A comparison of vowel normalization procedures for language variation research

Patti Adank[a)]
Center for Language Studies, Radboud University Nijmegen, PO Box 9103, 6500 HD Nijmegen, The Netherlands

Roel Smits
Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands

Roeland van Hout
Center for Language Studies, Radboud University Nijmegen, PO Box 9103, 6500 HD Nijmegen, The Netherlands

(Received 24 February 2004; revised 9 July 2004; accepted 26 July 2004)

An evaluation of vowel normalization procedures for the purpose of studying language variation is presented. The procedures were compared on how effectively they (a) preserve phonemic information, (b) preserve information about the talker's regional background (or sociolinguistic information), and (c) minimize anatomical/physiological variation in acoustic representations of vowels. Recordings were made for 80 female talkers and 80 male talkers of Dutch. These talkers were stratified according to their gender and regional background. The normalization procedures were applied to measurements of the fundamental frequency and the first three formant frequencies for a large set of vowel tokens. The normalization procedures were evaluated through statistical pattern analysis. The results show that normalization procedures that use information across multiple vowels (''vowel-extrinsic'' information) to normalize a single vowel token performed better than those that include only information contained in the vowel token itself (''vowel-intrinsic'' information). Furthermore, the results show that normalization procedures that operate on individual formants performed better than those that use information across multiple formants (e.g., ''formant-extrinsic'' $F2$-$F1$).
[DOI: 10.1121/1.1795335]

PACS numbers: 43.70.Jt, 43.71.Es, 43.70.Kv [AL] Pages: 3099–3107

I. INTRODUCTION

In their widely cited study on vowel perception, Ladefoged and Broadbent (1957) argue that three types of information are conveyed when a talker pronounces a vowel sound: (a) Phonemic information, i.e., the intended phonemic identity of the vowel sound; (b) anatomical/physiological information about the talker's vocal tract shape, gender, or physiology; and (c) sociolinguistic information, i.e., information about the talker's group characteristics, such as regional background or socioeconomic status. The first type of information is related to the linguistic message, whereas the second and the third types are talker-related. All three information types have been found to systematically affect formant frequencies (e.g., Peterson and Barney, 1952, for the first two information types and Hindle, 1978 and Labov, 2001, for the third type).

The influence of anatomical/physiological and sociolinguistic talker-related factors on formant frequencies has generally been treated as unwanted variation in research on vowel perception (Peterson and Barney, 1952; Pols *et al.*, 1973). Several studies aimed at eliminating the talker-related variation by designing procedures that can be subsumed under the heading of vowel, or talker, normalization (e.g., Gerstman, 1968; Lobanov, 1971; Nordström, 1976; Nearey, 1978; Syrdal and Gopal, 1986; Miller, 1989).

Traditionally, vowel normalization procedures are classified according to the type of information they employ. The procedures are defined as either vowel-intrinsic or vowel-extrinsic (Ainsworth, 1975; Nearey, 1989). Vowel-intrinsic procedures use only acoustic information contained within a single vowel token to normalize that vowel token. These procedures typically consist of a nonlinear transformation of the frequency scale (log, mel, bark), and/or a transformation based on a combination of formant frequencies (e.g., $F1$-$F10$). An example of an intrinsic procedure can be found in Syrdal and Gopal (1986). Vowel-extrinsic procedures, on the other hand, assume that information is required that is distributed across more than one vowel of a talker; e.g., the formant frequencies of the point vowels for that talker. Examples of extrinsic procedures can be found in Gerstman (1968), Lobanov (1971), Nordström (1976), and Nearey (1978). Generally speaking, vowel-intrinsic procedures were developed with the primary aim of modeling human vowel perception, while vowel-extrinsic procedures were developed with the purpose of obtaining higher percentages correctly classified vowel tokens for automatic speech recognition purposes.

