etc, rather than in the stimuli. Here again one can object that this does not just make the task more difficult but may also change the way the stimuli are processed internally (van den Broecke, 1976).

Another procedure in perceptual experiments is similarity judgments. Stimuli are presented in pairs and subjects have to say how similar the two stimuli are on, for instance, a seven-point scale. Or stimuli are presented in triads and subjects are requested to decide which pair out of the three possible ones is most similar, and which pair is least similar. In both procedures the stimuli are not judged along some predefined scale, as in semantic scaling, but the subject has to take into account all possible aspects of the stimuli. A similarity matrix is the result of this type of experiment. In such a matrix the cell values only have an ordinal relation. Through multidimensional scaling programs the information in these matrices can be represented as configurations of vowel points (Kruskal, 1964a, 1964b; Carroll and Chang, 1970; Harshman, 1970). Subsequently it is then often possible to make a comparison with the physical parameters of the stimuli.

1.5. REVIEW OF THE LITERATURE ON THE PERCEPTUAL DIFFERENCES BETWEEN VOWEL SOUNDS

After having discussed the spectral differences between vowel sounds, we will now consider the perceptual differences. Here again we can make the distinction between isolated vowel sounds, or vowels in a neutral or null context, the effect of different speakers, and the effect of different contexts.

1.5.1. Perception of isolated vowel sounds

Well-articulated undistorted isolated vowel sounds can reasonably well be identified (Fairbanks and Grubb, 1961), albeit that subjects score higher when some context is available (Shankweiler *et al.*, 1975; Strange *et al.*, 1976). Identification or confusion experiments can be used to get some insight into the internal vowel representation. Only Govaerts (1974) did such a vowel confusion experiment with undistorted, naturally spoken, isolated vowel sounds. Large numbers of listeners were necessary to get enough non-zero off-diagonal cell values in the confusion matrix. Therefore, most investigators prefer some type of distortion such as high-pass or low-pass filtering (Miller, 1956; Lehiste and Peterson, 1959; Carterette, 1967), bandpass filtering (Castle, 1964; Landercy and Renard, 1975), noise masking (Pickett, 1957; Nooteboom, 1968), stimulus shortening (van der Kamp and Pols, 1971), or clipping (Gupta *et al.*, 1971).

Sometimes semantic scales, like bright-dark, are used to study the perceptual dimensions of vowels (Fischer-Jørgensen, 1967), or short-term recall (Wickelgren, 1965). Subjects are also quite capable of indicating that for instance /ɑ/ and /ɔ/ are more similar than /i/ and /u/. This type of similarity judgments can be used to build up an internal vowel representation in terms of a points configuration. Every vowel is a point in such a space, and the distances between different vowels are a measure of their dissimilarities (Pols *et al.*, 1969).

Terbeek and Harshman (1971, 1972) found cross-language differences for such a perceptual vowel space and concluded that there is no universal perceptual space. Neither the number of dimensions nor their interpretation appeared to be consistent across the five languages tested (German, Thai, Turkish, Swedish, and English): "The patterns which emerge suggest that the function relating perceived distances between vowels to their position along underlying perceptual dimensions is non-Euclidean in two ways. First, the perceptual dimensions do not lie orthogonally to one another, implying that they are related in meaning. Second, they interact nonlinearly, producing a curved space, an effect which causes the extraction of an uninterpretable dimension when linear models are used". For the time being we nevertheless prefer linear models which are far more easy to use, since data are hardly ever detailed enough to test this non-linearity.

Some investigators correlated the dimensions of the linear perceptual space with general spectral measures of the stimuli, or with specific formant parameters (Hanson, 1967; Knops, 1967; Pols *et al.*, 1969). Fig. 1.5.1 is an example of

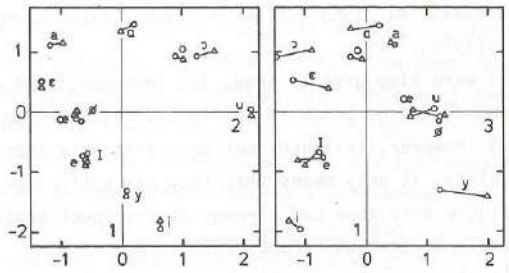


Fig. 1.5.1. Result of matching a four-dimensional perceptual configuration (o) with a four-dimensional principal-components representation of average vowel bandfilter spectra (Δ). The correlation coefficients along the three presented dimensions are .997, .995, and .907, respectively (Klein, Plomp, and Pols, 1970).

such a perceptual vowel representation derived from an identification experiment with 100-msec gated-out vowel segments from 50 male speakers, presented to 10 listeners (Klein *et al.*, 1970). The cumulative confusion matrix was symmetrized and processed by Kruskal's multidimensional scaling technique. The four-dimensional perceptual configuration was matched (Cliff, 1966) with the four-dimensional principal-components representation of the average bandfilter spectra. The first three dimensions are represented in Fig. 1.5.1.

The similarity judgments are also often correlated with some type of distinctive-feature system (Hemdal and Hughes, 1967; Mohr and Wang, 1968; Singh and Woods, 1971). We restricted ourselves to relations between the physical properties of the speech stimuli and speech perception, without making the steps from speech sound to speech producing mechanism and to speech commands.

Bernstein (1974, 1975) indicated a possibility to test the consistency of vowel differences in some physical data representation with the perceived distances between vowels. The underlying mathematics had been developed by Tversky and Krantz (1970). Bernstein used as stimuli 400-msec three-formant synthetic vowels which were presented in pairs for a magnitude judgment. Only in those parts of the formant space where F_1 was close to F_2, and F_2 close to F_3, was this representation not consistent with the perceived vowel distances.

Another approach is to use synthetic stimuli and to vary those specific parameters one believes to be important in vowel perception. In this way, Flanagan determined the difference limen for vowel formant frequencies (Flanagan, 1955) and for formant amplitudes (Flanagan, 1957). Stevens (1951) measured the just perceptual change in formant bandwidth. Flanagan (1961) and Rosenberg (1971) measured the influences of the glottal wave upon vowel quality. Interactions between fundamental frequency and vowel quality have been studied by Ainsworth (1971), Fujisaki and Kawashima (1968), Slawson (1968), Kakusho and Kato (1968), and Matsumoto *et al.* (1973).

Synthetic stimuli were also used to show, for instance, that with only two formants all vowel sounds can be produced (Miller, 1953; Cohen *et al.*, 1967; Carlson *et al.*, 1970). However, this does not mean that only those two parameters define vowel quality, it only means that they are sufficient under certain conditions. Even a single pure tone can already give a vowel sensation if subjects are asked to reply in that way (Fant, 1959).

Synthetic stimuli also make it possible to generate a series of vowel-like sounds which follow a continuous trajectory from one vowel to another in some physical vowel representation, such as the two-formant plane. Those stimuli are used in identification and/or discrimination experiments.

Kasuya *et al.* (1971) presented data and a model for vowel identification and discrimination in terms of what they call the Psychological Auditory Space (PAS). This PAS, constructed by using a multidimensional scaling method, is similar to what we have called the two-dimensional perceptual space (Pols *et al.*, 1969). In their model, a vowel, with certain formant frequencies in the physical space, is represented in the PAS by a normal density function with a certain mean value and a certain σ, being the psychological noise. They found from difference-limen (DL) measurements that the DL of phonetic quality, represented in the PAS, was a circle with equal radius for all vowels. In our opinion this is a predictable result, considering the meaning of such a representation. The DL in the F_1-F_2 space was smaller at the vowel boundaries than in the centre of a vowel region. One can also say that the perceptual distance between two vowels at a fixed distance in the F_1-F_2 plane, was greater at the vowel boundary than in the centre of a vowel region. Therefore, the discriminability at phoneme boundaries increases, leading to a more categorical perception.

Identification and discrimination of vowels are already more "central" processes of speech perception. Speech as an acoustic signal is first processed in the inner ear, this organ is capable of performing a temporal and spectral analysis. Our interests concern the question as to which physical parameters of the vowel sounds extracted by the inner ear are the most relevant attributes leading to perception of a speech sound. It is very difficult to study this question in an unprejudiced way. The experimental set-up of a perceptual experiment often implies a biased selection of certain models because of the type of synthesis used (Fant, 1970), the way in which certain stimulus parameters are varied (Mushnikov and Chistovich, 1973), the interpretation of the experimental results (Mohr and Wang, 1968), etc.

We can differentiate between a number of models, hypotheses, interpretations, or approaches by referring to the suggested basic parameters in vowel perception: (1) (articulatory) (distinctive) features; (2) (weighted) formant parameters, and (3) spectral weighting.

The feature approach is often used when no physical measurements of the stimuli are available, and perceptual results are interpreted in terms of known or supposed qualities of the stimuli. Examples of these are the already mentioned studies of Mohr and Wang (1968), and Singh and Woods (1971), who tried to explain their perceptual results in terms of different feature systems.

Hemdal and Hughes (1967) used the physical measures of formant frequencies to classify vowel sounds in terms of the distinctive features proposed by Jakobsen, Fant and Halle (1952).

In surveying the a-posteriori features of vowel perception, Singh (1974) concluded that the two most important features found in most studies (Hanson, 1967; Pols *et al.*, 1969; Singh and Woods, 1971; Anglin, 1971; Shepard, 1972; Terbeek and Harshman, 1971; Grant, 1971) are tongue height and tongue advancement, with the latter having the larger contribution. Less frequently recovered features are tenseness and retroflexion. Singh could only come to this general agreement by reinterpreting findings from certain authors in his own way. This is especially clear where he interprets F_1 and F_2 in terms of height and advancement, suggesting a one-to-one relationship. However, this presupposes that physical measurements of the stimulus, like F_1 and F_2, can be interpreted in terms of articulatory features. We prefer to make less stringent presuppositions by comparing the perceptual data directly with the physical parameters of the vowels.

The other two approaches mentioned have the same starting point: perceptual results are interpreted in terms of physical properties of the stimuli. The first two formant frequencies are traditionally considered to be very important factors in speech perception (Hanson, 1967; Knops, 1967; Pols *et al.*, 1969). Fant (1959) and Carlson *et al.* (1970) suggested expressing vowel quality in terms of F_1 and F_2' with F_2' matching F_2 for back vowels, representing a location intermediate between F_2 and F_3 for open front vowels, and a location close to F_3 or higher for high unrounded front vowels. Recently Carlson *et al.* (1975) replaced the simple F_2' formula by one relating F_2' to F_1, F_2, F_3, and F_4. This new F_2'-value corresponds fairly well with results from matching experiments where the higher formant frequency of two-formant vowels is matched with synthetic vowels with four formants. Carlson *et al.* also presented an operational model for deriving F_1 and F_2', based on the most prominent peaks in a histogram of the zero-crossing distribution in a number of frequency bands. This can only be an operational model, and not a model of the perceptual process. Karnickaya *et al.* (1975) suggest that "certain linear combinations of the formant frequencies rather than the formant frequencies themselves can be considered as useful features underlying the vowel identification", although their combinatory rules are not well specified. They present a rather detailed model for the perception of steady-state vowels (Karnickaya *et al.*, 1975; Chistovich, 1971). The part of the model dealing with peripheral signal processing is based on psycho-acoustical and electrophysiological results and the part dealing with parameter extraction and final identification is based on vowel-perception experiments. The model includes the following steps:

- spectral analysis of the input signal by a bank of parallel filters uniformly spaced along the bark scale at 0.1 bark distance. Bark is a unit for critical bandwidth;
- comparison of the level per filter with a threshold level;
- nonlinear transformation of the filter outputs into a loudness density function according to the model proposed by Zwicker and Feldtkeller (1967);
- simulation of lateral inhibition through a weighting function including negative terms, which causes a sharpening of the dominant spectral maxima. The final result is what the authors call the spectral equivalent curve;
- positions of the maxima are extracted with a peak-picking mechanism resulting in formant frequencies as the characteristic cues of the spectral equivalent curve. Spectral envelope and bandpass hypothesis are rejected by these authors as possible bases for the spectral equivalent (Chistovich and Mushnikov, 1971; Mushnikov and Chistovich, 1973). However, there is still the problem of how to describe the peak-picking mechanism in the first formant range for high F_0. Does a single harmonic, namely the one with the greatest amplitude (Mushnikov and Chistovich, 1973), or a weighted mean of adjacent harmonics represent the perceived F_1 (Carlson *et al.*, 1975)? The monotonous /i-e/ boundary shift in F_1 with increasing F_0 favours the latter view;
- vowel identification is finally based on simple decision rules applied to the formant frequencies.

In the description of our spectral-analysis method in Chapter 2, we will come back to the successive steps mentioned in this model, and see what modifications, additions, and simplifications may be made.

The use of (weighted) formant frequencies is just one way of representing the spectral vowel information. The spectral weighting based on a principal-components representation of the vowel bandfilter spectra is another one. The distances between vowel points in this spectral space correlate highly with the perceptual vowel differences, as we saw in Fig. 1.5.1 (Pols *et al.*, 1969; Klein *et al.*, 1970). Vowel identification scores derived from the positions of the vowel points in this spectral space can compete with identification scores based on formant parameters (Pols *et al.*, 1973; van Nierop *et al.*, 1973). Instead of the simple decision rules suggested by Karnickaya *et al.* (1975) maximum likelihood regions can be used, and, also, a speaker normalization can be introduced.

In this section we have seen that it is quite possible to evaluate the perceptual differences between isolated vowel sounds. Processing the experimental results in such a way that the vowel stimuli are represented as a configuration of points seems to be most promising since in that way a comparison with physical or other characteristics of the stimuli is possible. Experimental condi-

tions often confuse the interpretation of the data since they may interfere with the stimuli themselves or with their internal processing.

1.5.2. Perception of vowel sounds from different speakers

The most striking difference between vowel sounds from different speakers is the difference in pitch between males, females and children. In addition to this difference in pitch, there is also a difference in overall vowel quality, best known in terms of higher formant frequencies for female speakers than for male speakers. There is a general feeling that the fundamental frequency F_0 has an influence on the perceived vowel boundaries. Miller (1953), using a 100-component harmonic tone synthesizer to produce steady-state vowel sounds, found that when F_0 was doubled, the perceptual vowel boundaries only shifted notably for the centrally located vowels /ʊ, ʌ, æ/. In an identification experiment with two-formant stimuli, Fant (1970) found that the average vowel position shifted on the average 75 mels if F_0 was doubled from 110 to 220 Hz. Slawson (1968) extensively studied the effects of changing fundamental frequency and spectral envelope, on both vowel quality and musical timbre. Subjects were asked to judge differences in vowel quality between pairs of stimuli differing in spectral envelope or fundamental frequency to various degrees. The synthetic vowels used were /i, æ, ɑ, ɔ, o, u/. With a doubling of the fundamental frequency, F_1 and F_2 had to be increased by about 10% in order to minimize the change in vowel quality. This fits the actually found differences between male and female speech. In general, vowel quality and timbre depend more on the absolute frequency of the spectral envelope than on its position relative to the fundamental frequency.

Fujisaki and Kawashima (1968) studied the importance of pitch and higher formants by investigating the extent to which changes affect the perceptual boundaries between pairs of synthetic vowels /u-e/ and /o-a/, which share approximately the same ratio of F_2 to F_1. Combined changes in pitch and higher formants are necessary to counteract changes in F_1 and F_2. The shifts in vowel boundaries were proportional to F_0. This was also found by Carlson *et al.* (1975) for the vowel boundary between /i-e/ from an identification test with varying F_0 and F_1. Ainsworth (1974) found wide individual variations in the extent to which F_0 influences the vowel boundaries /u-ʊ-æ/ for English listeners. This suggests that F_0 alone is not uniformly used as a means of normalizing the vowels produced by men, women and children.

Under normal conditions, the human listener is very flexible in switching from one speaker to another, perceptual speaker normalization seems to be

almost instantaneous. Only under special experimental conditions can the listener be misled (Ladefoged and Broadbent, 1957). Verbrugge *et al.* (1976) did some experiments to find out what specific information enables a listener to map a talker's vowel space. Three vowels from one speaker presented as precursors of a to-be-identified vowel in an /h-d/, or a /p-p/ context, spoken by the same speaker, did not reduce the vowel identification errors relative to the condition without precursors. The vowels spoken by 30 speakers (men, women, and children) were presented in a mixed order. The three point vowels /i/, /a/, and /u/ as precursors did not give better results than three central vowels /I/, /æ/, and /ʌ/. Precursors mainly influenced listener's response biases, rather than effecting true improvements in vowel identifiability. Sentence context did aid vowel identification, but the authors claim that this is primarily because then adjustment to a talker's tempo is possible, rather than to that talker's vocal tract. This tempo is very important: results from Strange *et al.* (1976) offer strong evidence that dynamic acoustic information, as distributed over the temporal course of the whole syllable, is used often by the listener to identify the middle vowel.

These results suggest, somewhat to our surprise, that speaker-normalization procedures, in terms of adjusting perceptual vowel boundaries, may not be essential in human vowel perception. Perceptual speaker normalization seems to have more to do with tracking the dynamics of the ongoing articulation than with vocal-tract normalization as traditionally defined. I regard it as an omission that the possible role of the fundamental frequency is not even mentioned in these studies. Also, the fact that precursors and test words were spoken as independent utterances may have influenced the results.

In the light of these experimental results it is somewhat difficult to understand that statistical speaker-normalization procedures on formant data, like the one of Gerstman (1968), give such high scores. One has to realize, however, that such algorithms are not the same as perceptual strategies and that those calculations are based on single formant values, whereas human identification is based on the whole vowel. Recently van Balen (1977) seriously criticized different aspects of the experiments of Verbrugge *et al.* and Strange *et al.*

1.5.3. Perception of vowels in different contexts

In section 1.3.3 we have briefly described some of the spectral variations of vowel sounds caused by different consonantal environments. In this section we will consider what perceptual effects on vowel quality are described in the literature. It is clear from spectral measurements on vowels that there is spec-

tral undershoot and neutralization caused by consonantal environments and styles of speech. However, most of these vowel sounds are not misunderstood if presented in their original context. Misidentification can occur if the whole consonant or part of it is cut away and only the gated vowel part is presented to subjects. So either the spectrum of a vowel in a certain context is near enough to a target to be correctly understood, or, if it is not, the listener adapts himself to the situation. Does he extrapolate the unfinished spectral movement towards the target and then base his decision on this extrapolated target value, or does he adapt his perceptual vowel boundaries to the different consonantal environments, styles of speech, and talking rates? Present measurements do not yet allow a definite choice between these two concepts, as will become apparent from the investigations described below.

Consonant and vowel perception can be studied by presenting, for instance, CV syllables (spoken in isolation) from which more and more of the initial part of the utterance is removed ('t Hart and Cohen, 1964; Grimm, 1966). Kuwahara and Sakai (1973) found, as can be expected, that the perception of the initial consonant was heavily affected, as more and more of it was removed. The perception of the vowel, however, remained unaffected.

Perception of the vowel in CV syllables, not spoken in isolation but taken from a read radio news broadcast, was found to be seriously impaired if the vowel parts were presented in isolation. The scores for the five Japanese vowels /i, e, o, u, a/ ranged from 52% to 70%. These perceptual vowel confusions were reported to be highly correlated with the deviating positions of these vowel portions in the F_1-F_2 plane, but the experimental data were not very convincing (Kuwahara and Sakai, 1973). When complete CV syllables were presented, these were perceived only 42% correct; vowels were perceived 80% correct because of the existence of the, often misunderstood, preceding consonant. This vowel score increased further for two-syllable words and became 97% correct for the recognition of the vowels of the middle syllable in three-syllable words taken from connected speech. These results were supported by Strange *et al.* (1976), who also stressed the importance of dynamic acoustic information. However, the higher vowel scores for multisyllabic words have certainly also been influenced by the meaningfulness of these words.

Fujimura and Ochiai (1963) did a forced-choice identification study with 50-msec vowel segments gated out of Japanese words from various phonetic environments, see also Ochiai and Fujimura (1971). Many confusions were found; for instance, the /u/ from *yuyusi* was recognized as /i/. The implication of this study is according to Lindblom and Studdert-Kennedy (1967) that the assignment of symbols to vowels normally involves some sort of context-sensitive routine.

Lindblom and Studdert-Kennedy (1967) performed a detailed study on the perceptual boundary shift between /ʊ/ and /I/ in the contexts of /#V#/, /wVw/, and /jVj/. They used synthetic stimuli in which the vowel had 20 different possible positions between the locus positions for /ʊ/ and /I/ in the F_2-F_3 plane. F_1 and F_4 were fixed, as were the formant bandwidths. There was a considerable shift in boundary location for the /w/ context: in terms of F_2, about -185 Hz relative to the boundary for the neutral context. The average shift in boundary location for the /j/ context was in the other direction and amounted to about +75 Hz. The authors concluded that there was a tendency for perceptual vowel categorization to compensate for the formant frequency undershoot associated with vowel production. Not just the spectral pattern at the points of closest approach to target, but also the direction and rate of adjacent spectral transitions, determine the identity of a vowel sound in a certain context.

Kuwahara and Sakai (1976) found similar results for 26 synthetic vowels V positioned between the locus positions for /u/ and /e/, and presented in the context of /#V#/, /eVe/, and /uVu/. The location of the vowel boundary between /u/ and /e/ shifted towards the vowel areas of the surrounding vowels, compared with the boundary location for the isolated vowels. So here again perceptual vowel recognition compensates for the reduced vowel production; it is called the "complementary function" by the authors. In a subsequent fusion experiment with the same type of stimuli, it appeared that fusion of the middle vowels of two dichotically presented stimuli only occurred if the formant frequencies of the middle vowels were close to each other. Fusion did not occur for stimuli with different formant frequencies which had been identified as the same vowels. This shows, as could be expected, that the complementary function does not occur in the peripheral auditory system but that it more likely is a function of central processes.

Kanamori *et al.* (1971) describe the modification of the perceptual boundary location, caused by the vowel environment, in terms of a shift of the phoneme boundary in the "psychological auditory space" (PAS, see Kasuya *et al.*, 1971). Experiments were done with synthetic V_1bV_2 syllables, with V_1 and V_2 in the region of /i/, /e/, or /a/. The vowel boundary between /e/ and /a/ for V_2 shifted as a function of V_1, suggesting a contrast effect. The amount of shift of the perceptual vowel boundary was determined by the distance between the vowel boundary and the position of V_1 in the PAS. Very recently these authors suggested the same approach for the automatic recognition of vowels in connected speech (Kanamori and Kido, 1976).

The hypothesis that subjects change their categorization boundary between vowels in consonantal contexts to compensate for the undershoot of vowel target frequencies was not confirmed by a recent formant-matching experiment reported by Mermelstein (1975). His results suggest that, instead of indicating extrapolated target values, the matched formant values correspond to some appropriate time-average of the time-varying formant frequencies. However, the second-formant match also had to compensate for the higher formants. Furthermore, it is not quite clear that adjusting the first two formant frequencies of a synthesized vowel in such a way that it matches the colour of a vowel heard in a syllabic context, is similar to identifying that vowel. This task seems to be more of a psychophysical timbre-matching experiment, which says something about the peripheral signal processing only. Mermelstein (1976) in the meantime has found that for certain subjects his matching results are confirmed by categorization results. Other subjects appeared to behave more in line with the results reported by Lindblom and Studdert-Kennedy.

Lehiste and Shockey (1972) showed that for VCV words cut in two parts, the coarticulation in the remaining VC- or -CV part was not sufficient for subjects to identify what had been the deleted initial or final vowel. So, despite measurements (Öhman, 1966) which showed that formant transitions from the first vowel to the intervocalic consonant were influenced by the phonetic quality of the final vowel, these effects are not sufficient to have an influence on the perception of that (deleted) final vowel.

Benguerel and Adelman (1976), on the other hand, recently found that subjects could identify correctly the missing vowel well above chance in truncated segments. These segments were taken from utterances containing the consonant clusters /kstr/, /rstr/, or /rskr/ followed by one of the vowels /i/, /y/, or /u/. The positive results were achieved with segments including at least half of the final consonant of the cluster. The authors conclude that coarticulatory effects due to lip rounding and horizontal place of articulation provide perceivable information which may be used by the perceptual mechanism as an aid in speech-sound identification. However, they also state that "the fact that subjects can use subphonemic coarticulatory information to identify an upcoming vowel, does not mean that the perception process necessarily incorporates this ability". These cues can be redundant.

Consonant identification on the basis of the formant transitions in CV or VC syllables is possible, or at least the place of articulation of the consonant can be indicated, as was shown by Sharf and Hemeyer (1972). Labial, alveolar, and palatal stops and fricatives were used as consonants and the vowel was always the neutral vowel /ə/. The noise portions of the consonants were removed, thus leaving only the formant transitions. Identification of deleted voiced consonants in VC syllables was best. There was a significant advantage of VC transitions over CV transitions in consonant identification, suggesting a greater effect of forward coarticulation resulting in the assimilation of consonant features by vowels. This is in contrast with the locus concept in which the importance of the CV formant transition has been greatly stressed, see for instance Delattre *et al.* (1955).

In Chapter 4 we will describe our own experiments on the identification of vowel segments taken from monosyllabic words with different consonantal environments, without going into the details of a possible identification of the (deleted) consonants.

1.5.4. Concluding remarks

As we have seen in the preceding section, present knowledge of the perception of vowel sounds of which the spectral characteristics have been changed by surrounding consonants, is still rather vague. Neither is it very clear if, and in what way, coarticulatory information in the vowel is used for the recognition of preceding and/or following sounds. Instead of adding some new arguments to this discussion, we prefer to give in Chapter 4 a systematic description of the perceptual consequences in terms of the identification of different vowel-phoneme realizations in different consonantal environments. Perceptual consequences would be very difficult to measure if the vowels were presented in their original contexts as spoken by the subjects, since in that case errors in vowel identification would be marginal. Although we have warned, in section 1.5.1, against the use of experimental conditions which interfere with the stimuli, we saw no way around this danger ourselves. We used the method of having listeners identify vowel segments which were isolated from the original words. Identification errors can give information on the amount of consonant-specific variation in the vowel segments.

Discrimination tasks or similarity judgments would have been very difficult with those vowel segments since too many other variables, like duration and pitch contour, could have influenced the decisions of the subjects.

CHAPTER 2

METHOD

2.1. REPRESENTATION OF THE SPECTRAL INFORMATION OF VOWEL SOUNDS

As already briefly outlined in paragraphs 1.2 and 1.3, the common way to describe spectral differences between vowel sounds is in terms of the first two formant frequencies, sometimes supplemented with higher formant frequencies, formant levels, or formant bandwidths. Formants are here considered to be maxima in the envelope of the amplitude spectrum. Usually several maxima are detectable in a spectral analysis of a vowel, derived, for example, by means of a spectrograph or a wave analyzer. The formant representation is generally accepted and understood by speech scientists. The formant frequencies are a measure of the geometry of the vocal tract, and they can easily be used to control a formant synthesizer producing intelligible speech. A great deal of work has also been done on the dynamic variations of the formant positions, the so-called formant transitions. So, both from a historical and a practical point of view a formant representation seems to be highly recommendable.

Although Fant (1960, p. 26) states that the vowel spectrum is sufficiently specified by the "formant pattern", comprising about four formant frequencies, one may wonder whether spectral differences may not be represented in another, perhaps simpler and more general way.

There is a need for such a simpler approach because the (automatic) extraction of formant frequencies is notoriously difficult, especially for formants with highly asymmetric shapes, closely spaced formants, and when the voice fundamental frequency is high. For the following reasons one may wonder if it is at all necessary to determine the formant positions with great accuracy: (1) Slight variations in the fundamental frequency do not influence the formant position, although the harmonic distribution under the spectral envelope changes. It is

difficult to think of a mechanism that can take care of this. (2) In the reverberant conditions of everyday life there is an intrinsic variation of the amplitude spectrum caused by a point-to-point level variation, with a standard deviation of 5.7 dB of any harmonic of a complex tone (Plomp and Steeneken, 1973). (3) The just-noticeable perceptual difference for a single formant frequency of a vowel is three to five percent (Flanagan, 1957).

Although the formant representation is not strictly limited to vowel sounds, it is less suitable for most non-vowel sounds. Any audible sound entering the human ear has to be processed by the peripheral hearing mechanism before it leads to a sensation. The vowel sounds are just a subset of these sounds and only this group of sounds with peaked spectra gets a special treatment and is described in terms of formants. A more uniform representation would be attractive, especially in, for instance, automatic speech recognition and speech synthesis.

A spectral representation of vowel sounds should preferably meet the following requirements. It should be: (1) based on the peripheral processing of auditory stimuli; (2) easy to determine - if possible, automatically and in real time; (3) similar for all speech sounds and not specific for vowels only; (4) equivalent to the formant representation for vowels in terms of data representation and information content.

We believe we are able to present such an alternative approach. The basic principle of such a spectral representation was first outlined in Plomp, Pols, and van de Geer (1967). It involves two steps. The first one is a spectral analysis using a set of contiguous bandfilters with bandwidths similar to the bandwidths of the ear's auditory filters, called critical bandwidths (Plomp, 1964 and 1976). Although the auditory filters in the ear can be described as filters with running midfrequency, the variations in actual speech sounds as a function of frequency are such that overlapping filters are not necessary. Nevertheless, two institutes have built systems with about 80 overlapping critical bandwidth filters (Karnickaya *et al.*, 1975; Seidel, 1974). This greatly increases the amount of data without giving much new information. However, it must be said that Karnickaya *et al.* used the system for formant extraction, and that Seidel himself affirmed that half-overlapping filters are sufficient.

Perceptual experiments related to the general attributes of sounds, being loudness (Zwicker and Feldtkeller, 1967) and timbre (Plomp, 1976) also indicate that a very detailed spectral analysis, required for the determination of the formant parameters, is not necessary for explaining loudness and timbre differences between stimuli.

The second step in this spectral representation is an optimal data reduction

by means of a multivariate analysis of the bandfilter spectra. Just as the formant representation is a way of reducing spectral information which in principle is available in much more detail, for instance as a line spectrum, so the bandfilter spectrum can be represented with only a few parameters using the principal-components analysis. Any spectrum can be represented as a point in a multidimensional space. The coordinate values of such a point are equal to the, say, 17 levels of the bandfilter spectrum.

Principal-components analysis, being a type of factor analysis (Harman, 1967; Horst, 1965), is a mathematically well defined procedure to find a lower-dimensional subspace which explains, given the number of dimensions, as much of the original variation in the data points (or vowel spectra) as possible. Dimensionality reduction is possible because levels in neighbouring filterbands are not completely independent. In other words, not every theoretically possible spectrum actually occurs and therefore the 17-dimensional space is only partly filled with points. Furthermore, these points are concentrated in restricted regions. The principal-components analysis makes use of the variation,or more specifically variance, of these points along each of the original dimensions, and of the correlations between dimensions (expressed in a so-called variance-covariance matrix), in order to determine new dimensions. These new dimensions are linear combinations of the original dimensions and explain as much as possible of the original variance of the data points. The subspace is specified by a set of eigenvectors. The elements of these eigenvectors are called direction cosines; they are equal to the cosines of the angles between the new dimensions and the successive original dimensions. The number of dimensions of the subspace is a matter of choice. Sometimes there are good reasons to choose a two-dimensional representation. For display purposes this is very attractive. However, in most cases not all the relevant information in the speech spectra will be described by two dimensions. For instance, in automatic speech recognition three to five dimensions will probably be preferable, depending upon the information content of such a representation.

In an actual word-recognition experiment we opted for a three-dimensional representation (Pols, 1971). If dimensionality reduction is used as a tool for bit-rate reduction in speech communication, the available bandwidth and the minimally required intelligibility determine the dimensionality. Usually a compromise has to be chosen between the gain of information achieved by taking into account more dimensions and the increase in computation time and complexity involved in processing higher dimensional data. In speech-perception experiments the number of dimensions can be an experimental parameter. Further details of

dimensionality reduction will be discussed in paragraph 2.3.

The procedure takes into account the statistical variation in the data points and is not based on some predefined model like the formant representation. However, since the formant representation is also an optimal representation in many respects, one can imagine that, at least for vowel sounds, there must be a high correlation between both representations.

In the study by Plomp *et al.* (1967) there was only a rough indication of this correlation based on the representation of the average vowel positions in the two-dimensional principal-components representation. In Klein *et al.* (1970) the so-called canonical matching procedure (Cliff, 1966) was applied to calculate this correlation on the basis of actual data. However, formant frequencies were only roughly determined, based upon maxima in the average one-third octave vowel spectra. Much better data for comparison were obtained by Pols *et al.* (1973) and van Nierop *et al.* (1973) through actual formant measurements on all Dutch vowels of 50 male, and 25 female speakers, respectively. Correlation coefficients as high as 0.99 were achieved between the coordinate values of the average vowel points in the log F_1 - log F_2 representation and a two-dimensional representation of the same vowel segments based on bandfilter spectra. Also, vowel identification results based on maximum-likelihood regions in both representations were very similar.

These encouraging results were obtained by directly using the filter levels as basic parameters. In the system described by Karnickaya *et al.* (1975), the filter levels are modified in various ways: by comparison with a threshold level or a noise floor (Klatt, 1976) per filter, by non-linear transformation, and by the introduction of lateral suppression. This greatly complicates the processing of the data but does not necessarily result in a more representative spectral representation. With respect to the introduction of lateral suppression we have tested this in a pilot experiment. Lateral suppression was known for some time from electrophysiological experiments. Not untill Houtgast (1974a) introduced the condition of non-simultaneous stimulation in the so-called pulsation threshold paradigm, could lateral suppression be demonstrated also psychophysically. Houtgast (1974b) also showed the possible implications in terms of spectral sharpening of the vowel spectra. However, in some pilot experiments we found that for stimuli with another harmonic structure and lower fundamental frequency than Houtgast had used, the pulsation method did *not* give realistic results for the "internal spectral representation" of the stimuli. This is probably caused by the high crest factors of those signals when they are internally filtered. For the time being we do not feel the need to include the concept of lateral

suppression in every day speech perception, and thus use the bandfilter levels themselves in the 17-dimensional representation.

In order to demonstrate that the spectral data representation, based on a principal-components analysis of bandfilter spectra, is certainly not limited to vowel sounds, we also applied this approach successfully to the analysis and recognition of non-vowel sounds and words (Pols, 1971a, 1971b, 1972).

2.2. TECHNICAL DESCRIPTION OF THE ANALYSIS SYSTEM

As the analysis system in its present form has not been fully described in an earlier publication, we will give a short description of it below. Fig. 2.2.1 represents a block diagram of the analysis system. The speech signal from a tape-recorder forms the input to a parallel set of bandfilters. The 14 filters with midfrequencies from 400 to 8000 Hz are one-third octave filters. This bandwidth comes close to the critical bandwidth. To simulate the ear's constant critical bandwidth for low frequencies, three filters of about 90 Hz bandwidth and centre frequencies of 122, 215, and 307 Hz were used. To these 17 filters is added a lin-C filter (-3 dB points at 32 Hz and 8000 Hz) to measure the overall level. The parallel filter outputs are followed by integrator circuits. The integration is in fact an envelope peak detection after a one-way rectification, reset after sampling. The dynamic range of the peak detectors does not allow processing of the linear filter output. Therefore, the peak detectors are preceded by logarithmic amplifiers. A reliable 60-dB dynamic range could be achieved by putting certain temperature-sensitive components into an oven with constant temperature.

Because of the asymmetry of the speech waveform we added an inverter to the broad-band filter, which makes it possible to measure both the positive and negative overall level. In practice, the higher of these two levels is used as a measure of the overall level.

The 17 filters plus the two broad-band filters together make 19 channels; a twentieth channel was added in the form of a zero-crossings counter. This device measures the number of positive-going zero crossings in the original speech waveform during 10 msec. All the outputs of these 20 channels are sampled every 10 msec through a 20-channel multiplexer. This information is fed through a 10-bit analog-to-digital converter into a digital computer.

The 10-msec sampling time was chosen as a good compromise between following rapid envelope changes, and avoiding variations within a single pitch period. The filters used have frequency-dependent time constants because of their dif-

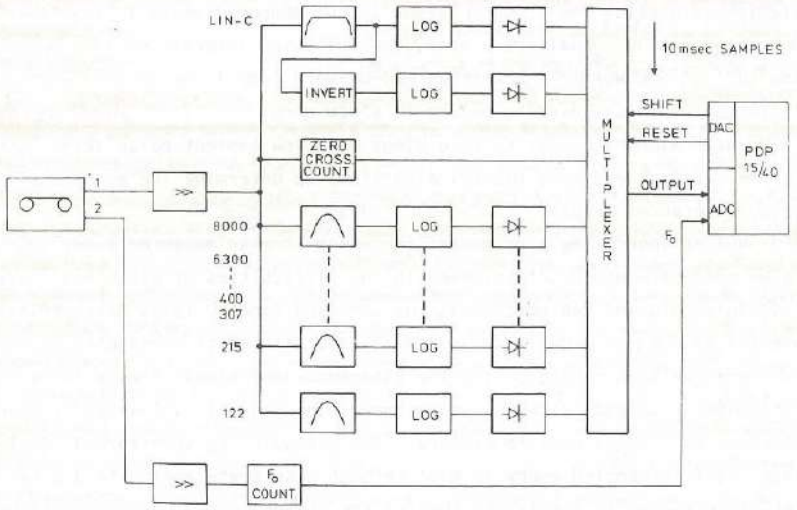


Fig. 2.2.1. Block diagram of the analysis system.

ferent bandwidths. The intrinsic time delay between maximum input and maximum output is only 0.4 msec for the 8000-Hz filter, but almost 10 msec for the 122-Hz filter. In order to compensate for this delay, the 122-Hz filter is sampled and reset almost 9.5 msec later than the 8000-Hz filter, and the other filters are reset accordingly. The shift pulses for the multiplexer and the reset pulses are given through digital-to-analog converters.

