UvA-DARE (Digital Academic Repository)

Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children

Pols, L.C.W.; Bakkum, M.J.; Plomp, R.

Publication date
1995

Published in
The Journal of the Acoustical Society of America

[Link to publication](#)

Citation for published version (APA):

Pols, L. C. W., Bakkum, M. J., & Plomp, R. (1995). Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children. *The Journal of the Acoustical Society of America*, 98(2).

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children

Mark J. Bakkum and Reinier Plomp

Department of Oto-Rhino-Laryngology, Free University Hospital, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands

Louis C. W. Pols

Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands

(Received 23 February 1994; revised 2 March 1995; accepted 13 March 1995)

Objective whole-spectrum and formant analyses have been performed on all 15 Dutch vowels pronounced in /C₁VC₂/ words by 24 deaf and 24 normal-hearing children, in order to develop a model of pronunciation quality for evaluating (deaf) speech; the results as obtained for adult males by Bakkum *et al.* [J. Acoust. Soc. Am. **94**, 1989–2004 (1993)] have been verified and extended. Spectral representations of the vowels were created by determining the output levels of a bank of 16 filters (90–7200 Hz), with 1/3-oct bandwidths and logarithmic spacing of their center frequencies. Spectral differences agree well with subjective differences in pronunciation quality obtained from magnitude estimation and identification experiments. Spectral differences not related to pronunciation quality judgments arise as a consequence of physiological interspeaker differences and variation in fundamental frequency, but these differences can be compensated for by speaker-normalization and F₀-compensation procedures. Using principal components analysis (PCA), the vowel spectra can be described by a limited number of dimensions, without losing much information; a description in a two-dimensional PCA subspace still agrees well with the subjective judgments and it also agrees with a description by the first two formants. The *whole-spectrum* approach provides a determinate, readily interpretable model of pronunciation quality for evaluating vowels. As a practical advantage, its computational requirements are modest and, in conjunction with PCA, the vowel dynamics can be visualized, which makes the approach suitable for vowel training and diagnostics. © 1995 Acoustical Society of America.

PACS numbers: 43.71.Es, 43.71.Gv, 43.72.Ar

INTRODUCTION

One of the issues in speech research is to obtain an adequate acoustic description of vowels. Such an objective description may clarify the relation between vowel acoustics and vowel production for different speakers, it may clarify the nature of vowel perception, and it can also produce an objective judgment of *pronunciation quality*, which is essential for vowel training and diagnostics. Here, pronunciation (or articulatory) quality is defined as the extent to which the ideal utterance is reached, regarding those aspects that the speaker is supposed to have under control by his/her articulators; speaker-specific aspects of voicing (voice quality) are thus disregarded. The adequacy of the acoustic description can be expressed by the extent of its agreement with subjective assessment, by its ability to deal, in a determinate way, with a wide variety of vowels and speakers, and by the degree to which it is clearly interpretable. Furthermore, it should preferably be in accordance with known properties of the human ear, and the analysis should require limited computational effort. A spectral analysis is generally accepted to be the basis of such an objective description, and the approaches found in the literature can roughly be divided in two main directions. Most commonly used, usually with satisfactory results, is *formant analysis* (see, among many others, Peterson and Barney, 1952; Strange, 1989a). However, a *whole-spectrum analysis* is preferred by other researchers for

its completeness and determinacy (cf. Plomp *et al.*, 1967; Bladon, 1982; Zahorian and Jagharghi, 1993). In an earlier paper (Bakkum *et al.*, 1993) we have shown promising results when applying the whole-spectrum approach to describe the pronunciation quality of vowels spoken by native, non-native, and deaf adult male speakers of Dutch. The objective analysis by 16 filter bands showed a good agreement with subjective assessment. The analysis was a first-order approximation to the concept of the ear's critical bands since the filter bandwidths were comparable with the equivalent rectangular bandwidth (ERB) functions of the auditory filter (cf. Patterson and Moore, 1986), though somewhat broader, especially at low frequencies. The original 16-dimensional description could be reduced by *principal components analysis* (PCA) to only two dimensions without losing too much information. The vowels can then be visualized in an effective way, comparable to the construction of the *vowel triangle* in the F₁–F₂ domain (cf. Pols *et al.*, 1969; Povel and Wansink, 1986). Advantages of the *whole-spectrum+PCA* approach are its robustness and the possibility to perform the analysis in real time, whereas correct and real-time formant extraction is known to be quite difficult. However, an actual formant analysis was not performed in the previous study.

In the present study, the whole-spectrum approach has been applied to objectively assess the pronunciation quality of the vowels pronounced by a large group of *children*: boys

TABLE I. Notations, keywords, durations, and frequencies of the fundamental, the first formant, and the second formant of 15 Dutch vowels (12 monophthongs and 3 diphthongs). Though Dutch vowel pronunciation differs strongly from English (with regard to articulatory as well as to durational properties), keywords are given with reasonable articulatory similarity. Means and standard deviations of measured values are given for two different groups of speakers: 24 norm and 24 deaf children. Also given are the values averaged over all 15 vowels.

			Monophthongs											Diphthongs				
IPA notation			a	ɑ	e	ɛ	i	ɪ	o	ɔ	u	œ	y	ø	au	ay	ɛi	
Dutch orthographic notation			aa	a	ee	e	ie	i	oo	o	oe	u	uu	eu	au/ou	ui	ei/ij	Average
Dutch keyword (as used)			ja	bad	zeep	pet	piet	wit	boot	sok	koe	bus	muur	neus	touw	huis	vijf	
English (<i>French</i>) keyword			fast	art	able	bet	eat	bit	no	law	book	cup	<i>vue</i>	<i>ceux</i>	out	<i>feuille</i>	(my)	
Duration (ms)	norm	mean	252	130	260	150	123	145	244	129	214	166	226	266	302	326	296	217
		s.d.	63	32	35	36	25	29	46	30	57	32	45	52	51	54	47	79
	deaf	mean	450	224	467	208	259	187	603	193	522	219	493	518	464	564	572	396
		s.d.	188	125	141	127	163	117	271	102	214	113	171	174	125	154	167	222
Fundamental frequency (Hz)	norm	mean	246	238	244	249	249	254	231	252	260	245	248	240	238	240	242	245
		s.d.	41	40	36	40	51	37	41	47	46	38	39	40	37	30	42	40
	deaf	mean	254	272	275	279	288	284	267	277	284	284	277	274	262	261	266	274
		s.d.	42	43	43	50	49	42	39	48	50	54	37	36	40	40	41	44
First formant (Hz)	norm	mean	1050	876	479	530	303	463	439	481	353	447	356	455	771	875	626	567
		s.d.	118	173	59	57	40	62	66	66	51	60	48	81	173	128	182	239
	deaf	mean	892	841	524	711	413	480	528	638	409	537	425	531	742	662	677	601
		s.d.	147	125	75	116	69	99	52	114	62	89	88	82	118	132	108	178
Second formant (Hz)	norm	mean	1931	1556	2190	2032	2367	2145	1152	1058	939	1875	1837	1818	1579	1961	1978	1761
		s.d.	171	149	174	171	239	253	79	99	84	113	146	136	182	127	196	441
	deaf	mean	1693	1764	2196	2073	2384	2215	1262	1339	1176	1696	1550	1845	1588	1795	2020	1773
		s.d.	179	188	246	254	282	311	96	158	125	310	373	252	140	226	199	417

as well as girls, normal hearing as well as deaf. It is interesting to find out whether the promising results for adult males (Bakkum *et al.*, 1993) could also be achieved for children. A comparison has been made between the whole-spectrum approach and formant analysis, based on a robust LPC algorithm (linear prediction coding; Markel and Gray, 1976). This comparison was meant to clarify theoretical as well as practical implications of both approaches. With regard to the whole-spectrum analysis, an important question was the influence of the high fundamental frequency in children's speech, resulting in the absence of harmonics in some of the lower filter bands.

Previous research on vowel production by deaf children was generally based on formant analysis (cf. Angelocci *et al.*, 1964; Monsen, 1978) and revealed a reduction of the $F1-F2$ range compared to that of normal-hearing children. For adult males, such a reduction of "deaf" vowels was also shown when applying the whole-spectrum+PCA approach (Bakkum *et al.*, 1993). Perceptual experiments (cf. Smith, 1975; Gold, 1980; Levitt and Stromberg, 1983) also revealed that deaf vowels often showed reduction: They are confused with vowels with more central positions in the $F1-F2$ domain.

The present study extends this work by considering the possible confusions, as obtained from an identification experiment on the children's vowels. A representation of all vowels in a perceptual space could be achieved by applying a multidimensional-scaling technique (cf. Kruskal, 1964a,b) to the confusion matrix. As Klein *et al.* (1970) have shown, such a representation in two dimensions may be compared with two-dimensional representations from objective analyses (formants, PCA). Our listeners also subjectively assessed

the children's vowels by a magnitude estimation of pronunciation quality. By comparing this subjective assessment with the objective results, the value of the objective approaches could be investigated.