In recent years, vowel normalization procedures have

[a)]Author to whom correspondence should be addressed. Electronic mail: p.adank@student.ru.nl

been applied in studies with a purpose other than modeling vowel perception or improving automatic vowel classification, i.e., in language variation studies describing the linguistic characteristics of vowel systems for specific languages or language varieties. These variation studies included vowel-intrinsic as well as vowel-extrinsic procedures. Labov (2001) used Nearey's (1978) logmean procedure for the description of the vowel system of Philadelphia. Most *et al.* (2000) used the procedure proposed by Syrdal and Gopal (1986) to describe the Hebrew vowel system. Watson *et al.* (2000) used Lobanov's (1971) procedure for their description of the vowels of New Zealand-English. Hagiwara (1997) transformed the formant values for the (Californian) American-English vowels to bark, as did Deterding (1997) for the vowels of Standard Southern British-English. Finally, Hillenbrand *et al.* (1995) transformed the vowels of American-English to $F1$-$F0$ and $F3$-$F2$ on a mel scale.[1]

However, using normalization procedures in language variation research is not without drawbacks. It has been reported that some normalization procedures introduce artificial variation patterns into the description when the vowel systems of the languages/dialects to be compared are not phonologically equivalent (Disner, 1980). Moreover, there are indications that applying normalization procedures reduces sociolinguistic variation in the acoustic representation along with the anatomical/physiological variation (Hindle, 1978). However, Labov (2001), evaluated the same two procedures as Hindle (1978), (i.e., Nearey, 1978 and Nordström, 1976) and stated that Hindle's conclusion was too strong and that most of the sociolinguistic variation was retained in the normalized data after normalization using Nearey's procedure.

The purpose of the present study is to establish to what extent vowel normalization procedures are suitable for use in language variation research and which (type of) procedure performs best. We attempted to extend earlier studies that compare vowel normalization procedures, such as Hindle (1978), Disner (1980), Syrdal and Gopal (1986), Nearey (1989), and Labov (2001) and to evaluate how well the procedures preserve sociolinguistic variation in normalized vowel data. Although the earlier studies can be said to have evaluated normalization procedures on how well they preserve sociolinguistic differences, they are limited in that a small number of talkers was used (Hindle, 1978; Labov, 2001), or in that the vowel systems that were compared were not phonologically equivalent (Disner, 1980).

We compared a set of eleven normalization procedures to a baseline condition (no normalization, i.e., formant frequencies in Hz) using measurements of the nine monophthongal vowels of Dutch, produced by 160 talkers of Dutch who were stratified for their gender and regional background. For each vowel token, the fundamental frequency and the frequencies of the first three formants were measured. Subsequently, we applied the procedures to the acoustic measurements, thus generating eleven normalized representations of the vowel data. These representations were compared on how well they preserved phonemic and sociolinguistic information and to what degree they succeeded in reducing anatomical/physiological information in each representation as compared to the other representations.

II. METHOD

A. Speech material

We used a database of measurements previously described in Adank *et al.* (2004) and in Adank (2003). These materials consist of recordings of 160 talkers of Dutch who were stratified for their regional background (speaking one of eight regional varieties of Standard Dutch) and their gender. The talkers can be regarded as professional language users, as they were all teachers of Dutch at secondary education institutes at the time the recordings were made. All 160 talkers produced two tokens of each of nine monophthongal vowels of Dutch, /ɑ a ɛ ɪ i ɔ u ʏ y/, in a neutral /sVs/ context.

Two speech communities were distinguished: The Netherlands and Flanders (Belgium). Two different varieties of Dutch can be identified: Northern Standard Dutch as spoken in the Netherlands, and Southern Standard Dutch as spoken in Flanders. The pronunciation of the two varieties has evolved differently from the time the Dutch area was split up in the 19th century. See Van de Velde *et al.* (1997) for a detailed overview. The 160 talkers were sampled across four regions per speech community: A central region, an intermediate region, and two peripheral regions. The central region is the economically and culturally dominant region in each speech community. For the Netherlands, the central region is the west, consisting of the provinces of North Holland, South Holland and Utrecht, also known as ''the Randstad.'' The cities Amsterdam, Rotterdam, Utrecht, and The Hague are part of the Randstad. In Flanders, the central region is ''Brabant.'' Brabant comprises the Belgian provinces Antwerpen and Flemish Brabant, with the cities of Antwerpen and Leuven, respectively. The intermediate region in the Netherlands encloses the southern part of the province Gelderland and part of the province Utrecht. The intermediate region in Flanders is the province East Flanders. In the Netherlands, the two peripheral regions are the province Limburg, in the south of the Netherlands, and the province Groningen, in the north of the Netherlands. The two peripheral regions for Flanders are the provinces (Belgian) Limburg and West Flanders. In each of the eight regions, recordings were made of twenty talkers, ten women and ten men.