With this system the spectral analysis of a word of say 0.5 sec duration results in fifty 10-msec samples. Every sample is a series of 20 numbers being the positive and negative overall level, the number of zero crossings, and the levels in the 17 filters. The filter levels are given in steps of 0.2 dB. These values are relative and defined in such a way that a fixed calibration signal results in a level of 60 dB. The dynamic range is also at least 60 dB. The accuracy and test-to-test reliability of the measurements is approximately 1 dB, but for the lower levels of 10-15 dB it is somewhat worse.

The output of the broad-band filter is used as a signal/no signal detector. As long as the higher overall level is below a predetermined trigger level, the 17 filters are not sampled. During that time the broad-band filter is sampled every 1 msec. Once the level is above the trigger value, and remains there for at least two successive msec, indicating that it was no interfering pulse, frequency analysis starts. The analysis stops when the overall level of a 10-msec

sample is again below the trigger value. If the measured pause is less than a predetermined value, usually 250 msec, the following information is supposed to belong to the same word. Otherwise an end-of-word code is generated, and following information is supposed to belong to a new word.

We always used high-quality recordings with low ambient-noise levels and therefore did not need more complex algorithms to determine the endpoints of isolated utterances (Rabiner and Sambur, 1975).

For voiced speech the fundamental frequency is also measured every 10 msec. Since we do not want to get involved in the difficulties of pitch extraction (Noll, 1967; Gold and Rabiner, 1969; Fujisaki and Tanabe, 1972; Markel, 1972; McGonegal *et al.*, 1975), we have chosen a practical solution: together with the audio signal we also record during the utterances the signal from a throat microphone onto the second track of the recorder. This signal is low-pass filtered, and the periodicity is easy to measure. This periodicity, transformed into a DC-voltage, is also sampled every 10 msec through a separate analog-to-digital converter. This system is similar to that of the laryngograph (Fourcin and Abberton, 1971), the electroglottograph, or the accelerometer (Stevens *et al.*, 1975).

The great advantage of this analysis system is that it can be used in real time. Furthermore, the data from the analysis are directly available in numerical form for further processing in the computer.

This analysis system appears to be an efficient method for spectral analysis of speech, not just for vowel sounds or sonorants, but for all speech sounds. It is particularly suitable for the spectral analysis of sounds varying over time as in vowel coarticulation, topic of the present research.

2.3. DIMENSIONALITY REDUCTION

In this paragraph we focus our attention on the procedures for dimensionality reduction already broadly discussed above.

In the first section we will specify the concept of variance and the dimensionality reduction procedure itself, illustrated with an example. In the second section we will discuss what this means for the processing of specific data sets, and how the resulting data representation can be interpreted, again illustrated with an example.

2.3.1. Concept of variance and dimensionality reduction procedure

In data-reduction procedures the concept of *variance* is very important. The variance v of n numbers x_i (i=1, n) is equal to

$$v = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

in which

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The spread σ is equal to the square root of the variance

$$\sigma = \sqrt{v}.$$

The standard deviation s is defined as

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \sigma \sqrt{\frac{n}{n-1}}.$$

For large n standard deviation s and spread σ are equal. In all our calculations we prefer to work with the variation in the actual data, and therefore use variance v and spread σ.

By using the subscript j for the j-th dimension (j=1, m, typically m=17), x_{ij} becomes the coordinate value of the i-th point in the j-th dimension, which in our case is the level of the i-th 10-msec speech sample in the j-th filter. Thus any bandfilter spectrum can be represented as a point in the m-dimensional spectral space, having coordinate values x_{ij} equal to the levels in the m bandfilters. In order to make the dimensional spectral representation independent of absolute sound level a simple level normalization was introduced by using the levels relative to overall level per sample as input data. If, for a certain application, one wants to preserve the overall level an extra dimension can be used. The variance v_j along the j-th dimension for all n points x_{ij} becomes:

$$v_j = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2,$$

and the total variance of all n points in the m-dimensional space is

$$v = \sum_{j=1}^{m} v_j = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2. \qquad [2.1]$$

Eq. [2.1] shows that the total variance is equal to the sum of the variances for each dimension and that, in terms of a geometric model, the total variance v of the n points is equal to the sum of the squares, divided by n, of the distances between the individual points x_{ij} to their "centre of gravity", $\bar{x}_j$.

The total variance in m dimensions is a good measure for the total variation in the data points. The variance per dimension represents the fraction of the total variance accounted for by that dimension. However, in a multidimensional space the average value, $\bar{x}_j$, and the variance per dimension, v_j, do not fully describe the data, the correlation of the data points along the dimensions is also important. This correlation is expressed by the covariance v_{jk}:

$$v_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k). \qquad j \text{ and } k = 1, m \qquad [2.2]$$

The variance, v_j, is then equal to v_{jk} for k=j.

All v_{jk}'s together represent a variance-covariance matrix. In fact it is more convenient to use a matrix notation. All n data points in m dimensions x_{ij} can be described in an (n x m) matrix X. The (m x m) variance-covariance matrix V is

$$V = \frac{1}{n} X'X.$$

The matrix X in fact has to represent the data points as deviations from the average value per dimension.

For our data reduction procedure we have mainly applied the principal-components analysis on the (variance-)covariance matrix. Principal-components analysis is one of the techniques of factor analysis (Horst, 1965; Harman, 1967; van de Geer, 1967; Anderson, 1958); it is also known as dispersion analysis (Li *et al.*, 1968) and Karhunen-Loève expansion (Watanabe, 1965).

This procedure yields new dimensions, or factors, which successively explain the greatest amount of variance left in the data. Each new dimension is a linear combination of the original dimensions and is specified by an eigenvector e_j. The explained variance per new dimension is equal to the eigenvalue λ_j. This can be expressed in the following way

$$VE = VD,$$

where E is an (m x m) matrix of which the columns are the m orthogonal eigenvectors e_j, and D is a diagonal matrix with the eigenvalues λ_j as the diagonal elements. The sum of all eigenvalues is equal to the total variance.

If for the original m-dimensional data all m eigenvectors were determined, all data points could be described in a new m-dimensional space now defined by the m eigenvectors. This new data representation is essentially the same as the original representation, apart from rotation of the axes. The specific result of this rotation is the fact that the data points now have their largest variance along the first new dimension, the second largest variance along the second independent new dimension, etc. Owing to correlations in the original data, the explained variance in the higher new dimensions will be very small. In terms of data reduction it is very interesting to work with just a few new dimensions which nevertheless explain most of the variance in the data.

In actual data the original dimensions are mostly not independent and the variance per dimension is not very high. However, the new dimensions or factors, defined by the orthogonal eigenvectors, are independent and the first few new dimensions explain much more of the total variance than any of the original dimensions. We will illustrate this with an example from one of the earlier studies (Klein *et al.*, 1970).

The level-normalized spectra of 12 Dutch vowels spoken in a context of h(vowel)t by 50 male speakers were measured with 18 one-third octave filters

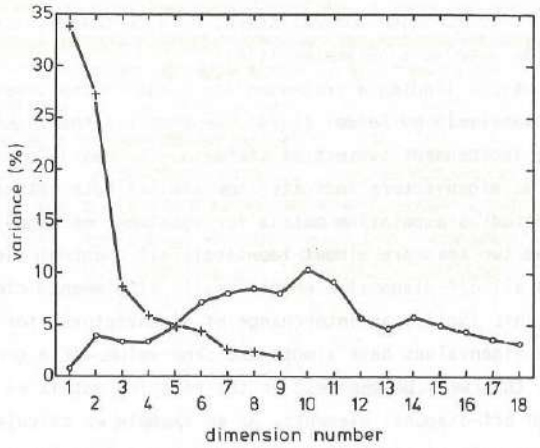


Fig. 2.3.1. Percentage of variance explained by each of the 18 original dimensions, and by the first nine new factors. (Adapted from Klein, Plomp, and Pols, 1970.)

with midfrequencies from 100 to 10,000 Hz. This resulted in a dimensional spectral representation of 600 points in an 18-dimensional space. The variance, as a percentage of the total variance (= 1182 dB^2), is given for each of the original 18 dimensions in Fig. 2.3.1. None of the dimensions explained more than 10.3% of the total variance. On the covariance matrix of these data a principal-components analysis was performed. The first four new dimensions, or factors, explained 33.7%, 27.2%, 8.7% and 5.8% of the total variance, respectively. After the ninth factor the extraction of new eigenvectors was stopped because only 8.3% of the total variance remained unexplained, and it was reasonable to assume that the higher factors represented nothing but noise in the data.

That only a few dimensions are important is even more evident when not the individual spectra, but the average vowel spectra are processed. Then the total variance is 715 dB^2, and the first four factors already explain 97.8% of the total variance. Thus we see that most of the spectral variation in the original data, distributed over many dimensions, can be expressed as variance along only a few new factors.

2.3.2. Processing of actual data sets

In the present procedure the eigenvectors are determined one after another and the extraction of new eigenvectors can be stopped according to different criteria. The number of new dimensions finally used, is determined by a sudden decrease in percentage of explained variance, or by the total amount of explained variance, or a percentage correct score, or a certain synthetic speech quality, or it can be used as a parameter itself.

An interesting way of finding a criterion for defining the dimensionality of the data is used intensively by Seidel (1974). He compares the eigenvalues and eigenvectors of two independent subsets of the data. The matrix of inner products of both sets of eigenvectors indicates how similar both sets are. This matrix can be interpreted as a rotation matrix for rotating one eigenvector space to the other. If the two sets are almost identical, all diagonal elements will be close to 1.0 and all off-diagonal elements small. If elements close to the diagonal are high, this implies an interchange of eigenvectors; for instance when two successive eigenvalues have almost the same value. If a group of eigenvectors is unstable this will be manifest in the rotation matrix as an area with many relatively high off-diagonal elements. As an example we calculated the rotation matrix for two sets of eight eigenvectors derived from two subsets of the speaker-normalized vowel data of 50 speakers (Klein *et al.*, 1970). The subdivision was achieved by considering the first 25 and the second 25 speakers separately. The rotation matrix is given in Table 2.3.1 and suggests that three common dimensions are a good choice.

		second 25							
		1	2	3	4	5	6	7	8
first 25	1	.99	-.10	.03	.19	-.16	.03	.06	-.03
	2	-.09	.72	.22	-.16	-.12	-.13	.01	.18
	3	-.03	.26	.94	-.21	-.23	-.03	-.02	.05
	4	.10	.02	.19	.38	.08	.04	-.11	.12
	5	-.00	-.04	-.12	-.13	.17	.52	-.22	.06
	6	-.05	.11	.14	-.19	.14	-.08	-.20	.14
	7	.02	.06	.07	-.09	.10	.06	.55	-.17
	8	.04	.27	.31	-.20	-.06	.03	.35	.10

Table 2.3.1. Rotation matrix of two sets of eight eigenvectors derived from two independent subsets of the speaker-normalized vowel data of 50 male speakers. The solid lines enclose values above 0.5, and the dashed lines values between 0.5 and |0.2|.

For every new data set new eigenvectors can be determined, or one could work with one general eigenvector base for all speech spectra. Throughout our research we have always preferred to work with specific eigenvector bases for specific data sets. In the development phase this is more attractive and leads to the best results. However, it means that each time a new covariance matrix has to be determined. Especially for the analysis of long speech utterances this appeared to be very time-consuming because, according to Eq. [2.2], first the 17-dimensional spectra of all 10-msec samples have to be stored. Subsequently, the average values have to be computed and subtracted from all these spectra. Thus only after completion of the analysis can the covariance be determined. However, the developed software now makes it possible to compute the covariance matrix in real time by updating with every new sample x_{ij} a matrix with elements c_{jk} and a vector with elements a_j:

$$c_{jk}^{i} = c_{jk}^{i-1} + x_{ij}x_{ik}, \qquad j \text{ and } k = 1,17$$

$$a_j^{i} = a_j^{i-1} + x_{ij}.$$

After completion of the analysis the actual elements of the covariance matrix are directly found by

$$v_{jk} = \frac{1}{n} c_{jk} - \bar{x}_j\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \bar{x}_j\bar{x}_k, \qquad [2.3]$$

where

$$\bar{x}_j = \frac{1}{n} a_j.$$

It can easily be shown that the elements, v_{jk}, determined in this way are equal to the earlier defined Eq. [2.2]:

$$v_{jk} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij}-\bar{x}_j)(x_{ik}-x_k), \qquad [2.2]$$

since

$$v_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \frac{1}{n}\sum_{i=1}^{n} \bar{x}_j x_{ik} - \frac{1}{n}\sum_{i=1}^{n} \bar{x}_k x_{ij} + \frac{1}{n}\sum_{i=1}^{n} \bar{x}_j\bar{x}_k$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \frac{\bar{x}_j}{n}\sum_{i=1}^{n} x_{ik} - \frac{\bar{x}_k}{n}\sum_{i=1}^{n} x_{ij} + \frac{1}{n}(n\,\bar{x}_j\bar{x}_k)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \frac{\bar{x}_j}{n}(n\,\bar{x}_k) - \frac{\bar{x}_k}{n}(n\,\bar{x}_j) + \bar{x}_j\bar{x}_k$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \bar{x}_j\bar{x}_k.$$

Using the eigenvectors e_j, with elements e_{jk}, we can easily compute for a specific data set the coordinate values y_{ik} in a lower-dimensional factor space from every level-normalized vowel spectrum x_{ij}.

$$y_{ik} = \sum_{j=1}^{17} x_{ij}e_{jk}, \qquad k=1,m \quad m\leq 17.$$

As an example we give in Fig. 2.3.2 the average vowel positions for one speaker in a I-II factor plane (m=2). The details of this analysis will be described in Chapter 3, but for the moment it is sufficient to know that each vowel point is the average position of a number of vowels in different contexts. After some rotation, this diagram can be easily recognized as the well-known vowel triangle or formant representation. This relationship is studied in more detail in Pols *et al.* (1973) and van Nierop *et al.* (1973).

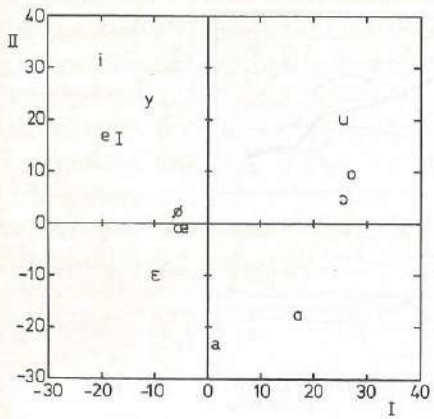


Fig. 2.3.2. Positions of the vowels of one speaker in the I-II factor plane. The plane is defined by the eigenvectors given in Fig. 2.3.3. Each vowel point is the average position of a number of the vowel samples in different consonant contexts.

The configuration of the points in Fig. 2.3.2 is such that the overall centre of gravity coincides with the origin of the coordinate system. We call data corrected in this way centred data. For the individual data points this means a speaker-dependent translation for all vowels of one speaker. In practice it means that from the individual level-normalized spectra the average spectrum per speaker is subtracted. This appears to be a very efficient form of speaker normalization, see also section 1.3.2.

This mathematically well-defined factor representation is not as easy to interpret as a formant representation. The eigenvectors defining the plane of Fig. 2.3.2 are given in Fig. 2.3.3. The direction cosines of these eigenvectors

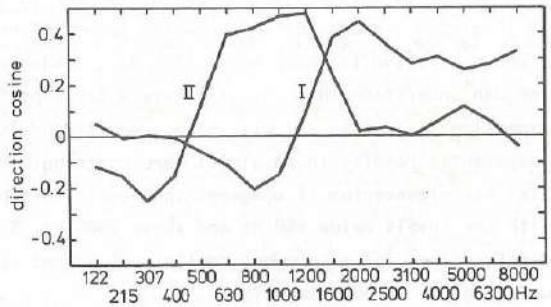


Fig. 2.3.3. Direction cosines of the first two eigenvectors defining the I-II factor plane. Data were 10,279 10-msec sample points of vowel segments from three male speakers. These first two new dimensions explained 37.1% and 30.1% of the total variance, respectively.

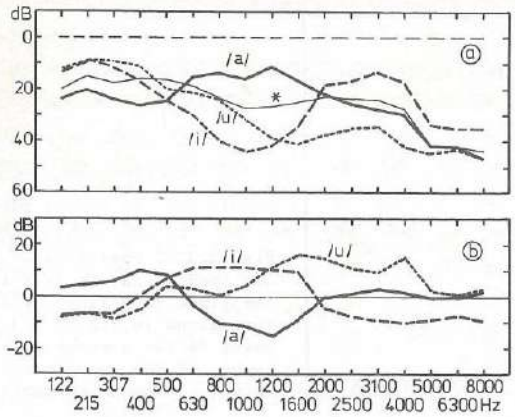


Fig. 2.3.4. Average one-third octave spectra for the three vowels /i/, /u/, and /a/ of speaker 1. The curve indicated with an asterisk represents the spectrum averaged over all vowels of that speaker. The filter levels are relative to overall level. The lower diagram gives speaker-normalized, or centred spectra, as deviations from the average vowel spectrum.

are factors weighting the contributions of the levels in the original filter bands to the final coordinate value along one of the new dimensions. Fig. 2.3.4a represents the average spectra for three vowels /i/. /u/, and /a/, as well as the spectrum averaged over all vowels. The filter levels are relative to the overall level. Fig. 2.3.4b gives the speaker-normalized or centred spectra, relative to the average vowel spectrum. Using the eigenvectors I and II, these spectra are represented as the corresponding vowel points in the I-II factor plane, see Fig. 2.3.2. Fig. 2.3.3 shows that eigenvector I can be interpreted as weighting the levels above relative to those below 1100 Hz. Looking at the spectra in Fig. 2.3.4b one can understand that /u/ will have a large positive coordinate value along dimension I, just as /i/ will have a negative value, whereas the spectrum of /a/ apparently results in an almost zero contribution along dimension I. In a similar way, eigenvector II compares the levels in a band between 450 Hz and 2000 Hz with the levels below 450 Hz and above 2000 Hz. So /i/ gets a large positive coordinate value, /u/ a somewhat smaller value, and /a/ a negative coordinate value along the second dimension.

For the third and higher dimensions the eigenvectors are increasingly complex; this makes it more difficult to interpret them directly. Fig. 2.3.5 gives the third, fourth, and fifth eigenvectors for the spectra of the vowel segments described in more detail in Chapter 3. We see that for these eigenvectors the

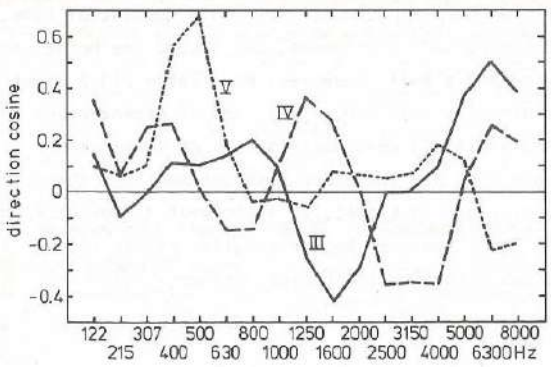


Fig. 2.3.5. Direction cosines of the third, fourth, and fifth eigenvectors, explaining 11.4%, 6.0% and 4.5% of the total variance, respectively. Data were the same as for the first two eigenvectors represented in Fig. 2.3.3.

spectral weightings become more and more specific. However, these higher dimensions become less important because they explain only small amounts of variance in the data, see Table 2.3.2. There is a tendency in the eigenvectors to have more oscillations as their number gets higher. This suggests some Fourier-like decomposition of the bandfilter spectra. A theoretical set of orthonormal eigen-

	calculated eigenvectors			theoretical eigenvectors		
	absolute	%	cum. %	absolute	%	cum. %
1	382.8 dB^2	37.1	37.1	222.4 dB^2	21.5	21.5
2	310.6	30.1	67.2	219.5	21.3	42.8
3	117.2	11.4	78.5	130.5	12.6	55.4
4	62.1	6.0	84.5	160.5	15.5	71.0
5	46.3	4.5	89.0	41.9	4.1	75.0
6	29.1	2.8	91.8	48.0	4.6	79.7
7	18.5	1.8	93.6	23.0	2.2	81.9
TOTAL	1032.5 dB^2	N = 10,279				

Table 2.3.2. Explained variance per new dimension (in absolute value, in percentage, and in cumulative percentage), both for the calculated eigenvectors, as well as for the theoretical set of sinewave eigenvectors. Data consisted of 10,279 10-msec sample points of vowel segments from three male speakers.

vectors, based on such a Fourier decomposition could consist of sine functions, the first eigenvector being half a sinewave, the second one being one sinewave, the third one being one and a half sinewaves, etc. Table 2.3.2 gives the percentages of variance explained by this theoretical set of eigenvectors for the same data. We see that they explain a reasonable amount of variance but their decline is less progressive and they are therefore less optimal than the calculated eigenvectors. It is quite possible that, on a somewhat different frequency scale, the sinefunctions will give somewhat better results. Yilmaz (1967, 1972) postulated this type of eigenvectors on theoretical ground.

2.4. SPECTRAL REGENERATION

We have shown that vowel spectra like those of Fig. 2.3.4b can be represented as points in a plane, see Fig. 2.3.2. This process of going from a 17-dimensional spectrum to a two-dimensional point representation is not simply reversible if nothing but the two-dimensional information is left. Nevertheless, there is a method to regenerate the original spectrum to some degree, starting from the coordinate values in two dimensions. One then uses as (unknown) coordinate values in dimensions 3 to 17 the average coordinate values along those dimensions. The smaller the variation in these higher dimensions for the original data, the smaller the error relative to the original spectrum. If we work with centred data it is not even necessary to know the eigenvectors 3 to 17, because the average coordinate values are then equal to zero. In vector notation the procedure is as follows: The i-th point in 17 dimensions x_{ij}, j=1,17 can be transformed into m dimensions (m≤17) by using the eigenvectors with elements e_{jk}

$$y_{ik} = \sum_{j=1}^{17} x_{ij} e_{jk}. \qquad k=1, m$$

The m-dimensional data y_{ik}, k=1, m can then be used to reconstitute 17-dimensional spectra z_{ij}

$$z_{ij} = \sum_{k=1}^{m} y_{ik} e_{jk} + \sum_{k=m+1}^{17} \bar{y}_k e_{jk}. \qquad [2.4]$$

If we subtract the overall average $\bar{y}_k$ from all y_{ik}, in other words: use centred data, then the second term in [2.4] becomes zero, which reduces computation, and makes it unnecessary to calculate the eigenvectors for k>m. Afterwards, the

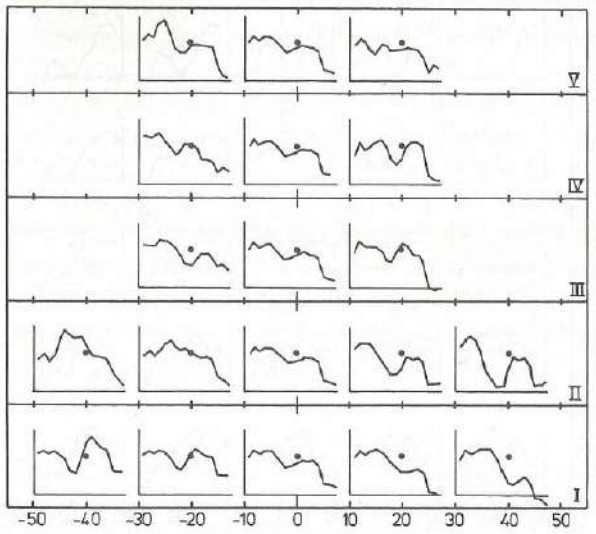


Fig. 2.4.1. Spectral variation represented by points in the subspace, of which the coordinate values vary only along one of the indicated dimensions I to V over a range representative of actual speech data. The coordinate values along all other dimensions are kept equal to the average.

overall average spectrum in 17 dimensions $\bar{x}_j$ has to be added to z_{ij} to obtain again filter levels relative to the overall level. The difference between this final spectrum and the original spectrum x_{ij} is a measure for the information loss caused by applying m instead of 17 dimensions.

In the next paragraph we will show how this procedure is used for resynthesis of speech using spectral information in less than 17 dimensions. In Fig. 2.4.1 the regeneration procedure is used to illustrate in successive rows what the variation in coordinate value along each of the first five new dimensions means in terms of renegerated spectra. The coordinate values along all other dimensions are then kept equal to the average. The coordinate value 0 always represents the average vowel spectrum. For example, the coordinate value along dimension I varies for actual data roughly from -40 to 40. A coordinate value of -40 along this dimension represents a spectrum as given in the lower left-hand corner of Fig. 2.4.1, whereas a value of +40 represents a spectrum as given in the lower right-hand corner. The variation in spectra along this first dimension reflects the weighting of eigenvector I as represented in Fig. 2.3.3. Fig. 2.4.2 shows what the spectral variation is if the first *two* coordinate values are sys-

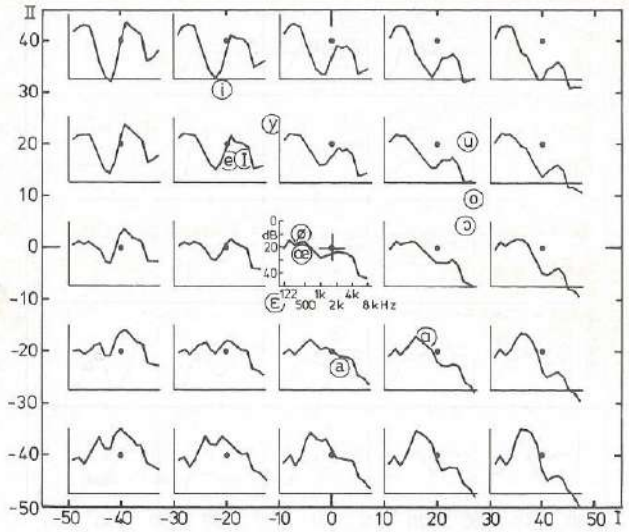


Fig. 2.4.2. Similar to Fig. 2.4.1, but now the coordinate values along the first two dimensions are systematically varied. The corresponding spectral variation in the I-II plane is shown. The middle horizontal section is equal to the bottom row in Fig. 2.4.1, and the middle vertical section is equal to the bottom row but one in Fig. 2.4.1. For reference the average vowel positions of Fig. 2.3.2 are also drawn.

tematically varied. For reference the average vowel positions of Fig. 2.3.2 are also drawn.

The dimensional spectral representation as described above was also used by Kulya (1964), Li *et al.* (1968, 1969), Boehm and Wright (1968), Seidel and Paulus (1971), and Wright (1972).

Li *et al.* (1969) preferred to work with the original spectra, without level normalization. The percentage of variance explained by their first eigenvector is very high, but the total variance in the data is much larger. This first eigenvector has almost equal weights for all filters, and consequently deals with the over-all level variations in the data. The second and third eigenvectors for the original data are similar to the first and second eigenvectors for the level-normalized data (Pols, 1970).

In view of the aim of the present investigation only vowel spectra are considered but it will be clear that the dimensional spectral approach is much more generally applicable. If not just single vowel sounds are analyzed, but words, or running speech, the first eigenvector always appears to represent a weighting

of low-frequency levels relative to high-frequency levels (Li *et al.*, 1969; Pols, 1971, 1972, 1974). In combination with the number of zero crossings, this eigenvector could be used very efficiently to separate sonorant and non-sonorant speech samples (Pols, 1974; Weinstein *et al.*, 1974). Separate eigenvector bases could be determined for these two groups of speech sounds, which is interesting both for segmentation and labelling (Pols, 1972), and for speech synthesis (Pols, 1974). Zurcher *et al.* (1976) even extended this to a further subdivision into eight two-dimensional subspaces.

2.5. SPEECH SYNTHESIS

Additionally to analyzing and processing speech data, it is important also to *listen* to these signals, to *hear* what the actual effect of, for instance, dimensionality reduction is. With a speech synthesis system it would furthermore be possible to listen to a specific word or to any part segmented out of it. The effect of adding or omitting small portions is then also under auditory control. Furthermore, it is useful to have a more flexible system for generating stimuli for intelligibility measurements under different conditions as well as for perceptual experiments. If unprocessed natural speech cannot be used directly, stimulus compilation is always very time consuming, because it has to be done by means of gating, or tape splicing, or waveform manipulation. Another reason for developing a speech synthesis system was the interesting question if it is possible, in combination with the speech analysis system, and on the basis of the dimensional spectral representation of speech, to develop a speech communication system with a low bit-rate.

It will be clear from the above remarks that a synthesis system is wanted which can be controlled by the output parameters of the speech analysis system. This brings us to a channel-vocoder-type system by adding a synthesis system to the analysis system, with the computer acting as intermediate storage or, in communication terms, as the transmission channel.

2.5.1. Technical description of the synthesis system

The core of the synthesizer is a parallel set of bandfilters identical to the analyzing filters. This means that we have 17 filters with midfrequencies ranging from 122 up to 8000 Hz with parallel inputs and parallel outputs. From 400 Hz on, the filters are 1/3-octave filters, the three lowest filters with midfrequencies of 122, 215, and 307 Hz, have a fixed bandwidth of about 90 Hz. The ordinary way to get voiced speech is to excite all filters with the same

periodic source with a constant average power but with a periodicity variable in time. For unvoiced sounds a noise generator is used as a source signal. The spectral variation in frequency and time is achieved by modulating the output of each filter separately. The control parameters for these modulators have to be delivered by the analysis system. Summation of the outputs of the modulators results in reconstructed speech (Flanagan, 1972).

Our synthesis system deviates at several points from this general concept. The computer is not just an intermediate storage of the synthesis parameters, it is also an active controller of the synthesizer itself. It is of course possible, through the analog outputs of a computer, to control the modulators and the frequency of the voiced source. However, our computer did not have enough digital-to-analog converters. Furthermore, inexpensive good quality modulators with a high dynamic range are not available. But most important of all, we wanted to have a digitally controlled synthesizer which did not always occupy the computer completely.

This finally resulted in a concept in which each filter was excited by means

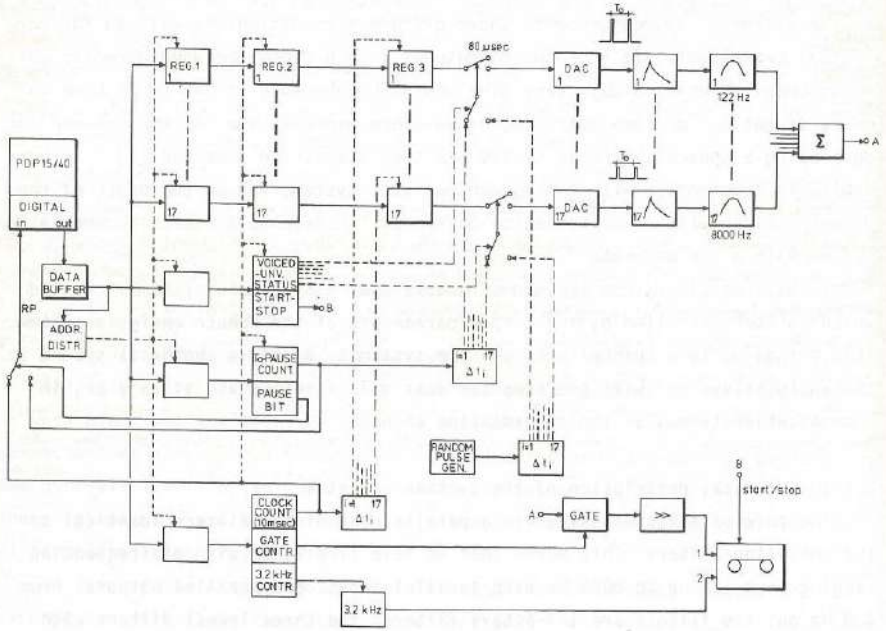


Fig. 2.5.1. Block diagram of the synthesis system. For an explanation see text.

of a separate source. Each source signal is in principle a pulse train of which amplitude and inter-pulse distance can be controlled. By changing the distance between the pulses, the periodicity varies, which is necessary for voiced signals. Poisson pulses are generated for unvoiced signals. The amplitude of the incoming pulse train is controlled separately for each filter, which results in variation of the spectrum in frequency and time. Finally all filter outputs are summated to give synthesized speech. In fact the pulse trains are sent through integrators to get a sawtooth-like signal more representative of the glottal source. For a block diagram of the synthesis system see Fig. 2.5.1.

In the upper part of that diagram we see the data and signal flow for the 17 channels. In the lower half of the diagram the different control functions are represented.

The pulse levels per filter are stored in three successive digital registers. By opening a gate during 80 µsec, followed by a digital-to-analog conversion, the pulses are produced. The moments of opening the gate are controlled by the F_0 information. For unvoiced sounds a switch is set such that the gate is controlled by a noise source. Each filter channel has such a separate switch and gate, both are logical AND and OR circuits. Since all information necessary for synthesis is stored in digital registers, the whole system can be digitally controlled.

Every 10 msec the synthesizer, as a peripheral apparatus, sends a "request pulse" (RP) to the computer for new information. In less than one msec all information of the 10-msec sample concerned is then sent to the digital registers of the synthesizer through the digital output of the computer.

This digital input/output system, developed and built in our institute, is based on 24-bit words, which for this application contain in 4 bits the address code and in the remaining 20 bits the actual information. *Nine* words are used to transmit the level information for the 17 filters. A single word holds the 10-bit information for two filters. A *tenth* word is used to give the status of each channel. This status bit indicates whether the channel concerned has to get voiced or unvoiced input. One of the remaining bits in this tenth word is used to indicate whether the built-in delay (Δt_i) for controlling the gates for the different channels i=1,17 has to be used or not. Another bit actually starts or stops the synthesizer and controls the pause function of a recorder. For voiced sounds the *eleventh* word contains the pitch information. When there is a pause within a word or between words, the duration of this pause is given instead in this word. One bit of the tenth word tells us in which of both possible ways (pitch or pause) this information has to be interpreted. Seven bits of the

twelfth word are used to set the clock interval at a value between 1 and 128 msec. Normally this interval is fixed at 10 msec, but for specific applications like lengthening or shortening (part of) an utterance without spectral distortion, another value can be chosen. Furthermore, one bit in this twelfth word is used to indicate if an overall gate has to be opened or closed. Another bit starts or stops the generation of a 3.2 kHz pure tone, available on a separate output of the synthesizer. This signal can be recorded on the second channel of a recorder, in parallel and timed with the onset and/or offset of the generated words. This is useful for perceptual experiments where gates, warning lights, or anything else has to be controlled synchronously with the words.

With every new clock pulse, normally every 10 msec, a request for data is sent to the computer. These data may have been read directly from disk, or generated or modified under program control. There is ample time for these operations since the actual data transfer through the digital output to the synthesizer takes only about one msec.

The address labels attached to the *twelve* 24-bit words are used to distribute the information over the different registers. At the same time the present content of these registers is transferred to the next layer of registers where it becomes available for the actual synthesis. In fact there is a third layer of registers where the same information as in the second layer is available but, if wanted, with a delay Δt_i that varies for the different filters in relation to their bandwidths. The delays, created with a multiple-output shift register with minimal steps of 100 µsec, vary from almost 10 msec for the highest filters to no delay for the lowest filters.

Asynchronously with the 10-msec clock, the level values in the third layer of registers are sent to the 10-bit digital-to-analog converters through gates which are open for 80 µsec. The moment of opening is determined by the periodicity of the sample concerned. This periodicity value is available in the t_0-register, which sets a t_0-counter. The step-size of this counter is 16 µsec. Every t_0 msec a pulse is generated which opens, for 80 µsec, the gates between the third layer of registers and the digital-to-analog converters of the 17 filter channels. Not all channels get this t_0-pulse at the same moment, but, if wanted, with different delays for the various filters. The pulse train from the digital-to-analog converters is then shaped into a sawtooth, by some sort of leaking peak detector with a charge time of 50 µsec and a much longer decharge time of about 1.5 msec. This signal is then filtered by the bandfilters, and summated.