The high fundamental frequencies of children's vowels may cause the problem that no harmonics are present in particular lower filter bands of the whole-spectrum analysis. Therefore, the efficacy of an $F0$ -compensation procedure (cf. Van Alphen, 1992) was investigated. Furthermore, an *extrinsic* normalization procedure was applied; extrinsic indicates that the vowels are normalized according to general properties of the speaker concerned, such as the speakers' average spectrum; this can be contrasted with *intrinsic* normalization, which implies that a vowel is normalized according to properties of the vowel itself, such as the fundamental frequency (cf. Disner, 1980; Nearey, 1989; Bakkum *et al.*, 1993). An attempt has also been made to describe the dynamic behavior of the vowels: Vowel-inherent spectral change is assumed to play a role in vowel perception and it presumably is of major importance for the diphthongs (cf. Pols, 1979). In formant analysis, intrinsic as well as extrinsic normalization procedures (cf. Traunmüller, 1981) and frequency transformations (in order to achieve an optimal agreement with the characteristics of auditory perception; cf. Syrdal, 1985; Miller, 1989) could be applied; the implications for our data were examined.

I. MATERIAL

All possible Dutch vowels were considered in this study. The 15 vowels (12 monophthongs and 3 diphthongs, as given in Table I) were pronounced in a meaningful conso-

nantal context (C_1VC_2 words), this in contrast with the vowels in the (mostly) meaningless /h-t/ context as used with the adult males (Bakkum *et al.*, 1993). It was regarded as necessary that the very young speakers were familiar with the words.

A. Recordings

The 48 native Dutch speakers ranged in age from 6 to 12 yr. Twenty-four of them, twelve boys and twelve girls, had normal hearing. They may be considered to be normal speakers of standard Dutch and they will be referred to as *norm* speakers. The other 24 speakers, 12 boys and 12 girls, were prelingually *deaf* pupils of an institute for the deaf. Their hearing loss, measured at the best ear and averaged over 500, 1000, and 2000 Hz, was 106 dB on the average, with a minimum of 83 dB. The 15 meaningful C_1VC_2 words, including all vowels, were presented to all 48 speakers in written form on cards. They were asked to read out the words at a normal level, without exaggerated stress or articulation. Each speaker contributed one token of each vowel target, so that the variation for a given vowel reflected interspeaker differences only.

All recordings took place in office rooms of the speakers' school or institute, which were neither completely insulated nor anechoic. Recordings were made on a DAT recorder (Sony 55-ES for the deaf; Casio DA-7 for the normal hearing), using high-quality microphones (Sennheiser MD-441 for the deaf; Bruel & Kjaer 4003 with amplifier Bruel & Kjaer 2812 for the normal hearing), followed by a high-pass filter (190 Hz, 12 dB/oct) to eliminate ambient low-frequency noise.¹ The words were sampled at 20 kHz by the 12-bit A/D convertor of a comprehensive audio signal processor, after low-pass filtering (7.4 kHz, 60 dB/oct).

B. Segmentation

The vowels were isolated from the consonantal context by an automatic segmentation procedure that made use of changes in overall level and spectrum, as measured in frames of 12.8 ms (see Sec. II A 1). In order to examine (and occasionally to correct) the segmentation, the oscillograms of the words were displayed and the segments chosen could be made audible. To avoid abrupt onsets and offsets, cosine-shaped rising and falling windows were applied to the first and last frames. Average durations of the different vowels are given in Table I. Within the norm and deaf groups no significant differences were found between boys and girls, so no separate results are given for both sexes. The vowels of the normal-hearing children show longer durations and greater variability than the vowels of the normal-hearing adult males, as reported by Bakkum *et al.* (1993). This is partly due to the different consonantal contexts in which the vowels have been pronounced: Vowels followed by a plosive (e.g., /t/) generally are shorter than vowels followed by fricatives or sonorants (cf. Nooteboom and Cohen, 1984, pp. 130–131). As was the case with the deaf adult males, the deaf children show substantially longer durations. On the average the lengthening is 82%, while in general lengthening for “long” vowels is greater than for “short” ones.

II. OBJECTIVE ANALYSIS

A. Whole-spectrum analysis

1. Description of analysis method

To perform the whole-spectrum analysis, a set of 16 contiguous bandpass filters has been developed, as argued for by Bakkum *et al.* (1989, 1993). The filter set covers a frequency range of 90–7200 Hz. The cutoff frequencies (–3 dB) of adjacent filters are coinciding and the slopes of the recursive, third-order, elliptic filters are such that attenuation at the center frequencies of the adjacent filters is more than 20 dB. The filter set was designed as a first-order approximation to the concept of the ear's critical bands, with filter bandwidths of 1/3 oct, except for the lowest three filters, which had a fixed bandwidth of 90 Hz (cf. Zwicker and Terhardt, 1980). To enable fast and accurate processing, the filter set was implemented on a real-time digital analysis system. Filter band and overall sound-pressure levels in decibels are calculated within a frame window of 12.8 ms.

2. F_0 compensation

Usually, vowels are voiced, which implies that they generally evoke a sensation of pitch, and that they are characterized by a harmonic structure. Because the lower-frequency filters have, in the absolute sense, small bandwidths, harmonics in a particular band may be missing, so that some filter band levels will be low. This is especially the case for high fundamental frequencies (F_0), as with most vowels pronounced by children (see Table I). The problem is illustrated by Fig. 1: For F_0 up to 370 Hz, frequencies of the different harmonics (f_1 – f_7) are indicated by straight lines; passbands (–6-dB boundaries) of the first nine filter bands are indicated by hatched horizontal areas. It is obvious that in the dark shaded areas no harmonics are present in the filter bands concerned and that this may play an important role for the first seven filter bands. It will result in F_0 -dependent gaps in the vowel spectrum that are related to the harmonic structure but not to the pronunciation quality. With regard to the latter we are interested in the modification of the harmonic structure by the vocal tract resonances, as represented by the spectral envelope. In order to avoid F_0 -dependent differences when comparing the spectra of various vowels, we applied an F_0 -compensation procedure without modifications of the filter set itself, as described by Van Alphen (1992): For every 12.8-ms spectrum, local spectral gaps (one or two filter bands “wide”) were smoothed by a linear interpolation between the levels of the neighboring filter bands (see Fig. 2); if a gap occurred at the first filter band, then its level was made equal to that of the second filter band. Care must be taken to avoid confusing the F_0 -dependent gaps with valleys related to the spectral envelope (formant structure) of the vowel: For that reason, the F_0 -compensation procedure was applied only to the six lowest filter bands. Though the procedure introduced a form of (local) distortion into the spectra, we believe that the comparison of the (global) spectral envelopes, which are of most interest for our analysis, did benefit from it.

FIG. 1. Schematic illustration of the absence of harmonics in the lower filter bands of the analysis system for fundamental frequencies up to 370 Hz. The harmonics (f_1 – f_7) are indicated by straight lines, and the filter bands by hatched horizontal areas. In the dark shaded areas no harmonics will be present for vowels with the accompanying fundamental frequency.

3. Normalization

All calculated spectra were level-normalized by subtracting the average level of all 16 bands from the separate band levels, since differences in overall level are not of interest in evaluating vowel pronunciation.

Differences between the norm and deaf groups in re-

FIG. 2. Original spectrum and resulting F_0 -compensated spectrum of the vowel /t/ of a deaf girl ($F_0=289$ Hz).

cording circumstances and apparatus, resulting in different average spectra, were removed by a group normalization procedure. The difference between the average spectra of both groups was calculated, and then the individual vowel spectra of the deaf speakers were corrected by subtracting this difference spectrum. Henceforth, the resulting individual spectra will be referred to as *basic*.

Pronunciation quality rather than vowel quality in general was under investigation, and therefore we were not very interested in the unchanging, speaker-specific characteristics determined by the size of the vocal tract and by the shape and functioning of the sound-production source. To compensate for the spectral effects of such nonarticulatory interspeaker variation, a simple extrinsic speaker-normalization procedure was applied which has been used successfully by Klein *et al.* (1970), Pols *et al.* (1973), and Bakkum *et al.* (1993). The average decibel spectrum of all vowels pronounced by one speaker was calculated and this speaker-specific spectrum was subtracted from the individual vowel spectra of that speaker. However, spectral interspeaker variation caused by speakers' fixed articulatory characteristics, such as nasalization, was also eliminated by this procedure. In the cases that these characteristics were relevant with respect to pronunciation quality, the compensation was too extreme: Overnormalization may have occurred.

4. A static and a dynamic approach

In the *static* approach the dynamic spectral information was eliminated by calculating the average spectrum over the complete vowel duration (which will be referred to as the static spectrum); by doing so, the number of variables to be processed was largely reduced. We assumed that, in assessing pronunciation quality, the vowel-inherent spectral change is less important than the average spectrum of the nearly steady-state part of the vowels (see Nearey, 1989). For male vowels, such an approach applied quite well, even for the diphthongs (Bakkum *et al.*, 1993).

However, vowel-inherent spectral change certainly plays a role in vowel perception (see, among others, Nearey and Assmann, 1986; Fox, 1989; Pols and Van Son, 1993). The spectral (formant) information, both at the nearly steady-state part [often called vowel nucleus or (formants') endpoint] and at the vowel onglide and offglide (following and preceding the consonantal transitions, respectively), is generally considered to be most important. Therefore, we applied a *dynamic* approach, comparable with the dynamic vowel specification by Strange (1989b): Every vowel was represented by the spectra (averaged over a time window of 25.6 ms) at three different points of time: at the *initial* (onglide) and at the *final* (offglide) part, and at the *center* (nucleus) of the vowel. The center was defined as the point of time at which the spectrum deviates maximally from both the initial and the final spectrum. By such a time alignment method the dynamic spectral behavior of the different vowels could be compared objectively, leaving the durational variation aside.

FIG. 3. Variance per filter band for vowels spoken by (a) 24 normal-hearing and (b) 24 deaf children. Given is the total variance, the variance of the 15 vowel spectra, averaged over all speakers, the variance of the speaker-specific spectra, averaged over their 15 vowels, and the residual variance.