The vowel tokens were recorded as a task in a so-called ''sociolinguistic interview'' in which vowels and consonants were elicited in a wide variety of tasks. All target vowels were produced in a carrier sentences task, which was repeated twice in the course of the interview. The vowels were available in three different consonantal contexts (CVC, CVCV, or V). The vowels in the CVC contexts (/sVs/) were selected for further processing. In total, 2880 vowel tokens were recorded: Two tokens of each of the nine monophthongal Dutch vowels, produced by 160 talkers.

Recording conditions were different for each of the talkers. Some were interviewed in an empty classroom and others were interviewed at their own home. Due to these differences in recording conditions, in rare cases, background

TABLE I. The selected procedures, divided according to whether they use vowel-intrinsic or vowel-extrinsic information.

	Vowel-intrinsic procedures
HZ	baseline condition, formant frequencies in Hz
LOG	log-transformation of the frequency scale
BARK	bark-transformation of the frequency scale
MEL	mel-transformation of the frequency scale
ERB	ERB-transformation of the frequency scale
S & G	Syrdal and Gopal's (1986) bark-distance model
	Vowel-extrinsic procedures
LOBANOV	Lobanov's (1971) z-score transformation
NEAREY1	Nearey's (1978) single logmean procedure
NEAREY2	Nearey's (1978) shared logmean procedure
GERSTMAN	Gerstman's (1968) range normalization
NORDSTRÖM	Nordström's (1976) vocal-tract scaling
MILLER	Miller's (1989) formant-ratio model

noises were audible. Whenever this was the case, the speech segment was excluded from further analysis.

$F0$, $F1$, $F2$, and $F3$ were extracted from each token's temporal mid point. $F0$ was extracted automatically with the speech-processing software program Praat using an autocorrelation-based procedure that was evaluated as the best option available in Praat (Boersma, 1993). The formant frequencies were obtained through a semiautomatic procedure developed by Nearey *et al.* (2002). For further details of the process through which the acoustic measurements were obtained, see Adank *et al.* (2004) and Adank (2003).

B. Selection of normalization procedures

Only normalization procedures that were described in previously published studies on acoustic vowel normalization were selected. A variety of studies evaluate the performance of procedures, either for use in language variation and change (Hindle, 1978; Disner, 1980), for a phonetic theory of vowel perception (Nearey, 1978, 1992; Syrdal, 1984; Nearey, 1978), or for automatic speech recognition (Deterding, 1990). We included all procedures described in these six studies that take formant frequencies as their input and that generate output in the form of normalized versions of those formant frequencies.[2] Table I lists the selected procedures.

Each procedure was implemented as follows. HZ, or the baseline condition, refers to the frequencies for the fundamental frequency $F0$ and formant frequencies $F1$ through $F3$. LOG refers to log-transformed $F0$ through $F3$ in Hz. BARK, the bark-transformation of the baseline, was implemented with Traunmüller's (1990) Eq. (1).[3] We decided to use this transformation, because Traunmüller (1990) shows that his equation fits Zwicker's (1961) table of critical bands better than Zwicker and Terhardt's (1980)

$$F_i^{\mathrm{B}} = 26.81 \times \left(\frac{F_i}{1960 + F_i} \right) - 0.53. \tag{1}$$

F_i in (1) is $F0$, $F1$, $F2$, or $F3$. The mel-transformed data, MEL, was obtained by transforming $F0$ through $F3$ using Stevens and Volkmann's (1940) equation as in (2)

$$F_i^{\mathrm{M}} = 2595 \times \ln\left(1 + \frac{F_i}{700} \right). \tag{2}$$

The ERB-transformation was implemented using Glasberg and Moore's (1990) Eq. (3).

$$F_i^{\mathrm{E}} = 21.4 \times \ln(0.004\,37 \times F_i + 1). \tag{3}$$

Syrdal and Gopal's bark-distance transformation (S & G) was implemented by first transforming $F0$ through $F3$ to bark using (1) and subsequently by applying Eqs. (4) and (5). Syrdal and Gopal (1986) originally used Zwicker and Terhardt's (1980) bark-transformation, while we used Traunmüller's (1990) for reasons stated above. We chose to use one type of bark-transformation in the present study; as a consequence Syrdal and Gopal's procedure was implemented with a bark-transformation different from the one they used in their 1986 paper.