If the status bits indicate that a sample is unvoiced, a random-pulse generator controls the 80-µsec gate and decides every 100 µsec whether the gate has

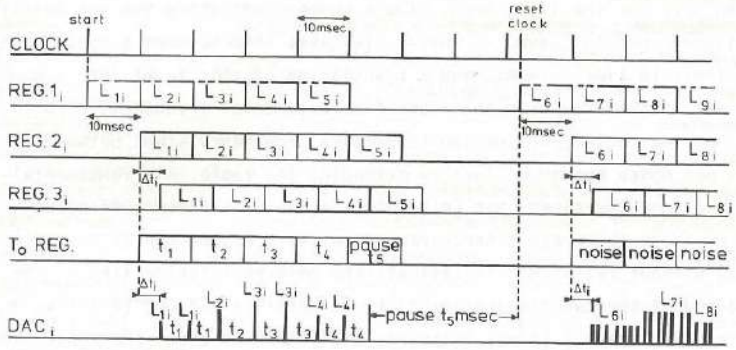


Fig. 2.5.2. Example of the timing sequence in the synthesis system. The artificial speech sound represented here, is assumed to consist of four voiced 10-msec samples, followed by a pause, and some unvoiced 10-msec samples. Only the data transfer for one of the 17 filters is represented.

to be open for 80 µsec or not. For a specific 10-msec unvoiced sample and a specific filter, all pulses during this 10 msec have the same amplitude but a pseudo-random pulse-distance distribution, thus generating a so-called poisson noise. Fig. 2.5.2 gives an example of the timing sequence in the synthesis system. It gives the timing for an artificial speech sound consisting of four voiced 10-msec samples,followed by a pause, and some unvoiced 10-msec samples. The data transfer for only one of the 17 filters is illustrated.

The present concept allows that some of the filters are excited with periodic pulses and some of the filters with noise. A separate bit for each filter, and a separate source signal for each filter make this possible. We had the impression that this could perhaps be a useful feature for generating voiced stops and voiced fricatives by means of a voiced signal in some lower filters and noise in the higher filters. The present analysis system, however, does not give us this detailed information: a specific sample is either voiced, or unvoiced. Somewhat to our surprise, quite natural voiced stops and fricatives can be produced, simply by generating some voiced samples (vocal murmur) followed by some (unvoiced) noise samples.

The possibility of generating both a voiced and an unvoiced signal, within one sample, was used in a selective vowel-masking experiment (Pols, 1975).

The actual control values for the synthesizer are obtained through a transformation table which has as input parameters: the label voiced or unvoiced, and if voiced, also the fundamental frequencies, plus the dB levels to be achieved in the 17 filters. Outputs from this transformation table are 17 numbers between

0 and 1023 for the 17 filters, plus a number indicating the periodicity in multiples of 16 μsec. This table takes into account a conversion from logarithmic to linear level, and a translation of this level into a number specific for that filter with that specific fundamental frequency, or with noise. At present the table contains 130 fundamental frequency steps between 71 and 200 Hz, plus one noise spectrum. Just by extending the table this fundamental-frequency range can be increased, but up to now the range has been sufficient for male voices. With the present hardware and software any length of text can be regenerated without real-time limitations. The only restriction lies in the storage capacity of the two fixed-heads disks (a total of 260,000 18-bit words). Time restrictions may of course intervene when, for instance, the clock time is made much shorter than 10 msec, or when too many operations have to be done on the data. These operations can for instance be: dimensionality reduction, or rule synthesis, or intonation modifications, or data interpolations.

In Pols (1974) the synthesizer was used for the first time to study the perceptual differences between vowel sounds in a set of 120 C_1VC_2 words; at that time all phonemes had to be voiced.

At present, both the analysis and the synthesis system are such that a detailed analysis of large sets of speech data is possible.

2.5.2. Intelligibility measurements with the synthesis system

In addition to the technical description of the synthesis system given above, some data on the speech intelligibility obtained with this system will be given. As we said before, the synthesis system was primarily developed as a research tool for flexible stimulus generation and not as an optimal vocoder system with low bit-rate and high naturalness of speech quality. Somewhat similar synthesis systems have been used by other researchers. Kramer and Mathews (1956) and Crowther and Rader (1966) suggested low-bit-rate vocoder systems based on linear transformations of the vocoder channel signals. Li *et al.* (1971, 1973a) reconstituted speech from spectra of reduced dimensionality. The analysis and subsequent dimensionality reduction was similar to our approach but the synthesis part differed. Their 35-dimensional spectral information is reduced to m dimensions applying a principal-components analysis. From this m-dimensional information a 35-dimensional spectrum is reconstituted, and from that the speech waveform, by means of inverse fast-Fourier-transform (FFT) computation. The step from a 35-dimensional bandpass spectrum to a line spectrum introduces some ambiguity since it is not unique. Moreover, an inverse FFT can, on successive samples, not be done in real time and gives trouble in concatenating successive

periods. Therefore, we preferred a synthesis which was a complete mirror-image of the one-third octave analysis. Li *et al.* (1973) mention consonantal intelligibility scores of 47% to 75% as the spectral dimensionality varies from 3 to 35 - these scores are lower than ours, as we will see in the next section. Cartier and Graillot (1974) use a 14-channel vocoder and almost the same dimensionality-reduction procedure (analysis of correspondence). The system was tested with logatoms spoken by two male and two female speakers. The reported recognition results for the five vowels and the 20 initial consonants using 2, 3, 5, or 14 factors are slightly better than our results, but not many details are given. Längle and Paulus (1974) use a similar data reduction technique (Karhunen-Loeve expansion) for low-bit-rate speech transmission but start from a fixed number of pitch-synchronous time samples instead of spectral information. Results are summarized by rate-distortion functions instead of intelligibility scores.

We also have used the number of dimensions (m) with which the spectral information is described as a variable in the intelligibility measurements. In paragraph 2.4 it was specified how the levels in the 17 filter bands can be reconstituted from the coordinate values in m dimensions and the average information in the remaining (17-m) dimensions. With this regenerated 17-dimensional information the synthesizer can be controlled.

From some preliminary experiments (Pols, 1973, 1974) it is clear that five conditions give a good insight into the effect of dimensionality on speech intelligibility. These five conditions are: the use of all 17 original filter levels, and regeneration of the bandfilter spectra using the coordinate values in 7, 4, 3, or only 2 dimensions. In order to make the intelligibility scores from these tests directly comparable with known scores for other communication channels, lists of phonetically balanced nonsense words (PB-words) were used.

Each of five male speakers pronounced a different list of 50 PB-words in a quiet room. Each list was preceded by 10 disyllabic words which were used to adapt the listener to the specific synthesis condition of that list. The 300 words from those five speakers were recorded and subsequently analyzed with our analysis system. All information of the, in total, 16,523 10-msec samples was stored on disk. The total variance of all level-normalized spectral points in 17 dimensions was 1864.8 dB^2. As outlined in paragraph 2.3 the optimal subspace for this material was determined, both for group 1 (unvoiced, mainly non-sonorants) and group 2 (voiced, sonorants) samples. The first eigenvector of all the samples is represented in Fig. 2.5.3. The number of zero crossings and the coordinate value along the first dimension, determined by this eigenvector, are used to make the group decision. 12,628 samples (76.4%) belonged to group 2, the so-

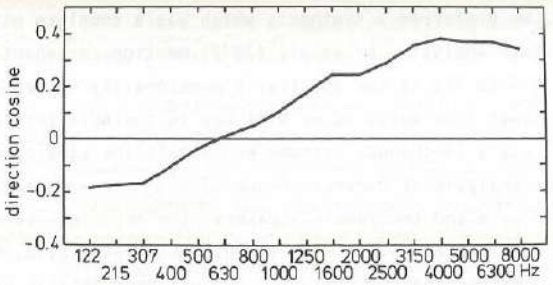


Fig. 2.5.3. Direction cosines of the first eigenvector explaining 54.5% of the total variance (1864.8 dB^2) of the data consisting of 16,523 10-msec sample points of 250 PB-words from five different male speakers.

norants, with a total variance of 1242.2 dB^2. The percentage explained variance per new dimension is given in Fig. 2.5.4. The right-hand diagram gives the same for the group-1 samples, with a total variance of 976.9 dB^2.

In an off-line calculation the coordinate values of all samples in the group-1, or group-2, 7-dimensional subspace were determined, and stored on disk together with the original 17 filter levels. We slightly smoothed the F_0-contour of each word to reduce the "creakiness" of the synthetic voice. This smoothing was done by taking into account the F_0-information of five neighbouring samples on both sides.

Subsequently, all five lists were resynthesized under each of the five synthesis conditions (via 17, 7, 4, 3, and 2 dimensions). Between the resynthesized words there was a silent interval of three seconds. To prevent possible familiarization of the listeners with a certain list, each list was generated under

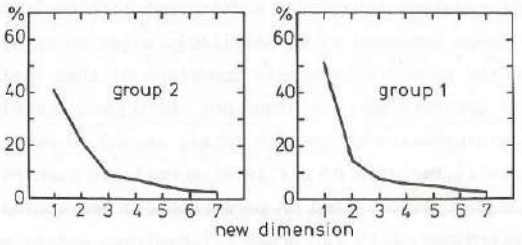


Fig. 2.5.4. Percentage explained variance per new dimension of the 12,628 group-2 samples (voiced, sonorants) with a total variance of 1242.2 dB^2 (left). The right diagram represents the percentage explained variance per new dimension of the 3,895 group-1 samples (unvoiced, non-sonorants) with a total variance of 976.9 dB^2.

each of the five conditions with a different order of the 50 words within that list. This synthetic speech was recorded on a magnetic tape recorder with the five lists per synthesis condition in a random order. Five experienced listeners listened to this material through headphones and wrote down the words they heard. These persons were experienced logatom listeners, which means that they are trained to write down what they really hear, whether or not it is a nonsense word, but they were not at all accustomed to this specific synthetic-speech quality. Therefore,we started each listening session in which one synthesis condition was presented, with a sixth list, from a sixth speaker, which was generated under the same synthesis condition. The listeners heard this list with prior knowledge of the words which were spoken, in order to get somewhat accustomed to that synthesis condition. Only after this presentation did the actual session start with the presentation of all five lists. Since each list had originally been spoken by a different person, we also allowed a short adaptation to this speaker by presenting first the 10 known disyllabic words before the 50 unknown PB words. The order of presentation of the five synthesis conditions was balanced for the five listeners. In order to get some idea of the learning effect every listener got the first synthesis condition he had listened to, for a second time in the last listening session.

2.5.3. Intelligibility scores

Although, for the present research project, we are mainly interested in the specifications of the synthesis system in terms of vowel scores, we also give the consonant and word data here, in order to have a complete system description.

The individual and average word scores, and the correct scores for the vowels and the initial and final consonants are given in Table 2.5.1. These scores have not yet been corrected for the different frequencies of occurence of the different phonemes. For a graphical representation of the average scores, see Fig. 2.5.5.

The correct score under the optimal synthesis condition (using the original 17-dimensional information) is an important figure, since it tells us how well this system can be used as a research tool for auditory speech segmentation, for detailed listening to (parts of) an utterance, or for stimulus generation. The vowel score is 96.8% which is, for our present experiments, an acceptable figure. The correct word score of 77.8% has to be considered with respect to the fact that nonsense words were used. A minimum demand for transmitting intelligible speech is a 30% PB-word score (Kryter, 1972), whereas almost 100% sentence

listener	synthesis condition 2	3	4	7	17	learning effect	
1	38.4	46.8	50.8	60.8	78.4	17	72.8
2	25.2	45.2	59.6	72.8	77.6	7	62.0
3	30.4	42.8	54.0	60.4	72.8	4	35.6
4	33.6	48.4	49.2	72.0	82.8	3	36.8
5	36.4	37.2	52.0	66.8	77.2	2	28.0
average	32.8	44.1	53.1	66.6	77.8		
listener LP	56.8	68.8	68.8	84.8	90.0	17	89.6
C_i	58.2	65.4	68.1	77.0	83.4		
V	67.6	78.1	85.4	92.0	96.8		
C_f	78.6	80.0	87.4	93.4	96.2		

Table 2.5.1. Individual and average percentages of correct word scores per synthesis condition, as well as the scores for the listener LP. Each listener had one identical synthesis condition in the first and last listening sessions. The last column gives the score for that condition in the first session. Comparison with the score for the same condition in the last session gives an indication of the learning effect. The lowest part of the table gives the average correct scores per synthesis condition for the initial consonants (C_i), vowels (V), and final consonants (C_f), separately.

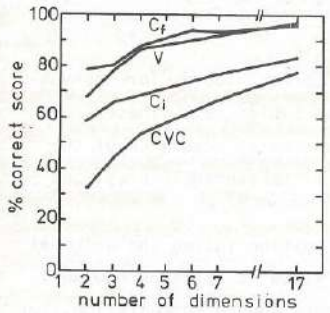


Fig. 2.5.5. Percentage-correct score per synthesis condition, averaged over the five listeners. Scores for initial consonants, vowels, final consonants, and whole words are given separately.

intelligibility is achieved with a 50% PB-word score (Kryter, 1962).

The individual word scores, as a function of the number of dimensions used for resynthesis, are represented in Fig. 2.5.6.

The word scores for the five listeners do not differ much, but there is a con-

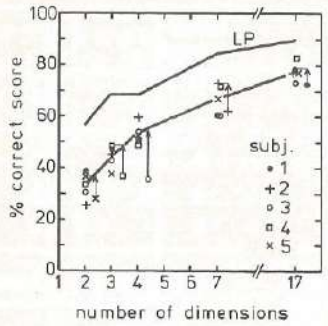


Fig. 2.5.6. Individual and average percentage-correct word scores per synthesis condition. The arrows show the improvement, in each case for one subject and one condition, between the first and the last listening sessions in which the synthesis condition was the same. The upper line gives the percentage correct score for a trained listener.

siderable learning effect, indicated by means of the arrows in Fig. 2.5.6. These arrows show the improvement between the first and the last listening sessions, with equal stimulus conditions per subject. Surprisingly enough, none of the listeners had the impression that their understanding of the synthetic speech improved with time. However, the author, best accustomed to this synthetic speech, came to still considerably higher scores in a separate listening session, see Table 2.5.1 and Fig. 2.5.6 for the data labelled LP.

A better insight into the qualities and peculiarities of the synthesis system can be achieved by studying not just how many errors, but also what kind of errors for C_i, V, and C_f are made under the different conditions. Such confusion matrices, accumulated over lists and listeners, are given in Table 2.5.2 for the synthesis condition using the original 17-dimensional filter levels. The tables are corrected in such a way that the row totals are 100. In the original lists there was of course considerable variation in the frequency of occurrence of different phonemes, since the lists were phonetically balanced (Huizing and Molenaar, 1944; Reyntjes, 1951). A significant amount of vowel confusion exists only between /œ/ and /I/, but, in every PB-word list there was only one word with /œ/. The percentage correct score for the final consonants is also very high (92%), but some nasals show some confusion favouring /n/, which is also the most frequently occurring consonant in the list. The group of 17 initial consonanta gives most errors especially for the short plosives and the low-intensity /f/. This is partly because these phonemes have already been lost during the analysis, their overall level being sometimes below the threshold established to exclude noise. After resynthesis they can only be identified by making use of coarticulatory features in the vowel, depending on each individual listener's experience. This can be seen in Table 2.5.3 where the individual responses to these phonemes are given. For instance, listener 3 makes many errors with /k/,

C_i	response																
stimulus	p	t	k	b	d	f	s	x	v	z	h	w	j	l	r	m	n
p (1)	36.0			32.0	8.0						16.0	8.0					
t (3)	14.7	44.0	1.3	8.0	32.0												
k (2)	8.0	14.0	72.0								6.0						
b (3)				78.7								17.3				4.0	
d (8)		0.5		15.5	69.5							12.5	1.5	0.5			
f (1)						72.0			4.0	24.0							
s (2)		10.0					90.0										
x (3)						1.3		97.4	1.3								
v (4)								2.0	81.0			17.0					
z (3)							1.3			98.7							
h (3)			2.7								96.0	1.3					
w (3)				5.3								92.1				1.3	1.3
j (1)													100.0				
l (2)														100.0			
r (6)											1.3	2.7		1.3	94.7		
m (2)																94.0	6.0
n (3)																6.7	93.3
TOTAL	58.7	68.5	76.0	139.5	109.5	73.3	91.3	103.4	108.5	98.7	119.3	150.9	101.5	101.8	94.7	106.0	100.6

V	response											
stimulus	ɑ	a	ɛ	ɪ	e	i	ɔ	o	u	œ	ʏy	ɛi	others
ɑ (9)	97.8						2.2						
a (6)		100.0											
ɛ (5)			96.8	2.4						0.8			
ɪ (5)				91.2		4.0				4.0			0.8
e (5)				0.8	98.4							0.8	
i (3)					2.7	97.3							
ɔ (5)							100.0						
o (4)								100.0					
u (2)									100.0				
œ (1)			4.0	20.0						76.0			
ʏy (1)											100.0		
ɛi (4)											10.0	90.0	
TOTAL	97.8	100.0	100.8	114.4	101.1	101.3	102.2	100.0	100.0	80.8	110.0	90.8	0.8

C_f	response										
stimulus	p	t	k	f	s	x	l	r	m	n	ŋ
p (1)	84.0		12.0					4.0			
t (10)		100.0									
k (3)			100.0								
f (1)				92.0		8.0					
s (5)					100.0						
x (4)				4.0		96.0					
l (5)							100.0				
r (5)								100.0			
m (2)									68.0	32.0	
n (13)									3.4	95.4	1.2
ŋ (1)									8.0	16.0	76.0
TOTAL	84.0	100.0	112.0	96.0	100.0	104.0	100.0	104.0	79.4	143.4	77.2

Table 2.5.2. Confusion matrices presented separately for the initial consonants, vowels, and final consonants in the PB-words. These are percentage scores, accumulated over lists and listeners for the synthesis condition using the original 17 filter values. The number of times every phoneme actually occurred in each list of 50 PB-words, is indicated between brackets.

	p	b	d	w	h	total errors
1	4			1		1
2	1	3			1	4
p 3	3			1	1	2
4	0	2	1		2	5
5	1	3	1			4
total	9	8	2	2	4	16

	t	p	k	b	d	total errors
1	10	2		1	2	5
2	5	1	1	1	7	10
t 3	6	3		2	4	9
4	5	2		2	6	10
5	7	3			5	8
total	33	11	1	6	24	42

	k	p	t	h	total errors
1	7	1	2		3
2	7	1	1	1	3
k 3	5		3	2	5
4	9	1			1
5	8	1	1		2
total	36	4	7	3	14

	b	w	m	total errors
1	14		1	1
2	14	1		1
b 3	7	7	1	8
4	15			0
5	9	5	1	6
total	59	13	3	16

	d	t	b	w	j	l	total errors
1	26	1	12	1			14
2	32		5	3			8
d 3	21		7	12			19
4	36		4				4
5	24		3	9	3	1	16
total	139	1	31	25	3	1	61

	f	X	w	total errors
1	3		2	2
2	0	1	4	5
f 3	5			
4	5			
5	5			
total	18	1	6	7

Table 2.5.3. Individual responses of the five listeners to the initial plosives /p, t, k, b, d/ and the low-intensity /f/. The 17-dimensional spectral information was used for synthesis.

/b/ and /d/, whereas listener 4 identifies these phonemes almost always correctly.

Table 2.5.4 gives the confusion matrices for the synthesis condition using the two-dimensional coordinate values. The neutral vowel /œ/ is only correctly recognized 16% of the time, the responses are scattered over /ɪ/, /ɛ/, /ɔ/, /i/, and /u/. Apart from a possible spectral similarity there is also the duration information which causes long vowels like /ɑ/, /e/, /o/, and /u/ to be seldom confused with short vowels like /ɑ/, /ɛ/, /ɪ/, /i/, /ɔ/, and /œ/, and vice versa.

Most consonant confusions can be described in terms of manner or place of articulation.

Certain phonemes loose more from deleting spectral information than others. This can be seen in Table 2.5.5, where the correct phoneme scores using 2-dimensional spectral information are compared with those using 3 and 17 dimensions. For the vowels a complete list over all synthesis conditions is given. The

C_i	response: p	t	k	b	d	f	s	x	v	z	h	w	j	l	r	m	n
p (1)	56.0			28.0							12.0	4.0					
t (3)	37.3	32.1	1.3	16.0	12.0						1.3						
k (2)	26.0	38.0	18.0	6.0							10.0	2.0					
b (3)				38.7	5.3				24.0				1.3	4.0		12.0	14.7
d (8)	0.5			15.5	43.0							22.5	12.5	0.5	0.5	2.0	3.0
f (1)						32.0		20.0	48.0								
s (2)		10.0				8.0	82.0										
x (3)						21.0		56.1	17.3						5.3		
v (4)							1.0	2.0	81.0	1.0		11.0	1.0		3.0		
z (3)							17.3		9.3	73.4							
h (3)	4.0		2.7					2.7	4.0		59.9	10.7	2.7	2.7	9.3	1.3	
w (3)				1.3							2.7	76.0		4.0		6.7	9.3
j (1)													76.0			12.0	12.0
l (2)												4.0	2.0	50.0		4.0	40.0
r (6)											1.3	17.3	2.7		78.7		
m (2)												12.0		6.0		44.0	38.0
n (3)														9.3		21.3	69.4
TOTAL	123.8	80.1	22.0	105.5	60.3	61.3	100.3	80.8	183.6	74.4	87.2	159.5	98.2	76.5	96.8	103.3	186.4

stimulus

V	response: ɑ	a	ɛ	I	e	i	ɔ	o	u	œ	ʌy	ɛi	others
ɑ (9)	48.0	4.0	7.6				36.0	3.6			0.4		0.4
a (6)		88.7	2.0				1.3	2.0			6.0		
ɛ (5)	6.4	1.6	74.4	4.0	0.8	0.8	4.8		0.8	5.6	0.8		
I (5)			4.0	52.8	2.4	34.4	1.6	0.8	3.2		0.8		
e (5)			0.8	4.0	85.6	6.4					0.8	1.6	0.8
i (3)				1.3	1.3	93.4			2.7				1.3
ɔ (5)	4.8		7.2	2.4			60.0	7.2	15.2	2.4		0.8	
o (4)	1.0	7.0	1.0		2.0		3.0	82.0			3.0	1.0	
u (2)						2.0	2.0	2.0	94.0				
œ (1)			16.0	48.0		4.0	12.0		4.0	16.0			
ʌy (1)			24.0					16.0			48.0	12.0	
ɛi (4)			10.0	2.0	6.0		1.0	1.0			30.0	48.0	2.0
TOTAL	60.2	101.3	147.0	114.5	98.1	141.0	121.7	114.6	119.9	24.0	89.8	63.4	4.5

stimulus

C_f	response: p	t	k	f	s	x	l	r	m	n	ŋ
p (1)	72.0	4.0	20.0					4.0			
t (10)	4.0	95.2	0.4					0.4			
k (3)	32.0	12.0	56.0								
f (1)				68.0		32.0					
s (5)				8.8	90.4	0.8					
x (4)	1.0		1.0	29.0	2.0	50.0		17.0			
l (5)							70.4	3.2		23.2	3.2
r (5)				0.8			2.4	93.6		3.2	
m (2)							8.0		12.0	78.0	2.0
n (13)							2.8		4.6	88.6	4.0
ŋ (1)							4.0		4.0	68.0	24.0
TOTAL	109.0	111.2	78.2	105.8	92.4	82.8	87.6	118.2	20.6	261.0	33.2

Table 2.5.4. Similar to Table 2.5.2 but now the two-dimensional coordinate values were used for resynthesis.

	initial consonants: p	t	k	b	d	f	s	x	v	z	h	w	j	l	r	m	n	all C_i
2 dim.	55.0	32.1	18.0	38.7	43.0	32.0	82.0	56.1	81.0	73.4	59.9	76.0	76.0	50.0	78.7	44.0	69.4	56.8
3 dim.	40.0	30.6	26.0	68.1	47.0	52.0	80.0	70.7	84.0	86.7	77.3	79.9	96.0	50.0	80.0	46.0	82.6	64.5
17 dim.	36.0	44.0	72.0	78.7	69.5	72.0	90.0	97.4	81.0	98.7	96.0	92.1	100.0	100.0	94.7	94.0	93.3	82.9

	vowels: ɑ	a	ɛ	I	e	i	ɔ	o	u	œ	ʌy	ɛi	all V
2 dim.	48.0	88.7	74.4	52.8	85.6	93.4	60.0	82.0	94.0	16.0	48.0	48.0	65.9
3 dim.	66.7	95.3	78.4	79.2	92.8	98.7	69.6	74.0	94.0	8.0	80.0	66.0	75.2
4 dim.	64.6	100.0	84.0	90.4	97.6	86.7	91.2	94.0	94.0	20.0	96.0	83.0	83.5
7 dim.	91.1	100.0	84.8	92.8	95.2	93.3	92.8	96.0	98.0	56.0	100.0	84.0	90.3
17 dim.	97.8	100.0	96.8	91.2	98.4	97.3	100.0	100.0	100.0	76.0	100.0	90.0	95.6

	final consonants: p	t	k	f	s	x	l	r	m	n	ŋ	all C_f
2 dim.	72.0	95.2	56.0	68.0	90.4	50.0	70.4	93.6	12.0	88.6	24.0	65.5
3 dim.	76.0	99.6	64.0	72.0	85.6	69.0	71.2	93.6	14.0	82.2	40.0	69.8
17 dim.	84.0	100.0	100.0	92.0	100.0	96.0	100.0	100.0	68.0	95.4	76.0	92.0

Table 2.5.5. Percentage-correct score per phoneme under different synthesis conditions. The last column gives the average scores.

average scores for C_i, V, and C_f slightly differ from those given in Table 2.5.1 because of the normalization to equal occurrence of all phonemes.

An analysis of variance (Riemersma and Burry, 1973) was performed on the vowel recognition data, see Table 2.5.6. The differences between synthesis con-

main effects and interactions	degrees of freedom	% variance	significance level
conditions	4	21.1	**
listeners	4	9.8	**
conditions x listeners	16	0.3	n.s.
vowels	11	47.0	**
conditions x vowels	44	13.6	**
listeners x vowels	44	5.2	**

Table 2.5.6. Results of an analysis of variance on the recognition scores for 12 vowels under five different synthesis conditions for five listeners. The percentage explained variance per variable and per interaction, as well as the significance level of the differences found are indicated (** significant at 1% level; n.s. not significant).

ditions, listeners, and vowels are all very significant and explain 21.1%, 9.8%, and 47.0% of the variance in the data, respectively. The effect of the synthesis conditions on intelligibility is the main effect we are interested in, see Fig. 2.5.5. The differences between the listeners are relatively small. Averaged over all conditions and vowels, the best listener has a correct score of 84.9%, and the worst listener 78.2%. The large main vowel effect, and the relatively large interactions, mean that one vowel under a certain condition is far better understood than another vowel. This is partly due to the properties of the synthesis system, as can be seen by the significant condition x vowel interaction, and partly caused by the listeners, as can be seen by the significant listener x vowel interaction. For example, under a certain synthesis condition, listener 3 did not make a single /ɑ-ɔ/ confusion, whilst listener 4 made 14 /ɑ-ɔ/ confusions out of a total of 45 presentations.

The results of the listening experiment can be used for further, more detailed, analyses. We tested, for instance, whether the listeners were not biased in favour of frequently occurring and/or meaningful words. If the responses of the subjects to the initial consonant, vowel, and final consonant per word are independent from each other, the word scores can be predicted from the C_i, V and C_f scores. Fig. 2.5.7 plots these predicted scores against the actually found scores, for the different listeners under different conditions. In general, the correspondence is excellent, but for the poorer conditions nearly all actually found word scores are slightly higher than the predicted scores. This means that phoneme errors are accumulated in words. In other words: once there is a phoneme wrong in a word, this raises the chance that a second phoneme in the same word will be wrong too. This favours the correct word score relative to the phoneme score, since for the word score it makes no difference whether one, two, or all three phonemes in one word are wrong.

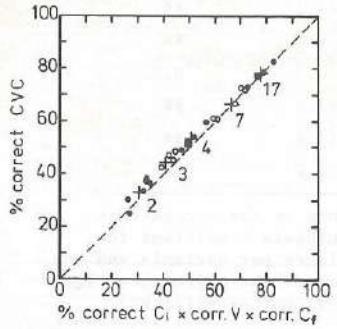


Fig. 2.5.7. Percentage-correct PB-word score plotted against the product of the percentage-correct initial-consonant, vowel, and final-consonant scores. Every point represents one listener under one synthesis condition. Crosses represent averages over listeners under the condition indicated.

The intelligibility results show, and the actual use in practice demonstrates, that this synthesis system is a very appropriate research tool.

2.6. CONCLUDING REMARKS

In this chapter we have described the three tools we will use in studying vowel coarticulation: analysis, data processing, and synthesis.

The real-time analysis system, based on bandfilter spectra, allows a fast, automatic, and reproducible way of spectral analysis. The successive 10-msec samples form the basis of a detailed description of both the average and the dynamic spectral variation. Although the one-third octave spectra only represent a limited spectral resolution, this representation is detailed enough to present most of the relevant spectral differences within and between vowel phonemes in different contexts.

The storage of all basic analysis parameters in the computer allows efficient data processing, the most important aspect of which is data reduction. The principal-components representation of the bandfilter spectra in two or more dimensions nicely shows the basic spectral differences between different vowel phonemes. As we will see in the next chapter, this also makes it possible to represent the dynamic spectral variation within a vowel segment as a trace of successive 10-msec sample points in the vowel subspace.

The synthesis system, finally, allows regeneration of the word, or any specific segment in it, for detailed listening. And once the vowel segments have been isolated, it also permits stimulus generation. The vowel-intelligibility scores given in section 2.5.3 give us confidence that the system is appropriate for stimulus generation. After all, if in an identification experiment, confusions are made by the subjects, one wants to be sure that these are related to the experimental condition and are no artefacts of the synthesis system.

CHAPTER 3

SPECTRAL ANALYSIS OF DUTCH VOWELS IN MONOSYLLABIC WORDS

3.1. INTRODUCTION

In Chapter 2 we have indicated that a bandfilter analysis together with a dimensional representation of the spectral information is a fast, objective and reproducible way of representing the relevant spectral differences between vowel sounds in a neutral context. Now we will apply this approach to study the spectral differences between vowels in different consonant contexts. Different consonant contexts, stresses, tempi, and styles of speech result in coarticulation and/or vowel reduction. Although all these aspects are important, in the present investigations attention is focussed on one, relatively simple, condition: the influence on vowel sounds of preceding and following consonants in isolated, stressed, monosyllabic words of the type consonant-vowel-consonant. We studied this influence both on the spectral and the perceptual level. The spectral analysis did not only result in average spectral information but we also tried to get some insight into the dynamic spectral characteristics of the vowel sounds. The results of this study may stimulate subsequent studies in the direction of non-vowel sonorants, other consonants and consonant clusters, multisyllabic words, stressed and unstressed words in sentences read aloud and conversational speech.

The present project can be seen as a natural extension of earlier measurements on average spectral differences between vowels in a context of h(vowel)t, first for only ten native speakers of Dutch (Plomp *et al.*, 1967), later for 50 male (Klein *et al.*, 1970; Pols *et al.*, 1973), and 25 female speakers (van Nierop *et al.*, 1973), including both bandfilter analysis and formant analysis. In those studies, the vowel sounds were supposed to be stationary with no dynamic characteristics. The perceptual differences between the 100-msec vowel segments of the

50 male speakers were also studied (Klein *et al.*, 1970), as was the relation between the perceptual and physical dimensional representations of a well-defined set of vowel sounds (Pols *et al.*, 1969; van der Kamp and Pols, 1971). Next we analyzed a set of 50 C_1VC_2 words from five speakers (Pols, 1971), later extended to a list of 270 C_1VC_2 words spoken twice by one talker (Pols, 1972). These last two studies were not specifically concerned with vowels, but were attempts in the direction of automatic segmentation and labelling of the phoneme-like stationary segments.

In this chapter we will specify our chosen word material (3.2), give a description of the spectral analysis of this list of Dutch words spoken by three native male speakers (3.3), discuss the procedure to isolate the vowel segments by using the synthesizer (3.4), and, finally, give a dimensional spectral representation of these vowel segments in paragraphs 3.5 and 3.6. In Chapter 4 experiments are described in which all vowel segments of one speaker were identified by a group of native listeners. The results of these perceptual experiments are related with the spectral data.

3.2. SPECIFICATION OF THE WORD LIST

In selecting the list of words to be analyzed, we did not want to limit ourselves too much in advance, so we decided to include in principle all vowels, all consonants, and all CV and VC pairs. Furthermore, we preferred to work with meaningful words in order to assure a more natural pronunciation.

In Dutch there are 18 initial consonants (C_i), including the "no initial consonant" condition (.):

/p t k b d f s x v z h w j l r m n ./.

Two fairly rare consonants, /ʃ/ and /ʒ/, *i.e.* the initial consonants in the Dutch words *sjaal* and *gelei*, are not included. The velar fricative /ɣ/ was also excluded since it can only occur in an intervocalic position. For a discussion of the /ɣ-x/ opposition in Dutch, see van den Broecke and van Heuven (1976).

All 18 C_i's had to be combined with the 15 Dutch vowels. These vowels include 12 monophthongs and three diphthongs:

/ɑ a ɛ ɪ e i ɔ o u œ ø y ɑu ʌy ɛi/.

This results in a list of 18 x 15 = 270 C_iV combinations to which final consonants (C_f) have to be added. Not all initial consonants can occur as final consonants. In Dutch there are no final voiced stops and fricatives, and /h/ cannot be final either, but the /ŋ/ is added, resulting in 14 final consonants:

/p t k f s x w j l r m n ŋ ./.

Since there are fewer final than initial consonants, some final consonants may occur twice or more in combination with the same vowel.

For certain C_iV and VC_f combinations, several acceptable Dutch words are available. However, it is interesting to note that many perfectly legitimate words are unique for that C_iV or VC_f combination. As far as we could check, unique C_iV combinations exist in the following Dutch words: *puur*, *touw*, *kuur*, *dauw*, *fooi*, *fat*, *fut*, *feut*, *fee*, *soos*, *set*, *safe*, *sier*, *suit*, *sein*, *vul*, *vuur*, *vouw*, *vuil*, *zuur*, *gek*, *nu*, *nauw*, *lauw*, *rauw*, *joel*, *jeuk*, *Jees*, *Juul*, *jous*, *juich*, *jij*, *uur*. Unique VC_f combinations in monosyllabic CVC words are two Christian names: *Huub* and *Guus*. The middle part of Table 3.2.1 shows which consonants, mainly /ŋ/, /w/, and /j/, cannot occur in final position after which vowels. Certain combinations of a vowel followed by /j/ are disputable, since these are exclamations (*hoj*, *aj*) which can also be seen as combinations with /i/. Many data on the distribution of Dutch phonemes can be found in Cohen *et al.* (1961). Our 270 C_iVC_f word list is mainly based on the tables given in that book, but we also consulted van de Berg (1969) and van Dale's Dutch dictionary.

A small number of C_iV combinations does not exist in initial position in meaningful Dutch words: /sɸ, nʌy, wɸ, wy, ji, .ɸ, sy/. In that situation we used (city) names or two-syllabic words: *masseur*, *Nuis*, *Weurt*, *Wuustwezel*, *jeep*, *Eufraat*, *regu*, respectively,to explain these C_iV combinations to the speakers and next asked them to pronounce the appropriate nonsense words: /sɸr, nʌys, wɸr, wys, jip, ɸf, sy/. The same holds for certain VC_f combinations which do not exist in final position in monosyllabic meaningful Dutch words: /œr, œŋ, œf, ɸl, ɸm, Ix, yx, yn/. We explained these combinations again with resembling Dutch words like: *Turk*, *dunk*, *wuft*, *veulen*, *reuma*, *jicht*, *fuga*, *immuun*. Apart from the above-mentioned exceptions, all words are meaningful, although some of them at a somewhat academic level. Some Christian names were also included (*Piet*, *Fien*, *Saul*, *Guus*, *Huub*, *Juul*, *Jees*, *Jet*). The final word list is given in Table 3.2.1, in matrix notation.

Despite the length of this list of 270 words, it is still a rather incomplete one for studying coarticulation in monosyllabic C_iVC_f words. The main reason is that the list does not include any repetitions of C_iV combinations and only a few repetitions of some VC_f combinations. Moreover, the C_i-C_f combinations per vowel are arbitrary and incomplete. A systematic study of right-to-left and left-to-right coarticulation is not possible (Ohde and Sharf, 1975). Only a few symmetrical consonant combinations are present. However, the list satisfies our main aim: to study whether the dimensional spectral representation is detailed enough to describe coarticulation effects.