5. Spectral variance

The total variances in all 16 filter bands, which are shown for *norm* and *deaf* speakers as bold solid curves in Fig. 3, express the information conveyed by the basic spectra as resulting from the static approach. No systematic difference between boys and girls could be observed. The total variance consists of three parts:

(1) *Vowel-dependent variance* is the variance of the 15 vowel spectra, averaged over all speakers (dotted curves). In general, the vowel-dependent variance is largest in the midrange. The norm speakers show larger vowel-dependent variances than the deaf speakers, indicating better vowel separation and, correspondingly, more clear vowel pronunciation.

(2) *Speaker-dependent variance* is the variance of the 24 speaker-specific spectra, averaged over all vowels (broken curves). For both groups, the speaker-dependent variance in the first two filter bands is relatively large. This is caused by the variation in F_0 values: The compensation procedure (see Sec. II A 2) cannot deal completely with filter bands below F_0 . However, for males, and most likely also for children, it has been shown before that the frequency region covered by these two filter bands (under about 260 Hz) is not important with regard to the pronunciation quality of vowels (Steeneken, 1992; Bakkum *et al.*, 1993). Therefore, these bands will be disregarded further on. Apart from the two lowest and two highest filter bands, the speaker-dependent variance of the norm speakers is small; their speaker-specific spectra agree well with each other. The deaf speakers show somewhat larger speaker-dependent variances, but their

speaker-specific spectra agree reasonably well with each other, too. It seems that, apart from variations in F_0 , the speakers did not show extremely large fixed articulatory or nonarticulatory deviations, at least not those deviations that would have resulted in large spectral variation.

(3) The *residual variance* is a consequence of differences in pronunciation of the same vowel by different speakers (thin solid curves). The residual variance of the norm speakers is certainly not negligible, although their pronunciation quality is supposed to be quite high: Obviously, a certain tolerance in the spectra does exist. However, the residual variance of the deaf speakers is larger, indicating that their pronunciation quality varies more, as expected.

B. Representation in a reduced spectral space

The different dimensions of vowel spectra (levels in the separate filter bands) are highly correlated so the spectral information can be reduced to a representative lower-dimensional space by a robust method of dimensionality reduction: *principal components analysis*, followed by a *Varimax rotation* to facilitate the interpretation of the new dimensions or components (cf. Harman, 1967; Bakkum *et al.*, 1993). The first new dimension, with directions given by cosines relative to the original dimensions, explains most of the original variance and thus is believed to convey most information. Then a second, orthogonal dimension explains most of the remaining variance, and so on. All vowel spectra can be transformed to their coordinates in a space defined by the first two dimensions; this results in a simple, but effective visualization of the vowels.

A PCA, followed by a Varimax rotation in two dimensions, was applied to the F_0 -compensated and speaker-normalized static vowel spectra of the norm speakers to obtain a two-dimensional *norm subspace* for Dutch vowel pronunciation by children. The vowel spectra of the deaf speakers were projected in the norm subspace to examine the degree of agreement with correctly pronounced vowels. The explained variance in the norm subspace is 76% for the norm and 70% for the deaf speakers (see Table II). Applying a PCA to the basic rather than to the speaker-normalized vowel spectra of the norm speakers results in less explained variance: 72% and 66%, respectively. When applying a PCA to the speaker-normalized vowel spectra of deaf speakers, the direction cosines determining the subspace differ in detail only. The explained variance in their own subspace is 72%, only slightly larger than in the norm subspace.

Figure 4 shows the direction cosines of the norm subspace. In Fig. 5 the various monophthongs are plotted in the norm subspace, for both groups. The centers of the ellipses represent the values averaged over all speakers of each group. The long axes correspond with the directions in which the interspeaker variance for each vowel is maximal. The orthogonal short axes represent the remaining variance. The lengths of the axes represent twice the standard deviation. Diphthongs are not included for their vowel-inherent spectral change is very large and varies greatly across different speakers, especially with respect to durational aspects. Therefore, describing them by "static" spectra seems to be

TABLE II. Variation of the vowel spectra for groups of 24 norm and 24 deaf children. Total variance is given for basic, as well as speaker-normalized 14-dimensional spectra of all 15 vowels. Explained variance is given for data projected in a three-dimensional subspace defined by a PCA on the vowel spectra of the norm speakers. Also given are the total and overlapping area and the resulting separation of the ellipses of 12 monophthongs as shown in Fig. 5. All data (except for the separation values) are given in dB².

Vowel spectra	Basic		Speaker normalized	
	norm	deaf	norm	deaf
Speakers				
Total variance	584	558	506	461
Explained variance after projection in PCA subspace				
first component	254	205	243	177
second component	165	161	142	145
third component	55	33	42	27
Ellipses in norm subspace				
total area	1044	2673	688	1883
overlapping area	569	1996	165	1138
separation (%)	57.3	25.3	76.0	39.5

inappropriate; it will result in large ellipses, which by nature overlap with the monophthongs.

The less the ellipses in Fig. 5 overlap, the better the monophthongs can be distinguished. A measure for separation, used by Disner (1980) to evaluate normalization procedures, is the nonoverlapping area as a percentage of the total area occupied by the ellipses. Separation is 76% for the norm speakers and only 40% for the deaf speakers. The efficacy of the speaker-normalization procedure can be illustrated by the fact that, without applying it, the separation is only 57% and 25%, respectively.

The dynamic approach (Sec. II A 4) gives some insight in the implications of time averaging. The initial, central, and final spectra of the vowels can be represented by three-point traces in the norm subspace. The average results per vowel are given in Fig. 6. It appears that the traces, which represent vowel-inherent spectral change, are short for most monophthongs of both the normal-hearing and the deaf children. The traces of the three diphthongs (/au/, /ay/, and /ei/) are sub-

FIG. 5. Projections in the norm subspace of average, F_0 -compensated, speaker-normalized spectra of the monophthongs spoken by (a) 24 normal-hearing and (b) 24 deaf children. The lengths of the axes of the ellipses are twice the standard deviations in principal directions.

stantially longer. Three monophthongs show diphthongization effects; in general, /e/ tends to /i/, /o/ to /u/, and /ɤ/ to /y/. This is in agreement with the results for adult male speakers (Bakku m *et al.*, 1993). The fact that /y/ tends to /œ/ is a coarticulation effect caused by the following consonant in the keyword /myr/ (cf. Pols, 1977, pp. 94–99).

C. Formant analysis

1. Description of analysis method

Formant analysis is the second approach in objectively assessing the vowels. The formant frequencies were estimated by an LPC analysis, using the Split-Levinson algorithm (see Willems, 1986). The fundamental frequency was measured as well, using an algorithm given by Reetz (1989), which is based on the detection of glottal pulses. The complete vowel segments as described in Sec. I were analyzed with a 25.6-ms Hamming window that was shifted in steps of 10 ms. The 720 vowels were downsampled to 10 kHz, after digital low-pass filtering at 4.5 kHz, thereby decreasing the number of formants that can be estimated. However, the formants below 4.5 kHz, which certainly are the most important ones with regard to pronunciation quality, could be estimated with greater accuracy. Especially for children's speech, characterized by high fundamental frequencies, the choice of the LPC order has to be made deliberately. For instance, an LPC

FIG. 4. Direction cosines per filter band of the first two dimensions, determining the norm subspace; a PCA, followed by a Varimax rotation, was applied to the 15 F_0 -compensated and speaker-normalized vowel spectra of 24 norm speakers.

FIG. 6. Projections in the norm subspace of the initial, center, and final spectra of the 15 vowels, averaged for (a) 24 normal-hearing and (b) 24 deaf children. The initial part of each vowel is indicated by a dot.

analysis with the order eight results in four estimated formants, irrespective of the number of formants actually present; however, in cases with high F_4 (near or above 4.5 kHz, which may be the case for vowels with high F_2), underestimations of F_2 and F_3 may arise: The order six should be applied. The high F_0 values quite often result in F_1 estimates that “switch” between the real first formant and F_0 or one of the lower harmonics; if the F_1 estimate coincides with the F_0 value, the F_2 and F_3 estimates possibly represent the first and second formant. The problems as described do particularly show up for the initial and final parts of the vowel segments, when the vowel intensity is relatively low. To obtain reliable formant estimates, the analysis was performed automatically by a computer program, for three different LPC orders: six, eight, and ten. Then we have compared (by eye) the resulting formant traces for the three different orders with the F_0 traces and with the spectra as obtained before (see Sec. II A). For each segment, the optimal order was chosen and obviously erroneous formant values, as indicated by abrupt formant changes, were corrected by hand. For the vowels with high F_2 values (e.g., /i/, /ɪ/, /e/, and /ɛ/, but also most samples of /y/, /œ/, /φ/, /a/, and /ɑ/), the order six proved to be most appropriate. For the vowels /u/, /o/, and /ɔ/, the order eight resulted in correct estimates, generally. In most cases of coinciding F_1 estimates and F_0 values, the order eight or ten was chosen. For many of the

diphthongs, especially for /au/ and /ɔy/, we did not succeed in choosing a single correct LPC order for the complete segment; in those cases the optimal order for the initial part was chosen, and we estimated the formants of the final part (the glide) by closely examining the displays of the different LPC spectra. Furthermore, errors in the F_0 values (caused by octave jumps or by the fact that the algorithm generated a value zero whenever the pulse detection failed) were corrected by interpolation or by estimating the fundamental frequency from the vowel’s oscillogram. Obviously, the formant analysis procedure was very time consuming.