$$F_1^{\mathrm{S\&G}} = F_1^{\mathrm{B}} - F_0^{\mathrm{B}}, \tag{4}$$

$$F_2^{\mathrm{S\&G}} = F_3^{\mathrm{B}} - F_2^{\mathrm{B}}. \tag{5}$$

Gerstman's (1968) normalization (GERSTMAN) was calculated for $F0$ through $F3$ as in (6)

$$F_{ti}^{\mathrm{Gerstman}} = 999 \times \frac{F_{ti} - F_{ti}^{\min}}{F_{ti}^{\max} - F_{ti}^{\min}}, \tag{6}$$

where $F_{ti}^{\min}$ is the minimum value of F_i for all nine vowels for talker t and $F_{ti}^{\max}$ is the maximum of F_i for the nine monophthongal vowels for that talker. Lobanov's (1971) z-score transformation was calculated for $F0$ through $F3$ as in Eq. (7)

$$F_{ti}^{\mathrm{Lobanov}} = \frac{F_{ti} - \mu_{ti}}{\delta_{ti}}, \tag{7}$$

where μ_{ti} is the average formant frequency across the nine monophthongal vowels for talker t and δ_{ti} refers to the standard deviation for average μ_{ti}. Nearey's (1978) single logmean (NEAREY1) was calculated for $F0$ through $F3$ as in Eq. (8)

$$F_{ti}^{\mathrm{Nearey1}} = F_{ti}^{\mathrm{L}} - \mu_{D_{ti}^{\mathrm{L}}}, \tag{8}$$

where F_{ti}^{L} is the log-transformed value of F_i for talker t and $\mu_{D_{ti}^{L}}$ is the average across the log-transformed formant frequencies across the nine vowels for that talker t. NEAREY1 uses a separate scale factor for each formant. Nearey's (1978) shared logmean (NEAREY2) uses a scale factor that is identical across formants. NEAREY2 was calculated for $F0$ through $F3$ as in (9).

$$F_{ti}^{\mathrm{Nearey2}} = F_{ti}^{\mathrm{L}} - (\mu_{D_{0t}^{\mathrm{L}}} + \mu_{D_{1t}^{\mathrm{L}}} + \mu_{D_{2t}^{\mathrm{L}}} + \mu_{D_{3t}^{\mathrm{L}}}). \tag{9}$$

The shared logmean $F_{ti}^{\mathrm{Nearey2}}$ is thus based on the four logmeans for $F0$, $F1$, $F2$, $F3$ ($\mu_{D_{0t}^{L}}$, $\mu_{D_{1t}^{L}}$, $\mu_{D_{2t}^{L}}$, and $\mu_{D_{3t}^{L}}$) in Eq. (9). Each log-transformed $F0$ or formant frequency is expressed as its distance to the shared logmean for a given talker t. Nordström's (1976) vocal-tract scaling, or NORDSTRÖM, was calculated as in (10) and (11)

$$F_i^{\mathrm{Nordström}} = kF_i^{\mathrm{female}}, \tag{10}$$

$$k = \frac{L^{\text{male}}}{L^{\text{female}}} = \frac{\mu_{F_3}^{\text{male}}}{\mu_{F_3}^{\text{female}}}, \tag{11}$$

where the scaling factor k in (11) expresses the ratio of the length L^{female} of the average female vocal tract to the length L^{male} of the average male vocal tract. k is calculated across all vowel tokens with an $F1$ greater than 600 Hz (across all 160 talkers), $\mu_{F_3}^{\text{male}}$ is the average $F3$ for all male talkers calculated across all vowel tokens with $F1 > 600$ Hz and $\mu_{F_3}^{\text{female}}$ is the average $F3$ for all female talkers calculated across all vowel tokens with $F1 > 600$ Hz. All values of $F0$ through $F3$ for the female talkers were subsequently transformed using (10). Finally, Miller's (1989) formant-ratio model was implemented using Eqs. (12)–(15).