	α	a	ɛ	I	e	i	ɔ	o	u	œ	ɸ	y	αu	ʌy	ɛi
p	n	r	p	n	r	t	p	.	n	s	l	r	k	.	p
t	l	k	r	l	r	n	r	m	t	r	x	t	w	s	t
k	n	s	n	n	r	m	m	r	n	n	.	r	s	p	f
b	.	r	f	t	k	r	n	r	t	s	r	r	t	t	t
d	m	x	n	m	r	.	f	r	k	ŋ	r	k	w	n	k
f	t	m	l	t	.	n	p	j	j	t	t	x	n	k	n
s	r	j	t	p	f	r	m	s	s	l	r	.	l	t	n
x	f	p	k	ŋ	f	r	ŋ	t	t	n	r	s	t	t	t
v	ŋ	n	r	s	x	s	l	r	x	l	l	r	w	l	f
z	x	t	s	ŋ	m	k	t	n	t	.	r	r	t	l	s
h	k	n	.	ŋ	r	f	j	r	s	p	p	p	t	t	x
w	s	n	x	r	t	l	n	x	f	f	r	s	t	f	n
j	s	.	t	x	s	p	.	l	l	f	k	l	w	x	.
l	x	t	s	k	w	x	k	p	r	s	n	r	w	t	m
r	t	r	m	f	p	t	t	f	p	x	m	w	w	m	n
m	l	r	n	s	l	r	r	r	.	m	x	n	t	s	n
n	p	f	l	s	r	w	x	r	m	k	s	.	w	s	t
.	j	l	ŋ	n	n	p	s	k	r	k	f	r	t	t	s
not C_f	w	ŋ	w	w	ŋ	ŋ	w	ŋ	ŋ	w	ŋ	ŋ	ŋm	n	n
		w	j	j	j	j		w	w	j	w	j	jx	w	w
											j	f	rp	j	j
												m	f.	r	r
extra C_f	t	t	t	t	f	p	p	6r	3t	k	x	s	6t	5t	2t
	s	2n	s	2s	5r	t	t		s	f	l	6r	6w	2s	f
	x	3r	2n	2n		n	m		n	2s	5r	.		l	s
	n		l	2ŋ		3r	n		r	n					4n
	l		r				r			l					

Table 3.2.1. List of the 270 CVC words, represented in the form of a (18 x 15) matrix. The rows specify the initial consonants, the columns specify the vowels, and the cell values specify the final consonants. The middle part of the table indicates which of the 14 possible final consonants could not be used in combination with that vowel. Since there are 18 positions per vowel and only 14, or fewer, final consonants, some final consonants were used more than once. This is indicated in the lower part of the table.

3.3. SPECTRAL ANALYSIS OF THE WORDS SPOKEN BY THREE SPEAKERS

The 270 words were spoken by three native speakers of Dutch, in different orders, in a silent room, and were recorded on tape. The signal of a throat microphone was recorded simultaneously on the second track of the tape to be used later for pitch information. To have some running speech, each speaker also read the story of the north wind and the sun in Dutch (IPA, 1967). The data per speaker were processed separately, and only combined after segmentation of the vowel segments.

	speaker 1	speaker 2	speaker 3	total
total number of 10-msec samples	14111	12378	10853	
total variance in dB^2	1745.6	1988.4	1579.1	
variance explained by the first eigenvector in %	55.1	63.6	53.9	
number of group-1 samples (non-sonorants)	3293	2796	3007	
variance of group-1 samples in dB^2	745.3	558.9	1014.2	
number of group-2 samples (sonorants)	10818	9582	7846	
variance of group-2 samples in dB^2	1173.1	1385.2	874.3	
explained variance by successive factors of group 2 in % 1	40.9	48.5	37.9	
2	23.7	23.2	26.4	
3	11.3	9.8	9.2	
4	8.1	5.2	6.9	
5	5.3	4.0	5.3	
6	2.5	1.8	3.0	
7	1.8	1.5	2.7	
number of vowel samples	3721	3229	3332	10282
total variance in dB^2	951.5	867.0	763.9	1032.5
explained variance by successive factors of the vowel subspace in % 1	40.3	42.0	40.6	37.1
2	33.0	33.8	30.5	30.1
3	8.3	7.1	9.6	11.4
4	5.7	4.4	4.8	6.0
5	3.2	3.1	3.8	4.5
6	2.4	2.4	2.6	2.8
7	1.4	1.5	1.6	1.8

Table 3.3.1. Number of samples in the different groups and the related total variances, as well as the variances explained by the different sets of new dimensions for each of the three speakers individually. Under the heading "total", this information is given for the vowel space for the three speakers combined.

The recordings of the utterances per speaker were analyzed with the analysis system described in paragraph 2.2. The trigger level was chosen some 40 dB below the loudest passages in the recording and well above the tape noise. Pauses shorter than 250 msec, were supposed to be part of the word. The raw data per 10-msec sample were stored on disk, and copied on DEC-tape, for later data processing. The level-normalized samples were used to determine the covariance matrix.

The number of samples and the total variance are given in the first two rows of Table 3.3.1. A principal-components analysis on the covariance matrix per speaker resulted in a first eigenvector which explained for each speaker more than 50% of the total variance; see the third row. These first eigenvectors for all three speakers are represented in Fig. 3.3.1. As was said before (pag. 54), the coordinate value along this dimension is very efficient as information additional to the number of zero crossings (ZCC), for labelling every sample as group 1 (unvoiced, mainly non-sonorants), or group 2 (voiced, mainly sonorants). If ZCC < 10, or ZCC ≥ 15, the sample gets the group code 2, or 1, respectively. Only if 10 ≤ ZCC < 15, the coordinate value along the first dimension is decisive; in such a way that if this value is smaller than 50, the sample gets group code 1, otherwise 2. This shows that the time-consuming derivation of the group code from the coordinate value along the first dimension is only used if the number of zero crossings is not a clear enough indication for the group code. This

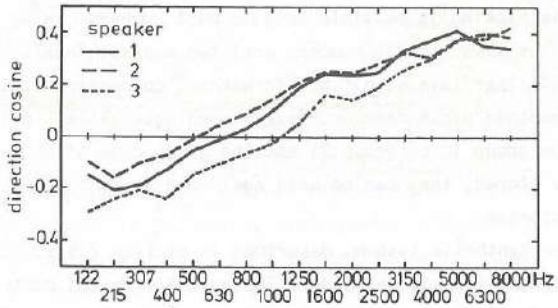


Fig. 3.3.1. Direction cosines of the first eigenvector for the data points representing all 10-msec samples of the 270 CVC words per speaker.

operational group definition was thoroughly tested and appeared to work reliably for almost all cases.

There are two reasons why every sample was labelled as 1 or 2:

- it makes it possible to determine two different subspaces for the group-1 and group-2 samples;
- for resynthesis, the group label is used as an unvoiced/voiced label.

Subsequently, separate covariance matrices for the group-1 samples and the group-2 samples were determined. Rows 4-7 of Table 3.3.1 give the total variance and the number of samples for both data sets. Next we determined the first seven eigenvalues and eigenvectors of both covariance matrices. The middle part of Table 3.3.1 gives the percentage of variance explained per new dimension for the group-2 data.

The group-2 space is for the present study the more interesting of the two because its samples include the vowel sounds. By using, for instance, the first *two* group-2 eigenvectors, we can represent the voiced part of a word as a series of successive sample points in a plane. The unvoiced part of the word, if present, is not represented in such a plane; but if we wish, it can be represented in a group-1 plane. The display in the group-2 plane was used to define the vowel segments in the successive words.

For every sample of the 270 words, the seven coordinate values in the group-1, or group-2, space were stored on disk together with the original data. A smoothed pitch contour per word was calculated by averaging some neighbouring values. In addition, the missing pitch information, on the basis of continuity constraints, was filled in for the few group-2 samples which failed to have pitch information. This correction of the original pitch information was also stored on disk. Together with the group code, this finally resulted in 30 numbers per 10-msec sample which were stored on disk in fifteen 18-bit words in a "packed" form. This packing is possible because most numbers can be described as integers in 9 bits or less. The 30 numbers are: two overall levels, number of zero crossings, 17 filter levels, pitch information, correction of this pitch information for a smoothed pitch contour, seven coordinate values in the 7-dimensional subspace for group 1, or group 2, and the group code of this sample. Once all these data are stored, they can be used again and again for data processing, displays, or resynthesis.

By means of the synthesis system, described in section 2.5.1, (parts of) the words could be resynthesized in any order. The voiced/unvoiced control is dictated by the group code, but can be overruled if necessary. One can listen to (part of) the word repeatedly, and by changing the clock of the synthesizer

(normally 10 msec), the utterance can also be stretched out for closer listening. All these features can be used, together with a numerical display of the original data, or a trace display on a CRT screen of the word in a subplane, to isolate the vowel segment of the word. The next paragraph gives the details of this segmentation procedure.

3.4. ISOLATION OF VOWEL SEGMENTS IN THE WORDS

The purpose of segmentation is to isolate the vowel parts from the words in order to study the spectral characteristics. We first of all studied the *average* spectral vowel characteristics. This means that we isolated the vowel segment as well as possible and, next, determined the average spectral information of that segment in order to see whether there was any systematic influence of the consonant context. These vowel segments were also presented to subjects in an identification experiment, see Chapter 4. It became clear quite soon that the average characteristics are often insufficient to explain the results of the perceptual experiments and that the dynamic spectral variation within the vowel segment is often more important than the average spectral information. The spectral analysis occurring every 10 msec allows such a detailed description.

We did not succeed in developing a purely objective and automatic segmentation procedure. The segmentation algorithm only gave an indication of the possible vowel segments. By changing the segment boundaries and subsequently listening to the synthesized vowel part and studying the word trace, a final decision was made by the experimenter about the sample numbers which define the segment.

The segmentation algorithm first traced the group-2 sample in the word with the highest linear level, which for our data was always in the vowel part. From that sample on, the word was scanned to the left and to the right until a group-1 sample, or a word pause, or the beginning or the end of the word was encountered. If the spectral distance between two successive group-2 samples was larger than 12 dB, this was also supposed to be a segment boundary. The preliminary vowel segment found in this way was displayed on a CRT screen as a more intense part of the word trace in the two-dimensional group-2 space. Then the whole resynthesized word was made audible, as well as this isolated segment. Two other sample numbers could be typed in to modify the segment. Also, the push buttons of the display control could be used to step through the word, sample by sample. In this way the beginning, and/or the end, of the segment could be changed systematically under visual and auditory control. That part of the word which clearly sounded like the intended vowel, containing no audibly different vocalic or

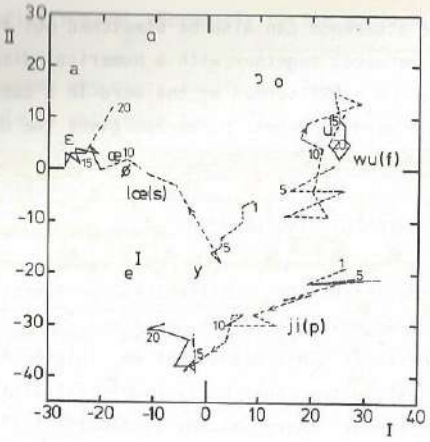


Fig. 3.4.1. Examples of word traces represented in the group-2 subspace of speaker 1. The group-2 samples of the words /lœ s/, /wuf/, and /jip/ are represented as labelled points and are connected to form traces. Those parts of the traces drawn as solid lines were chosen as the vowel segments from these words. For reference, the average vowel positions for this speaker are also given.

consonantal components, was ultimately defined as the vowel segment. This could also be checked by listening to the removed parts of the word. In case of uncertainty as to whether any trace of surrounding sounds was audible, a shorter vowel was favoured.

When the experimenter was satisfied, he typed in the name of the word, which name then was stored in a "library", together with the starting address of that word on the disk, and the initial and final sample numbers of the selected vowel segment.

Fig. 3.4.1 gives, as an example, the positions of the group-2 samples in the words /lœ s/, /wuf/, and /jip/. The vowel segments which were finally chosen are represented as solid lines. For general orientation the average vowel positions for this speaker are also given.

To summarize the procedure, we see that segmentation has been based on three types of information:

- looking at the number display of the raw data;
- looking at the display of the word trace in a group-2 subspace;
- listening to the resynthesized word and vowel segment.

Since this segmentation procedure is not completely objective, an extra control step was included, by presenting all words and vowel segments to a colleague who independently gave his opinion about the segmentation. In the few cases of disagreement, a compromise was made.

Once the segmentation is completed, all vowel segments can be made visible and/or audible in any order just by scanning the library in a predefined way. For instance *A * gives all /ɑ/ segments, *OE* all /u/ segments, ***R all vowel segments of words which end on an /r/, etc.

3.5. DIMENSIONAL SPECTRAL REPRESENTATION OF THE VOWEL SEGMENTS

3.5.1. Vowel subspace

In the segmentation procedure described in the preceding paragraph, for each speaker an individual group-2 subspace was used as the optimal subspace. As we now have isolated all *vowel* segments of the three speakers, we can next define one *vowel subspace*. The individual group-2 subspaces had been based on all group-2 samples including not just vowels, but also nasals, liquids, glides, and vocal murmurs. One should remember that it had been relatively easy to label every sample 1 or 2; however, it would have been far more difficult to label every sample "vowel" or "non-vowel". After having isolated all vowel segments, we next determined a covariance matrix on the basis of the samples of all these vowel segments from the three speakers. The lower part of Table 3.3.1 gives the number of samples and the total variance. Again the first seven eigenvalues and eigenvectors were determined through a principal-components analysis of this covariance matrix. The percentages of explained variance are also given in the table.

The first two dimensions together explain 67.2% of the total variance. The direction cosines of the first five eigenvectors were represented in Fig. 2.3.3 and Fig. 2.3.5.

The first two eigenvectors found by this straightforward principal-components analysis show a remarkable correspondence with a pair of eigenvectors found earlier (Pols *et al.*, 1973), even without an allowable rotation. Those eigenvectors defined a plane in which the 12 Dutch vowels from 50 male speakers were best identified on the basis of maximum-likelihood regions. Furthermore, this plane was rotated in such a way that there was an optimal match between the average vowel points in this plane and in the log F_1-log F_2 plane. The direction cosines of both sets of eigenvectors are plotted in Fig. 3.5.1. The only difference is that in this diagram the order of both eigenvectors from the earlier study has been reversed.

We have also tested how specific or how general the eigenvectors of this vowel subspace are. One must consider the possibility that they are optimal only

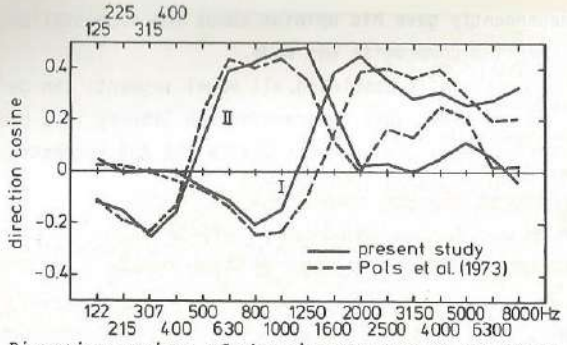


Fig. 3.5.1. Direction cosines of the eigenvectors of the first two dimensions of the subspace determined by the 270 vowel segments of three speakers. The dashed curves are from an earlier study (Pols *et al.*, 1973) based on the constant vowel parts in a neutral context uttered by 50 male speakers.

for the present vowel segment data in which, for instance, the long vowels have more influence than the short vowels because of their greater number of samples. We repeated the calculation of the eigenvectors on the basis of the *average* segment positions only, in that way excluding duration effects. The resulting eigenvector base was practically the same as the original one. So, it seems appropriate to conclude that the present eigenvector base specifies a general vowel subspace.

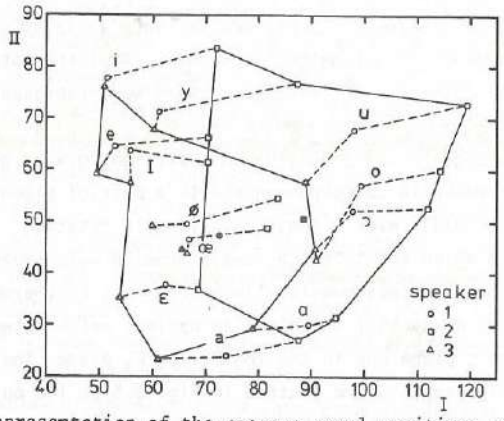


Fig. 3.5.2. Representation of the average vowel positions per speaker in the two-dimensional vowel subspace. The open points have been obtained by averaging the coordinate values of all samples in the 18 segments per vowel. The filled points represent the overall average positions per speaker.

3.5.2. Average vowel positions

After having defined the eigenvector base, we can now investigate how the actual level-normalized vowel spectra are represented in this space. Fig. 3.5.2 gives a representation of the average vowel positions per speaker in the two-dimensional vowel subspace. These average positions have been obtained by averaging per vowel the coordinate values of all samples in the 18 segments. The vowel triangle /i-u-a/ is easily recognized for all three speakers, although there are large individual differences. These differences stem largely from overall speaker-dependent characteristics. As was already discussed in section 2.3.2, we can apply a simple speaker-dependent correction by centring the data; this means that the average spectrum per speaker (see Fig. 3.5.3) is subtracted from all his 10-msec spectra. This correction is applied before data reduction. One ar-

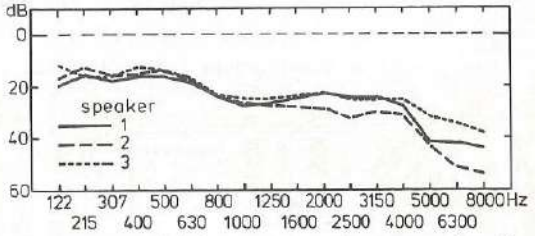


Fig. 3.5.3. Average one-third octave vowel spectra for the three speakers. The levels in the bandfilters are represented relative to the overall level.

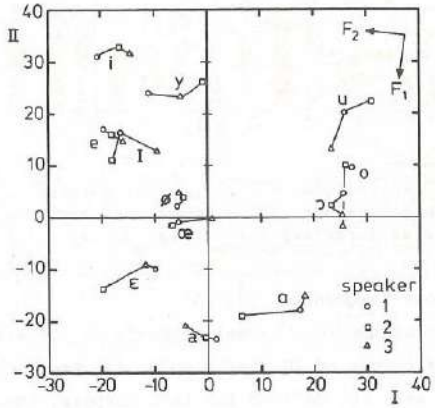


Fig. 3.5.4. Similar to Fig. 3.5.2, but now the average *centred* vowel positions per speaker are represented. The analogy with the F_1-F_2 representation is indicated by the two arrows.

rives at the same result by subtracting the overall average coordinate values per speaker in the subspace from the coordinate values per sample, so after data reduction. The resulting average centred vowel positions are given in Fig. 3.5.4. The between-speaker variation is considerably reduced but of course not completely eliminated. This may be due partly to the relatively simple uniform transformation (Fant, 1975), partly to the dynamic variation within the vowel not having the same effect on the average vowel position of different speakers.

If one prefers describing the spectral differences between vowel sounds in terms of formant parameters, the dimensional spectral representation of Fig. 3.5.4 can easily be "interpreted" as an F_1-F_2 representation. Without going into a matching procedure (Pols *et al.*, 1973), we roughly indicate in Fig. 3.5.4 how the F_1 and F_2 axes are oriented. Fig. 3.5.5 gives the average centred vowel positions in the plane of the third and fourth dimensions of the vowel subspace. The variation between the vowel points becomes larger relative to the overall variation, but still some vowels or vowel groups can be distinguished.

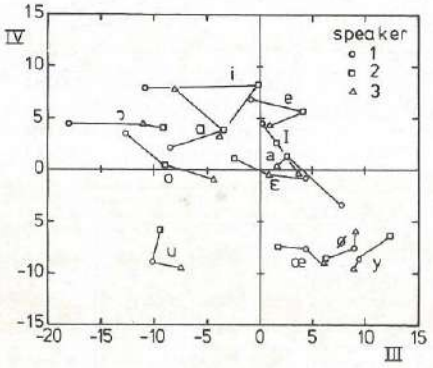


Fig. 3.5.5. Average centred vowel positions per speaker in the plane of the third and fourth dimensions of the vowel subspace. The scale of this figure is twice as large as that of Fig. 3.5.4.

3.5.3. Duration of the vowel segments

In Table 3.5.1 the durations of all vowel segments as isolated from the context are given in terms of number of 10-msec samples for the three speakers. Since the vowel segments were not defined for this purpose, these numbers have to be regarded as mere indications of vowel duration, not as accurate values. Nevertheless, the average durations of the vowel segments per speaker, given in Fig. 3.5.6, clearly separate the diphthongs plus the long vowels /a/, /e/, /o/,

		ɑ	a	ɛ	ɪ	e	i	ɔ	o	u	œ	ø	y	ɑu	ei	ʌy
p	1	9	33	6	10	22	8	3	15	8	10	18	13	12	21	32
	2	9	21	7	8	14	6	7	20	8	9	18	15	13	13	21
	3	9	10	8	7	12	7	9	14	5	9	14	11	16	16	22
t	1	12	16	9	5	18	7	11	14	8	4	21	6	21	24	27
	2	11	18	11	8	13	8	12	14	7	8	19	11	19	21	17
	3	8	13	15	6	14	7	10	10	9	9	12	11	18	15	17
k	1	6	23	9	10	18	9	5	19	6	6	26	20	20	23	18
	2	7	21	12	10	13	7	8	17	10	8	13	13	13	16	15
	3	11	13	7	11	12	11	9	11	9	8	16	10	15	18	16
b	1	8	28	10	8	18	21	9	19	8	9	11	16	17	19	22
	2	7	23	6	6	11	12	11	19	7	9	16	15	15	13	15
	3	12	15	9	9	16	10	11	12	8	11	15	10	17	17	16
d	1	9	24	10	7	15	11	10	16	7	10	15	7	28	19	23
	2	8	20	8	8	16	8	9	16	4	9	16	8	16	12	16
	3	9	16	9	8	15	14	10	11	9	8	10	9	18	16	19
f	1	3	19	5	7	15	5	7	19	8	7	17	9	20	24	18
	2	5	16	8	6	9	8	7	16	7	6	11	7	15	19	12
	3	10	18	10	8	23	7	11	14	9	8	14	10	16	19	19
s	1	8	24	7	7	21	21	10	21	8	10	18	11	24	21	18
	2	13	21	8	7	13	15	10	12	7	9	16	9	17	19	14
	3	10	18	7	7	16	10	8	10	10	9	12	13	18	17	16
x	1	9	19	4	9	17	14	10	17	8	6	22	11	19	22	19
	2	9	13	5	5	14	9	7	11	6	10	16	8	13	15	13
	3	11	17	8	8	16	12	10	13	9	10	16	11	17	15	18
v	1	10	14	8	9	17	9	14	16	7	10	21	17	22	25	21
	2	11	18	13	7	18	7	8	16	8	7	20	13	18	17	20
	3	13	17	16	11	16	13	13	12	10	10	15	13	19	18	18
z	1	4	22	10	10	21	8	6	16	5	10	17	24	16	21	22
	2	7	18	10	-	18	6	7	13	7	6	14	13	14	17	18
	3	9	13	9	7	14	9	7	13	7	12	14	14	18	17	18
h	1	5	25	10	9	17	6	7	23	8	5	14	7	16	21	17
	2	7	17	10	7	14	9	7	17	7	6	9	7	16	16	14
	3	7	16	7	9	15	9	9	9	10	8	14	11	15	18	15
w	1	9	21	9	6	15	11	9	20	7	8	15	9	19	21	22
	2	8	18	10	6	10	7	9	17	7	8	15	9	12	19	19
	3	9	16	12	9	16	6	11	15	10	9	13	8	17	21	17
j	1	5	29	3	8	18	8	3	14	10	7	14	6	19	26	14
	2	7	19	6	8	14	6	9	18	7	6	10	6	14	18	17
	3	9	19	8	8	13	10	14	10	8	8	13	8	17	19	18
l	1	7	25	4	7	14	9	7	16	11	9	24	21	25	21	22
	2	7	16	10	6	14	9	7	12	12	9	17	14	15	16	14
	3	10	16	12	7	15	10	8	13	11	8	17	13	17	17	17
r	1	7	27	12	7	15	6	7	21	5	10	19	7	22	-	16
	2	9	22	9	9	14	6	7	16	6	9	14	9	16	14	14
	3	10	15	8	7	15	9	12	13	10	8	17	10	20	18	20
m	1	9	30	13	8	19	17	6	19	6	10	20	11	19	25	26
	2	9	24	11	6	15	13	8	23	11	11	18	10	15	17	19
	3	9	14	12	7	13	12	8	14	12	9	16	12	17	18	19
n	1	6	20	9	9	18	7	9	14	10	5	21	11	12	16	22
	2	8	16	7	8	13	8	8	21	10	7	15	13	17	14	17
	3	7	15	9	9	15	7	10	11	7	8	17	17	17	16	17
.	1	6	10	7	6	22	5	4	16	8	7	24	15	17	26	18
	2	8	20	8	10	14	10	9	11	18	8	16	13	13	17	14
	3	10	17	7	9	15	9	8	14	15	7	13	9	16	18	14
AVERAGE	1	7.3	22.7	8.1	7.9	17.8	10.1	7.6	17.5	7.7	7.9	18.7	12.3	19.3	22.1	20.9
	2	8.3	18.9	8.8	7.4	13.7	8.6	8.3	16.1	8.3	8.1	15.2	10.7	15.1	16.3	16.1
	3	9.6	15.4	9.6	8.2	15.1	9.6	9.9	12.2	9.3	8.8	14.3	11.1	17.1	17.4	17.6

Table 3.5.1. Durations of the isolated vowel segments in the 270 CVC words in terms of number of 10-msec samples. All values are given for the three speakers. The individual as well as the average values per vowel per speaker are given. The two dashes indicate mispronunciations.

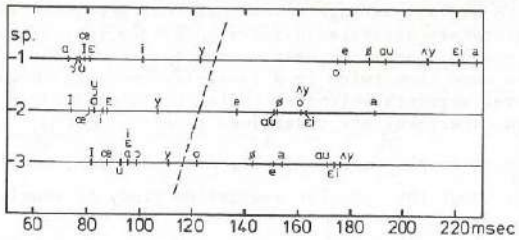


Fig. 3.5.6. Average durations of the vowel segments per speaker. The dashed line separates diphthongs and long vowels from short vowels.

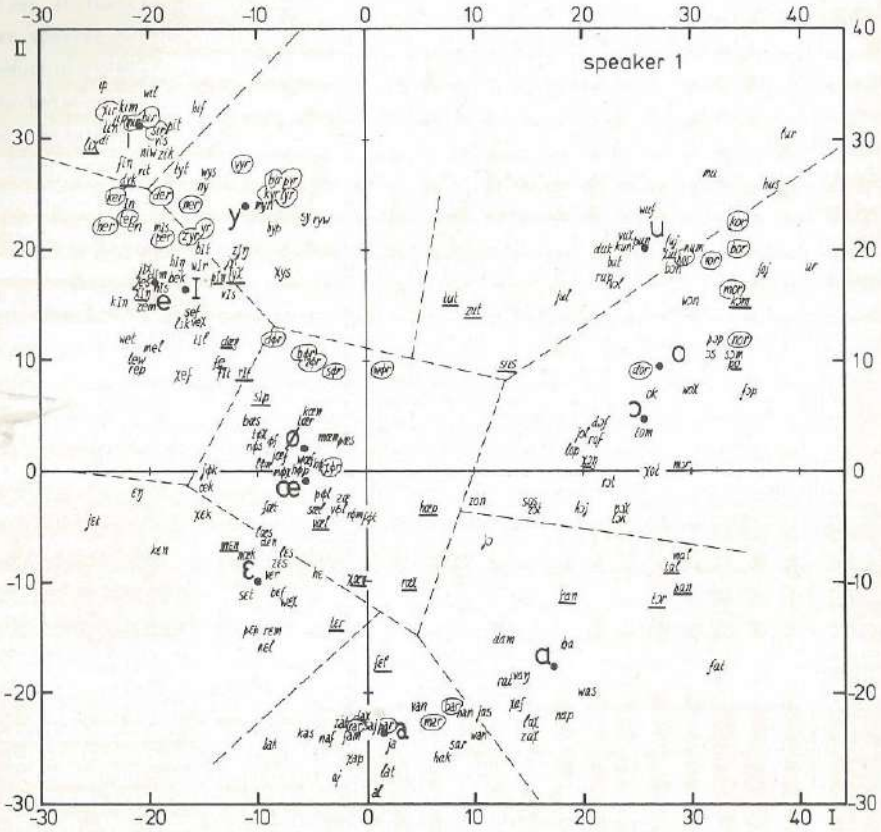


Fig. 3.5.7. Average centred positions of all individual vowel segments from speaker 1 in the two-dimensional vowel subspace. In order to specify the consonant context in which every vowel occurred, the average positions of the vowel segments are indicated by the names of the CVC words. Also the overall average vowel positions are given. The different vowel regions are separated by straight lines, which have been drawn by eye and only serve an easy search for specific words. If a vowel occurred more than twice in a final-/r/ context, those words are encircled. Vowel segments which were often misidentified in a subsequent identification experiment are underlined (see p. 115).

and /ø/ from the short vowels /ɑ/, /ɛ/, /ɪ/, /i/, /ɔ/, /u/, and /œ/ with perhaps some ambiguity about the /y/. For a detailed study of vowel duration in Dutch, see Nooteboom (1972).

3.5.4. Average positions of the vowel segments

Fig. 3.5.4 only gave the average vowel positions per speaker; Figs. 3.5.7, 3.5.8, and 3.5.9 give the average centred positions of all the individual vowel segments for each of the three speakers. [Because in this chapter, and in the next one, I will often refer to Fig. 3.5.7, this diagram is also enclosed as a separate figure inside the back cover.] The different vowel regions are separated by straight lines. These lines have been drawn by eye and only serve an easy search for specific words; they have no statistical significance in contrast with the equal-probability lines used earlier (Klein *et al.*, 1970).

We used the positions averaged over the whole available dynamic pattern of the vowel segments, and not the middle and/or the outer positions of the vowel segments as done by most researchers (see section 1.3.3). This choice was made in order to get a single measure which is more stable than one or two values at

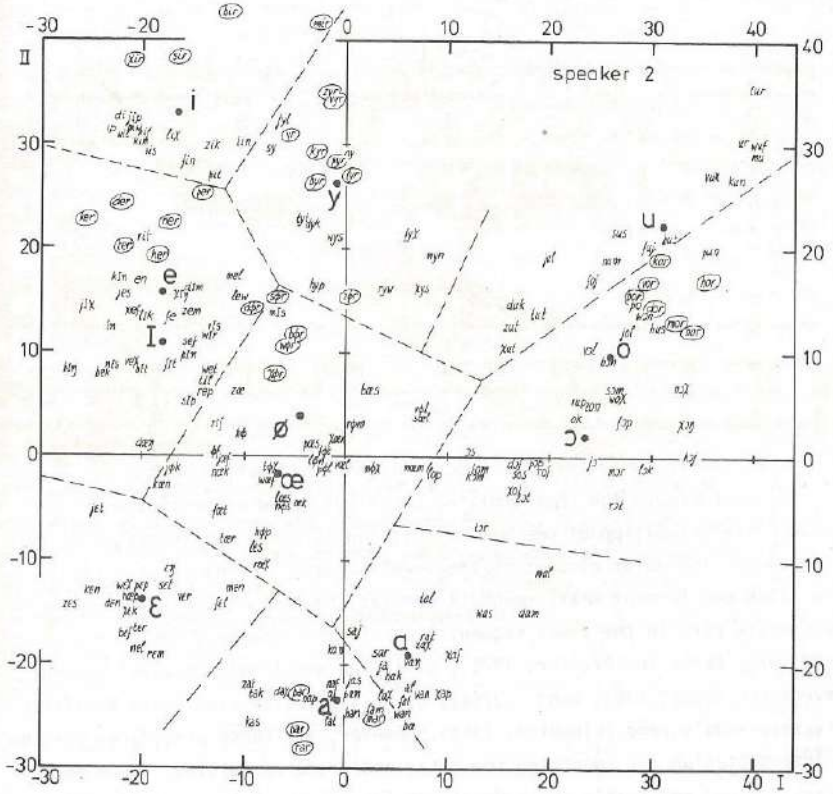


Fig. 3.5.8. Similar to Fig. 3.5.7, but now for speaker 2.

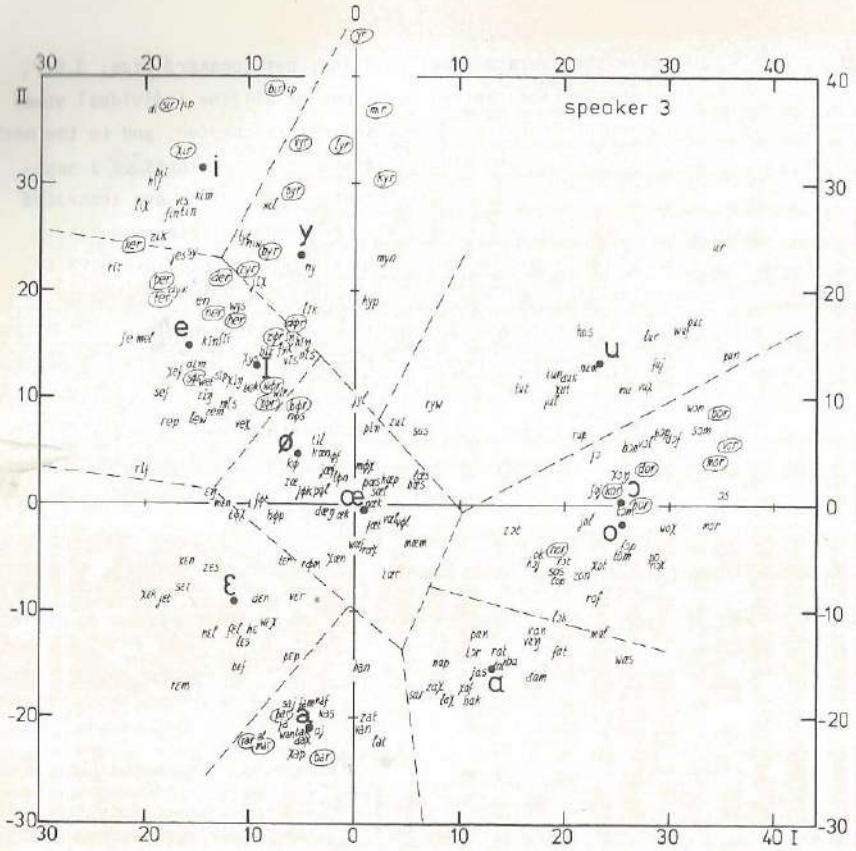


Fig. 3.5.9. Similar to Fig. 3.5.7, but now for speaker 3.

specific points in the utterance. Especially the information at the outer points strongly depends upon the choice of the position of the segment boundaries. Moreover, the definition of one middle position is quite arbitrary: it certainly is not always the point closest to the vowel target (Stevens *et al.*, 1965). Quite often one formant measurement is done at the most stationary part or the steady-state part in the vowel segment (Lehiste and Peterson, 1961; Ohde and Sharf, 1975; Earle and Pfeifer, 1975), or in the middle of the vowel segment (Stevens and House, 1963; Bond, 1976a), or at a point in time where the first derivative equals zero (Lindblom, 1963). However, all these procedures used human interpretation for smoothing the data, and/or curve fitting. Since we had already defined relatively short durations for our vowel segments (p. 80) it

seemed to be better to use the average spectral information over the whole vowel segment. This information is perhaps best comparable with that at the midway or steady-state position in other studies. Apart from this average spectral information, the dynamic spectral variation over the whole segment is still available for a more detailed look. We will use this dynamic information when discussing the spectral representation of the diphthongs (3.6) and the identification results (4.3) in more detail.