2. Normalization

An advantage of the present analysis program was that an intrinsic normalization could be applied in which the measured F_0 served as a reference (cf. Traunmüller, 1981). In certain intrinsic normalization procedures, F_0 values are subtracted from F_1 and F_2 values.² Transformations of the frequency values from a linear hertz scale to a logarithmic or Bark scale have been suggested by many authors for being in optimal agreement with the characteristics of auditory perception (cf. Zwicker and Terhardt, 1980). To calculate auditory distance measures, Syrdal (1985) used Bark-transformed values; Miller (1989) used a logarithmic scale and transformed F_0 to the reference $RF_0 = 168 * (F_0/168)^{1/3}$. All of these normalization and transformation procedures have been applied to our data. An equivalent extrinsic speaker-normalization procedure as used with the filter band analysis has been applied, too: The average F_0 , F_1 , and F_2 values of all vowels pronounced by one speaker were calculated and subtracted from the separate F_0 , F_1 , and F_2 values of that speaker.

3. Results

Means and standard deviations of estimated F_0 , F_1 , and F_2 are given in Table I, for both groups of speakers. In Fig. 7 the various monophthongs are plotted in the non-normalized F_1 – F_2 domain. Comparison with reported formant data for children’s vowels (Peterson and Barney, 1952; Angelocci *et al.*, 1964; Weenink, 1985) showed a reasonable agreement, except that in our data the average F_2 values of the front vowels were somewhat lower. For some vowels (e.g., /ɛ/ and /ɔ/ of the norm speakers) the F_1 estimates are rather low, probably due to interactions with the lower harmonics. The same measure for vowel separation as in Sec. II B was used to indicate the value of the different frequency transformations and of the normalization procedures. The underlying assumption was that the separation of the norm speakers’ monophthongs should be maximal, since subjectively they could be distinguished well (see Sec. III B). The results are given in Table III. The separation is 65% for the non-normalized F_1 – F_2 configuration, expressed on a linear scale [No. (1)], of the norm speakers. Transformation to a Bark [(9)] or logarithmic [(5)] scale does not induce great changes. Results after intrinsic normalization [(2), (6), and (10)] are systematically better, with an optimal separation of 77% for the linear scale. Extrinsic (speaker) normalization leads to about the same results: The separation is 78% for the

FIG. 7. Average values of the first and second formants of the monophthongs spoken by (a) 24 normal-hearing and (b) 24 deaf children. The lengths of the axes of the ellipses are twice the standard deviations in principal directions.

logarithmic scale. The combination of both normalization procedures results in the highest separation values [(4), (8), and (12), maximally 79% on the logarithmic scale], though these values are not substantially higher than those after only

one of the normalization procedures. For the deaf speakers, the separation values are significantly lower in all cases (varying from 39% to 57%).

Comparable with the dynamic approach in the reduced spectral space (Sec. II B), the initial, central, and final formants of the vowels can be represented by three-point traces in the $F1-F2$ domain. The average results per vowel are given in Fig. 8. It appears that the formant traces are short for most monophthongs of both the normal-hearing and the deaf children. The traces of the three diphthongs (/au/, /ɛy/, and /ei/) are longer. The diphthongization effects of /e/, /o/, and /ɔ/ can still be observed.

D. Discussion

The configuration for the norm speakers in the PCA subspace resembles the vowel triangle, given by the configuration of the average $F1$ and $F2$ values of the norm speakers [compare Figs. 5(a) and 7(a)], though both configurations are slightly rotated with respect to each other. To indicate the correspondence, the canonical correlation between average values in both configurations was calculated. The correlation coefficients were very high: 0.99 and 0.90. The composing direction cosines in Fig. 4 already indicate this correspondence: The first dimension is most sensitive (as indicated by the steepest slope) to spectral variations in the region of 1000–2500 Hz, corresponding with variations in $F2$, and the second dimension is most sensitive to variations in the region of 400–1000 Hz, corresponding with $F1$.

The vowels pronounced by the normal-hearing children were well separated in the norm PCA subspace, though their internal variation and overlap were somewhat higher than they proved to be for norm adult males (Bakkum *et al.*, 1993). In general, the deaf children made use of a limited part of the subspace. They showed a reduction of many vowels; projections of their vowels on the subspace tended to take more central positions, approaching /œ/. The reduction was yet less apparent than it proved to be for the deaf adult

TABLE III. The separation of the monophthongs for various configurations based on the first two formants. Formants are expressed on a linear, log, and Bark scale; corresponding units for the area values are Hz^2 , oct^2 , and Bark^2 , respectively. Both intrinsic (based on the fundamental frequency $F0$) and extrinsic (based on the average formants per speaker and indicated by the SN subscript) normalization procedures have been investigated.

No.	Configuration	Transformation	Normalization	Norm children			Deaf children		
				total area	overlapping area	separation (%)	total area	overlapping area	separation (%)
(1)	$F1$ vs $F2$	linear	none	351 000	124 000	64.7	789 000	461 000	41.6
(2)	$F1-F0$ vs $F2-F0$	linear	intrinsic	295 000	68 000	77.0	883 000	542 000	38.6
(3)	$F1_{SN}$ vs $F2_{SN}$	linear	extrinsic	292 000	74 000	74.7	660 000	320 000	51.5
(4)	$(F1-F0)_{SN}$ vs $(F2-F0)_{SN}$	linear	intrinsic+extrinsic	301 000	70 000	76.7	707 000	362 000	48.8
(5)	$\log F1$ vs $\log F2$	log	none	0.852	0.280	67.1	1.611	0.867	46.2
(6)	$\log(F1/RF0)$ vs $\log(F2/RF0)$	log	intrinsic	0.791	0.225	71.6	1.854	1.052	43.3
(7)	$(\log F1)_{SN}$ vs $(\log F2)_{SN}$	log	extrinsic	0.639	0.142	77.8	1.313	0.562	57.2
(8)	$[\log(F1/RF0)]_{SN}$ vs $[\log(F2/RF0)]_{SN}$	log	intrinsic+extrinsic	0.658	0.141	78.6	1.413	0.632	55.3
(9)	$B1$ vs $B2$	Bark	none	10.9	3.61	66.9	23.2	12.5	46.1
(10)	$B1-B0$ vs $B2-B0$	Bark	intrinsic	10.4	2.61	74.9	29.1	17.2	40.9
(11)	$B1_{SN}$ vs $B2_{SN}$	Bark	extrinsic	8.69	2.03	76.6	19.1	8.35	56.3
(12)	$(B1-B0)_{SN}$ vs $(B2-B0)_{SN}$	Bark	intrinsic+extrinsic	9.20	2.05	77.7	21.8	10.3	52.8

FIG. 8. The initial, center, and final formant values of the 15 vowels, averaged for (a) 24 normal-hearing and (b) 24 deaf children. The initial part of each vowel is indicated by a dot.

males. The deaf children's vowel configuration in the $F1$ – $F2$ domain showed a reduction, too, so it appears that they are less capable of realizing extreme formant frequencies (cf. Gold, 1980). The range of mean $F1$ values reduced from 747 Hz (see Table I) for the normal hearing to only 483 Hz for the deaf children; the $F2$ range reduced from 1423 to 1208 Hz. In a comparison between normal-hearing and deaf boys, Angelocci *et al.* (1964) reported a reduction of the $F1$ range from 655 to 330 Hz and a reduction of the $F2$ range from 1715 to 1148 Hz. Especially, the reduction of the $F2$ range was more apparent than in our research; however, this is not that much a consequence of a larger $F2$ range of the deaf speakers in our study, than rather of the smaller $F2$ range of the norm speakers, due to the lower $F2$ values we obtained for their front vowels (see Sec. II C). Furthermore, both in the norm subspace as well as in the $F1$ – $F2$ domain, the vowel ellipses of the deaf children were large, compared to those of the normal-hearing children; this indicates an enlarged overlap and a substantial variation in pronunciation quality.

In displaying vowels, the whole-spectrum+PCA approach and the formant approach were quite comparable. This conclusion is supported by the separation values for norm speakers' vowel ellipses which were of the same order for both approaches (see Tables II and III). Applying speaker normalization resulted in better separation values. This was

also true for intrinsic ($F0$) normalization in the formant approach, but a combination of both normalization procedures did not significantly improve the separation: Applying speaker normalization only seems to be sufficient. Suggested frequency transformations (to logarithmic and Bark scales) resulted in only slightly higher separation values, but they surely led to a more uniform spacing of the vowels along the $F1$ and $F2$ axes.

III. SUBJECTIVE ASSESSMENT

Subjective experiments can provide more specific evidence about the relation between pronunciation quality of the vowels and their objective descriptions by the whole spectrum, by the PCA dimensions, or by the formants. They may also shed light on the implications of normalization procedures and time averaging for the objective assessment of vowel pronunciation. Therefore, two simple but essential subjective experiments have been carried out, which could be performed with untrained listeners. In the *magnitude estimation* experiment, listeners were asked to assess the vowels; they were emphatically instructed to judge pronunciation quality, not vowel quality in general, and to disregard durational differences as far as possible. In the *identification* experiment, listeners were asked to name the vowels.