$$F_{1t}^{\text{Miller}} = \left(\frac{F_{1t}^{\text{L}}}{SR} \right), \tag{12}$$

$$F_{2t}^{\text{Miller}} = \left(\frac{F_{2t}^{\text{L}}}{F_{1t}^{\text{L}}} \right), \tag{13}$$

$$F_{3t}^{\text{Miller}} = \left(\frac{F_{3t}^{\text{L}}}{F_{2t}^{\text{L}}} \right), \tag{14}$$

$$SR = k \left(\frac{\mu_{F_{0t}^{\text{L}}}}{k} \right)^{1/3}, \tag{15}$$

SR in (15) expresses Miller's talker-specific ''Sensory reference,'' which was calculated using the geometric average of all values of $F0$ for talker t, expressed by $\mu_{F_{0t}^{\text{L}}}$. The constant k reflects the geometric average of the overall average $F0$ across the 80 male (148 Hz) and 80 female talkers (234 Hz) and was set to 186 Hz for the present study.

III. RESULTS

A. Preserving phonemic variation

A series of discriminant analyses was carried out to establish how well the normalization procedures preserved information about the vowel token's intended phonemic identity in the normalized acoustic variables (two variables for S & G, three for MILLER, and four for all other methods). The acoustic variables served as predictors, while the intended vowel category, having nine possible values, was the dependent variable. A high percentage correctly classified vowel tokens indicates that the procedure succeeded at preserving phonemic variation.

Discriminant analysis (DA) is a standard pattern recognition technique that uses the pooled within-groups covariance matrix of the acoustic variables to classify cases. Linear discriminant analysis (LDA) assumes that the within-groups covariance matrices are equal across categories. If the data do not meet this assumption (which often holds for vowel formant frequencies), Quadratic discriminant analysis (QDA) is the appropriate analysis. However, although QDA theoretically models the individual vowel distributions more accurately, it has the drawback that it requires much larger numbers of parameters to be estimated than LDA, thus risking overfitting the data. Therefore, LDA as well as QDA were carried out. The results are presented in Table II.

TABLE II. Percentages correctly classified vowel tokens for LDA 1 and QDA 1 on the pooled data from 160 talkers. The dependent variable for each analysis is vowel category and $F0$ through $F3$ served as predictors. For LDA 1, all percentages higher than 81%, indicated by "$^{\uparrow}$," or lower than 77%, "$^{\downarrow}$," (all percentages are rounded off to the nearest whole number) are significantly different from the baseline condition (HZ). For QDA 1, this is 83% and 79%, respectively.

		LDA 1	QDA 1
	HZ	79	81
	LOG	80	81
Vowel-intrinsic	BARK	80	82
	ERB	80	82
	MEL	80	82
	S & G	$69^{\downarrow}$	70
	LOBANOV	$92^{\uparrow}$	$93^{\uparrow}$
	NEAREY1	90	$91^{\uparrow}$
Vowel-extrinsic	NEAREY2	$82^{\uparrow}$	83
	GERSTMAN	$84^{\uparrow}$	$86^{\uparrow}$
	NORDSTRÖM	$82^{\uparrow}$	$84^{\uparrow}$
	MILLER	$76^{\downarrow}$	77

Table II shows, first, that the percentages correctly classified vowel tokens for QDA 1 are only 1% to 2% higher than those for LDA 1. Given the parsimony of the LDA model relative to QDA, we decided to use LDA instead of QDA in the rest of this study. Second, it appears for LDA 1 that five procedures performed better than the baseline (HZ) and two procedure performed worse. LOBANOV (92%) and NEAREY1 (90%) preserved the phonemic variation in the data best of all procedures, followed by GERSTMAN (84%), NORDSTRÖM (82%), and NEAREY2 (82%), while MILLER (76%) and S & G (69%) performed poorest of all. No significant improvement over the baseline was found for the scale transformations LOG, BARK, ERB, and MEL.

Disner (1980) compared four procedures with raw data in Hz: Gerstman's range normalization (1968), Lobanov's z-transformation (1971), Nearey's logmean procedure (1978), and Harshman's (1970) PARAFAC model (not discussed in the present study). She applied these procedures to vowel data from six Germanic languages: English, Norwegian, Swedish, German, Danish, and Dutch. Disner calculated the percentage of scatter reduction of the formant frequencies per vowel in an $F1/F2$ plot per procedure. Her results show, although no specific procedure is the most effective for all the languages, that Nearey's procedure is generally the most effective (especially for Danish and Dutch). Lobanov's procedure is slightly less effective than Nearey's, followed by Gerstman's. Overall, our results seem compatible with Disner's.