It is of interest to know how well the vowels can be recognized on the basis of only the average positions of the vowel segments in the two-dimensional vowel subspace. For that purpose we calculated for the average position of all vowel segments the Euclidean distances between those points and the 12 average vowel positions for the same speaker. A vowel was supposed to be correctly classified

stimulus	response									number of errors			
										speaker			total
	ɑ	a	ɛ	I,e	i	ɔ,o	u	œ,ɸ	y	1	2	3	
ɑ	43/42	8/9				3/3				6/5	4/6	1/1	11/12
a	1/6	53/48								0/0	1/6	0/0	1/6
ɛ		0/1	54/50					0/3		0/1	0/1	0/2	0/4
I				49/46	0/1			4/5	0/1	1/4	1/1	2/2	4/7
e				54/51	0/2				0/1	0/1	0/0	0/2	0/3
i				2/1	52/51				0/2	0/0	1/1	1/2	2/3
ɔ	3/4					50/47	0/3	1/0		2/4	1/0	1/3	4/7
o	2/0					51/41	1/11	0/2		2/7	1/5	0/1	3/13
u						7/8	45/43	2/2		0/1	7/7	2/2	9/10
œ			2/5	0/1				52/48		1/4	1/2	0/0	2/6
ɸ				4/8				50/56		0/1	0/2	4/5	4/8
y				0/4	1/5			2/2	51/41	1/4	0/0	2/7	3/11
								Total number of errors		13/32	17/31	13/27	43/90
								% correct 17 dim		94.0	92.1	94.0	93.4
								% correct 2 dim		85.2	85.6	87.5	86.1

Table 3.5.2. Confusion matrix, accumulated over the data of the three speakers. Identification is based on shortest distance to the average vowel positions per speaker in the original 17-dimensional space (left numbers), and in the two-dimensional subspace (right numbers). Overall number of errors per vowel and per speaker are also given, as well as the total number of errors and the total percentage-correct scores.

if the distance to the corresponding average vowel position was the shortest of all twelve distances. Confusions between /I-e/, /ɔ-o/, and /œ-ø/ were not counted, since durational information was not taken into consideration. The "confusion" matrix obtained in this way, accumulated over the three speakers, is given in Table 3.5.2. The overall correct score is 86.1%. Confusions based on the distances in the original 17-dimensional space are also given; the correct score then rises to 93.4%. Some improvement may be expected by using, instead of the shortest-distance concept, maximum likelihoods.Especially for those vowels, like /u/, for which the points scatter in one specific direction, and not uniformly in all directions, this could give a major improvement. However, the number of data per vowel is rather limited and an independent test set is not available. For comparison, we may mention that the correct score for average positions of vowel segments from a fixed context of h(vowel)t, from 50 different male speakers, represented in a two-dimensional vowel subspace, was 88.0% (Klein *et al.*, 1970). This score rose to 97.7% when it was based on maximum likelihoods and speaker-normalized data in six dimensions. Although the data sets are not directly comparable, and the score calculations are not the same, one gets the impression that one speaker in different contexts causes at least as much variation as, but probably more variation than, different speakers in a fixed context,after speaker normalization.

3.5.5. Vowel-coarticulation effects

In the preceding sections we have described how we derived the vowel subspace, the average vowel positions in this subspace, and the average position of each vowel segment. In this section we should like to see how the positions of these vowel segments vary with context, and whether there is some systematic effect related to the vowels, and/or the consonant environment.

Figs. 3.5.7 to 3.5.9 show that there is indeed considerable variation between the different average positions of the segments per vowel. The total variation can be represented by the spread σ, equal to the square root of the variance, per dimension per vowel. Fig. 3.5.10 gives these σ's for the three speakers for the first three dimensions.

One must realize that these values do not just represent the spread in the 18 average positions of the segments per vowel, plotted in Figs. 3.5.7 to 3.5.9, but also include the spread of the individual samples per vowel segment. Therefore, a small σ means that both the variation within the samples of all 18 segments, as well as the variation between the average positions of the segments, is small for that vowel. If we look at the total σ in the first two dimensions

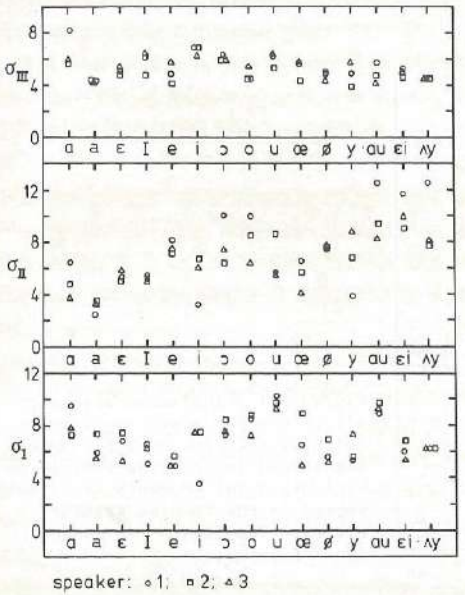


Fig. 3.5.10. The total spread σ per vowel along the first three dimensions for the three speakers.

($\sqrt{\sigma_I^2+\sigma_{II}^2}$) we see that for all three speakers this σ is small for vowel /a/, for speaker 1 also for /i/ and /y/, for speaker 3 also for /œ/. The vowels /ɑ/, /ɔ/, /o/, and /u/, and of course also the diphthongs, have a relatively large σ for all three speakers. The third dimension does not give much specific information.

As was said above, the total spread is built up from two sources. These two values, and the total spread, which is the same as the one represented in Fig. 3.5.10, are given per vowel in Figs. 3.5.11 to 3.5.13, for the three speakers, respectively. One source is the spread in the 18 average positions of the segments per vowel. This spread is always smaller than the total σ, the more so for vowels which show a large variation in the segments themselves. This second source, the actual spread within the segment traces per vowel, can be calculated by applying some sort of vowel normalization (it is not simply equal to the difference between total variance and variance between averages, because the number of samples per segment is not equal): the differences between the average positions of the different segments per vowel are eliminated by translating all segment traces in the vowel subspace in such a way that these average positions coincide. One could call the result the "pooled segments". The variance in these

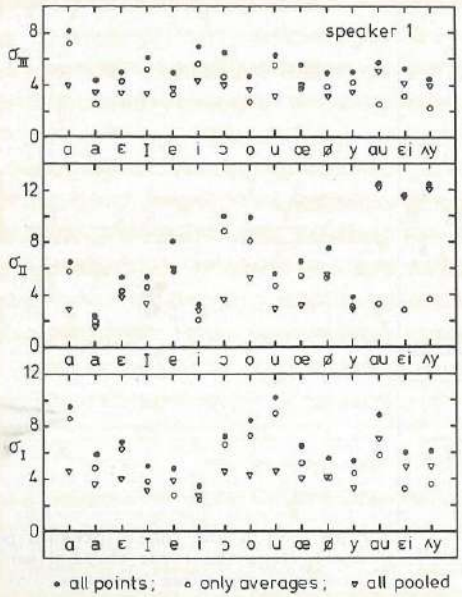


Fig. 3.5.11. The total spread σ for all points per vowel, per dimension, split up into two sources, namely the spread in only the average positions of the segments, and the spread in the "pooled segment" positions. The figure represents the data of speaker 1.

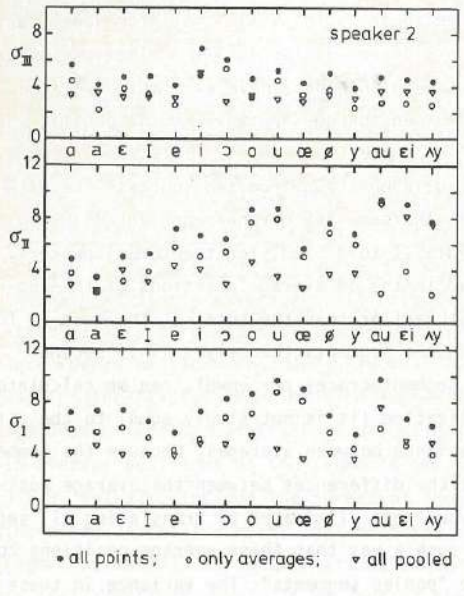


Fig. 3.5.12. Similar to Fig. 3.5.11, but now for speaker 2.

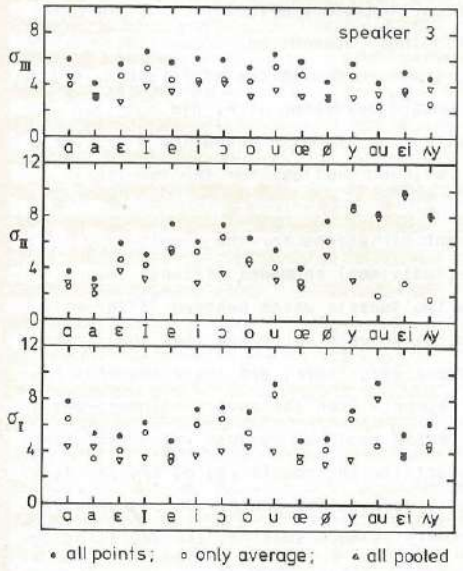


Fig. 3.5.13. Similar to Fig. 3.5.11, but now for speaker 3.

pooled sample points per vowel is then equal to the variance within the segments only. The square root of this variance is the spread σ, represented in Figs. 3.5.11 to 3.5.13 by triangles. For vowel sounds with strong spectral variations within the segments this σ will be large. This effect is strongest for the three diphthongs /ɑu, ɛi, ʌy/, although for the latter two diphthongs the variation is mainly along the second dimension. The diphthongs are discussed in more detail in paragraph 3.6. The next largest segment variation is in the vowels /e, o, ø/, again along the second dimension. This variation would have been even larger if the durations of these isolated segments had been made longer. In that case the diphthong-like nature of these three Dutch vowels, which could be phonetically transcribed as /Ii/, /Ɔu/, and / œy/, respectively, would have been even more pronounced.

Large spectral variation is not simply related to duration of the vowel, since the long vowel /a/ has very little spectral variation both within and between the different vowel segments.

Not just /a/, but also /ɛ, I, i/ show relatively little variation, representing, as they do, stable vowels with little coarticulation. The same can be said of /œ , y/, but not for all three speakers. The variation in the average vowel positions of speaker 1 is largest for /ɑ, Ɔ, o, u/.

Ohde and Sharf (1975) found greater coarticulatory effects on /u/ than on /i/, which is not contradictory to our findings. Stevens and House (1963) arrived at the same result by looking at the standard deviations of F_1 and F_2 of all vowels in all different contexts. The standard deviation of F_1 did not change markedly from one vowel to another. However, the standard deviation of F_2 was greatest for the rounded vowels /u/ and /ʊ/, and smallest for the vowels /i/, /æ/, and /ɑ/.

Because of the relatively large segment variations for the vowels /e, o, ø/ we looked in somewhat more detail at the individual segments of these vowels. Apparently the data could be split up into two subsets which behaved differently. The underlying parameter was the presence or absence of /r/ as the final consonant. Almost all average positions of the /-er/, /-or/, and /-ør/ segments have extreme positions relative to the other segments; see the encircled positions in Figs. 3.5.7 to 3.5.9. This effect could become manifest because /r/ quite often occurred in the word list as final consonant for the vowels /e, o, ø/. /a, i, y/ also have a number of segments with final /r/, but, for these vowels the variation in the average positions of the segments is much smaller, the deviating positions for /-ar/, /-ir/, /-yr/ being somewhat less pronounced.

Fig. 3.5.14 illustrates this *"r-effect"* for speaker 1. His vowel segment-positions, given in Fig. 3.5.7, have been split up into non-/-r/- and /-r/-segments. The positions averaged over these two groups of segments are indicated per

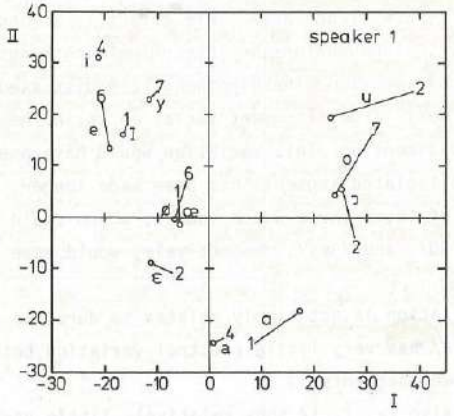


Fig. 3.5.14. Illustration of the r-effect for speaker 1. The circles represent the average vowel positions for those segments which were isolated from contexts *without* a final /r/. The numbers, on the other hand, represent the segments *with* final /r/. The numbers themselves indicate how many segments were available to determine these average positions.

vowel. The circles represent the non-/-r/ average vowel positions and the numbers the /-r/ average vowel positions. The numbers themselves indicate how many /-r/-segments were available to determine this average position. The stable parts of the /-r/-vowel segments for /e/, /o/, and /ø/ clearly have a tendency to take extreme vowel positions, away from the neutral vowel. For speaker 1 this tendency is less pronounced for the vowels /i/ and /y/, but the other two speakers show the effect clearly, also for these two vowels, see Figs. 3.5.8 and 3.5.9. The final /r/ has hardly any influence on the position of the stable part of the vowel /a/. Although for the other vowels the number of words with an /r/ was too small to give reliable results, the effect of /r/ on /u/ and /ɔ/ also appears to be consistent for all three speakers.

The tendency of the stable vowel parts, segmented from C_1Vr words, to take extreme vowel positions has, up to now, only been illustrated in the two-dimensional vowel subspace. Although, statistically, this plane represents most of the variation in the data, it is wise to test now and then if not too much specific information is lost by the reduction from 17 to two dimensions. For that purpose we calculated for the average positions of all vowel segments the distances between those points and the 12 average vowel positions, not only in two dimensions but also in the original 17 dimensions. The distances in two dimensions give in numbers what was already visible in the two-dimensional spectral representation of Fig. 3.5.7. The distances in 17 dimensions can tell us, for the original data, how every segment is positioned relative to the average vowel positions. Without any doubt, the so-called r-effect is also present in the original 17-dimensional representation. This means that in 17 dimensions, too, the vowel segments from the /-r/-context do have extreme positions relative to the same vowel segments in other contexts.

Table 3.5.3 is an illustration of this effect for the vowel /e/: all /-r/-segments have shorter distances to the /i/, which means more extreme positions, than all non-/-r/-segments. Results for /o/ and /ø/ are similar.

For the two subsets of /e, o, ø/ segments, namely those originating from words with and without final /r/, the average segment traces were determined. These patterns are represented in Fig. 3.5.15. That the /-r/-segments are indeed very stable is nicely illustrated here. This stability is not just apparent for the average segments, but it is equally present in the individual segments; this can be seen from the σ's given in Table 3.5.4 for the /-er/-segments. For other vowels these σ's are similarly small.

The much larger variation in the non-/-r/-segments (thick lines in Fig. 3.5.15) shows that these vowels are of a diphthong-like nature. In the next

word	in 17 dimensions			in 2 dimensions		
	I	e	i	I	e	i
per	12.0	8.9	16.5	6.0	4.9	9.5
ter	13.1	8.8	16.2	8.7	6.6	8.3
ker	16.3	12.4	13.5	11.0	9.0	6.3
der	13.7	11.5	15.9	8.6	7.7	6.8
her	13.2	9.3	17.7	9.4	6.9	9.6
ner	15.7	14.4	15.1	7.8	8.0	8.4
bek	5.4	4.4	24.3	1.8	1.6	14.0
fe.	11.6	14.9	35.0	7.8	9.9	23.1
sef	5.5	7.1	27.9	2.2	4.3	17.6
vex	4.8	7.0	27.3	2.6	4.7	18.0
zem	9.1	7.6	27.4	4.3	2.5	16.4
xef	10.3	11.5	32.6	7.8	8.7	22.9
mel	7.9	9.2	27.7	6.2	5.9	20.1
lew	10.1	9.8	30.8	8.0	7.3	21.2
rep	10.0	9.3	29.1	8.2	7.4	21.3
wet	9.0	7.7	28.4	7.0	5.6	19.2
jes	9.4	5.5	23.5	3.9	1.1	14.6
.en	13.3	9.2	19.2	7.7	5.7	8.9

Table 3.5.3. Distances between the average positions of the segments of the indicated words and the overall average vowel position of /I, e, i/. Distances are given both in the 17-dimensional and the two-dimensional space. Words with and without final /r/ are grouped together. For the first group of segments, the distance to /i/ is always shorter than for the second group of segments.

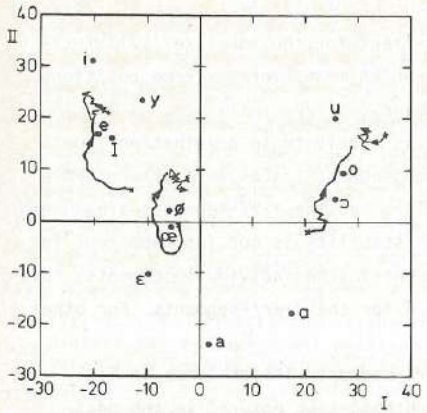


Fig. 3.5.15. Average traces for all linearly time-normalized /e/, /o/, and /ɸ/ segments which were isolated from contexts without a final-/r/. The clustered traces (thin lines), on the other hand, represent averages of the stable segments isolated from the final-/r/ contexts. The data are from speaker 1. The overall average vowel positions are also indicated.

word	nr of samples	σ_I	σ_{II}
per	22	2.2	2.6
ter	18	1.9	1.7
ker	18	1.5	2.4
der	15	1.6	1.0
her	17	1.8	2.4
ner	18	4.0	1.3

Table 3.5.4. Variation in the initial stable parts of the vowel segments isolated from /-er/ words. The variation is expressed in the spread σ along the first and second dimensions in the vowel subspace.

chapter we will see that this also influences identification.

In the literature, one finds little information about the r-effect. In a recent article, Kameny (1975) gives the positions of retroflexed and nonretroflexed American-English vowels in the formant space. However, the formant measurement was only done at one point in the vowel. Furthermore, the individual differences were quite large. Kameny gives suggestions for constructing a retroflexed vowel space from a nonretroflexed vowel space, but as long as dynamic vowel behaviour is not taken into account, this does not make much sense. Koopmans (1969) qualitatively described the Dutch vowels followed by /r/ as having the characteristics of being longer, and being fairly constant in quality until a change to the neutral vowel /ə/ begins. In our original segmentation procedure we did not include this change to /ə/ in the vowel segments followed by /r/. What then remained was a stable representation of the vowel which happened to be quite extreme, see Fig. 3.5.14.

In order to get a better idea of what actually happens in the complete vowel segments with a final /r/, we now looked at the spectral pattern of the *total* vowel segments. As was said before, the vowel segments isolated originally in these words with final /r/ only contained the initial stable part of the vowel. These vowel parts of the /-er/ words are represented as circles in Fig. 3.5.16. The "tails" attached to these circles represent the successive samples, following the initial stable part, until the rattle of the /r/ is reached in the word. All /-er/ words reveal a similar characteristic pattern. By using a linear time normalization to the average number of 15 samples per "tail", and an averaging of these traces, we get the /-er/ average trace plotted in Fig. 3.5.17. In the same way the dynamic traces for the vowels /i, y, ɸ, a, o/ were processed, re-

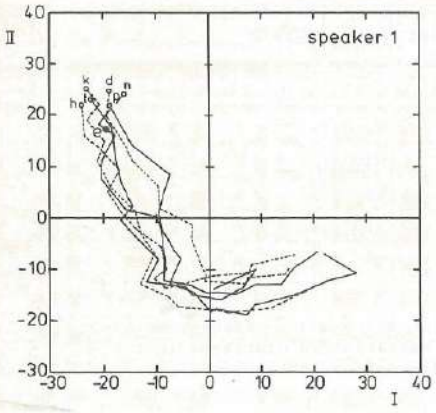


Fig. 3.5.16. Representation of the complete vowel traces from /-er/ words of speaker 1. The originally isolated stable vowel segments in these words are represented by circles; they are the same as the encircled /-er/ positions in Fig. 3.5.7. The successive samples, following these initial stable parts represent the rest of the total vowel segment until the rattle of the /r/ is reached.

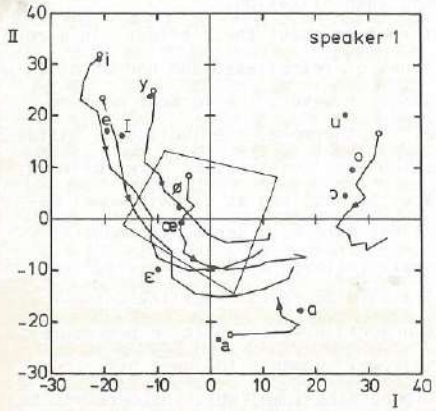


Fig. 3.5.17. Average vowel traces for the /-ar/, /-er/, /-ir/, /-or/, /-ϕr/, and /-yr/ words of speaker 1. Average traces are achieved after a linear time normalization of the individual samples to a number of 15. For reference, the overall average vowel positions and the /œ-ϕ/ region from Fig. 3.5.7 are also indicated.

sulting in the other average traces in this figure. The number of words for the remaining vowels was too small to do the same. The traces show that the description given by Koopmans (1969), in terms of a stationary vowel part followed by a change to the /ə/, is too simple for this speaker. The first part is indeed stationary but at an extreme vowel position, and with a duration of about 160 msec for /i, e, ϕ, o/; /a/ is somewhat longer, and /y/ somewhat shorter, see Table 3.5.1. Then, quite abruptly, a spectral change starts in the direction of something like an "/r/-locus". This is not simply a movement to /ə/ but goes for /i/ via /I/ and /ɛ/; for /e/ via /ɛ/; for /ϕ/ via /œ/; for /o/ via /ɔ/; and for /a/ in the direction of /ɑ/. This is necessarily a speaker-specific description related to the speaking habits of the speaker. The patterns for speaker 2 are very

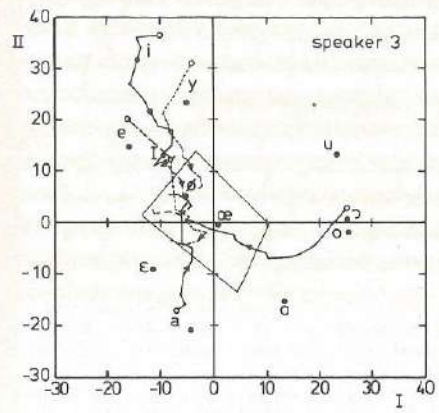


Fig. 3.5.18. Similar to Fig. 3.5.17, but now for speaker 3.

similar to those given in Fig. 3.5.17. However, the patterns for speaker 3 are quite different, see Fig. 3.5.18. He has no clear tongue-tip or rattle /r/ but he uses the back of his tongue. His traces are more like the stylized ones given by Koopmans (1969), namely all directly pointing to the central /œ/.

It is not possible to study the dynamic spectral effect of other preceding or following consonants on the vowels in a similarly systematic way as with the vowels followed by /r/, because not enough replications are available in this study.

Therefore, we only want to point out here some possible coarticulation effects which are suggested by the data, but which have to be tested in greater detail. For all three speakers, the /u/'s from /tut/, /zut/, and /sus/ have a centralized position. This is in agreement with a higher F_2 found by Stevens and House (1963) midway in the /u/ for symmetric contexts of postdentals /t, z, s/. For all three speakers, the /ɔ/ segment from /tɔr/ has a position more similar to /ɑ/, which may very well be the r-effect again. The /ɑ/ from /mɑl/ has a position close to /ɔ/. The influence of final liquids and glides seems in general to be large, see for instance also /y/ from /ryw/, and /ɛ/ from /fɛl/. This evidently also depends on the chosen vowel boundary. The nasals also seem to have a strong effect on the average vowel positions, see for instance /mu/, /mœm/, /myn/; this effect will be much stronger in running speech, where vowels can actually become nasalized.

Because of the extensive statistical study on formant trajectories for the vowel /I/ in CVC syllables by Broad and Fertig (1970), we have also studied in some more detail our data concerning this vowel. The average number of samples

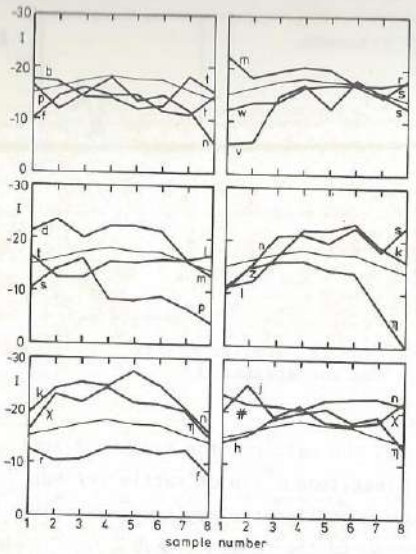


Fig. 3.5.19. Variation as a function of sample number along the first dimension of the vowel subspace, of the 18 different /I/ segments. The initial and final consonants surrounding each vowel segment are indicated. All individual segments are linearly time-normalized to a number of eight samples. In each graph, the overall average variation in the /I/ segments is also drawn as a thin line for reference.

in our isolated /I/-segments is eight. Therefore, we time-normalized in a linear way all these segments to eight samples for easier comparison. Fig. 3.5.19 gives the spectral variation over these samples, in terms of the coordinate value along the first dimension, for the various consonantal contexts. The average spectral variation is always drawn in for reference. This first dimension can roughly be interpreted as F_2 (compare Fig. 3.5.4). The general variation has a convex form which is also found for the F_2 transition functions by Broad and Fertig, although they indicate a minor concaveness in the final transition function for certain consonants. Our data show that the transition function generally goes down if the vowel is preceded or followed by a nasal. Stevens and House (1963) did not include nasal consonants "because of the difficulties of measuring formant frequencies for nasalized vowels". The effect found by these authors for the other consonants at the midpoint of the vowel appeared to be strongly related to the place of articulation of the consonants. They differentiated the labials (/p, b, f, v/), postdentals (/θ, ð, s, z, t, d, č, ǰ/), and velars (/k, g/). The effect on F_2 of the labials and the postdentals for the front vowels /I, ɛ, æ/ was considerable, as was the effect of F_2 by the postdentals for the back vowels /u, ʊ, ʌ/. These systematic effects are not so clearly reflected in the study by Broad and Fertig on /I/, nor are they in our data. There is also some divergence between their results and those of Stevens *et al.* (1966) regarding F_2 at beginning and end of the vocalic portions of C_1VC_1 syllables. Our approach, if applied to an optimally chosen data set, could give more information about these divergences, but the present data are too limited to be conclusive in this respect.

Only measurements on more replications of the same context, and more different C_1-C_2 combinations to separate the effects of initial and final consonants, will make it possible to specify clear effects like those with the final /r/. It will be clear that the analysis system used in this study can be an efficient tool for studying vowel coarticulation effects and dynamic vowel characteristics. Additionally, there is a great need for perceptual data, like those described in Chapter 4, to evaluate the significance of variations found in the vowel spectra.

3.6. DIMENSIONAL SPECTRAL REPRESENTATION OF THE DIPHTHONGS

Apart from the 12 Dutch monophthongs in our list, we also included the three Dutch diphthongs /ɑu/, /ʌy/, and /ɛi/ in our word material. The isolated vowel segments from these words can of course not be represented by a single average position in the spectral subspace since diphthongs essentially are non-stationary by nature. As with the monophthongs, these dynamic patterns are influenced to some extent by the surrounding consonants. However, the variation in any of these diphthongs is so large that we could hardly identify consonant-specific spectral effects in this limited material. Gay (1968) found no spectral effects either, only a duration effect for voiced final consonants.

Instead we wanted to study the average diphthong behaviour. The average num-

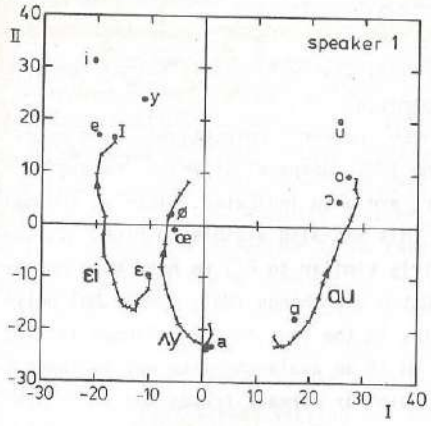


Fig. 3.6.1. Average traces for the diphthongs /ɑu/, /ʌy/, and /ɛi/ of speaker 1 in the two-dimensional vowel subspace. All individual segments are linearly time-normalized to 18 samples. For reference, also the overall average vowel positions of this speaker are indicated.

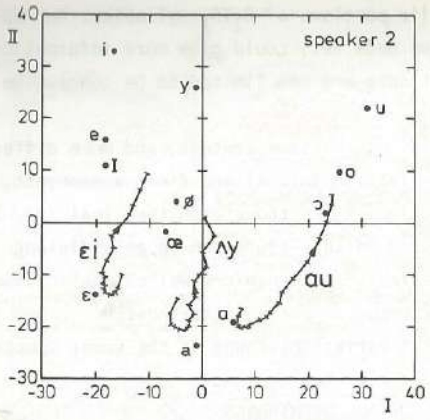


Fig. 3.6.2. Similar to Fig. 3.6.1, but now for speaker 2.

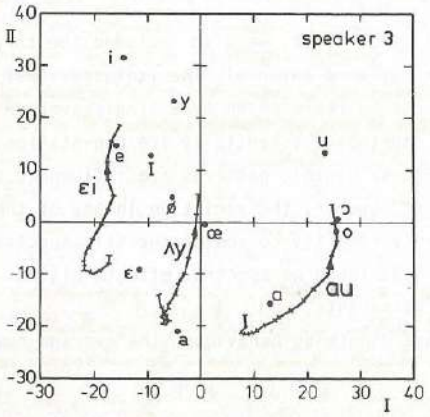


Fig. 3.6.3. Similar to Fig. 3.6.1, but now for speaker 3.

ber of samples in the diphthong segments was 18, so we linearly normalized all segments to 18 samples for an easier comparison.

Figs. 3.6.1 to 3.6.3 give, for the three speakers, respectively, the average traces for /ɑu/, /ʌy/, and /ɛi/ in the I-II subspace, in which the average vowel positions, which differ per speaker, are also indicated. Clearly, the main variation is along the second dimension; this was also visible in Figs. 3.5.10 to 3.5.13. As the second dimension is fairly similar to F_1, we have good agreement with known formant measurements on Dutch diphthongs (Mol, 1969). Mol only gives stylized formant traces, but explains in the text that F_2 changes little, whereas F_1 changes from a certain moment on in an avalanche-like way to lower values. Slis and van Katwijk (1963) found similar formant traces for their syn-

thesized two-formant diphthongs judged best by subjects. Cohen (1971) used these data in a discussion on the nature of diphthongs. He suggests accepting the Dutch diphthongs as a separate vowel class, recognizable as such and distinguishable from other long and short vowels, on account of their peculiar dynamic character. The fact that they can be synthesized by setting up two steady state vowel segments does not mean that a natural diphthong is biphonemic. Gay (1970) showed that good synthetic diphthongs can also be obtained by only generating a formant transition, without steady-state parts. He furthermore concluded that the rate of change of the second formant for the American-English diphthongs /ɔi, ai, au/ was more important than the onset and offset values. Spectrographic measurement on naturally spoken diphthongs showed that the formant position of the onset target is also a fixed feature of the diphthong formant movement, together with the second-formant rate of change (Gay, 1968). He did not make measurements of first-formant rates of change "because of possible measurement error effects". Gay supports the description of diphthongs as unit phonemes rather than as sequences of vowel plus semi-vowel, or vowel plus vowel.

Our data suggest that a diphthong can be described as quite a long steady-state onset part followed by a fast specific transition to an offset area where no steady-state part is necessary. The diphthong /ɑu/ starts at /ɑ/ and terminates at /ɔ, o/; /ɛi/ starts at /ɛ/ and goes to /I, e/; and /ʌy/ starts at /a/ and goes to /œ, ø/. So, none of the three Dutch diphthongs reaches the vowel position indicated in its phonetic transcription. 't Hart (1969) also observed this by listening to gated diphthongs. It has also been found for the American-English diphthongs /ɔi, ai, au/ by Gay (1968), and Holbrook and Fairbanks (1962). Lehiste and Peterson's (1961) analysis shows /au/ termination closer to, but not at /ʊ/. But since Gay showed the importance of speaking rate, the final position can be strongly influenced by that parameter too. We do not believe that speaking rate varied much over the isolated words spoken by our three speakers. Nevertheless there exist durational differences between the diphthongs in our material, see Table 3.5.1. Whether this is personal variation or influence of surrounding consonants cannot at this moment be concluded from our data. If a diphthong segment is shorter, this results most of the time in a shorter tail with still an indication of the direction to be taken. This supports Gay's (1968) idea that a diphthong is mainly determined by its onset area in combination with the rate of change of the dynamic part, whereas it is less important whether the final position is actually reached. Our data suggest that it is perhaps better to consider the direction of change rather than the rate of change. Fig. 3.6.4 gives three examples of /ʌy/ traces varying in duration from 180 to 260 msec, but all three

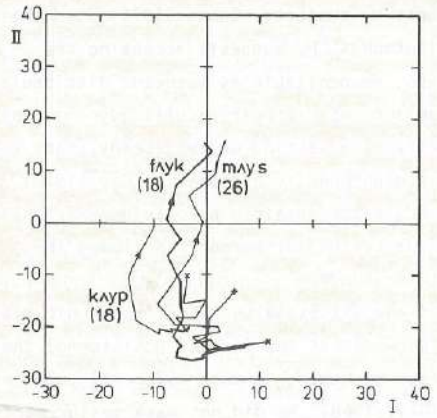


Fig. 3.6.4. Three examples of /ʌy/-traces varying in duration from 180 to 260 msec, but all three unanimously identified as /ʌy/ (see paragraph 4.3).

being almost unanimously identified as /ʌy/, as we will see in the next chapter. The segments from the words /fʌyk/ and /mʌys/ have a similar tail but a different number of samples in the onset area, whereas the tail of /kʌyp/ is shorter but apparently distinctive enough in its direction also to cause a clear /ʌy/ impression. The above description of the dynamic spectral variation in Dutch diphthongs does not give really new facts, but it confirms many "facts" which had not been measured systematically, or published, before. I refer to the stable starting position, to the fast specific transition, and to the offset area, as well as to the individual differences. The analysis and synthesis system used and the efficient data representation in a two-dimensional vowel space are especially useful here.

3.7. CONCLUDING REMARKS

In this chapter we have given a fairly elaborate description of the spectral differences between vowels as represented in the principal-components vowel subspace derived from the bandfilter spectra. This description was in terms of the average positions of the vowel segments as isolated from the 270 monosyllabic CVC words, often extended to a presentation of the dynamic spectral behaviour within the vowel segment in terms of a trace in the vowel subspace.

Purely on the basis of the average positions of the isolated segments, the vowels were recognized 93.4% correct in the original 17-dimensional space, and 86.1% correct in the two-dimensional subspace when a simple shortest-distance measure was used (Table 3.5.2). These scores concern the data on all three speakers. Corresponding long and short vowels /I-e/, /ɔ-o/, and /œ-ɸ/ are grouped together. One would expect that, in the identification experiment which will be described in the next chapter, at least similar scores will have to be derived since listeners will also have the dynamic variations at their disposal. On the other hand, it is quite possible that dynamic variation is the more important factor of the two.

As a measure of the variation per vowel, caused by different consonant contexts, we introduced the spread σ per dimension. A large σ for a certain vowel could represent a large coarticulation effect and hence cause many identification errors. A rank order of the vowels with respect to their spread in the first two dimensions is: /i, a, y, I, ɛ, œ, ɸ, e, ɑ, u, ɔ, o, ɛi, ʌy, ɑu/. One should, however, keep in mind that σ is not the only important parameter, but that the position of a vowel with respect to neighbouring vowels is also important.

A clear consonant-specific effect on the spectral vowel positions was found for those vowel segments that had been taken from words with final /r/. This may very well have an influence on the identification of these segments. It will also be interesting to see how the subjects react to the diphthong-like character of the /e, o, ɸ/ segments.

With respect to the diphthongs, we have the impression that these sounds have such a typical dynamic spectral behaviour, represented as traces in the vowel subspace, that identification will not be very difficult for the native listeners. This does not mean that we can simply present an algorithm for recognizing these segments automatically. Procedures for pattern matching of such traces are not readily available.

CHAPTER 4

IDENTIFICATION OF THE ISOLATED VOWEL SEGMENTS

4.1. INTRODUCTION

After having considered in the preceding chapter the spectral differences between isolated vowel segments, we shall discuss in this chapter how these vowel segments were perceived by subjects in an identification experiment.

There were certain reasons for choosing the identification paradigm for the perceptual task. First of all, it is a natural task which needs few instructions and is relatively easy to carry out even with phonetically naive subjects if orthographic instead of phonetic symbols as response categories can be used. Furthermore, the stimuli can be presented in their original form without the equalization of stimulus parameters like pitch, intensity, and duration that is necessary in discrimination experiments or similarity judgments. The human identification is, moreover, not unrelated to the problem of automatic speech recognition and the subjects' responses may shed some new light on that technical problem too.

It is interesting to know whether the isolated vowel segment in itself is still identifiable as the intended vowel, or whether it is modified so much by the context that identification becomes difficult. Once vowel segments have been misidentified, it is interesting to see whether the types of error correspond to the dimensional spectral representation of those segments.

The influence of one phoneme on the recognition of adjacent phonemes is usually studied in other ways, like:

(a) Identifying (partly) deleted phonemes, as a function of the cutting point in naturally spoken utterances (*cf*, *e.g.*, Ostreicher and Sharf, 1976);

(b) Identifying one phoneme in a context of other phonemes, in complete, synthesized, syllables of which the characteristics are systematically varied (*cf*, *e.g.*, Lindblom and Studdert-Kennedy, 1967);

(c) Identifying isolated vowels or vowels spoken in different consonant environments (*cf*, *e.g.*, Strange *et al.*, 1976).