A. Magnitude estimation

1. Procedure

In the magnitude estimation experiment the isolated vowels (gated from CVC words as described in Sec. I B) were presented individually to listeners, who assigned a value, corresponding with his/her perceptual judgment of the pronunciation quality of the vowel. The method permitted many vowels to be assessed quickly, and it has proved to be accurate and reliable when a sufficient number of listeners is used (Bakkum *et al.*, 1993). Therefore, the final assessment was derived from the combined estimations of 16 listeners. In order to limit the required time per session, the boys and girls were judged separately; every listener had to judge the vowels of 24 speakers. The 15 vowels of one speaker were presented in succession to the listener. This corresponds with actual diagnostic procedures, but a consequence of this may have been that a speaker was "classified" as good or poor after the first vowels, so that the subsequent vowel judgments may have been biased: In that case, even poor vowels of a "good" speaker might have been estimated as quite good whereas good vowels of a "poor" speaker might have been estimated only as moderate. For each listener, the order of the 15 successive vowels was fixed for the 24 speakers; getting accustomed to this order simplified the task and prevented any confusion with regard to the intended vowel. To eliminate order effects, this vowel order varied over the 16 listeners in a Latin-square design; since the number of vowels was odd and not equal to the number of listeners, digram-balancing could only be approximated (see Wagenaar, 1969). The order of the 24 speakers also varied over the 16 listeners, again as prescribed by a Latin square.

FIG. 9. Quality assessment of 15 individual vowels and the complete vowel set ("overall") by 16 listeners, for two subjective methods: (a) magnitude estimation and (b) identification. Average values and standard deviations are given for both groups of speakers.

2. Listeners

The listeners were 25 normal-hearing native speakers of Dutch, 18 of whom were students and 7 colleague researchers. Seven of these listeners judged the vowels of both the boys and the girls, but in two different sessions; the remaining eighteen listeners judged the vowels of only one of these groups. The listeners participated individually, in a soundproof room. The vowels were presented to the listeners binaurally by headphone (Sony MDR-CD999) at a sensation level of 70 dB. All 24 lists of 15 vowels that had to be assessed (one speaker in every list) were alternated with the 24 lists of 15 vowels to be named (identification experiment; see below Sec. III B). In the magnitude estimation experiment, the listeners were instructed to circle a number between 1 and 7, corresponding with the perceived pronunciation quality (1 stands for very poor, 7 for excellent). On the response sheet, beside the numbers, the intended (target) vowel was given. For familiarization with the task, each listener first had to judge the vowel set of one additional deaf child. The total experiment took about one hour per listener; the students were paid for their participation; the others were unpaid volunteers.

3. Results

Individual differences in this way of assessing pronunciation could be shown by comparing the distributions of magnitude estimations. It appeared that the magnitude estimations were distributed quite well (from 1 to 7) for the majority of listeners; however, a not inconsiderable minority tended to be not too critical once they had understood the vowel as corresponding with the target: Both in the group that judged the girls as well as in the group that judged the boys, 7 out of the 16 listeners estimated more than half of the norm vowels as "excellent" (value 7). Combining the estimations of the 16 listeners compensated for these individual differences and led to aggregate magnitude estimations for

every vowel and speaker. No obvious differences occurred between the ratings for boys and girls; therefore, no distinction will be made further on. Per vowel, the average value and the standard deviation of the aggregate magnitude estimations of the 24 speakers per group (norm versus deaf) are given in Fig. 9(a). The results of the assessment of the complete vowel set (the average rating of all 15 vowels of one speaker) are labeled as "overall."

A two-way repeated measures analysis of variance with speaker groups as the between-cells factor and vowels as the within-cells factor showed that group effects on the ratings were highly significant [$F(1,46)=212.0$, $p<0.001$]. The group by vowel interaction was also significant [$F(14,644)=4.96$, $p<0.001$]. We applied a one-way analysis of variance to the assessment of the individual vowels, and of the complete vowel set ("overall"). Group effects proved to be highly significant [see Table IV; vowel /a/: $F(1,46)=16.54$, $p<0.001$; all other vowels revealed higher F ratios]. The norm and deaf groups can be considered to be separated for all vowels.

B. Identification

1. Procedure

In the subjective identification experiment the isolated vowels were presented to the listener, who had to name them, choosing from the 15 vowels possible in Dutch. The task was simple and very essential with regard to daily communication, but a disadvantage of the method was that the pronunciation quality could only be assessed roughly: Once vowels were correctly identified, no perceptual differences could be revealed. Nevertheless, the identification scores accumulated over 16 listeners are assumed to give at least an indication of the pronunciation quality of the vowels. To limit the time per session, the 15 vowels of the 24 boys and the 24 girls were named by two different groups of 16 listeners. The vowels were presented in random order, with the

TABLE IV. Significance levels of group effects in assessing vowel pronunciation by two subjective methods [groups of 24 norm and 24 deaf children; one-way analysis of variance for the assessment of separate vowels, given per vowel and for all vowels of all speakers together (tot), and for the assessment of the complete vowel set per speaker (overall)]. Also given are the correlation coefficients between the vowel ratings, obtained from both methods, for both groups separately and for all speakers together.

	Monophthongs												Diphthongs			Together	
	a	α	e	ε	i	ɪ	o	ɔ	u	œ	y	φ	au	ay	ei	tot	overall
Significance level of group effects																	
Method of magnitude estimation	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Method of vowel identification	0.889	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	0.002	0.041	<0.001	<0.001	<0.001	<0.001	<0.001
Correlation coefficient magnitude estimations vs identification rates																	
Norm speakers only	0.79	0.62	0.26	0.83	0.51	0.57	0.69	0.63	0.86	0.76	0.78	0.75	0.47	0.24	0.80	0.72	0.84
Deaf speakers only	0.79	0.76	0.89	0.87	0.90	0.35	0.81	0.69	0.87	0.71	0.87	0.88	0.83	0.85	0.84	0.84	0.93
Norm and deaf speakers	0.69	0.83	0.95	0.89	0.92	0.34	0.81	0.88	0.85	0.82	0.82	0.73	0.88	0.91	0.90	0.86	0.98

restriction that two vowels of one speaker never succeeded each other. To eliminate order effects, the vowel order varied for the first eight listeners and was reversed for the remaining eight listeners.

2. Listeners

The identification experiments were performed in the same sessions as the magnitude estimation experiments, so the same listeners participated (see Sec. III A 2). They were instructed to write down the perceived vowel on the response sheet. For familiarization with the task, each listener first had to name a list of 15 vowels of some additional deaf and normal-hearing children.

3. Results

As expected, the vowels of the norm speakers were identified correctly more often than those of the deaf (83% and 43%, respectively, for all vowels of boys and girls). The ability to identify the vowels varied somewhat over the listeners. The individual correct-identification scores, for all vowels of norm and deaf speakers, varied between 56% and 69%, but these differences were compensated for by aggregating the results of 16 listeners: For every vowel utterance, the identification rate was defined as the percentage of the 16 listeners that named it correctly. No obvious differences occurred between the identification rates for boys and girls. The averaged identification rates are given per vowel in Fig. 9(b), for the norm and deaf speakers separately.

A two-way repeated measures analysis of variance with speaker groups as the between-cells factor and vowels as the within-cells factor showed that group effects on the identification rates were highly significant [$F(1,46)=139.7$, $p<0.001$]. The group by vowel interaction was also significant [$F(14,644)=4.28$, $p<0.001$]. We applied a one-way analysis of variance to the individual vowels. Group effects proved to be highly significant for all vowels but the /a/ [see Table IV; vowel /a/: $F(1,46)=0.02$, $p=0.89$; vowel /φ/: $F(1,46)=4.44$, $p<0.05$; all other vowels revealed higher F ratios]. Therefore, the norm and deaf groups can be considered to be separated for all vowels but the /a/.

4. Representation in a two-dimensional perceptual space

Using the results of the identification experiment, all vowels could be projected in an n -dimensional perceptual space by a procedure adopted from Kruskal (1964a,b). The configuration in such a space provides information about the degree of similarity of the various vowels and, if the number of dimensions is limited to two, it can be compared directly with the configurations in the PCA subspace and in the $F1-F2$ domain (as obtained in Sec. II). The Kruskal procedure was based on the (576×576) mutual dissimilarity matrix of all individual monophthongs, which was calculated from the (576×12) confusion matrix of the identification experiment in a way described by Klein *et al.* (1970). The procedure started by representing all 48×12 vowels as points in a two-dimensional space, in an arbitrary start configuration. Iteratively, the configuration of points was changed in such a way that, finally, the mutual distances did optimally agree with the calculated dissimilarities. A computer program has been developed to apply the procedure to such large numbers. It resulted in vowel representations for the norm and deaf as given by Fig. 10. The stress factor, expressing the extent of agreement, is 14.0%, which, in view of the large number, indicates a fair agreement. The configuration for the norm speakers shows a clear resemblance with their configurations in the objectively defined spaces. Calculating the canonical correlation between average values in the (norm) perceptual space and the (norm) PCA subspace [Fig. 5(a)] resulted in correlation coefficients of 0.98 and 0.86. Comparing the Kruskal configuration with the $F1-F2$ configuration [Fig. 7(a)] resulted in correlation coefficients of 0.95 and 0.76. The same measure for vowel separation as in Sec. II B yielded separation values of 99% for the norm speakers, and 34% for the deaf speakers. The separation of the norm speakers' monophthongs proved to be very high, confirming that, subjectively, they could be distinguished well.³

C. Discussion

Two assessment methods have been applied to the vowels, both of which revealed differences in pronunciation quality, especially when comparing norm and deaf speakers.