Syrdal (1984) compared eight normalization procedures with raw data in Hz: The log-transformation, the bark-transformation, Syrdal's bark-difference model (1984), two versions of Miller (1980), two versions of Nearey's (1978) procedure, and Gerstman (1968). She applied them to Peterson and Barney's (1952) data set and calculated the percentage correctly classified vowel tokens from LDA. Overall, our results in Table II show a pattern similar to Syrdal's. Syrdal reports that Nearey's procedure (similar to NEAREY1) per-

TABLE III. Percentages correctly classified vowel tokens for LDA 2–4 on the pooled data from 160 talkers. The dependent variable for each analysis is gender (chance level 50%). For LDA 2, all percentages lower than 92% differ significantly from the baseline (HZ). For LDA 3, this is 87%, and for LDA 4, this is 78%. For all LDAs, percentages lower than 53% indicate performance at chance level (labeled with "*"). LDAs 3 and 4 were not carried out for S & G and MILLER; these procedures do not use $F0$, or $F1$-$F3$ in the same way as the other procedures [cf. Eqs. (4–5) and (12–15)].

	Predictor variables	LDA 2 $F0$, $F1$, $F2$, $F3$	LDA 3 $F0$	LDA 4 $F1$, $F2$, $F3$
	HZ	93	89	80
	LOG	93	89	80
Vowel-intrinsic	BARK	93	89	80
	ERB	93	89	80
	MEL	92	89	80
	S & G	53*	⋯	⋯
	LOBANOV	50*	51*	51*
	NEAREY1	50*	51*	49*
Vowel-extrinsic	NEAREY2	81	78	69
	GERSTMAN	53*	53*	51*
	NORDSTRÖM	83	82	52*
	MILLER	79	⋯	⋯

formed best, while we found that NEAREY1 performed second best, after LOBANOV (not evaluated by Syrdal). One major difference between Syrdal's results and our results is that Syrdal reports that the bark-difference procedure (nearly identical to S & G) performed better (85.9%) than her baseline condition (82.3%), while we found that S & G performed poorer than the baseline. This discrepancy may be partly attributed to differences in the implementation of the bark-transformation: Syrdal used Zwicker and Terhardt's (1980) and we used Traunmüller's (1990). Furthermore, we used talkers of Dutch and Syrdal's talkers spoke American English. Dutch may be one of the languages that cannot be described adequately by S & G's second dimension [cf. Eq. (5)]. Syrdal and Gopal (1986) stated that the critical distance for the front-back dimension [cf. Eq. (4)] is language-specific and that this distance is not a language-universal measure reflecting front-back vowel distinctions.

B. Reducing anatomical/physiological variation

Three LDAs were carried out (LDA 2–4) to establish to what extent anatomical/physiological gender-related variation was eliminated from the transformed data. LDA 2 evaluated whether information on the talker's gender was present in all four procedures' output. For LDA 2, the procedures' output variables served as predictors. LDA 3 and LDA 4 were carried out to investigate whether differences between the procedures found for LDA 2 could be attributed for the most part to gender-specific $F0$-differences, or to differences in the formant frequencies. In LDA 3, $F0$ served as the sole predictor, and $F1$, $F2$, and $F3$ served as predictors in LDA 4. For all three LDAs, it is assumed that a procedure is successful at eliminating gender-related anatomical/physiological variation when performing at chance level (50%).