In the literature review (section 1.5.3) we have seen that the problem of identifying vowel segments isolated from CVC syllables has hardly been studied. Only Kuwahara and Sakai (1973) investigated vowel identification in isolated CV syllables and in CV syllables taken from connected speech with more and more of the consonant part removed. Fujimura and Ochiai (1963) did an identification experiment with 50-msec vowel segments gated out of 100, quite arbitrarily chosen, multi-syllabic words (Ochiai and Fujimura, 1971).

Since, in the present study, we have available a detailed spectral description of the vowel segments presented to the subjects, a confrontation of the identification results with the spectral data will be possible.

4.2. EXPERIMENTAL PROCEDURE

The 270 vowel segments of speaker 1 were used for this experiment. A stimulus tape was made, containing nine groups of 30 segments which were resynthesized by means of the speech synthesis system using the original 17 filter values per 10 msec. The duration of each segment was as indicated in Table 3.5.1, the time between two segments was 3.5 sec, with an extra pause after every 30 stimuli. All 15 vowels occurred twice in random order within such a block of 30 segments. For every three listeners, the listening session was started at a different block. Since there were nine blocks of 30 segments, and also nine groups of three listeners, we could balance the order effect.

While recording the synthesized segments, a 3.2 kHz tone, also generated by the synthesizer under program control, was recorded on the second track of the

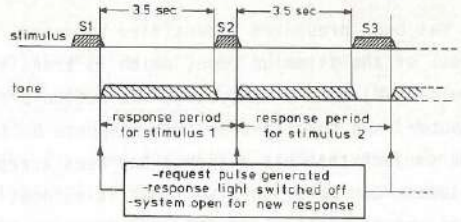


Fig. 4.2.1. Timing sequence of the stimulus presentation. The vowel stimuli are on one track of a tape; the 3.2 kHz tone, which controls the computer timing, is on the other track.

tape. During the segment this tone was interrupted, see Fig. 4.2.1. The moment of onset of this tone was used in the listening session to reset response lights, and to trigger the computer for the next response period. Subjects had to respond by pressing one of 15 push buttons showing the names of the 15 vowels in orthographic form (with our group of phonetically naive subjects we could not use phonetic symbols). The push buttons were ordered alphabetically. The symbols are given below.

orthographic: A AA AU E EE EI EU I IE O OE OO U UI UU

phonetic: ɑ a ɑu ɛ e ɛi ɸ I i ɔ u o œ ʌy y

Fifteen response categories is quite a large response set, and now and then it took the subjects some time to find the correct response button, once the stimulus itself had been identified. From pilot experiments the 3.5 sec interstimulus time appeared to be the best compromise between too short intervals which cause missed responses and too long intervals which are boring.

Each listener was first presented with three blocks of 30 segments to make him familiar with the synthetic stimuli, the response categories, and the positions of the response buttons. Next, the 270 actual stimuli were presented in one session. In this series, a response was seldom missed.

The response buttons and control lights were connected with the computer via the digital input-output system; for a block diagram see Fig. 4.2.2.

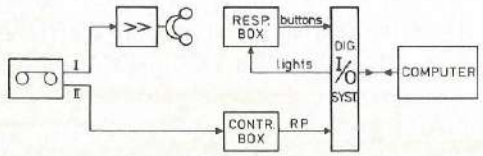


Fig. 4.2.2. Block diagram of the stimulus presentation.

After a stimulus has been presented (identified by means of the 3.2 kHz onset on the second track of the stimulus tape, which is translated by the control box into a request pulse (RP) to the computer), the button pressed first is identified by the computer, and the light on that response button is switched on. The light informs the subject that his response has been accepted by the system and that no other response can be given. The light is automatically switched off by the presentation of the next stimulus. All responses are stored in the computer. The stimuli are presented through headphones in a quiet room at a comfortable loudness level. A total of 27 naive subjects participated in the experiment,

with a single listening session for each subject. The listeners were informed that all vowels had the same probability of occurrence.

4.3. EXPERIMENTAL RESULTS

In Fig. 4.3.1 we give the percentage-correct score averaged over the 27 listeners, per block of 30 stimuli including the learning blocks (L_1, L_2 and L_3). So this figure gives the percentage correct score as a function of temporal order of presentation. It suggests that the three learning blocks have been very useful, and that learning during the actual experiment has been small and can be corrected for, as we did, by changing the order for the different subjects. The average correct score over all 270 vowel segments was 74.4%, with a standard deviation of 7.3% over the 27 subjects.

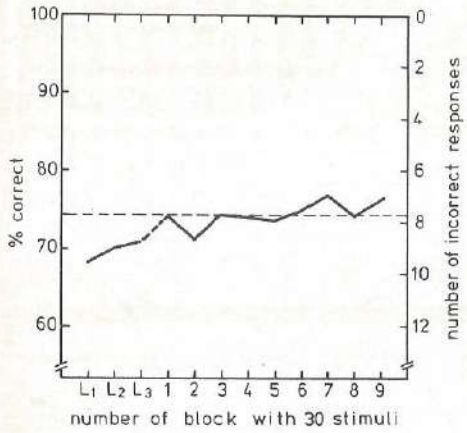


Fig. 4.3.1. Percentage-correct score per block of 30 stimuli. Scores are averages over 27 listeners. The right-hand scale gives the number of errors per block of 30 stimuli.

The confusion matrix is given in Table 4.3.1. Apparently, some vowels are much better identified than others. The long vowel /a/ is only seldom misidentified, and is then only confused with short /ɑ/. Vowel /u/ and diphthong /ɑu/ have the next highest correct scores. Some vowels with a lower correct score are confused with only one or two other vowels, like /ɑ/ with /ɔ/; /i/ with /I/ and /y/; /ɔ/ with /ɑ/ and /u/; and /o/ with /ɔ/. Most vowel confusions are certainly based on auditory similarity. However, there is always a small chance that visual similarity of the orthographic symbols, and/or nearness of response buttons may have had some influence.

Most response categories were used regularly; see last row of Table 4.3.1.

stimulus \ response	ɑ	a	ɛ	I	e	i	ɔ	o	u	œ	ø	y	ɑu	ɛi	ʏy	no resp.
ɑ	370	3	1				100			4			7		1	
a	12	474														
ɛ	30	3	288	13	2	1	63	1	1	60	10			2	10	2
I			41	320	5	32	2		9	50	2	23			1	1
e			33	152	248	15				9	12			14	1	2
i				54	1	399				3	1	24			1	3
ɔ	17						423	7	36	3						
o					1		201	266	16	1						1
u				1		4			443	12	1	23		1	1	
œ	2		45	27	2		120	1	15	246	21	6				1
ø	1		10	8	3	1	3	6	3	109	324	2		2	7	7
y			1	18		31			3	31	4	395			3	
ɑu		1					1	3					472	3	5	1
ɛi			7				2				24		1	313	111	1
ʏy	1	9						1			24	2	22	8	439	
TOTAL	433	490	426	593	262	483	915	285	526	528	423	475	502	343	560	19

Table 4.3.1. Vowel confusion matrix, accumulated over 27 listeners. Each listener identified 18 segments, from 18 different consonant contexts, per vowel. Responses add up to a total of 27 x 18 = 486 per vowel. For /ɛi/ one stimulus was missing, so for this diphthong the row total becomes 459.

However, the vowels /e/ and /o/ were used much less often as a response than the other vowels. We will come back to this point after the responses have been discussed in more detail. Vowel /ɔ/ was used far more often than any other vowel. The vowels /ɛ/, /I/, /œ/ and /ø/ were confused with quite a large number of other vowels. The most probable reasons for this are: the effect of coarticulation on these vowels, which makes them more difficult to identify, and the central position of these vowels in the vowel space which results in a greater number of neighbouring vowels. From the results of the intelligibility measurements described in section 2.5.3 we may conclude that the synthesis system itself did not cause these misidentifications. The quality of vowels /ɛ/ and /I/ was not worse than that of most of the other vowels. Vowel /ø/ did not occur in that test material, and vowel /œ/ did indeed give a lower score; there were confusions with vowels /ɛ/ (4%) and /I/ (20%), see Table 2.5.2. However, the type of confusions in the present experiment is in general of a different nature, which means that they are no artefacts of the synthesis system but are caused by vowel segmentation and coarticulation effects.

Tables 4.3.2 and 4.3.3 represent the results of the vowel segment identification experiment in another way. Here the number of vowel confusions, accumulated over the 27 listeners, is given as a function of the initial and final consonants, respectively, in the original word from which the vowel segment was isolated. Fig. 4.3.2 gives a graphical representation of the average number of

initial consonant \ vowel	ɑ	a	ɛ	I	e	i	ɔ	o	u	œ	ø	y	ɑu	ɛi	ʏy	total	average
p	17	0	9	20	24	1	1	23	0	6	9	0	1	9	0	120	8.0
t	21	0	18	12	24	1	11	15	8	16	1	9	0	23	1	160	10.7
k	23	0	5	14	19	2	11	17	0	8	1	0	5	9	1	115	7.7
b	8	0	8	1	2	1	6	15	2	12	25	0	2	7	1	90	6.0
d	6	0	14	2	25	2	0	26	2	20	25	17	0	9	2	150	10.0
f	1	0	19	7	18	2	3	15	0	9	2	22	0	3	2	103	6.9
s	4	1	16	23	1	6	4	0	15	13	15	1	0	5	5	109	7.3
v	0	2	9	8	11	7	6	20	1	22	6	0	0	11	13	116	7.7
z	0	0	10	11	2	4	0	2	8	14	26	1	1	16	17	112	7.5
x	3	0	5	23	14	2	10	12	1	13	7	4	0	4	2	100	6.7
h	4	0	14	3	27	6	1	15	1	14	7	9	0	16	2	119	7.9
m	8	0	23	4	12	1	0	18	0	9	6	1	1	5	1	89	5.9
n	3	0	16	2	27	6	1	26	0	16	3	4	3	8	3	118	7.9
l	1	0	5	1	25	16	0	0	2	8	0	0	0	5	0	63	4.2
r	6	0	15	23	5	8	4	0	3	20	4	4	0	-	5	98	7.0
w	4	1	5	3	1	10	4	0	0	13	17	7	0	9	1	75	5.0
j	4	0	7	8	0	9	3	15	0	10	6	11	1	4	8	84	5.6
.	3	8	0	1	0	3	0	1	0	17	2	1	0	3	3	42	2.8
TOTAL	116	12	198	166	238	87	63	220	43	240	162	91	14	146	67	1863	6.9

Table 4.3.2. Number of errors made in the identification of vowel segments taken from words with the indicated initial consonants. Errors are accumulated over the 27 listeners and, therefore, have a maximum value of 27. Total and average number of errors are also given. The dash represents one missing stimulus.

final consonant \ vowel	ɑ	a	ɛ	I	e	i	ɔ	o	u	œ	ø	y	ɑu	ɛi	ʏy	total	n	average
p	3	0	9	23	6	$^{2}12$	$^{2}4$	0	3	14	7	9	-	9	1	100	16	6.3
t	$^{2}7$	$^{2}0$	$^{2}23$	$^{2}8$	1	$^{2}9$	$^{2}4$	12	$^{4}19$	9	2	9	$^{7}4$	$^{3}19$	$^{6}13$	139	37	3.8
k	4	0	5	1	2	4	0	1	2	$^{2}33$	6	17	1	9	2	87	16	5.4
f	3	0	8	23	$^{2}15$	6	0	0	0	$^{2}23$	2	-	-	$^{2}20$	1	101	16	6.3
s	$^{2}8$	0	$^{2}15$	$^{3}14$	0	7	0	0	$^{2}16$	$^{3}26$	3	$^{2}11$	5	$^{2}19$	$^{3}5$	129	26	5.0
x	$^{2}1$	0	5	8	11	16	1	0	1	20	$^{2}7$	22	-	16	8	116	16	7.3
w	-	-	-	-	25	6	-	-	-	-	-	4	$^{7}4$	-	-	39	10	3.9
j	3	1	-	-	-	-	1	15	0	-	-	-	-	-	-	20	5	4.0
l	$^{2}29$	8	$^{2}35$	12	12	10	6	15	0	$^{2}35$	$^{2}15$	11	0	23	$^{2}30$	241	20	12.1
r	4	$^{4}0$	$^{2}27$	3	$^{6}146$	$^{4}10$	$^{2}11$	$^{7}137$	$^{2}2$	16	$^{6}115$	$^{7}2$	-	-	-	473	43	11.0
m	6	0	15	2	2	2	$^{2}15$	15	0	9	4	-	-	5	5	80	14	5.7
n	$^{2}40$	$^{3}3$	$^{3}42$	$^{3}35$	0	$^{2}3$	$^{2}10$	2	$^{2}0$	$^{2}21$	0	1	0	$^{5}22$	2	181	30	6.0
n	0	-	0	$^{3}37$	-	-	10	-	-	20	-	-	-	-	-	67	7	9.6
	8	0	14	-	18	2	1	23	0	14	1	$^{2}5$	-	4	0	90	14	6.9
TOTAL	116	12	198	166	238	87	63	220	43	240	162	91	14	146	67	1863	270	6.9

Table 4.3.3. Number of errors made by the 27 listeners in the identification of vowel segments taken from words with the indicated final consonants. The dashes indicate non-existing VC_f combinations. If a final consonant occurred more than once with the same vowel this is also indicated in the matrix as an upper left index. The total number of errors is also given, and so are the total number (n) of VC_f combinations per final consonant. Dividing these two gives the average number of vowel errors (last column).

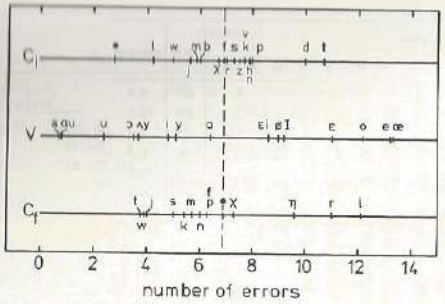


Fig. 4.3.2. Number of errors in the identification of the vowel segments. The errors are accumulated over the 27 listeners, and therefore have a maximum value of 27. The errors are given as a function of the initial consonant (upper line), the vowel (middle line), and the final consonant (lower line) in the original word from which the vowel segment was isolated. The dot represents the condition of no initial or final consonant in the original word.

errors taken from Tables 4.3.1, 4.3.2 and 4.3.3. The presentation of the data in this figure on separate lines perhaps suggests independent effects of the initial consonants, the vowels, and the final consonants on the vowel score. However, the structure of the present stimulus material does not allow such a general conclusion. For instance, half of the identification errors in the 15 vowel segments originating from words with an initial /d/ are caused by only three words with a final /r/ (/der, dor, dœr/), see Table 4.3.2. So one should not attach too much importance to the extreme position of the initial /d/ in Fig. 4.3.2. Also, the extreme position of the /t/ is probably mainly due to the final consonants in these words, since six of these words have /l/ or /r/ as final consonants. On the other hand it is clear from the position of the dot in this figure that *no* initial consonant favours a correct score of the vowels.

It seems that the effect of certain *final* consonants is somewhat clearer. Let us first skip the uncertain cases: /w/, /j/, and /ŋ/ only occur in combination with four or five vowels, see Table 3.2.1, so the present data make it impossible to draw conclusions concerning the positions of these consonants on the lower line of Fig. 4.3.2. However, the extreme positions of /r/ and /l/ on one hand, and /t/ on the other, are based on enough data to suggest a clear effect. It will be noted that /t/ causes little coarticulation in the preceding vowel, whereas /r/ and /l/ have a relatively large influence on the preceding vowel.

Before discussing in more detail not just how many incorrect identifications per vowel have been made, but also the type of confusions made in the various

contexts, we will first confront the identification results discussed so far with the spectral data.

As can be seen from the identification data in Table 4.3.1 and in Fig. 4.3.1, a score of 71.9% correct is derived for the 12 monophthongs. The confusion matrix based on distances between the average positions of the vowel segments and the average vowel positions in the spectral vowel space, showed for speaker 1 a score of 94.0% correct using all 17 dimensions, and 85.2% correct using two dimensions, see Table 3.5.2. The /ɔ-o/, /I-e/, and /œ-ɸ/ confusions were disregarded because confusions were based only on spectral distances and duration was not taken into account. If we apply the same restriction to the perceptual confusion data, the score rises to 80.4% correct, which is still much lower. If we order the vowels from low to high correct scores, this rank order is quite different for the spectral data and the identification data, as can be seen below (vowels between parentheses cannot be discriminated):

identification, /I-e/, /ɔ-o/, /œ-ɸ/ confusions disregarded	/œ, ɛ, I, ɑ, y, i, e, ɔ, ɸ, u, o, a/
speaker 1, spectral distances, 17 dim.	/ɑ, (ɔ, o), (œ, I, y), (a, ɛ, e, i, u, ɸ)/
idem, 2 dimensions	/o, ɑ, (I, ɔ, œ, y), (ɛ, e, u, ɸ), (a, i)/

The types of confusions show only little correspondence as can be seen upon comparison of Tables 4.3.1 and 3.5.2.

As was discussed in paragraph 3.7, another way of ordering the vowels is on the basis of the total variance in the spectral space over the segments per vowel, suggesting that a large variance for a certain vowel will cause many confusions. The rank order of the vowels of speaker 1 with regard to their spread in the first two dimensions is, from high to low:

/ɑu, Ʌy, ɛi, o, ɔ, u, ɑ, e, ɸ, œ, ɛ, I, y, a, i/

The only real correspondence between all these different rank orders is the far right position of /a/. All these data strongly suggest that a spectral representation of the vowel segments in terms of average positions in the vowel subspace is too simple to explain in all details the identification results. One reason for this is that durational information is not present in such an average spectral representation. An even more important reason is the dynamic spectral variation within the segments, which has greatly influenced many identifications by the subjects, as will be seen below.

We will now discuss the types of confusions made per vowel in each of the 18 contexts. These data are presented in Tables 4.3.4 to 4.3.18. In the first table we see, for instance, that the 100 /ɑ-ɔ/ confusions are not randomly distributed

over all 18 /ɑ/'s taken from 18 different contexts, but that 59% are caused by the /ɑ/'s taken from only three words: /pɑn/, /tɑl/, and /kɑn/. 78% of the /e-I/ confusions (Table 4.3.8) are determined by six /e/'s all taken from words with a final /r/. In the same way, 61% of the /o-ɔ/ confusions (Table 4.3.11) are determined by the seven /o/'s which were taken from words with a final /r/. The six /-ør/ segments account for 80% of the /ø-œ / confusions (Table 4.3.14).

The /i/, /y/, and /a/ segments from words with final /r/ did not give rise to specific confusions. For the other vowels there were no extra C_1Vr words in the list to test this final-/r/ effect. The r-effect must be of a spectral nature and certainly is no duration effect, since on the average the isolated vowel segments from the /-r/-contexts are even longer than those from the non-/-r/-contexts.

		ɑ	a	ɔ	ɑu	others
9	pɑn	10		17		
12	tɑl	6		20	1	
6	kɑn	4		22	1	
8	bɑ.	19		7	1	
9	dɑm	21		6		
3	fɑt	26		1		
8	sɑr	23	1		1	1
10	vɑn	27				
4	zɑx	27				
9	xɑf	24		1	2	
5	hɑk	23				4
9	mɑl	19		8		
6	nɑp	24		2	1	
7	lɑx	26		1		
7	rɑt	21		6		
9	wɑs	23		4		
5	jɑs	23		4		
6	.ɑj	24	2	1		
7.3		370 76.1%	3	100	7	6

Table 4.3.4. In this and the following tables, the types of confusions are given per vowel segment. The row totals are 27, being the number of subjects. The first column gives the duration of each segment in terms of the number of 10-msec samples. The percentage-correct score is also given.

		a	ɑ
33	par	27	
16	tak	27	
23	kas	27	
28	bar	27	
24	dax	27	
19	fam	27	
24	saj	26	1
14	van	25	2
22	zat	27	
19	xap	27	
25	han	27	
30	mar	27	
20	naf	27	
25	lat	27	
27	rar	27	
21	wan	26	1
29	ja.	27	
10	.al	19	8
22.7		474 97.5%	12

Table 4.3.5.

Some r-effect is not too surprising if one remembers what was pointed out on p. 94, namely that the stable vowel segments, isolated from words with a final /r/, have extreme positions in the vowel space. This can easily result in deviating identifications. On the basis of the positions of these segments in the vowel space (see Figs. 3.5.7 and 3.5.14), one might, however, predict that /e/ would become /i/ instead of /I/, /o/ would become /u/ instead of /ɔ/, and /ø/ would become /y/ instead of /œ /. A possible explanation for this unexpected result is that the average vowel positions in Fig. 3.5.7 are not necessarily also the optimal vowel positions in terms of best perceptual representation of that vowel. In fact, the identification results of the vowel segments from C_1Vr words suggest that the ideal vowel positions for /I/, /ɔ/, and /œ / in terms of perception are more like the /-r/-segment positions of /e/, /o/, and /ø/, respectively, see Fig. 3.5.14. Another possibility is that Dutch subjects are so strongly accustomed to the diphthong-like character of /e/, /o/, and /ø/, to be described as /Ii/, /ɔu/, and /œ y/, respectively, that if they are confronted with the stable initial parts of these vowels only, as in the /-r/-segments in this experiment, they identify these as /I/, /ɔ/, and /œ /, respectively. A third possible explanation is that the two-dimensional spectral representation does *not* give a realistic view of the vowel positions in this respect, because of too much data reduction. But this possibility was already excluded in the discussion on p. 95.

Let us finally discuss, systematically per vowel, the confusions made by our 27 subjects. In order to discuss the main effects, and not to go into too much detail, we will in general only draw attention to vowel segments which had been misidentified by 17 or more of the 27 subjects. For those vowel segments, the types of confusions will be compared with the average positions of these segments in the vowel subspace, as represented in Fig. 3.5.7. In this figure, the most pronounced misidentifications are underlined. This holds for all segments with 17 or more misidentifications, except the /-er/, /-or/, and /-ør/ segments which had already been encircled. The vowel segments /tut/, /sus/, /zut/, /xœ n/ and /hœ p/ are also underlined, although more than ten subjects identified them correctly. But with respect to the other segments of these same vowels they clearly differ. The dynamic spectral behaviour within the vowel segments will also be considered, whenever necessary.

Table 4.3.4 shows that most /ɑ-ɔ/ errors are caused by only four words /pɑn, tɑl, kɑn, mɑl/, all with an alveolar final consonant. Fig. 3.5.7 reveals that these four words are exactly those which have a position near the /ɔ/ area. The vowel /a/ gives very few confusions, see Table 4.3.5. The /a/ area in Fig. 3.5.7 is also very restricted.

		ɛ	ɑ	I	ɔ	œ	ɸ	Ay	others
6	pɛp	18			7	2			
9	tɛr	9	2	1	14	1			
9	kɛn	22		1		3			1
10	bɛf	19			7	1			
10	dɛn	13	1		3	10			
5	fɛl	8	3		8	4		2	2
7	sɛt	11			1	13	1		1
8	vɛr	18			1	8			
10	zɛs	17			4	4			2
4	xɛk	22	1	4					
10	hɛ.	13	4		10				
13	mɛn	4			1	6	8	8	
9	nɛl	11	11			2			3
4	lɛs	22			1	4			
12	rɛm	12	7		3	2	1		2
9	wɛx	22	1		3				1
3	jɛt	20		7					
7	.ɛn	27							
8.1		288 59.3%	30	13	63	60	10	10	12

Table 4.3.6.

We see in Table 4.3.6 that the confusions made with the /ɛ/ segments are quite diverse. All short vowels in the neighbourhood of /ɛ/ in the vowel space get responses, like /ɑ/, /I/, /ɔ/, and /œ/. The average positions of the segments, as represented in Fig. 3.5.7, do not always explain too well the confusions made. However, the dynamic behaviour of these vowel segments makes certain confusions more understandable, see Fig. 4.3.3. For instance, the stable middle part of the /ɛ/-segment from /mɛn/ is not sufficient to result in a unanimous /ɛ/ response, the tail pointing to the /œ/-area causes many /œ/, /ɸ/, and

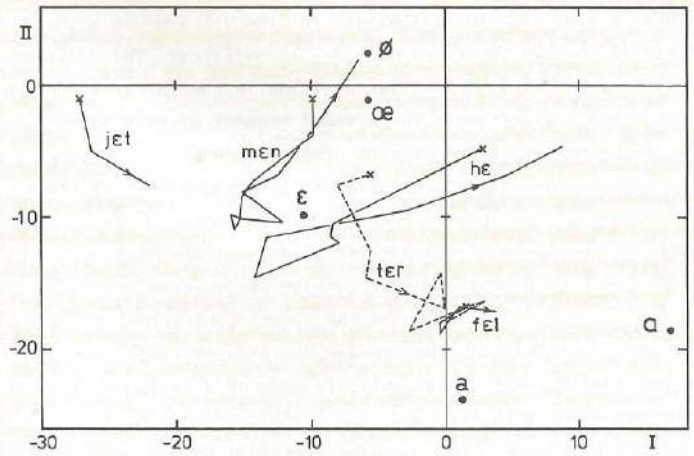


Fig. 4.3.3. Traces in the I-II vowel subspace of some /ɛ/ segments which cause many identification errors. The average vowel positions in this neighbourhood are indicated.

		I	ɛ	i	œ	y	others
10	pIn	7		3	3	6	8
5	tIl	15	1		10		1
10	kIn	13		10	1	2	1
8	bIt	26			1		
7	dIm	25			2		
7	fIt	20		3	4		
7	sIp	4			10	11	2
9	vIs	19	5		3		
10	zIŋ	16	3	4		3	1
9	xIŋ	4	6	11	1	1	4
9	hIŋ	24	2		1		
8	mIs	23	2		2		
9	nIs	25	1		1		
7	lIk	26			1		
7	rIf	4	13		8		2
6	wIr	24			2		1
8	jIx	19	8				
6	.In	26		1			
7.9		320 65.8%	41	32	50	23	20

Table 4.3.7.

		e	I	ɛ	i	ɸ	ɛi	others
22	per	3	18	1	3	1		1
18	ter	3	21	2				1
18	ker	8	17	1	1			
18	bek	25				2		
15	der	2	23	1				1
15	fe.	9	4	10		1		3
21	sef	26		1				
17	vex	16	6	3		1	1	
21	zem	25		1		1		
17	xef	13	2	4		2	6	
17	her		24				1	2
19	mel	15	2	2		2	5	1
18	ner		16		11			
14	lew	2	19	3				3
15	rep	21		4		1	1	
15	wet	26				1		
18	jes	27						
22	.en	27						
17.8		248 51.0%	152	33	15	12	14	12

Table 4.3.8.

even /ʌy/ confusions. A similar situation exists for the /ɛ/ from /hɛ/, here the tail points to the /ɔ/ area, which causes many /ɛ-ɔ/ confusions. The /ɛ/ segment of /fɛl/ is very stable, but at an unfamiliar position in the /a/ area. The long vowel /a/ cannot be given as a response for a vowel segment of only 50 msec duration, therefore, we probably get the various other responses.

Some extreme positions for /I/ segments lead to corresponding confusions, like /I/ in /rIf/ becoming /ɛ/ and /œ/, and in /sIp/ becoming /œ/ and /y/, see Table 4.3.7. Other confusions are less clear; for instance, why is the vowel segment from /xIn/ so often confused with /i/, whilst its average position is very close to the average /I/ position? Again the dynamic characteristics of such a segment are very important and the average position of the segment is not a sufficient clue by itself. The trace of this /I/ segment clearly moves to /i/, thus explaining the confusion with /i/.

With respect to the next vowel: /e/ we have already discussed on p. 114

		i	I	y	others
8	pit	26		1	
7	tin	26		1	
9	kim	25	1		1
21	bir	26		1	
11	di.	25		1	1
5	fin	25	1	1	
21	sir	21	3	1	2
9	vis	20	6	1	
8	zik	23	3		1
14	xir	25	1	1	
6	hif	21	4		2
17	mir	26		1	
7	niw	21		5	1
9	lix	11	15		1
6	rit	19	7	1	
11	wil	17	4	6	
8	jip	18	7	2	
5	.ip	24	2	1	
10.1		399 82.1%	54	24	9

Table 4.3.9.

the identification of the /-er/ segments. Most of the other /e/ segments, like those from the words /bek/, /sef/, /zem/, /wet/, /jes/, and /en/ are identified correctly by almost all subjects, see Table 4.3.8. These are the segments which are represented in the vowel space with dynamic traces pointing to /i/. So our subjects preferred an /e/ with some diphthongal character. The average dynamic pattern was illustrated in Fig. 3.5.15. This trace would have been even more pronounced if it had been constructed from only those segments which showed clear /e/ responses. Segments which did not show this clear tail-like pattern gave much less correct /e/ scores, like in the words /fe/, /vex/, /xef/, /mel/, and /lew/, see Table 4.3.8. The segment from this last word had, moreover, a duration of only 140 msec; this also favours an /I/ identification. /i/ gives few systematic errors (see Table 4.3.9), only /i/ from /lix/ becomes /I/ which is not explained by the dimensional spectral representation of that segment in Fig. 3.5.7, nor from its dynamic behaviour. Looking at the average spectral position of

		ɔ	ɑ	u	others
3	pɔp	26		1	
11	tɔr	16	11		
5	kɔm	16		11	
9	bɔn	21		6	
10	dɔf	27			
7	fɔp	24		3	
10	sɔm	23		2	2
14	vɔl	21		1	5
6	zɔt	27			
10	xɔn	17	1	8	1
7	hɔj	26			1
6	mɔr	27			
9	nɔx	26	1		
7	lɔk	27			
7	rɔt	23	4		
9	wɔn	23		4	
3	jɔ.	26			1
4	.ɔs	27			
7.6		423 87.0%	17	36	10

Table 4.3.10.

		o	ɔ	u	others
15	po.	4	22		1
14	tom	12	15		
19	kor	10	7	10	
19	bor	12	13	1	1
16	dor	1	26		
19	foj	12	14	1	
21	sos	27			
16	vor	7	19	1	
16	zon	25	1		1
17	xot	15	12		
23	hor	12	15		
19	mor	9	17	1	
14	nor	1	25	1	
16	lop	27			
21	rof	27			
20	wox	27			
14	jol	12	14	1	
16	.ok	26	1		
17.5		266 54.7%	201	16	3

Table 4.3.11.

/ɔ/ in /tɔr/ in the same figure it will be quite clear that this segment is often confused with /ɑ/. Most /ɔ-u/ confusions occurred in the words /kɔm/, /bɔn/, /xɔn/, and /wɔn/ (Table 4.3.10), all of them having nasal final consonants. The /ɔ-o/ confusion only happened for /vɔl/ and is probably a duration effect: this segment was 140 msec, the longest of all /ɔ/ segments. The /-or/ segments, causing /ɔ/ identification, have been discussed. Some other /o/ segments also caused many /o-ɔ/ confusions, like /po/, /tom/, /foj/, /xot/, and /jol/, see Table 4.3.11. Other ones like those from /sos/, /zon/, /lop/, /rof/, /wox/, and /ok/ were almost always correctly identified. Neither average segment positions in Fig. 3.5.7 nor segment durations can explain this. Here again, like with /e/, an explanation can be found in the dynamic behaviour of those segments. For the first group of words, the spectral information in the segments is fairly stable. The second group of words, on the other hand, shows segmental traces with tails pointing to /u/. So our group of naive, native Dutch listeners is only willing

		u	œ	y	others
8	pun	27			
8	tut	19	2	5	1
6	kun	27			
8	but	25	1	1	
7	duk	25	1		1
8	fuj	27			
8	sus	12	2	11	2
7	vux	26	1		
5	zut	19	2	4	2
8	xut	26			1
8	hus	26	1		
6	mu.	27			
10	num	27			
11	lur	25	1	1	
5	rup	24	1	1	1
7	wuf	27			
10	jul	27			
8	.ur	27			
7.7		443 91.2%	12	23	8

Table 4.3.12.

		œ	ɛ	I	ɔ	u	ɸ	others
10	pœs	21	3				3	
4	tœr	11	1	5	1	7	1	1
6	kœn	19	1	6		1		
9	bœs	15	2	7		3		
10	dœn	7	3				15	2
7	foet	18	3	2	3		1	
10	sœl	14		1	10		1	1
10	vœl	5			22			
10	zœ.	13	5		7	1		1
6	xoen	14			12			1
5	hœp	13	1		12			1
10	mœm	18			4	3		2
5	nœk	11	8	2	4			2
9	lœs	19	3	2	3			
10	rœx	7	1		18			1
8	wœf	14	3		10			
7	jœf	17	4	2	4			
7	.œk	10	7		10			
7.9		246 50.6%	45	27	120	15	21	12

Table 4.3.13.

		ɸ	œ	ɛ	others
18	pɸl	18	6	2	1
21	tɸx	26			1
26	kɸ.	26			1
11	bɸr	2	21	2	2
15	dɸr	2	17	1	7
17	fɸt	25	1		1
18	sɸr	12	7		8
21	vɸl	21	3		3
17	zɸr	1	22	2	2
22	xɸr	20	5		2
14	hɸp	20	4	1	2
20	mɸx	21	2		4
21	nɸs	24			3
24	lɸn	27			
19	rɸm	23	3		1
15	wɸr	10	15	1	1
14	jɸk	21	3	1	2
24	.ɸf	25			2
18.7		324 66.7%	109	10	43

Table 4.3.14.

to identify an isolated /o/ segment as an /o/ if the segment has a diphthongal nature; see also Fig. 3.5.15. The /u/ segments have a high correct score (Table 4.3.12), except those from the words /tut/, /sus/, and /zut/ which show some confusion with /y/ in line with their position in the vowel space. The large spectral variation between the different /u/ segments has no influence on their identification, probably because in this part of the spectral space there are no competing other vowels. As can be seen in Table 4.3.13, the vowel /œ/ was often confused with quite a number of other vowels, like /ɛ, I, ɔ, u, ɸ/; this only emphasizes the central position of this vowel. However, *most* confusions occurred with /ɔ/, despite the quite large gap between the /œ/ area and /ɔ/ area in the vowel space. 80% of the /ɸ-œ/ confusions are caused by the six /-ɸr/ segments, see Table 4.3.14. As was the case with vowels /o/ and /e/, all correctly recognized /ɸ/'s have a diphthong-like nature, see Fig. 3.5.15. Most pronounced errors for /y/ are in the words /dyk/ and /fyx/ (Table 4.3.15), in line with the

		y	I	i	œ	others
13	pyr	27				
6	tyt	18	1	4	4	
20	kyr	27				
16	byr	27				
7	dyk	10	5	11	1	
9	fyx	5	6		12	4
11	sy.	26		1		
17	vyr	27				
24	zyr	26				1
11	xys	23		2	1	1
7	hyp	18	1		7	1
11	myn	26	1			
11	ny.	23		4		
21	lyr	27				
7	ryw	23		1	3	
9	wys	20	1	4		2
6	jyl	16	3	4	2	2
15	.yr	26			1	
12.3		395 81.3%	18	31	31	11

Table 4.3.15.

average positions of these segments close to /i/ and to /œ/, respectively.

Table 4.3.16 shows that the diphthong /ɑu/ was hardly ever misinterpreted. When the diphthong /ɛi/ was incorrectly identified, it mostly became /ʌy/, and now and then /ɸ/. Fig. 3.6.1 shows that the average traces for the diphthongs /ɛi/ and /ʌy/ are indeed relatively close. The /ʌy-ɸ/ confusion is understandable from the preferred diphthong nature of /ɸ/, which can be written as /œy/, coming close to /ʌy/; compare Fig. 3.5.15 with Fig. 3.6.1. The /ɛi/ segment from /rɛin/ was, by accident, not presented to the subjects. The many errors for the /ɛi/ segment from /tɛil/ (Table 4.3.17) are caused by the tail of the transition part which first goes in the correct direction but then turns away under the influence of the following /l/. We see the same effect in the /ʌy/ segments from the words /vʌyl/ and /zʌyl/ (Table 4.3.18), again causing many identification errors. These deviating traces are shown in Fig. 4.3.4. These data show that both in monophthongs and diphthongs the final /l/ has a large effect

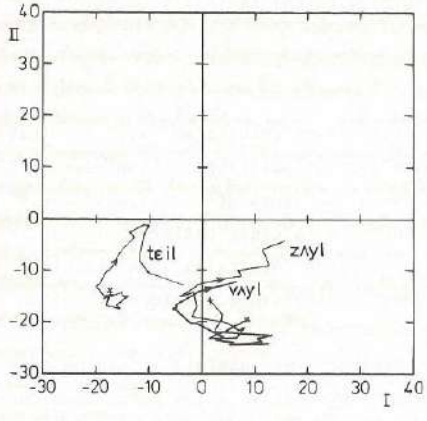


Fig. 4.3.4. Diphthong traces from final-/l/ contexts which cause many misidentifications of the isolated diphthong segments.