FIG. 10. The results of the identification experiment illustrated by a representation in a Kruskal space of the monophthongs spoken by (a) 24 normal-hearing and (b) 24 deaf children. The lengths of the axes of the ellipses are twice the standard deviations in principal directions.

To demonstrate the extent of agreement between both methods, the magnitude estimations of all monophthong utterances are plotted against the identification rates (Fig. 11). The fact that correctly identified vowels show quite a large spread in magnitude estimation scores indicates that moderate deviations from the reference pronunciation quality do not per definition affect intelligibility. The regression line of

FIG. 11. Magnitude estimations vs identification rates for the monophthongs of 24 normal-hearing and 24 deaf children.

the norm speakers obviously lies higher than the regression line of the deaf speakers: Though a deaf vowel might have been well intelligible, the pronunciation quality was apparently considered as lower than the pronunciation quality of a norm vowel with the same intelligibility. This group effect in the magnitude estimation results may have been due partly to the appearance of a *bias* (as discussed in Sec. III A 1), and partly to the fact that other quality aspects have been taken into consideration: Deviations in vowel duration and voice quality, which appeared most often for the deaf vowels, had a greater influence on the magnitude estimation than on the identification rates. The agreement between the identification rates and the magnitude estimations is given quantitatively in Table IV: Correlation coefficients were calculated per vowel, for all vowels of all speakers together (“tot”), and for the assessment of the complete vowel set per speaker (“overall”). For the norm and deaf speakers together the correlation coefficients were high. Even when considering those groups separately, the correlation proved to be significant in all cases but two (vowels /e/ and /ay/ of the norm speakers: As can be seen in Fig. 9, the variance between the ratings of the different utterances of these vowels is very small, while the average values are very high; this almost inevitably results in a low correlation).

From these results it can be concluded that both magnitude estimation and vowel identification can be used to assess vowel pronunciation. Magnitude estimation provides quantitative information about the extent to which the utterance has approximated the (ideal) target vowel, whereas identification provides qualitative information about the relation of the utterance with the complete set of (Dutch) vowels [as can be visualized in a perceptual space (Fig. 10)].

IV. COMPARISON OF OBJECTIVE ANALYSIS AND SUBJECTIVE ASSESSMENT

A. Reference vowels and spectral distances

We investigated how differences between filter band spectra, PCA levels, and formant positions of the various utterances were related to the judgments of pronunciation quality. For each vowel category, objective assessment of pronunciation quality of individual utterances could be derived from distances to a *reference*, being an utterance of good quality. Obvious choices were the utterances that obtained the highest aggregate magnitude estimations for the particular category. Their representations in the PCA subspace and in the $F1-F2$ domain proved to be located near the averages of the norm speakers, but in most cases they were somewhat more extreme (cf., for synthesized and natural male vowels, Cohen *et al.*, 1963, 1967; Bakkum *et al.*, 1993). The average vowel representations of the norm speakers might also serve as references, assuming that their pronunciation quality was high, compared with the deaf speakers’ pronunciation quality. Further on, these two reference sets will be referred to as *best judged* and *average norm*, respectively.

Spectral distances between (static or dynamic) vowel representations can be calculated in a variety of ways. An overview is given by Bakkum *et al.* (1993), who concluded

TABLE V. Comparison between various objective distance measures and subjective distances, obtained from magnitude estimation. The six right-most columns give the average correlation coefficients for each monophthong, the monophthongs without significant correlation, and the correlation coefficient between objective and subjective overall assessment per speaker, respectively, for two different reference sets.

No.	Measure based on	No. of dimensions	Speaker normalization	"Best judged" reference set			"Average norm" reference set		
				Average correlation coefficient	Nonsignificant correlations	Overall correlation coefficient	Average correlation coefficient	Nonsignificant correlations	Overall correlation coefficient
(1)	<i>F0</i> -compensated spectrum	14	no	0.62	e	0.83	0.59	...	0.78
(2)	<i>F0</i> -compensated spectrum	12	no	0.64	...	0.83	0.61	...	0.79
(3)	<i>F0</i> -compensated spectrum	12	yes	0.69	...	0.89	0.65	...	0.84
(4)	original spectrum	12	yes	0.54	o	0.82	0.55	a, o	0.76
(5)	PCA levels	2	no	0.45	a, e, ε	0.78	0.48	a, e	0.77
(6)	PCA levels	2	yes	0.54	a, e, ε	0.85	0.54	a	0.81
(7)	PCA levels	3	yes	0.62	a	0.88	0.60	a	0.83
(8)	formants (linear)	2	no	0.50	ε, i, u	0.81	0.49	a, e, φ, i, ɪ	0.75
(9)	formants (linear)	2	yes	0.51	e, ε, i	0.87	0.48	a, e, φ, i, ɪ	0.82
(10)	formants (log)	2	no	0.47	ε, i, u	0.84	0.55	α, i	0.82
(11)	formants (log)	2	yes	0.54	ε	0.92	0.56	α, φ	0.89
(12)	three-point spectral trace	12	yes	0.68	...	0.90	0.65	a	0.84
(13)	three-point PCA trace	2	yes	0.55	a, e, ε	0.88	0.56	a	0.83
(14)	three-point formant trace	2	yes	0.53	ε, i	0.86	0.50	a, α, e, i, ɪ	0.83
(15)	duration	1	no	0.44	a, α, i, ɔ	0.69	0.46	a, α, i, ɔ	0.69

that the Euclidean distance measure was the best metric for predicting paired comparison judgments of pronunciation quality:

$$D_{i,\text{ref}} = \left(\sum_{j=1}^m \sum_{k=1}^n [L_{i,j,k} - L_{\text{ref},j,k}]^2 \right)^{1/2}, \tag{1}$$

where $L_{i,j,k}$ is the value in the k th dimension of the representation of vowel i at point of time j . $L_{\text{ref},j,k}$ is the corresponding value of the reference vowel. The n dimensions are 14 or less filter band levels, two or three principal dimensions after PCA, or two formants, and all dimensions equally contribute to the distance measure. The total number of spectral representations per utterance is denoted by m ; $m=1$ for a static approach, and $m>1$ for a dynamic approach.

B. Results

1. Individual vowels: The static approach

For each vowel category, the correlation has been calculated between objective spectral distances and subjective perceptual deviations of the vowels of the 24 norm and 24 deaf speakers; the latter have been defined as distances to the ideal score (7) in the *magnitude estimation* experiment; the objective distances have been derived for both reference sets. In the case of the "best judged" reference set, the references themselves have been excluded from the calculations; since they were defined to have zero objective distances, they might have altered the correlations without a clear meaning. In the case of the "average norm" reference set, the spectral distance of a particular norm utterance has been calculated with the average of the remaining 23 norm utterances as the reference. Incorporating the particular utterance in the reference would unintentionally reduce its spectral distance, and therefore favor the norm speakers.

The results are summarized in Table V. For each measure the average correlation coefficient of the 12 monophthongs is given. The correlation proves to be significant (at a

strict 1% level) in most monophthongs; given are the monophthongs without significant correlation. One-way repeated measures analyses of variance have been conducted for all pairs of measures (and for both reference sets) to calculate the significance levels of the differences between the correlations obtained by these measures. A different choice of the reference set results in different spectral distances and thus in different correlation values, but none of the reference sets can be preferred purely on the basis of our data. For the best judged reference set, the *F0*-compensated basic spectral measure (1) yields an average correlation coefficient of 0.62, significant in all monophthongs but the /e/. Analysis of adult male vowels (Bakkum *et al.*, 1993) revealed that filter bands 3–13 (260–3600 Hz) were most important with regard to pronunciation quality. In the present experiments, concerning vowels with high fundamental frequencies, the first two bands were excluded by definition (see Sec. II A 5). Excluding the highest filter bands from the distance calculations revealed that filter bands 3–14 (260–4500 Hz) were the most important ones: Measure (2) yields an average correlation coefficient of 0.64, significant in all monophthongs. Speaker normalization results in higher correlation coefficients [compare (3) with (2)], but the effect is nonsignificant [$F(1,11)=1.953$, $p=0.190$]. Measures based on only the first two principal components [(5) and (6)] yield significantly lower correlation coefficients than the measures based on the whole spectrum [(2) and (3)] [$F(1,11)=13.507$, $p=0.004$, and $F(1,11)=15.692$, $p=0.002$, respectively], but the correlations themselves are still significant for most vowels. Speaker normalization improves the results substantially [compare (6) with (5), $F(1,11)=4.406$, $p=0.060$; for the average norm reference set the improvement is significant: $F(1,11)=6.250$, $p=0.030$]. Including the third principal component results in significantly higher correlation coefficients [compare (7) with (6): $F(1,11)=12.587$, $p=0.005$], though they are significantly lower than for the whole spectrum [compare (7)

FIG. 12. Subjective distances vs objective distances for 15 vowels, spoken by 24 normal-hearing and 24 deaf children. Objective distances are derived from weighted spectral differences, after F_0 compensation and speaker normalization [measure (3), Table V]. Subjective distances are derived from the magnitude estimation experiment. Per vowel, the correlation coefficient r is given.

with (3): $F(1,11)=5.667$, $p=0.036$]. In Fig. 12 the results of measure (3) are given for all vowels. The agreement between objective distances and subjective deviations, as indicated by the regression lines, is apparent for most vowels. Somewhat lower, but still significant correlations are found for the monophthongs /a/, which is characterized by relatively small subjective deviations of the deaf samples, and /ɔ/, which is characterized by relatively large objective distances of the norm samples. The correlation is nonsignificant for the diphthong /ei/. Also apparent is the gross separation of the two groups of speakers. So far, the spectra used for the measures have all been F_0 -compensated (see Sec. II A 2). The usefulness of this can be shown by comparing measure (4), in which filter bands 3–14 of the *original* spectrum have been used, with (3): The average correlation coefficient increases significantly from 0.54 to 0.69 after F_0 compensation [$F(1,11)=15.147$, $p=0.003$].