		ɑu	others
12	pɑuk	26	1
21	tɑuw	27	
20	kɑus	22	5
17	bɑut	25	2
28	dɑuw	27	
20	fɑun	27	
24	sɑul	27	
22	vɑuw	27	
16	zɑut	26	1
19	xɑut	27	
16	hɑut	27	
19	mɑut	26	1
12	nɑuw	24	3
25	lɑuw	27	
22	rɑuw	27	
19	wɑut	27	
19	jɑuw	26	1
17	.ɑut	27	
19.3		472 97.1%	14

Table 4.3.16.

		ɛi	ʌy	ɛ	ɸ	others
21	pɛip	18	5		4	
24	tɛil	4	8	5	6	4
23	kɛif	18	8		1	
19	bɛit	20	7			
19	dɛik	18	9			
24	fɛin	24	3			
21	sɛin	22	4		1	
25	vɛif	16	10		1	
21	zɛis	11	14		2	
22	xɛit	23	3		1	
21	hɛix	11	15		1	
25	mɛin	22	3		2	
16	nɛit	19	5	2	1	
21	lɛim	22	4		1	
	rɛin					
21	wein	18	8		1	
26	jɛi.	23	4			
26	.ɛis	24	1		2	
22.1		313 68.2%	111	7	24	4

Table 4.3.17.

on the preceding vowel. Table 4.3.3 and Fig. 4.3.2 confirm this finding.

4.4. CONCLUDING REMARKS

In the foregoing discussion of the identification results we have seen that there is only a limited relation between the average positions of the vowel segments in the two-dimensional vowel subspace and the types of confusions made. Especially when the final part of the segment shows a clear pattern pointing in a specific direction, this part of the segment often "overrules" the (stable) position of the central portion of the segment in the subject's identification decision.

Clearly, the identification experiment has been very useful; it can be considered as a supplement to the spectral data. The relatively low percentage-correct score in the identification experiment is in apparent contrast with our ex-

		ʌy	a	ɸ	ɑu	ɛi	others
32	pʌy.	27					
27	tʌys	26		1			
18	kʌyp	26		1			
22	bʌyt	26		1			
23	dʌyn	25		2			
18	fʌyk	25		1		1	
18	sʌyt	22		2	1	2	
21	vʌyl	14	7	1	4		1
22	zʌyl	10	2	4	11		
19	xʌyt	25		1			1
17	hʌyt	25			1		1
26	mʌys	26		1			
22	nʌys	24		1		2	
22	lʌyt	27					
16	rʌym	22		2	1	2	
22	wʌyf	26		1			
14	jʌyx	19		4	3	1	
18	ʌyt	24		1	1		1
20.9		419 86.2%	9	24	22	8	4

Fig. 4.3.18.

perience from daily life that vowel phonemes in well pronounced syllables are rarely misunderstood. This experience is confirmed by the data presented in section 2.5.3 (Table 2.5.1) for vowel phonemes in PB-words. Using the original 17 filter levels in the resynthesis, as we did in the identification experiment with the vowel segments, the correct score found was 96.8% averaged over 250 words, spoken by five native speakers, and presented to five native listeners. It is reasonable to assume that the vowel score would be close to 100% for meaningful monosyllabic words in conversational speech.

We nevertheless found very much lower scores for isolated vowel segments. We must assume that the extra information available in the coarticulated vowel about preceding and/or following phonemes, becomes a disambiguating factor in identifying the isolated vowel segments. The misidentifications give valuable information about the perceptual relevance of dynamic spectral variations in the vowel segments.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

5.1. INTRODUCTION

In this final chapter we will "draw up the balance sheet", *i.e.* see what we have gained and at what points improvements and/or more research are necessary.

One can recognize two main aspects in this research project: one is the methodological approach, the other is the actual experiment on vowel coarticulation. In the following two paragraphs we will discuss both points, and come to some conclusions.

5.2. METHODOLOGY

The present analysis, data processing, and synthesis systems allow the processing of many speech utterances in an efficient and objective way. Although the one-third octave bandfilters used have a limited spectral resolution, there is no indication that any important spectral variations in the vowel segments are left unmeasured. Because of the analogy between this system and the peripheral processing of acoustic signals in the ear, the reverse would have been unexpected. Even after a considerable data reduction from the original 17-dimensional bandfilter representation to a two-dimensional principal-components vowel subspace, this spectral representation is still a highly appropriate one.

For the average positions of vowel sounds from a neutral context, this principal-components representation strongly resembles the formant representation, see Fig. 1.3.2. The average positions of, and the dynamic spectral variation in, the vowel segments in different consonant contexts can be nicely displayed in the vowel subspace.

Contrary to a formant analysis using spectrograms, this analysis is automa-

tic and under computer control. This allows real-time processing, but even more important is the availability of all numerical data for subsequent data processing. This makes it easier to process the spectral data in a systematic way without just testing preliminary specified expectations.

The processing is done in two steps, the first is looking at the average positions of the isolated vowel segments; then, if necessary, the dynamic spectral variation within the segments can be studied in more detail. Apart from a graphical representation of the average positions of the vowel segments, which already shows large (consonant-specific) variations (see Fig. 3.5.7), distances to the average vowel positions (see Table 3.5.2), and the variance per vowel per dimension (see Fig. 3.5.10) can be calculated. This total variance per vowel can be split up into two sources: the variance between the 18 average positions of the segments, and the variance within the segment traces (see Fig. 3.5.11). For specific subsets, like the diphthongs, or the /-r/-segments, average traces can be calculated. The use of some time-normalization procedure is then indispensable. For our purpose a linear one was sufficient. To compare the data from various speakers, a speaker normalization is necessary. As was discussed in sections 2.3.2 and 3.5.2, centring the data gives good results.

Along with the numerical and the graphical display of the data per 10-msec sample, the speech synthesis was also an efficient research tool. It allowed repeated listenings to a specific word, or to any part of this word. For detailed listening, the utterance can also be stretched. The synthesis system was also used for compiling the stimulus tape. For that purpose, the intelligibility of lists of phonetically-balanced (PB) monosyllabic words was determined first. The vowel-recognition score (96.8%) of these resynthesized words was found to be good enough for the synthesis system to be used for stimulus generation.
This, when compared to tape splicing by hand, made it much easier to vary the duration of the isolated segments, to optimize the interstimulus time, and to randomize the stimulus order on the tape. Another advantage is that the acoustic characteristics of the stimuli are known exactly, because they are based on stored analysis parameters and a stable synthesis system. This also makes it easier to compare spectral data with perceptual results. Repeatedly spoken natural utterances would inevitably have introduced extra variation. On the other hand, one should always be aware that synthetic speech is being used and that this can introduce artefacts. An even better quality of the resynthesized speech would be desirable.

Since the identification experiment was also conducted under computer control, stimulus generation and processing of the responses of as many as 27 sub-

jects, each identifying 270 and more segments, could effectively be done. The presentation of the identification results in terms of confusion matrices which concentrate on different aspects of the stimuli (preceding consonant, see Table 4.3.2; the vowel itself, see Table 4.3.1; or following consonant, see Table 4.3.3), seems to be a good procedure to isolate the main effects, as illustrated in Fig. 4.3.2. With more data, this diagram would be even more representative than it is now because of the interaction in the data. Unfortunately, a more objective procedure could not be found to evaluate the results in the confusion matrices per vowel given in Tables 4.3.4, ff. For multidimensional scaling procedures the matrices are too empty and not sufficiently related.

5.3. EXPERIMENTAL RESULTS CONCERNING VOWEL COARTICULATION

Our interest in the spectral characteristics of coarticulated vowels, and the way in which isolated vowel segments are identified, stems from the view that the vowels are very important anchor points in the perception of speech utterances. It is reasonable to assume that the speech perception process makes use of optimally realized syllables as starting points for subsequent processing. One may furthermore suggest that, although speech perception probably is not based on phoneme-by-phoneme processing, certain phonemes in the speech stream are more readily recognizable than others, and that these units form the basis for subsequent recognition of the rest of a speech utterance. The vowels in stressed syllables are the best candidates for such anchor points. This is also a regularly used concept in systems for automatic understanding of speech.

But even the acoustic realizations of vowels in stressed syllables are (systematically) influenced by the surroundings in which they occur. With respect to vowel duration, one finds in the literature that the position of the syllable in the word (group) is very important, and so is, to a lesser degree, the consonant environment. With respect to the spectral composition of the vowel, the consonant environment seems to be the most important factor.

The questions formulated in the first paragraph of Chapter 1 are all directly related to these problems; we will now structure the present discussion in accordance with these questions.

(a) *How large is the spectral variation in the vowel due to different consonant environments?*

That the spectral variation in the different vowel-phoneme realizations is large, is already apparent from the average positions of the vowel segments in the two-dimensional principal-components vowel subspace, see Figs. 3.5.7 to

3.5.9. This variation can of course not be attributed completely to consonant-specific vowel coarticulation, since non-specific idiosyncratic variations are also included. Replicated measurements would be necessary to specify all those deviations which are specific.

Fig. 3.5.10 represents in another way the spectral variation per vowel by means of the spread σ per dimension. Ways have to be found to interpret this variation with respect to the distance between the vowel concerned and neighbouring vowel areas in the spectral vowel subspace.

Another aspect of the spectral variation per vowel is the variation within the vowel segments. Certain vowels show considerable dynamic variation, irrespective of the consonant context. This is, of course, the case with the diphthongs /ɑu, ʌy, ɛi/, but also for /e, o, ø/.

(b) *Can the spectral variation in the vowel be predicted from the particular consonant environment?*

Context-specific effects on the vowels were clearly found in our spectral data, especially those caused by the final /r/. Both the starting positions and the spectral transitions of the vowel segments are different for /r/ and non-/r/ final consonants. Especially the difference in starting positions is somewhat surprising, as this effect is hardly noticed in the literature, although 't Hart (1969) mentions it. It involves a coarticulation effect from the consonant on the vowel, not just in the vowel-consonant transition part, but already in the very stable initial part of the preceding vowel. The duration of this part is so long that the deviating position cannot be a matter of "undershoot", but must be some "preprogrammed" articulatory action. Incidentally, the influence of other consonants could also be demonstrated, such as the effect of the postdentals /t, z, s/ on /u/, and the effect of liquids, glides and nasals on various vowels. However, there were not enough replications of a single context available in our present data to draw firm conclusions; except for the "r-effect".

(c) *Can vowel segments be recognized, ignoring the consonant environment?*

We have shown in section 3.5.4 that vowels can indeed be recognized automatically on the basis of the average position of the vowel segments in the vowel space. Using all 17 dimensions the score, averaged over the three speakers and the 12 monophthongs, is 93.4% correct. Confusions between corresponding long-short vowels (/e-ɪ/, /o-ɔ/, and /ø-œ/) were disregarded. This score decreased to 86.1% when only two-dimensional spectral information was used.

Subjects in the identification experiment certainly have more trouble identifying these isolated vowel segments correctly, as we have seen in paragraph 4.3. At that point we assumed that the extra information available in the coar-

ticulated vowel segment about preceding and/or following phonemes becomes a disambiguating factor in identifying these vowel segments. So, under this specific condition, the straightforward procedure of averaging spectral information over the whole segment followed by a shortest-distance measure, is more efficient than the human recognition procedure, which takes into account all of the dynamic information, but which has to work without the normally available context.

(d) *Can misidentifications of vowel segments be understood from the acoustic properties of the vowel segments?*

With respect to numbers and types of confusions made per presented vowel, the relationship between misidentifications and acoustic properties of vowel segments is poor, if we only take into account the average positions of the vowel segments; compare Tables 3.5.2 and 4.3.1. This is not too surprising if one realizes that the listeners had other information available to base their decisions on; *e.g.* duration of the vowel segment and dynamic spectral variation within the segment.

The durational information generally prevents perceptual confusions between spectrally related long and short vowels. That nevertheless a number of /e, o, ø/ segments are identified as /I, ɔ, œ/, respectively, has to do with the final-/r/ context from which these vowels were isolated. We will discuss this point in more detail below.

The dynamic spectral variation within the segments turns out to be very important for (mis)identification of these segments as vowels. Although this information is also available in the spectral representation of the vowel segments, it is not easy to use it to actually predict the vowel confusions. Once certain confusions have been found, these can indeed more often than not be understood on the basis of the dynamic spectral behaviour of the presented vowel segments, the more so if we assume that a small onset in a certain direction is sufficient to change completely the identity of the vowel segment, especially if this onset is in the final vowel part.

The low scores for /e/ and /o/, and to a lesser degree also those for /ø/, are due to the diphthong-like nature of these vowels which was preferred by our subjects, see Fig. 3.5.15. It was not always present in our stimuli, because of coarticulation and because of the way of isolating the segments. A further complicating factor is the deviating characteristic of these vowels when they are followed by /r/, see Fig. 3.5.17. In the perceptual experiment, only the initial stable parts of these /-r/-segments were presented for identification. This results in many confusions of the type /e-I/, /o-ɔ/, and /ø-œ/. These confusions do not correspond too well with the average positions of these segments in the

I-II vowel subspace, see Fig. 3.5.7. A straightforward explanation is not easy to give for this effect, but it seems reasonable to assume that the average vowel positions in the spectral vowel subspace are not necessarily the ideal vowel positions. Moreover, the listeners are not accustomed to recognizing a sound as, for instance, the vowel /e/ if this sound lacks the characteristic spectral transition. This effect is most pronounced for /e, o, ø/, although also the total /-ir/ and /-yr/ vowel segments show large transitions, see Fig. 3.5.17. The identification of the initial segments of these words gives less confusion, see Tables 4.3.9 and 4.3.15.

The great difference between /e, o, ø/ segments in /-r/ and non-/-r/ words is best illustrated in Figs. 3.5.17 and 3.5.15.

The three Dutch diphthongs /ɑu, ʌy, ɛi/ combined have a correct score of 84.1%. Of these three, /ɑu/ causes hardly any confusions, /ʌy/ only a few /ʌy-ø/ and /ʌy-ɛi/ confusions, but there are many /ɛi-ʌy/ confusions. Although the two average traces are relatively close, see Fig. 3.6.1, this amount of confusion (22.8%) is somewhat unexpected.

From a comparison of the identification results and the spectral analysis of correct and false diphthong responses it became clear that the starting position is very important for the diphthong identification, together with an indication of the direction of the spectral transition. It is not necessary to actually reach a specific endpoint. On the other hand, a deviation of the transition part can easily result in misidentification of the isolated diphthong segment. This deviation can be caused by a following consonant, as was clearly visible in our data for the final /l/.

A stable, constant, vowel segment is not always the ideal vowel representation at all. Especially for the Dutch monophthongs /e, o, ø/ a spectral transition is almost a "must" for correct identification of those vowel segments. This spectral transition must go in the direction of /i, u, y/ respectively, but has no need at all to reach those target positions. The other Dutch monophthongs do not need such a clear spectral transition for a correct identification. Even a minor transition, if present in those vowels, can lead to a different identification. Such a transition mostly occurs as a consequence of the consonant environment.

The final consonants in the context seem to have a greater influence on the vowel characteristics than the initial consonants.

5.4. FUTURE RESEARCH

The main goal of this thesis work was to find out whether a bandfilter analysis, followed by a dimensional representation of the spectral data, could effectively be used for studying the dynamic spectral characteristics of vowel sounds and, also, for studying the effects of coarticulation on the spectral qualities of vowel sounds. It seems justified to conclude from the results that this is indeed possible. This means that this approach may be very useful in further research on spectral analysis and coarticulation of time varying speech.

Broad and Fertig (1970), for instance, say in the conclusion to their article that "The extension of (their) results to other vowels (than /I/), to more general types of utterances, and to other speakers and dialects will obviously require a rather massive body of data. An efficient automatic procedure for obtaining accurate formant-frequency measurements would aid such a program of study greatly". We believe that we have an alternative for, what they call, an automatic procedure for obtaining accurate formant-frequency measurements. The interesting results already achieved with the present data on the 270 CVC words justify an extension and deepening of these results. In a recently started project, supported by the Netherlands Organization for the Advancement of Pure Research (ZWO), the influence of consonantal context on the dynamic spectral characteristics of vowels, diphthongs, and glides is studied in more detail (Pols and Schouten, 1975). In that project, only a limited choice is made from all possible vowels and consonants but more attention is paid to the replicability within and between speakers by using utterances from a certain number of speakers and, also, a number of repetitions per speaker. Both CVC words spoken in isolation and the same words embedded in a story read aloud from paper will be used. Presumed regularities in vowel coarticulation effects will be verified by means of perceptual experiments.

We expect that the results will afford a better understanding of the dynamic processes of speech production and speech perception. It will present information for specifying the coarticulation rules necessary, for instance, in synthesis-by-rule. It will also ease the interpretation of the dynamically varying acoustic-phonetic information in automatic speech recognition systems, not only for isolated words but also for short phrases and continuous speech, uttered by different speakers.

SUMMARY

Vowel phonemes in carefully spoken monosyllabic words can be recognized very easily by native subjects in an identification experiment in which the complete syllables are presented. The results given in paragraph 2.5 of this study confirm this finding, even for resynthesized speech.

On the other hand, it is known that large differences exist between acoustic realizations of the same vowel phoneme in different contexts. This is for instance apparent in speech spectrograms, but it is just as clearly visible in a dimensional representation of vowel bandfilter spectra, see paragraph 3.5.

The aim of the present research project was to study this spectral variation in the vowel segments of 270 monosyllabic Dutch words of the type initial consonant-vowel-final consonant, spoken by three male native speakers. Other acoustic aspects of the vowel sounds, like duration and pitch, were disregarded. All isolated vowel segments were also presented to 27 native listeners for identification. The results show to what extent the vowel phoneme can be uniquely identified on the basis of acoustic information alone. A comparison with the spectral data reveals to what extent misidentifications can be understood from the average and/or dynamic spectral properties of the vowel segments.

A detailed spectral analysis of a large set of data, like the present set, calls for an efficient and automatized analysis and data-processing system. However, automatic and fast formant extraction procedures are not readily available. We furthermore are of the opinion that the formant parameters are not necessarily the most relevant parameters for describing speech perception.

As an alternative we used a bandfilter analysis, followed by a dimensional spectral representation of the data in a low-dimensional subspace, somewhat similar to the formant plane. The one-third octave bandfilter analysis simulates the limited spectral resolution of the human hearing mechanism. The dimensional spectral representation takes full account of the statistical variation and correlation in the bandfilter levels of many 10-msec samples of analyzed speech.

Paragraphs 2.1 to 2.3 give a detailed description of this approach.

In vowel segmentation and speech identification experiments, use of a programmable speech synthesis system is indispensable. Therefore, a computer-controlled speech synthesis system was developed; it is in principle the mirror-image of the speech analysis system, resulting in a vocoder-like total system. Paragraph 2.5 gives the details, together with intelligibility results. When the synthesis system is used as a research tool, these results are satisfactory and sufficient.

The total system made it possible to analyze and process in detail the set of 270 CVC words; see Chapter 3. Spectral information is available every 10 msec, giving all the data necessary for specifying the dynamic spectral behaviour. This allowed us to describe quantitatively the dynamic spectral characteristics of the Dutch diphthongs /ɑu, Ay, ɛi/; it furthermore demonstrated that the Dutch vowels /e, o, ø/ are also produced in a diphthong-like way, and behave differently when they are followed by /r/. There were other indications of coarticulatory effects in the spectral data. However, not enough replications were available in the word list to draw firm conclusions.

By looking at the spectral data, and at the two-dimensional spectral representation of the word as a trace, and by listening to (part of) the synthesized word, we could isolate the vowel segments. The vowel segments, isolated in this way, were then resynthesized and presented to subjects for identification. Confusions made by the 27 subjects were often clear indications of coarticulatory effects on the vowel segment. For certain vowel segments the confusions could be understood from the average position of such a segment in the two-dimensional spectral subspace. However, the dynamic spectral behaviour had to be taken into account regularly in order to understand the misidentifications. The direction of the final transitional part in the vowel segment was often more important than the, often much longer, stable initial part.

The final consonant in the context seems to be of greater influence on the vowel characteristics than the initial consonant.

Chapter 4 gives the details of the identification experiment, and a discussion of the results. In Chapter 1 a review is given of what is found in the literature about spectral and perceptual differences between vowel phonemes under various conditions.

We have shown that the analysis, data processing, and synthesis system makes it possible to study vowel coarticulation effects as well as the dynamic spectral behaviour of vowel sounds in general. An extension to non-vowel sonorants like nasals, liquids, and glides, is in preparation; it seems to be very promising.

The results achieved so far with the data from the list with 270 CVC words justify more detailed research with more speakers and more replications.

SAMENVATTING

Klinkerfonemen in zorgvuldig uitgesproken eenlettergrepige woorden kunnen gemakkelijk door proefpersonen in een benoemingsexperiment worden geidentificeerd, wanneer althans komplete lettergrepen worden aangeboden. De verstaanbaarheidsresultaten zoals gegeven in paragraaf 2.5 bevestigen dit, zelfs voor synthetische spraak. Het is evenwel ook bekend dat er grote verschillen bestaan tussen akoestische realisaties van hetzelfde klinkerfoneem in verschillende kontexten. In spraakspektrogrammen is dit duidelijk zichtbaar, evenals in een dimensionele representatie van bandfilterspektra van klinkers, zie paragraaf 3.5.

Doel van dit onderzoek is de spektrale variaties te bestuderen in de klinkersegmenten van 270 eenlettergrepige Nederlandse woorden van het type beginmedeklinker-klinker-eindmedeklinker. De woorden werden uitgesproken door drie mannelijke sprekers. Andere dan spektrale, akoestische aspekten van de klinkersegmenten, zoals duur en grondtoon, werden in dit onderzoek niet in detail bekeken. Alle 270 geïsoleerde klinkersegmenten van één spreker werden tevens ter identifikatie aangeboden aan 27 luisteraars. De resultaten tonen in hoeverre de klinkerfonemen eenduidig kunnen worden benoemd op basis van uitsluitend de akoestische informatie in de segmenten. Vergelijking met de spektrale data laat zien in hoeverre de foutieve benoemingen kunnen worden verklaard uit de gemiddelde en/of dynamische spektrale eigenschappen van de klinkersegmenten.

Voor een gedetailleerde spektrale analyse van een grote set data, zoals de onderhavige, is een efficient en automatisch analyse- en dataverwerkingssysteem gewenst. Automatische en snelle procedures voor het bepalen van de formanten zijn tot nu toe zo goed als niet beschikbaar. Bovendien zijn wij van mening dat formantparameters ook niet noodzakelijk de meest relevante parameters zijn om de spraakperceptie te beschrijven.

Wij gebruiken een bandfilteranalyse, gevolgd door een dimensionele spektrale representatie van de filternivo's in een laag-dimensionele subruimte, enigszins vergelijkbaar met het formantvlak. De analyse met behulp van een-derde oktaaf bandfilters simuleert de beperkte spektrale resolutie van het menselijk gehoororgaan. De dimensionele spektrale representatie maakt optimaal gebruik van de statistische variatie in, en de onderlinge korrelatie tussen de bandfilternivo's van vele 10-msec bemonsteringen van te analyseren spraak. Een gedetailleerde beschrijving van deze benadering wordt gegeven in paragraaf 2.1 tot 2.3.

Het gebruik van een programmeerbaar spraaksynthesesysteem is onmisbaar in klinkersegmentatie- en benoemingsexperimenten. Daarom werd een komputer-bestuurd spraaksynthesesysteem ontwikkeld. Het is in principe het spiegelbeeld van het spraakanalysesysteem, hetgeen resulteert in een vocoderachtig totaalsysteem. De details van het synthesesysteem, alsmede de ermee verkregen verstaanbaarheidsresultaten staan beschreven in paragraaf 2.5. De resultaten zijn ruim voldoende gezien het gebruik van dit synthesesysteem als onderzoekinstrument.

Het komplete systeem maakt een gedetailleerde analyse en dataverwerking van de set van 270 eenlettergrepige woorden mogelijk, zie hoofdstuk 3. Door de spektrale analyse per 10 msec is alle informatie beschikbaar om het dynamische spektrale verloop nauwkeurig te specificeren. Dit maakt het bijvoorbeeld mogelijk de dynamische spektrale eigenschappen van de Nederlandse tweeklanken *au*, *ui* en *ei* (/ɑu, ʌy, ɛi/) kwantitatief te beschrijven; het toont ons bovendien dat de Nederlandse klinkers *ee*, *oo* en *eu* (/e, o, ϕ/) op een tweeklank-achtige wijze worden geproduceerd, en een ander gedrag vertonen wanneer deze klinkers voorafgaan aan de *r*. Alhoewel er tevens aanwijzingen zijn van andere klinkerartikulatie-effekten in the spektrale data,bevatte onze woordlijst niet voldoende replikaties om deze effekten eenduidig aan te tonen.

Door de oorspronkelijke spektrale data te bekijken, maar meer nog door het verloop van het woord te bezien als spoor in een twee-dimensionele spektrale representatie, en door bovendien te luisteren naar (gedeelten uit) het gesynthetiseerde woord, konden de klinkersegmenten worden geïsoleerd. Deze geïsoleerde en gesynthetiseerde klinkersegmenten zijn vervolgens individueel aan proefpersonen aangeboden ter identifikatie. De door de 27 luisteraars gemaakte verwarringen waren vaak duidelijke aanwijzingen van koartikulatie-effekten in de klinkersegmenten. Een aantal foutieve benoemingen van klinkersegmenten kon begrepen worden vanuit de gemiddelde positie van die segmenten in de twee-dimensionele spektrale subruimte. Het dynamische spektrale verloop in de segmenten moest echter ook herhaaldelijk in de beschouwingen worden betrokken om de gemaakte foutieve benoemingen te kunnen begrijpen. De richting waarin het laatste, meest dy-

namisch verlopende, deel van het klinkersegment zich bewoog was meestal belangrijker dan het, vaak veel langere, stabiele, beginstuk van het klinkersegment.

De eindmedeklinker lijkt een grotere invloed te hebben op de eigenschappen van de klinker dan de beginmedeklinker. In hoofdstuk 4 zijn alle details van het benoemingsexperiment beschreven en worden de resultaten bediskussieerd. In hoofdstuk 1 wordt een overzicht gegeven van hetgeen in de literatuur gevonden wordt over de spektrale en perceptieve verschillen tussen klinkerfonemen onder diverse omstandigheden.

We hebben in dit onderzoek laten zien dat het goed mogelijk is met het gebruikte analyse-, dataverwerkings- en synthesesysteem klinkerkoartikulatie-effekten te bestuderen. Ook een studie van het dynamische spektrale verloop van klinkers in het algemeen is goed mogelijk. De in paragraaf 3.6 gegeven gedetailleerde beschrijving van de Nederlandse tweeklanken is hiervan een voorbeeld. Een uitbreiding in de richting van niet-klinker sonoranten, zoals nasalen en halfklinkers is in voorbereiding; de eerste resultaten lijken veelbelovend. De tot nu toe verkregen resultaten met de 270 eenlettergrepige woorden rechtvaardigen meer gedetailleerd onderzoek met meer sprekers, meer replikaties en meer kontexten.

REFERENCES

AINSWORTH, W.A. (1971). "Perception of synthesized isolated vowels and h-d words as a function of fundamental frequency", J. Acoust. Soc. Amer. *49*, 1323-1324 (L).

AINSWORTH, W.A. (1974). "Influence of fundamental frequency on perceived vowel boundaries in English", SCS-74, Stockholm.

ANDERSON, T.W. (1958). *An introduction to multivariate statistical analysis*, Wiley, New York.

ANGLIN, M.N. (1971). "Perceptual space of English vowels in word-context", M.A. thesis, Howard University Washington, D.C.

ATAL, B.S. and HANAUER, S.L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Amer. *50*, 637-655.

BALEN, C.W. van (1977). "Different views on problems of normalization". Progress Report Institute of Phonetics. Utrecht *2* (1), 32-46.

BENGUEREL, A.P. and ADELMAN, S. (1976). "Perception of coarticulated lip rounding", Phonetica *33*, 113-126.

BENINGHOF, W.J. and ROSS, M.J. (1970). "Investigation of an efficient representation of speech spectra for segmentation and classification of speech sounds", IEEE Trans AU-*18*, 33-42.

BERG, B. van den (1969). *Foniek van het Nederlands*, Van Goor Zonen, Den Haag.

BERNSTEIN, J. (1974). "Similarity structures among vowels", MIT QPR *114*, 173-179.

BERNSTEIN, J. (1975). "Quantitative phonetic theory", Paper 33 at the 8th Int. Congress of Phonetic Sciences, Leeds.

BOEHM, J.F. and WRIGHT, R.D. (1968). "Dimensional analysis and display of speech spectra", J. Acoust. Soc. Amer. *44*, 386 (A).

BOEHM, J.F. and WRIGHT, R.D. (1971). "Speaker normalization for automatic word recognition", J. Acoust. Soc. Amer. *49*, 133 (A).

BOND, Z.S. (1976a). "Identification of vowels excerpted from neutral and nasal contexts", J. Acoust. Soc. Amer. *59*, 1229-1232 (L).

BOND, Z.S. (1976b). "Identification of vowels excerpted from /l/ and /r/ contexts", J. Acoust. Soc. Amer. *60*, 906-910.

BROAD, D.J. and FERTIG, R.H. (1970). "Formant-frequency trajectories in selected CVC-syllable nuclei", J. Acoust. Soc. Amer. *47*, 1572-1582.

BROAD, D.J. and SHOUP, J.E. (1975). "Concepts for acoustic phonetic recognition", in *Speech Recognition, Invited papers presented at the 1974 IEEE Symposium*, D.R. Reddy, Ed., Academic Press, New York, 243-274.

BROECKE, M.P.R. van den (1976). "Hierarchies and rank orders in distinctive features", Doctoral thesis, University Utrecht.

BROECKE, M.P.R. van den,and Heuven, V.J.J.P. van (1976). "One or two velar fricatives in Dutch?", Progress Report Institute of Phonetics, Utrecht *1* (2), 21-25.

CARLSON, R., FANT, G. and GRANSTRÖM, B. (1975). "Two-formant models, pitch and vowel perception", in *Auditory analysis and perception of speech*, G. Fant and M.A.A. Tatham, Eds., Academic Press, London, 55-82.

CARLSON, R., GRANSTRÖM, B. and FANT, G. (1970). "Some studies concerning perception of isolated vowels", STL-QPSR *2-3*, 19-35.

CARROLL, J.D. and CHANG, J.J. (1970). "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition", Psychometrika *35*, 283-319.

CARTERETTE, E.C. (1967). "A simple linear model for vowel perception", in *Models for the perception of speech and visual forms*, W. Wathen-Dunn, Ed., MIT Press, Cambridge, 418-427.

CARTIER, M. and GRAILLOT, P. (1974). "Reduction and reconstitution of spectral data", SCS-74, Stockholm.

CASTLE, W.E. (1964). "The effect of narrow-band filtering on the perception of certain English vowels", Janua Linguarum, Series Practica XIII, Mouton & Co.

CHISTOVICH, L.A. (1971). "Auditory processing of speech stimuli - Evidence from psychoacoustics and neurophysiology", Proc. 7th ICA, Budapest, paper 21G1.

CHISTOVICH, L.A. and MUSHNIKOV, V.N. (1971). "Auditory measurements of the first formant", Proc. 7th ICA, Budapest, paper 24C17.

CLIFF, N. (1966). "Orthogonal rotation to congruence", Psychometrika *31*, 33-42.

COHEN, A. (1971). "Diphthongs, mainly Dutch", in *Form and substance*, L.L. Hammerick, R. Jakobson and E. Zwirner, Eds., Odense, 277-289.

COHEN, A., EBELING, C.L., FOKKEMA, K. and HOEK, A.G.F. van (1961). *Fonologie van het Nederlands en het Fries*, M. Nijhoff, 's-Gravenhage.

COHEN, A., SLIS, I.H. and 't HART, J. (1967). "On tolerance and intolerance in

vowel perception", Phonetica *16*, 65-70.

COHEN, P.S. and MERCER, R.L. (1975). "The phonological component of an automatic speech-recognition system", in *Speech Recognition, Invited papers presented at the 1974 IEEE Symposium*, D.R. Reddy, Ed., Academic Press, New York, 275-320.

CROWTHER, W.R. and RADER, C.M. (1966). "Efficient coding of vocoder channel signals using linear transformations", Proc. IEEE *54*, 1594-1595.

CULLER, G.J. (1970). "An attack on the problems of speech analysis and synthesis with the power of an on-line system", Proc. Int. Conf. Artificial Intelligence, Washington, D.C., 41-49.

DALE, van (1976). *Groot woordenboek der Nederlandse taal*, 10e druk, C. Kruyskamp, M. Nijhoff, 's-Gravenhage.

DANILOFF, R.G. and HAMMARBERG, R.E. (1973). "On defining coarticulation", J. Phonetics *1*, 239-248.

DELATTRE, P. (1969). "An acoustic and articulatory study of vowel reduction in four languages", Int. rev. appl. linguistics in language teaching IRAL *7*, 295-325.

DELATTRE, P.C., LIBERMAN, A.M. and COOPER, F.S. (1955). "Acoustic loci and transitional cues for consonants", J. Acoust. Soc. Amer. *27*, 769-773.

DE MORI, R. (1971). "A new method for representing a spoken word by a planar graph and for its automatic description", Proc. 7th ICA, Budapest, paper 25C11.

DERKACH, M., FANT, G. and DE SERPA-LEITÃO, A. (1970). "Phoneme coarticulation in Russian hard and soft VCV-utterances with voiceless fricatives", STL-QPSR *2-3*, 1-8.

EARLE, M.A. and PFEIFER, L.L. (1975). "Some acoustic characterisics of syllable nuclei in conversational speech", J. Acoust. Soc. Amer. *58*, S96 (A).

EEK, A. (1970). "Some coarticulation effects in Estonian", Soviet Fenno-Ugric Studies VI, 81-85.

EGUCHI, S. and HIRSH, I.J. (1969). "Development of speech sounds in children", Acta Oto-Laryngologica, Supplement 257.

FAIRBANKS, G. and GRUBB, P. (1961). "A psychophysical investigation of vowel formants", J. Speech Hear. Res. *4*, 203-219.

FANT, G. (1959). "Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Techniques *1*, 1-106.

FANT, G. (1960). *Acoustic theory of speech production*, Mouton, 's-Gravenhage.

FANT, G. (1966). "A note on vocal tract size factors and non-uniform F-pattern scaling", STL-QPSR *4*, 22-30.

FANT, G. (1970). "Sound, features, and perception", in *Proc. of the sixth international congress of phonetic sciences, Prague, 1967*, Academia, Prague. Also in Fant (1973).

FANT, G. (1973). *Speech sounds and features*, Current studies in linguistics series no. 4, the MIT Press, Cambridge.

FANT, G. (1975). "Non-uniform vowel normalization", STL-QPSR *2-3*, 1-19.

FANT, G., HENNINGSSON, G. and STÅLHAMMAR, U. (1969). "Formant frequencies of Swedish vowels", STL-QPSR *4*, 26-31.

FERRERO, F.E. (1968). "Diagrammi di esistenza delle vocali italiane", Instituto Elettrotecnico Nazionale Galileo Ferraris XLIII No. 1058-1059, 1-13.

FERRERO, F.E. (1972). "Caratteristiche acustiche dei fonemi vocalici italiani", Parole e Metodi *3*, 9-31.

FERRERO, F.E., MAGNO-CALDOGNETTO, E., VAGGES, K. and LAVAGNOLI, C. (1975). "Some acoustic and perceptual characteristics of the Italian vowels", Paper 91 at the 8th Int. Congress of Phonetic Sciences, Leeds.

FISCHER-JØRGENSEN, E. (1967). "Perceptual dimensions of vowels", in *To honor Roman Jakobson, Essays on the occasion of his seventieth birthday*, Mouton, The Hague, 667-671.

FLANAGAN, J.L. (1955). "A difference limen for vowel formant frequency", J. Acoust. Soc. Amer. *27*, 613-617.

FLANAGAN, J.L. (1957). "Difference limen for formant amplitude", J. Speech and Hearing Disorders *22*, 207-212.

FLANAGAN, J.L. (1961). "Some influences of the glottal wave upon vowel quality", 4th ICA, Helsinki.

FLANAGAN, J.L. (1972). *Speech analysis synthesis and perception*, 2nd expanded edition, Springer Verlag, Berlin.

FLANAGAN, J.L., ISHIZAKI, K. and SHIPLEY, K.L. (1975). "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell System Techn. J. *54*, 485-503.

FOULKES, J.D. (1961). "Computer identification of vowel types", J. Acoust. Soc. Amer. *33*, 7-11.

FOURCIN, A.J. and ABBERTON, E. (1971). "First applications of a new laryngograph", Medical & Biological Illustrations *21*, 172-182.

FRØKJÆR-JENSEN, B. (1967). "Statistic calculations of formant data", ARIPUC *2*, 158-170.

FUJIMURA, O. and OCHIAI, K. (1963). "Vowel identification and phonetic contexts", J. Acoust. Soc. Amer. *35*, 1889 (A).

FUJISAKI, H. and KAWASHIMA, T. (1968). "The roles of pitch and higher formants

in the perception of vowels", IEEE Trans. AU-*16*, 73-77.

FUJISAKI, H. and TANABE, Y. (1972). "A time-domain technique for pitch extraction of speech", Annual Report of the Engineering Research Institute *31*, 259-266.

FUJISAKI, H., YOSHIDA, M. and SATO, Y. (1974). "Formulation of the coarticulatory process in the formant frequency domain and its application to automatic recognition of connected vowels", SCS-74, Stockholm.

GAY, T. (1968). "Effect of speaking rate on diphthong formant movements", J. Acoust. Soc. Amer. *44*, 1570-1573.

GAY, T. (1970). "A perceptual study of American English diphthongs", Language and Speech *13*, 65-88.

GAY, T. (1975). "Some electromyographic measures of coarticulation in VCV utterances", Haskins SR-*44*, 137-145.

GEER, J.P. van de (1967). *Inleiding in de multivariate analyse*, Van Loghum Slaterus, Arnhem.

GERSTMAN, L.J. (1968). "Classification of self-normalized vowels", IEEE Trans. AU-*16*, 78-80.

GOLD, B. and RABINER, L.R. (1969). "Parallel processing techniques for estimating pitch periods of speech in the time domain", J. Acoust. Soc. Amer. *46*, 442-448.

GOVAERTS, G. (1974). "Psychologische en fysische strukturen van perceptueel geselekteerde klinkers. Een onderzoek aan de hand van Zuidnederlandse klinkers", Doctoral dissertation, University of Leuven.

GRANT, M. (1971). "INDSCAL analysis of Hindi and English vowels", M.A. Thesis, Howard University Washington, D.C.

GRIMM, W.A. (1966). "Perception of segments of English-spoken consonant-vowel syllables", J. Acoust. Soc. Amer. *40*, 1454-1461.

GUPTA, J.P., AGRAWAL, S.S. and AHMED, R. (1971). "Perception of (Hindi) vowels in clipped speech", J. Acoust. Soc. Amer. *49*, 567-568.

HANSON, G. (1967). "Dimensions in speech sound perception. An experimental study of vowel perception", Ericsson Technics *23*, 3-175.

HARMAN, H.H. (1967). *Modern factor analysis*, The University of Chicago Press, Chicago.

HARSHMAN, R.A. (1970). "Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis", UCLA Working Papers in Phonetics *16*.

HART, J. 't (1969). "Fonetische steunpunten", De Nieuwe Taalgids *62*, 168-174.

HART, J. 't and COHEN, A. (1964). "Gating techniques as an aid in speech analysis", Language and Speech *7*, 22-39.

HEMDAL, J.F. and HUGHES, G.W. (1967). "A feature-based computer recognition program for the modelling of vowel perception", in *Models for the perception of speech and visual form*, W. Wathen-Dunn, Ed., MIT-Press, Cambridge, 440-452.

HESS, W. (1972). "Digitale grundfrequenzsynchrone Analyse von Sprachsignalen als Teil eines automatischen Spracherkennungssystems", Doctoral dissertation,Un. of München.

HOLBROOK, A. and FAIRBANKS, G. (1962). "Diphthong formants and their movements", J. Speech Hear. Res. *5*, 33-58.

HORST, P. (1965). *Factor analysis of data matrices*, Holt, Rinehart and Winston, New York.

HOUSE, A.S. and FAIRBANKS, G. (1953). "The influence of consonant environment upon the secondary acoustical characteristics of vowels", J. Acoust. Soc. Amer. *25*, 105-113.

HOUTGAST, T. (1974a). "Lateral suppression in hearing. A psychophysical study on the ear's capability to preserve and enhance spectral contrasts", Doctoral thesis, Vrije Universiteit Amsterdam.

HOUTGAST, T. (1974b). "Auditory analysis of vowel-like sounds", Acustica *31*, 320-324.

HUGHES, G.W., HOUSE, A.S. and LI, K.-P. (1969). "Research on word spotting", Report AFCRL-69-0240.

HUIZING, H.C. and MOLENAAR-BIJL, A. (1944). "De betekenis der klankfrequentie in het Nederlands voor de oorheelkunde", N.T.G. *88*, 435.

International Phonetic Association, (1967). "The principles of the International Phonetic Association", Dept. of Phonetics, University College, London.

ITAKURA, F. (1975). "Minimum prediction residual principle applied to speech recognition", IEEE Trans. ASSP *23*, 67-72.

JAKOBSEN, R., FANT, G. and HALLE, M. (1952). "Preliminaries to speech analysis. The distinctive features and their correlates". Acoust. Lab., MIT Techn. Rep. No. 13, Cambridge; 4th printing published by the MIT Press, Cambridge, 1963.

JASSEM, W. (1968). "Vowel formant frequencies as cues to speaker discrimination", in *Speech analysis and synthesis, I*, W. Jassem, Ed., Warsaw, 9-41.

JÖRGENSEN, H.P. (1969). "Die gespannten und ungespannten Vokale in der norddeutschen Hochsprache mit einer spezifischen Untersuchung in der Struktur ihrer Formantfrequenzen", Phonetica *19*, 217-245.

KAKUSHO, O. and KATO, K. (1968). "Just discriminable change and matching range of acoustic parameters of vowels", Acustica *20*, 46-54.

KAMENY, I. (1974). "Comparison of the formant spaces of retroflexed and nonretroflexed vowels", IEEE Trans. ASSP-*23*, 38-49.

KAMP, L.J.Th. van der and POLS, L.C.W. (1971). "Perceptual analysis from confusions between vowels", Acta Psychologica *35*, 64-77.

KANAMORI, Y. (1975). "On the characteristics of individual vowels and the statistical characteristics of formant frequency patterns in connected speech", Systems Computers Controls *6*, 22-30.

KANAMORI, Y., KASUYA, H., ARAI, S. and KIDO, K. (1971). "Effect of context on vowel perception", Proc. 7th ICA, Budapest, paper 20C4.

KANAMORI, Y. and KIDO, K. (1976). "Recognition of vowels in connected speech by use of the characteristics on perception of vowel", Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Philadelphia, 174-177.

KARNICKAYA, E.G., MUSHNIKOV, V.N., SLEPOKUROVA, N.A. and ZHUKOV, S.J. (1975). "Auditory processing of steady-state vowels", in *Auditory analysis and perception of speech*, G. Fant and M.A.A. Tatham, Eds., Academic Press, London, 37-53.

KASUYA, K., KANAMORI, Y. and KIDO, K. (1971). "Psychological auditory space representing vowel quality", Proc. 7th ICA, Budapest, paper 20C5.

KIDO, K., KASUYA, H. and SUZUKI, H. (1968). "Discrimination of Japanese vowels in connected speech", Proc. 6th ICA, Tokyo, paper B-4-6.

KLATT, D.H. (1976). "A digital filter bank for spectral matching", Conference Record IEEE ASSP, Philadelphia, 573-576.

KLEIN, W., PLOMP, R. and POLS, L.C.W. (1970). "Vowel spectra, vowel spaces, and vowel identification", J. Acoust. Soc. Amer. *48*, 999-1009.

KNOPS, L. (1967). "De auditieve struktuur der spraaklanken", Psychol. Belg. *7*, 59-65.

KOOPMANS-van BEINUM, F.J. (1969). "Nog meer fonetische zekerheden", Nieuwe Taalgids *62*, 245-250.

KOOPMANS-van BEINUM, F.J. (1971). "Vergelijkend fonetisch klinkeronderzoek", IPA publication Nr. *32*, 1-59.

KOOPMANS-van BEINUM, F.J. (1973). "Comparative phonetic vowel analysis", J. of Phonetics *1*, 249-261.

KOOPMANS-van BEINUM, F.J. (1975). "Vowel reduction in Dutch", Paper 156 at the 8th Int. Congress of Phonetic Sciences, Leeds.

KOZHEVNIKOV, V.A. and CHISTOVICH, L.A. (1965). "Speech: articulation and perception", translated from Russian by Joint Publications Research Service, U.S. Dept. Commerce, Washington, D.C. *30*, 543.

KRAMER, H.P. and MATHEWS, M.V. (1956). "A linear coding for transmitting a set of correlated signals", IRE Trans. IT-*2*, 41-45.

KRUSKAL, J.B. (1964a). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", Psychometrika *29*, 1-27.

KRUSKAL, J.B. (1964b). "Nonmetric multidimensional scaling: a numerical method", Psychometrika *29*, 115-129.

KRYTER, K.D. (1962). "Methods for the calculation and use of the articulation index", J. Acoust. Soc. Amer. *34*, 1689-1697.

KRYTER, K.D. (1972). "Speech Communication", Chapter 5 in *Human Engineering Guide to Equipment Design*, H.P. Van Cott and R.C. Kinkade, Eds., U.S. Government Printing Office, Washington, D.C.

KULYA, V.I. (1964). "Experimental investigation of the correlation relations in the speech spectrum and a comparison of some variants of orthogonal vocoder", Telecommunications (USSR) *18*, 39-50.

KUWAHARA, H. and SAKAI, H. (1973). "Perception of vowels and C-V syllables segmented from connected speech", Bulletin of the NHK Broadcasting Science Research Laboratories *7*, 36-45.

KUWAHARA, H. and SAKAI, H. (1976). "Fusion and identification of synthetic vowels in dichotic listening", Conference Record IEEE ASSP, Philadelphia, 186-189.

LADEFOGED, P. and BROADBENT, D.E. (1957). "Information conveyed by vowels", J. Acoust. Soc. Amer. *29*, 98-104.

LADEFOGED, P., KAMENY, I. and BRACKENRIDGE, W.A. (1976). "Acoustic effects of style of speech", J. Acoust. Soc. Amer. *59*, 228-231 (L).

LANDERCY, A. and RENARD, R. (1975). "Zones fréquentielles et reconnaissance de voyelles françaises", Revue de Phonétique Appliquée, *33-34*, 51-79.

LÄNGLE, D and PAULUS, E. (1974). "Efficiency and limitations of linear transformations in digital speech transmission", SCS-74, Stockholm.

LEA, W.A., MEDRESS, M.F. and SKINNER, T.E. (1973). "Prosodic aids to speech recognition: III. Relationships between stress and phonemic recognition results", Sperry Univac Report No. PX 10430, Sept.

LEHISTE, I. and IVIĆ, P. (1963). "Accent in Serbocroation. An experimental study", Michigan Slavic Materials No. 4.

LEHISTE, I. and PETERSON, G.E. (1959). "The identification of filtered vowels", Phonetica *4*, 161-177.

LEHISTE, I. and PETERSON, G.E. (1961). "Transitions, glides, and diphthongs", J. Acoust. Soc. Amer. *33*, 268-277.

LEHISTE, I. and SHOCKEY, L. (1972). "On the perception of coarticulation effects in English VCV syllables", J. Speech and Hear. Res. *15*, 500-507.

LI, K.-P., HOUSE, A.S. and HUGHES, G.W. (1968). "Vowel classification using a dispersion-analysis method", J. Acoust. Soc. Amer. *44*, 390 (A).

LI, K.-P., HUGHES, G.W. and HOUSE, A.S. (1969). "Correlation characteristics and dimensionality of speech spectra", J. Acoust. Soc. Amer. *46*, 1019-1025.

LI, K.-P., HUGHES, G.W. and HOUSE, A.S. (1971). "Intelligibility of speech reconstituted from reduced spectral data", J. Acoust. Soc. Amer. *49*, 134 (A).

LI, K.-P., HUGHES, G.W. and HOUSE, A.S. (1972). "Effect of context on vowel recognition by dispersion analysis", J. Acoust. Soc. Amer. *51*, 130 (A).

LI, K.-P., HUGHES, G.W. and HOUSE, A.S. (1973a). "Speech reconstituted from spectra of reduced dimensionality: a study of intelligibility", J. Acoust. Soc. Amer. *53*, 329 (A).

LI, K.-P., HUGHES, G.W. and SNOW, T.B. (1973b). "Segment classification in continuous speech", IEEE Trans AU-*21*, 50-57.

LIBERMAN, A.M., COOPER, F.S., HARRIS, K.S., MACNEILAGE, P.F. and STUDDERT-KENNEDY, M. (1967). "Some observations on a model for speech perception", in *Models for the perception of speech and visual form*, W. Wathen-Dunn, Ed., MIT-Press, Cambridge, 68-87.

LIIV, G. and REMMEL, M. (1970). "On acoustic distinctions in the Estonian vowel system", Soviet Fenno-Ugric Studies VI, 7-23.

LINDBLOM, B. (1963). "Spectrographic study of vowel reduction", J. Acoust. Soc. Amer. *35*, 1773-1781.

LINDBLOM, B.E.F. and STUDDERT-KENNEDY, M. (1967). "On the rôle of formant transitions in vowel recognition", J. Acoust. Soc. Amer. *42*, 830-843.

LOBANOV, B.M. (1971). "Classification of Russian vowels spoken by different speakers", J. Acoust. Soc. Amer. *49*, 606-608 (L).

MACNEILAGE, P.F. and DECLERK, J.L. (1969). "On the motor control of coarticulation in CVC monosyllables", J. Acoust. Soc. Amer. *45*, 1217-1233.

MAJEWSKI, W. and HOLLIEN, H. (1967). "Formant frequency region of Polish vowels", J. Acoust. Soc. Amer. *42*, 1031-1037.

MAKHOUL, J.I. and WOLF, J.J. (1972). "Linear prediction and the spectral analysis of speech", BBN Report No. 2304.

MARKEL, J.D. (1972). "The SIFT algorithm for fundamental frequency estimation", IEEE Trans. AU-*20*, 367-377.

MARKEL, J.D. (1973). "Application of a digital inverse filter for automatic formant and F_0 analysis", IEEE Trans. AU-*21*, 154-160.

MARKEL, J.D. and GRAY, A.H. (1974). "A linear prediction vocoder simulation based upon the autocorrelation method", IEEE Trans. ASSP-*22*, 124-134.

MATSUMOTO, H., HIKI, S., SONE, T. and NIMURA, T. (1973). "Multidimensional representation of personal quality of vowels and its acoustical correlates", IEEE Trans. AU-*21*, 428-436.

MCCANDLESS, S.S. (1974a). "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. ASSP-*22*, 135-141.

MCCANDLESS, S.S. (1974b). "Use of formant motion in speech recognition", Proc. IEEE Symp. on Speech Recognition, Pittsburgh, 211.

MCGONEGAL, C.A., RABINER, L.R. and ROSENBERG, A.F. (1975). "A semi-automatic pitch detector (SAPD)", IEEE Trans. ASSP-*23*, 570-574.

MERMELSTEIN, P. (1975). "Vowel perception in consonantal context", J. Acoust. Soc. Amer. *58*, S56 (A).

MERMELSTEIN, P. (1976). Personal communication.

MILLER, R.L. (1953). "Auditory tests with synthetic vowels", J. Acoust. Soc. Amer. *25*, 114-121.

MILLER, G.A. (1956). "The perception of speech", in *For R. Jakobson: Essays on the occasion of his sixtieth birthday*, M. Halle *et al.*, Eds., Mouton & Co, 's-Gravenhage, 353-359.

MILLER, G.A. and NICELY, P.E. (1955). "An analysis of perceptual confusion among some English consonants", J. Acoust. Soc. Amer. *27*, 338-352.

MOHR, B. and WANG, W.S.I. (1968). "Perceptual distances and the specification of phonological features", Phonetica *18*, 31-45.

MOL, H. (1969). "Fonetische zekerheden", Nieuwe Taalgids *62*, 161-167.

MUSHNIKOV, V.N. and CHISTOVICH, L.A. (1973). "Experimental testing of the band hypothesis of vowel perception", Sov. Phys. Acoust. *19*, 250-254.

NEWMAN, R. (1971). "A syntactic approach to the recognition of certain dynamic speech sounds", Report TR-EE 71-18. Purdue University.

NIEDERJOHN, R.J. (1975). "A mathematical formulation and comparison of zero-crossing analysis techniques which have been applied to ASR", IEEE Trans. ASSP-*23*, 373-380.

NIEROP, D.J.P.J. van, POLS, L.C.W. and PLOMP, R. (1973). "Frequency analysis of Dutch vowels from 25 female speakers", Acustica *29*, 110-118.

NOLL, A.M. (1967). "Cepstrum pitch determination", J. Acoust. Soc. Amer. *41*, 293-309.

NOOTEBOOM, S.G. (1968). "Perceptual confusions among Dutch vowels presented in noise", IPO Annual Progress Report *3*, 68-71.

NOOTEBOOM, S.G. (1972). "Production and perception of vowel duration. A study of durational properties of vowels in Dutch", Doctoral thesis, University Utrecht.

NORDSTRÖM, P.-E. (1975). "Attempts to simulate female and infant vocal tracts

from male area functions", STL-QPSR *2-3*, 20-33.

NORDSTRÖM, P.-E. and LINDBLOM, B. (1975). "A normalization procedure for vowel formant data", Paper 212 at the 8th Int. Congress of Phonetic Sciences, Leeds.

OCHIAI, K. and FUJIMURA, O. (1971). "Vowel identification and phonetic contexts", Rep. Univ. Electro-Comm. *22-2*, (Sci. & Techn. Sect.), 103-111.

OHDE, R.N. and SHARF, D.J. (1975). "Coarticulatory effects of voiced stops on the reduction of acoustic vowel targets", J. Acoust. Soc. Amer. *58*, 923-927 (L).

ÖHMAN, S. (1964). "Note on palatalization in Russian", MIT QPR *73*, 167-171.

ÖHMAN, S.E.G. (1966). "Coarticulation in VCV-utterances: spectrographic measurements", J. Acoust. Soc. Amer. *39*, 151-168.

ÖHMAN, S.E. (1967). "Numerical model of coarticulation", J. Acoust. Soc. Amer. *41*, 310-320.

OSHIKA, B.T., ZUE, V.W., WEEKS, R.V., NEU, H. and AURBACH, J. (1975). "The role of phonological rules in speech understanding research", IEEE Trans. ASSP-*23*, 104-112.

OSTREICHER, H.J. and SHARF, D.J. (1976). "Effects of coarticulation on the identification of deleted consonant and vowel sounds", J. Phonetics *4*, 285-301.

PETERSON, G.E. and BARNEY, H.L. (1952). "Control methods used in a study of the vowels", J. Acoust. Soc. Amer. *24*, 175-184.

PFEIFER, L.L. (1972). "Isolated-word phoneme recognition using features derived from wave function parameters", Proc. Conf. on Speech Comm. and Proc., Boston, paper C2.

PICKETT, J.M. (1957). "Perception of vowels heard in noises of various spectra", J. Acoust. Soc. Amer. *29*, 613-620.

PLOMP, R. (1964). "The ear as a frequency analyzer", J. Acoust. Soc. Amer. *36*, 1628-1636.

PLOMP, R. (1976). *Aspects of tone sensation. A psychophysical study*, Academic Press, London.

PLOMP, R., POLS, L.C.W. and GEER, J.P. van de (1967). "Dimensional analysis of vowel spectra", J. Acoust. Soc. Amer. *41*, 707-712.

PLOMP, R. and STEENEKEN, H.J.M. (1973). "Place dependence of timbre in reverberant sound fields", Acustica *28*, 50-59.

POLS, L.C.W. (1970a). "Perceptual space of vowel-like sounds and its correlation with frequency spectrum", in *Frequency analysis and periodicity detection in hearing*, R. Plomp and G.F. Smoorenburg, Eds., A.W. Sijthoff, Leiden, 463-473.

POLS, L.C.W. (1970b). "Speech analysis and recognition", report IZF 1970-16.

POLS, L.C.W. (1971a). "Real-time recognition of spoken words", IEEE Trans. C-*20*, 972-978.

POLS, L.C.W. (1971b). "Dimensional representation of speech spectra", Proc. 7th ICA, Budapest, paper 25C7.

POLS, L.C.W. (1972). "Segmentation and recognition of mono-syllabic words", Proc. Conf. on Speech Comm. and Proc., Boston, paper C5.

POLS, L.C.W. (1974). "Intelligibility of speech resynthesized by using a dimensional spectral representation", SCS-74, Stockholm.

POLS, L.C.W. (1975). "Dominant spectral regions in vowel perception", Paper 347 at the 8th Int. Congress of Phonetic Sciences, Leeds.

POLS, L.C.W., KAMP, L.J.Th. van der, and PLOMP, R. (1969). "Perceptual and physical space of vowel sounds", J. Acoust. Soc. Amer. *46*, 458-467.

POLS, L.C.W. and SCHOUTEN, M.E.H. (1975). "Invloed van de kontekst op de dynamische spektrale eigenschappen van klinkers, tweeklanken en glijklanken", ZWO-projekt no. 30-50.

POLS, L.C.W., TROMP, H.R.C. and PLOMP, R. (1973). "Frequency analysis of Dutch vowels from 50 male speakers", J. Acoust. Soc. Amer. *53*, 1093-1101.

POTTER, R.K. and STEINBERG, J.C. (1950). "Toward the specification of speech", J. Acoust. Soc. Amer. *22*, 807-820.

RABINER, L.R. and SAMBUR, M.R. (1975). "An algorithm for determining the endpoints of isolated utterances", The Bell System Technical J. *54*, 297-315.

REDDY, D.R., ERMAN, L.D. and NEELY, R.B. (1973). "A model and a system for machine recognition of speech", IEEE Trans. AU-*21*, 229-238.

REITSMA, M.H. (1974). "The influence of the diminutive suffix on the preceding vowel", IPA report, Amsterdam.

RETZ, D.L. (1973). "Signal processing techniques based upon a wave packet representation", J. Acoust. Soc. Amer. *53*, 321 (A).

REYNTJES, J.A. (1951). *Spraakaudiometrie*, Doctoral thesis, Groningen University.

RIEMERSMA, J.B.J. and BURRIJ, S. (1973). "A general program for the analysis of variance of factorial and hierarchical designs", Institute for Perception TNO, report IZF 1973-C4.

ROSENBERG, A.E. (1971). "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Amer. *49*, 583-590.

SCARR, R.W.A. (1968). "Zero crossings as a means for obtaining spectral information in speech analysis", IEEE Trans. AU-*16*, 247-255.

SEIDEL, H. (1974). "Vorverarbeitungstechniken zur Datenreduktion bei Kurzzeitfrequenzspektren von Sprachsignalen", Doctoral thesis, TU MÜnchen.

SEIDEL, H. and PAULUS, E. (1971). "Hauptkomponentenanalyse von Sprachdaten", in

Papers in interdisciplinary speech research, Proceedings of the Speech Symposium, Szeged.

SHANKWEILER, D., STRANGE, W. and VERBRUGGE, R. (1975). "On accounting for the poor recognition of isolated vowels", Haskins Labs SR-*42/43*, 285-296.

SHEPARD, R.N. (1972). "Psychological representation of speech sounds", in *Human Communication: a Unified View*, D.E. David and P.B. Denes, Eds., McGraw-Hill, New York, 67-113.

SINGH, S. (1974). "A step towards a theory of speech perception", SCS-74, Stockholm.

SINGH, S.S. (1975)., Ed., *Measurement procedures in speech, hearing and language*, University Park Press, Baltimore.

SINGH, S. and WOODS, D.R. (1971). "Perceptual structure of 12 American English vowels", J. Acoust. Soc. Amer. *49*, 1861-1866.

SLAWSON, A.W. (1968). "Vowel quality and musical timbre as functions of spectral envelope and fundamental frequency", J. Acoust. Soc. Amer. *43*, 87-101.

SLIS, I.H. and KATWIJK, A.F.V. van (1963). "Onderzoek naar Nederlandse tweeklanken", IPO report nr. 31.

STÅLHAMMAR, U., KARLSSON, I. and FANT, G. (1973). "Contextual effects on vowel nuclei", STL-QPSR *4*, 1-18.

STEVENS, K.N. (1951). "Perception of sounds shaped by resonant circuits", MIT QPR, Cambridge.

STEVENS, K.N. and HOUSE, A.S. (1963). "Perturbation of vowel articulations by consonantal context: an acoustical study", J. Speech Hear. Res. *6*, 111-128.

STEVENS, K.N., HOUSE, A.S. and PAUL, A.P. (1966). "Acoustical description of syllabic nuclei: an interpretation in terms of a dynamic model of articulation", J. Acoust. Soc. Amer. *40*, 123-132.

STEVENS, K.N., KALIKOW, D.N. and WILLEMAIN, T.R. (1975). "A miniature accelerometer for detecting glottal waveforms and nasalization", J. Speech Hear. Res. *18*, 494-499.

STRANGE, W., VERBRUGGE, R.R., SHANKWEILER, D.P. and EDMAN, T.R. (1976). "Consonant environment specifies vowel identity", Haskins SR-*45/46*, 37-61; also J. Acoust. Soc. Amer. *60*, 213-224.

SUZUKI, H., KASUYA, H. and KIDO, K. (1967). "The acoustic parameters for vowel recognition without distinction of speakers", Proc. Conf. on Speech Comm. and Proc., Bedford, paper B5.

TANAKA, H. (1972). "Hadamard transform for speech wave analysis", Stanford Artificial Intell. Proj., Memo AIM-175, STAN-LS-307.

TARNÓCZY, T. and RADNAI, J. (1971). "Eine Möglichkeit automatischer Erkennung von Vokalen", Proc. 7th ICA, Budapest, paper 20C12.

TERBEEK, D. and HARSHMAN, R. (1971). "Cross-language differences in the perception of natural vowel sounds", UCLA Working Papers in Phonetics *19*.

TERBEEK, D. and HARSHMAN, R. (1972). "Is vowel perception non-Euclidean?", J. Acoust. Soc. Amer. *51*, 81 (A).

TIFFANY, W.R. (1959). "Non-random sources of variation in vowel quality", J. Speech and Hear. Res. *2*, 305-317.

TVERSKY, A. and KRANTZ, D.H. (1970). "The dimensional representation and the metric structure of similarity data", J. of Math. Psych. *7*, 572-596.

VERBRUGGE, R.R., STRANGE, W., SHANKWEILER, D.P. and EDMAN, T.R. (1976). "What information enables a listener to map a talker's vowel space", Haskins SR-*45/46*, 63-94; also J. Acoust. Soc. Amer. *60*, 198-212.

WAKITA, H. (1973). "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans AU-*21*, 417-427.

WAKITA, H. (1975). "An approach to vowel normalization", Techn. report SCRL, Santa Barbara.

WATANABE, S. (1965). "Karhunen-Loëve expansion and factor analysis. Theoretical remarks and applications", Proc. 4th Conf. Information Theory, Prague.

WEINSTEIN, C.J., MCCANDLESS, S.S., MONDSHEIN, L.F. and ZUE, V.W. (1974). "A system for acoustic-phonetic analysis of continuous speech", Proc. IEEE Symp. on Speech Recognition, Pittsburgh, 89-100.

WELCH, P.D. and WIMPRESS, R.S. (1961). "Two multivariate statistical computer programs and their application to the vowel recognition problem", J. Acoust. Soc. Amer. *33*, 426-434.

WHITE, G.M. (1976). "Speech recognition: A tutorial review", Computer *9*, 40-53.

WICKELGREN, W.A. (1965). "Distinctive features and errors in short term memory for English vowels", J. Acoust. Soc. Amer. *38*, 583-588.

WRIGHT, R.D. (1972). "Dimensional analysis of a full vowel set", J. Acoust. Soc. Amer. *52*, 182 (A).

YILMAZ, H. (1967). "A theory of speech perception", Bull. Math. Biophysics *29*, 793-824.

YILMAZ, H. (1972). "Statistical theory of speech perception", Proc. Conf. on Speech Comm. and Proc., Boston, paper F9.

YING, Y., SHUM, F., ELLIOTT, A.R. and OWEN BROWN, W. (1973). "Speech processing with Walsh-Hadamard Transforms", IEEE Trans AU-*21*, 174-179.

ZURCHER, J.F., GRAILLOT, P., CARTIER, M., DAVID, G., BREANT, P. and UFFELEN, J.P. van (1976). "Speech digitization with channel vocoders", Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Philadelphia, 95-97.

List of phonetic vowel symbols, together with the orthographic symbols and Dutch key words.

/ɑ/	a	gaf
/a/	aa	naaf
/ɛ/	e	ken
/I/	i	fit
/e/	ee	beek
/i/	ie	vies
/ɔ/	o	som
/o/	oo	zoon
/u/	oe	doek
/œ/	u	rug
/ø/	eu	heup
/y/	uu	tuut
/ə/		de
/ɑu/	au	pauk
/ʌy/	ui	muis
/ɛi/	ei	lijm

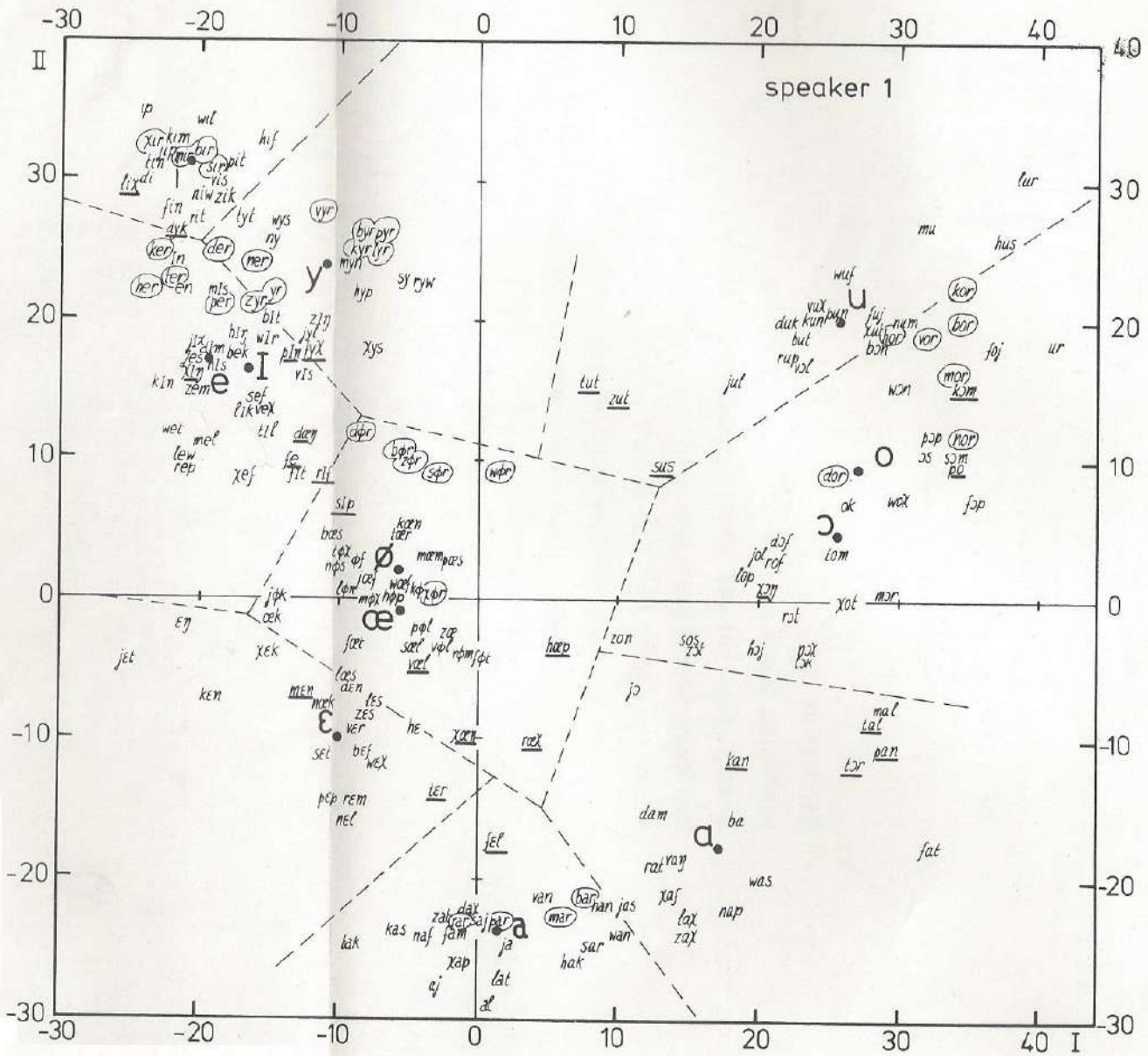


Fig. 3.5.7. For caption, see p. 86.

ZWICKER, E. and FELDTKELLER, R. (1967). *Das Ohr als Nachrichtenempfänger*, S. Hirzel Verlag, Stuttgart.

STELLINGEN

I

De dimensionele spektrale representatie van klinkers, gebaseerd op bandfilterniveau's, is in vrijwel alle opzichten gelijkwaardig aan de met meer moeite, en minder objektief, te verkrijgen formantrepresentatie.

II

De aanwezigheid van de /r/ als eindmedeklinker in zorgvuldig uitgesproken woorden van het type beginmedeklinker-klinker-eindmedeklinker, beinvloedt in hoge mate de spektrale eigenschappen van de in die woorden voorkomende klinker. Dit effekt is het duidelijkst voor *ee*, *oo*, en *eu* (/e, o, ϕ/).

III

Voor een korrekte benoeming van de uit woorden geisoleerde Nederlandse tweeklanken *au*, *ui* en *ei* (/ɑu, ʌy, ɛi/) is de spektrale startpositie van deze klanken zeer belangrijk, alsook een indikatie van de richting waarin de spektrale overgang zich beweegt. Het ook werkelijk bereiken van een specifieke eindpositie is hierbij niet noodzakelijk.

IV

Bij het psychoakoestisch aftasten van het lijnenspektrum van een komplex signaal (cosinus-optelling) met behulp van de „pulsatiemethode", wordt bij grondfrekwenties f lager dan ca. 160 Hz voor frekwenties hoger dan ca. $15f$, geen reëel beeld verkregen van de „interne representatie" van dit signaal.

T. Houtgast. Lateral suppression in hearing. Proefschrift. Amsterdam, 1974.

V

Grote voorzichtigheid is geboden bij het, in de naaste toekomst ook in Nederland te verwachten gerechtelijke gebruik van spraakspektrogrammen ("visible speech") voor het identificeren van personen op grond van opgenomen spraak.

VI

De recente ontwikkeling van "speech understanding systems" benadrukt sterk de syntaktische, semantische en zakelijke informatie in gesproken tekst. De tevens in het signaal aanwezige akoestisch-fonetische informatie wordt hierbij ten onrechte naar de achtergrond geschoven.

VII

Indien een komponist prijs stelt op een goede verstaanbaarheid van de tekst van vokale partijen dan dient hij geen noten boven F5 (f", 700 Hz) te laten zingen.

VIII

De in een recent ontwerp van aanbeveling van de ISO voorgestelde methode om het nivo van een gesproken tekst te meten door maximale wijzeruitslagen op een dB(A)-meter op het oog te middelen, gaat geheel voorbij aan reeds lange tijd beschikbare modernere meettechnieken.

Second draft proposal ISO/DP 4870 for "Acoustics-Recommended methods for the construction and calibration of speech intelligibility tests", augustus 1976..

IX

Het feit dat leerlingen op school nieuwe woorden nog al eens verkeerd aanleren, vindt mede zijn oorzaak in de vaak slechte akoestische omstandigheden in leslokalen.

X

Alhoewel de Nederlandse klinkers *ee*, *oo* en *eu* (/e, o, ϕ/) niet tot de tweeklanken worden gerekend, is het dynamische spektrale verloop binnen deze klinkers van wezenlijk belang voor hun korrekte benoeming.

XI

Het is een misvatting te menen dat bejaarden, vanwege hun ouderdomsslechthorendheid, kunnen worden blootgesteld aan meer achtergrondlawaai. Voor een optimale kommunikatie zijn juist stille en akoestisch goed behandelde ruimten gewenst.

XII

Behalve veiligheids- en komforteisen zouden aan het gebruik van een valhelm in het verkeer ook eisen gesteld moeten worden betreffende de waarneembaarheid van geluiden.

XIII

Uit frekwentie-selektieve maskeerproeven met klinkers blijkt dat de meest prominente maxima in de omhullende van het amplitudespektrum (formanten) niet altijd ook de meest dominante spektrale gebieden zijn voor een korrekte benoeming van die klinkers.

XIV

Van wetenschappelijke onderzoekers kan niet worden verlangd dat zijzelf jaar in jaar uit goed doordachte projektaanvragen produceren en bovendien die van anderen beoordelen, wanneer steeds weer vele goede aanvragen door de subsidie-verstrekkende instanties niet (kunnen) worden gehonoreerd.

Stellingen bij: Spectral analysis and identification of Dutch vowels in monosyllabic words.

L.C.W. Pols

Amsterdam, 16 juni 1977