In the formant domain [measures (8)–(11)], the correlation coefficients are comparable with those in the PCA domain [measures (5) and (6)]. The results do not significantly depend on the scale used [compare the linear (8) and logarithmic (10) measures]. Nevertheless, the linear frequency

scale yields nonsignificant correlations for more vowels than the logarithmic scale does. As was the case in the whole-spectrum domain, extrinsic (speaker) normalization does result in higher correlations [compare (9) with (8) and (11) with (10)], but the improvement is nonsignificant. Furthermore, the intrinsic normalization procedures appear not to change the results substantially (N.B. not given in Table V).

The same objective distances have also been compared with the subjective deviations, obtained from the *identification* experiment. The results of the various measures showed the same tendency, but, in general, the average correlation coefficients were lower. The correlation was nonsignificant for more vowels.

2. Individual vowels: The dynamic approach

Applying the dynamic approach does not significantly change the results for the monophthongs: Compare measure (12), based on the spectral trace (a three-point representation in 12 filter band dimensions) with (3), (13), based on the PCA trace (a three-point representation in two PCA dimensions) with (6), and (14), based on the formant trace (a three-

TABLE VI. Comparison between various objective distance measures and subjective distances, obtained from magnitude estimation. Given are the average correlation coefficients of the three diphthongs and the diphthongs without significant correlation, for two different reference sets. Speaker normalization has been applied in all cases; the number of dimensions is 12 for the spectral, and 2 for the PCA and formant measures.

Reference set:	Best judged		Average norm	
	Average correlation coefficient	Nonsignificant correlations	Average correlation coefficient	Nonsignificant correlations
Static spectrum	0.50	ei	0.61	...
Three-point spectral trace	0.47	...	0.55	...
Two-point spectral trace	0.63	...	0.67	...
Static PCA levels	0.49	ei	0.56	...
Three-point PCA trace	0.36	ei, ay	0.40	ei
Two-point PCA trace	0.54	ay	0.58	...
Static formants	0.22	au, ei	0.10	au, ei
Three-point formant trace	0.33	au, ei	0.23	au, ei
Two-point formant trace	0.48	...	0.37	au

point representation in the $F1-F2$ domain) with (10); this could be expected since their vowel-inherent spectral change was small (see Figs. 6 and 8). The vowel-inherent spectral change of the three diphthongs (/au/, /ei/, and /ay/) was much larger. A description of the diphthongs by a static, time-averaged spectrum resulted in rather low correlations with the subjective scores for the best judged reference set (an average value of 0.50; see Table VI); the probably non-consistent vowel-inherent spectral change may easily have induced a large uncertainty in the position of the average spectrum along the dynamic trace. Averaging over all norm vowels should reduce this uncertainty: Not surprisingly, the average norm reference set resulted in higher correlations for the three diphthongs (0.61 on the average). The dynamic approach in three stages resulted in somewhat lower correlations: an average value of 0.55 for the three-point spectral trace. If, however, only the *initial* and the *final* parts of the diphthongs were included in the dynamic approach, the correlations proved to be higher: 0.67 on the average for the two-point spectral trace. The *center* spectrum seems to be less important, which may be due to its uncertain position. As can be seen in Table VI, the results based on two PCA dimensions show the same tendency. Though the results based on formants are obscured by the difficulties in estimating the diphthongs' formants (see Sec. II B), in this case the agreement is also optimal for the two-point traces.

For completeness, we also calculated the correlations between the subjective scores and the absolute deviations from the reference vowels' durations [measure (15), Table V]. Though the listeners were instructed to disregard those deviations, the values are still quite high. Nevertheless, we believe that such a duration measure is confounded with the spectral measures to a large extent: Durational deviations mainly appeared for the deaf children, who also showed the largest deviations in pronunciation quality. Therefore, the results cannot lead to conclusions about the effect of durational deviations in themselves.

3. Complete vowel sets (overall assessment per speaker)

For each speaker, the *objective overall distance* was defined as the average objective distance of all monophthongs

of that speaker to a certain reference set. In a comparable way the *subjective overall deviation* was expressed as the average subjective deviation, based on the magnitude estimation experiment. Correlation coefficients between objective and subjective overall distances for 48 speakers are given in Table V, for both reference sets. All overall correlation coefficients appear to be quite high; measures based on the speaker-normalized spectral and PCA levels and those based on the formants yield comparable results. Only measure (15), based on duration, yields lower correlations. In Fig. 13 diagrams are given representing the subjective versus the objective overall distance, based on measures (3) and (6) with the best judged reference set. Along both axes, the separation between the norm and deaf speakers is almost complete, which largely explains the high correlation coefficients. Only one of the deaf speakers seems to be classified, subjectively as well as objectively, in the norm group. Within the subgroups, however, the relation is not as good as it was for the adult male vowels (cf. Bakku et al., 1993).

V. GENERAL DISCUSSION

As stated in the Introduction, the adequacy of an objective description of vowels depends on a number of important factors. The analysis by a static whole-spectrum approach properly fulfills those factors. Most important is the *agreement with subjective assessment*. For the children's complete vowel sets, the objective assessment of the pronunciation quality, derived from the spectral distances to a set of reference spectra, correlates well with the subjective assessment, derived by magnitude estimation. For most of the individual vowels, the correlation also proves to be significant, though the spread around the regression lines is substantial (Fig. 12): Reliable quality predictions can only be derived for quite large differences between the utterances. The choice of reference utterances has to be made deliberately. Even between well-judged utterances of norm speakers, the spectral differences are far from negligible, but these differences can be reduced considerably by speaker normalization.

As could be expected for children of these ages, no significant general differences can be shown between the vowel pronunciation of boys and girls.

FIG. 13. Subjective overall distances vs objective overall distances for 24 normal-hearing and 24 deaf children. Calculations are based on the 12 monophthongs. The subjective distance is derived from the magnitude estimation experiment; the objective distance is derived from weighted spectral differences [(a), measure (3), Table V], and from distances after projection in the norm subspace [(b), measure (6)]. The correlation coefficient r is given.

It must be realized that the results of the magnitude estimation cannot be considered as an indisputable assessment of pronunciation quality: The fact that all vowels of one speaker are judged consecutively, corresponding with actual diagnostic procedures, may have biased the ratings (see Sec. III A 1). Furthermore, though the listeners were instructed to judge pronunciation quality, voice quality as well as durational aspects will probably have had some influence on the judgments.

The whole-spectrum approach may be considered to be a first-order approximation of the concept of the ear's critical bands, which is a *basic model of auditory perception*, and the filter band spectra are *readily interpretable*. A data reduction by PCA enables a display of the vowels in a two-dimensional subspace. The resulting vowel configuration is similar to the configuration of the same vowels in a perceptual space, as obtained from the confusion matrices of the identification experiment (compare Figs. 5 and 10). Agreement between objective distances in such a two-dimensional PCA subspace and the subjective assessment by magnitude estimation is quite high for the children's complete vowel sets, and still significant for most of the individual vowels. Furthermore,

the representation of the vowels in the PCA subspace resembles the representation in the $F1-F2$ space (see Sec. II D). In both the PCA as well as in the formant domain (Figs. 5 and 7, and Tables II and IV) the different monophthongs of the norm speakers are well separated, whereas the monophthongs of the deaf speakers show considerable overlap; many utterances are reduced towards œ . However, the reduction is less striking for deaf children than it proved to be for deaf adult males (Bakkum *et al.*, 1993). This may be due to a deterioration of vowel pronunciation after the male voice mutation; it could also be caused by the recently improved vowel training procedures at the Institute for the Deaf. On the other hand, the separation of the normal-hearing children is less apparent than it was for the normal-hearing males; apart from the larger spread in pronunciation quality, an important factor is the great variation in fundamental frequencies of the children's vowels. An $F0$ -compensation procedure (Sec. II A 2) proved to be very useful. Speaker normalization also appeared to improve the separation of the monophthongs. In the comparison between the whole-spectrum analysis and subjective assessment, both $F0$ compensation and speaker normalization improved the agreement. The fact that some modifications are needed indicates that the whole-spectrum approach in itself is no perfect model of peripheral auditory analysis. These modifications may reflect underlying perceptual processes, such as adaptation to a speaker, but we believe that such perceptual processes are far more subtle than the straightforward procedures we used. In the formant domain, only a slight improvement of the correlation with subjective scores could be achieved by either extrinsic or intrinsic normalization; the separation of the monophthongs, however, did take benefit of both normalization procedures. First, speakers' average formant values are a limited cue to normalize individual vowels and, second, there is a relevant, but not dominating relation between the fundamental frequency and the formants. Apart from that, it appears that in expressing formants a logarithmic or Bark scale is preferable to a linear scale.

We also studied the dynamic aspects of the vowels. Average results (Figs. 6 and 8) show that vowel-inherent change is minimal for the monophthongs [though three of them ($/e/$, $/o/$, and $/\phi/$) show diphthongization effects], but large for the diphthongs. Not surprisingly, applying a distance measure based on such dynamic traces does not yield an improvement for the monophthongs. For the diphthongs, however, an improvement can be achieved if only the initial and final spectra are incorporated in the measure. This is in agreement with results of Pols (1979) and Bakkum *et al.* (1993). They concluded that the quality of the initial part determines to a large extent the perceived quality of the complete diphthong; only a part of the glide, pointing in the right direction, needs to be present.

The whole-spectrum approach will not account for all aspects of vowel perception, but it provides a useful model of pronunciation quality for evaluating (deaf) vowels. A model based on two formants yielded reasonable, but lower results, comparable to the results of a model based on two PCA dimensions. The concept of formant analysis as an auditory model appears to be comparable with the concept

based on a whole-spectrum analysis followed by PCA, provided that the formant estimation is performed accurately. However, whole-spectrum analysis, possibly followed by PCA, is a more robust and *determinate* approach than formant analysis; the latter may result, in the case of poorly defined formants, in errors of omission, and, if an automatic extraction procedure is applied, in underestimations of both $F1$ and $F2$ (see Sec. II D). For children's vowels, the formant analysis procedure proved to be not suitable for a real-time application. The whole-spectrum analysis, on the other hand, needs *limited computational effort*; it can be performed in real time, with as well as without projection in the PCA subspace. It can therefore easily be implemented in a speech training and observation system, which can be utilized by the deaf or by second language students to improve their pronunciation. Moreover, when using normalization and $F0$ -compensation procedures, it can handle a *wide variety of vowels and speakers*. The fact that children's vowels can be handled too is very promising, since (deaf) children make up one of the main target groups in possible applications of objective analysis in vowel training [cf. the phoneme identification and display procedure by Arends (1993)].

ACKNOWLEDGMENTS

This research was supported partly by the Institute for the Deaf in Sint-Michielsgestel, and partly by the Foundation for Linguistic Research, which is funded by the Netherlands Organization for Research NWO. We thank Dick Van Bergem and Rob van Son for their cooperation to the formant analysis.

¹The gently sloping high-pass filter, which was primarily meant to eliminate the 50-Hz noise caused by the power mains, somewhat reduced the output of the first filter band. However, the amplitude of a vowel's fundamental is always rather high, so that if the fundamental frequency occurred within the first band then it was still present (though slightly reduced) after high-pass filtering; in the cases of a higher $F0$ the high-pass filter did not make any difference at all. Anyway, as argued for in Sec. II A 5, the first two filter bands were not included in most analyses.

²In our case, using third formant ($F3$ or $F3-F2$) values as the second dimension, as suggested by, among others, Trauttmüller (1981) and Syrdal (1985), resulted in inferior separation of the vowels.

³The separation values calculated according to our definition represent a quite arbitrary measure, not an absolute one, for they are based on group ellipses, not on individual data. Highly deviating utterances may lie outside the vowel ellipses; such utterances may have been identified incorrectly, whereas the separation may still be 100%. This is the case for the norm speakers: Defining a vowel to be correctly identified if the majority of listeners assigned it to the correct category results in a correct identification of only 94% of the norm vowels, whereas the calculated separation was 99%. Using the same definition, 50% of the deaf vowels has been identified correctly, whereas the separation (averaged over girls and boys) was 34%.

Angelocci, A. A., Kopp, G. A., and Holbrook, A. (1964). "The vowel formants of deaf and normal-hearing eleven- to fourteen-year-old boys," *J. Speech Hear. Disord.* **29**, 156-170.

Arends, N. J. M. (1993). "The visual speech apparatus: An aid for speech training," Dissertation, Instituut voor Doven, Sint-Michielsgestel, The Netherlands.

Bakku, M. J., Plomp, R., and Pols, L. C. W. (1989). "Objective evaluation of word pronunciation by filter-band analysis," in *Proceedings EuroSpeech '89*, edited by J. Tubach and J. Mariani (CEP, Edinburgh, UK), Vol. 1, pp. 425-428.

- Bakku, M. J., Plomp, R., and Pols, L. C. W. (1993). "Objective analysis versus subjective assessment of vowels pronounced by native, non-native, and deaf male speakers of Dutch," *J. Acoust. Soc. Am.* **74**, 1989-2004.
- Bladon, R. A. W. (1982). "Arguments against formants in the auditory representation of speech," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström (Elsevier, Amsterdam, The Netherlands), pp. 95-102.
- Cohen, A., Slis, I. H., and 't Hart, J. (1963). "Perceptual tolerance of isolated Dutch vowels," *Phonetica* **9**, 65-78.
- Cohen, A., Slis, I. H., and 't Hart, J. (1967). "On tolerance and intolerance in vowel perception," *Phonetica* **16**, 65-70.
- Disner, S. F. (1980). "Evaluation of vowel normalization procedures," *J. Acoust. Soc. Am.* **67**, 253-261.
- Fox, R. A. (1989). "Dynamic information in the identification and discrimination of vowels," *Phonetica* **46**, 97-116.
- Gold, T. (1980). "Speech production in hearing impaired children," *J. Commun. Disord.* **13**, 397-418.
- Harman, H. H. (1967). *Modern Factor Analysis* (University of Chicago, Chicago).
- Klein, W., Plomp, R., and Pols, L. C. W. (1970). "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Am.* **48**, 999-1009.
- 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.
- Levitt, H., and Stromberg, H. (1983). "Segmental characteristics of the speech of hearing-impaired children: Factors affecting intelligibility," in *Speech of the Hearing Impaired*, edited by I. Hochberg, H. Levitt, and M. J. Osberger (University Park, Baltimore), pp. 53-73.
- Markel, J. D., and Gray, A. H., Jr. (1976). *Linear Prediction of Speech* (Springer-Verlag, Berlin).
- Miller, J. D. (1989). "Auditory-perceptual interpretation of the vowel," *J. Acoust. Soc. Am.* **85**, 2114-2134.
- Monsen, R. B. (1978). "Toward measuring how well hearing-impaired children speak," *J. Speech Hear. Res.* **21**, 197-219.
- Nearey, T. M., and Assmann, P. F. (1986). "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297-1308.
- Nearey, T. M. (1989). "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.* **85**, 2088-2113.
- Nooteboom, S. G., and Cohen, A. (1984). *Spoken en Verstaan: Een Nieuwe Inleiding tot de Experimentele Fonetiek* (Van Gorcum, Assen, The Netherlands).
- Patterson, R. D., and Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, London), pp. 123-177.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175-184.
- Plomp, R., Pols, L. C. W., and Van de Geer, J. P. (1967). "Dimensional analysis of vowel spectra," *J. Acoust. Soc. Am.* **41**, 707-712.
- Pols, L. C. W. (1977). "Spectral analysis and identification of Dutch vowels in monosyllabic words," Dissertation, Acad. Pers BV, Amsterdam, The Netherlands.
- Pols, L. C. W. (1979). "Some spectral and perceptual measurements on Dutch diphthongs," in *Studia Gratulatoria Dedicated to Hendrik Mol* (I.F.A., Amsterdam, The Netherlands), pp. 270-285.
- Pols, L. C. W., Tromp, H. R. C., and Plomp, R. (1973). "Frequency analysis of Dutch vowels from 50 male speakers," *J. Acoust. Soc. Am.* **53**, 1093-1101.
- Pols, L. C. W., Van der Kamp, L. J. T., and Plomp, R. (1969). "Perceptual and physical space of vowel sounds," *J. Acoust. Soc. Am.* **46**, 458-467.
- Pols, L. C. W., and Van Son, R. J. J. H. (1993). "Acoustics and perception of dynamic vowel segments," *Speech Commun.* **13**, 135-147.
- Povel, D., and Wansink, M. (1986). "A computer-controlled vowel corrector for the hearing impaired," *J. Speech Hear. Res.* **29**, 99-105.
- Reetz, H. (1989). "A fast expert program for pitch extraction," in *Proceedings EuroSpeech '89*, edited by J. Tubach and J. Mariani (CEP, Edinburgh, UK), Vol. 1, pp. 476-479.
- Smith, C. R. (1975). "Residual hearing and speech production in deaf children," *J. Speech Hear. Res.* **18**, 795-811.
- Steeneken, H. J. M. (1992). "On measuring and predicting speech intelligibility," Dissertation, University of Amsterdam, The Netherlands.

- Strange, W. (1989a). "Evolving theories of vowel perception," *J. Acoust. Soc. Am.* **85**, 2081–2087.
- Strange, W. (1989b). "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.* **85**, 2135–2153.
- Syrdal, A. K. (1985). "Aspects of a model of the auditory representation of American English vowels," *Speech Commun.* **4**, 121–135.
- Trautmüller, H. (1981). "Perceptual dimension of openness in vowels," *J. Acoust. Soc. Am.* **69**, 1465–1475.
- Van Alphen, P. (1992). "HMM-based continuous-speech recognition," Dissertation, PTT Research, Leidschendam, The Netherlands.
- Wagenaar, W. A. (1969). "Note on the construction of digram-balanced Latin squares," *Psychol. Bull.* **72**, 384–386.
- Weenink, D. (1985). "Formant analysis of Dutch vowels from 10 children," in *Proceedings of the Institute of Phonetic Sciences 9* (University of Amsterdam, The Netherlands), pp. 45–52.
- Willems, L. F. (1986). "Robust formant analysis," in *IPO Annual Progress Report 21* (IPO, Eindhoven, The Netherlands), pp. 34–40.
- Zahorian, S. A., and Jaghargi, A. J. (1993). "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Am.* **94**, 1966–1982.
- Zwicker, E., and Terhardt, E. (1980). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.* **68**, 1523–1525.