Table III shows the results for LDA 2–4. For LDA 2, 93% of the vowel tokens were categorized correctly (i.e., as spoken by a male or female talker) for HZ, indicating that the raw measurements display considerable anatomical/physiological variation. Only LOBANOV, NEAREY1, and GERSTMAN performed at chance level for LDA 2, the other procedures did not eliminate all gender-specific variation. In particular, the scale transformations did not remove any gender-related variation. The results for LDA 3 show first that $F0$ displays a lot of gender-specific variation; for HZ 89% of the vowel tokens could be classified correctly when only $F0$ was entered as a predictor variable. The variation in $F0$ stems most likely from differences in the anatomy and physiology of the larynx of males and females. The pattern in the results for LDA 3 is similar to the pattern found for LDA 2: LOBANOV, NEAREY1, and GERSTMAN performed best (at chance level), while all the other procedures perform above chance level. Finally, $F1$, $F2$, and $F3$ display anatomical/physiological gender-related variation as well, although less than $F0$. This variation probably originates from differences in vocal-tract length between males and females. NORDSTRÖM, a procedure designed to account for vocal-tract length differences, eliminated gender-related variation completely. Recall that LDA 3 showed that NORDSTRÖM was not successful at eliminating the (larynx-related) anatomical/physiological variation in $F0$.

Syrdal (1984) carried out an LDA that classified the data as having been produced by a man, woman, or a child. The results in our Table III are compatible with the results in Syrdal's (1984) Table II. For the procedures that are common to our study and Syrdal's study, Syrdal found that Nearey's and Gerstman's procedures performed best (at chance level), while the other procedures performed above chance level.

C. Preserving sociolinguistic variation

The 160 talkers were stratified for regional background (eight regional varieties). LDA 5 served to establish to what extent regional (sociolinguistic) variation was preserved in the transformed acoustic representations of the vowel data. $F0$ through $F3$, transformed through each normalization procedure, were entered as predictors. Region served as the dependent variable, having eight levels. The analysis was repeated for each of the nine vowels, to eliminate the effect of the vowel token's category. If a certain procedure brought a classification level down from a value above chance level

TABLE IV. Results for LDA 5: Percentages of vowel tokens that were classified into the correct region, for each vowel category, for each normalization procedure. The number of cases per vowel category is 320. Percentages higher than 18% (rounded) are significantly higher than chance level (12.5%), percentages at chance level are indicated with ''*''.

		/ɑ/	/a/	/ɛ/	/ɪ/	/i/	/ɔ/	/u/	/ʏ/	/y/	Average
Vowel-intrinsic	HZ	27	23	36	35	29	29	33	38	26	31
	LOG	26	20	37	33	26	31	33	36	26	30
	BARK	27	22	35	34	26	29	33	37	27	30
	ERB	26	22	35	34	26	30	33	37	27	30
	MEL	27	22	35	33	26	29	33	37	25	30
	S & G	22	19	32	30	20	25	25	28	22	25
Vowel-extrinsic	LOBANOV	26	18	35	31	28	27	32	25	31	28
	NEAREY1	23	19	34	31	29	29	33	31	28	28
	NEAREY2	28	20	27	35	31	31	30	32	25	30
	GERSTMAN	25	22	36	34	19	26	25	31	26	27
	NORDSTRÖM	27	21	37	33	29	30	33	34	27	30
	MILLER	23	17*	35	31	31	25	29	32	23	27
	Average	26	20	35	33	26	28	31	33	26	29

(12.5%), it must be concluded that the procedure reduces systematic sociolinguistic variation related to the talker's regional background.

Table IV shows the results for LDA 5. It can first be observed, that the percentages correctly classified vowel tokens are generally above chance level across all procedures, indicating that none of the investigated procedures eliminated all sociolinguistic variation. Second, some differences between procedures can be observed: S & G eliminated more sociolinguistic variation than the other procedures, followed by GERSTMAN and MILLER, LOBANOV, and NEAREY1. Procedures that reduce anatomical/physiological variation most effectively show a larger reduction of the sociolinguistic variation. Furthermore, this reduction is not uniform across vowels for a given procedure (e.g., LOBANOV shows a large reduction for /a/ and a small reduction for /ɛ/). Table IV shows finally that /ɛ/, /ɪ/, and /ʏ/ display the most regional variation. The point vowels /a/ and /i/ show little regional variation, while /u/ shows slightly more variation.

D. Comparing the sources of variation

The LDA-based analyses presented in the previous section treat the normalization issue as a pattern recognition problem: How accurately can vowel identity, talker-gender, and regional background be recognized from the normalized acoustic data. The present analysis is based on the reverse approach: how much of the variation in the normalized data can be explained from the three factors vowel, talker-gender, and regional background. Several Multivariate Analyses of Variance (MANOVA) were carried out to reveal how the procedures deal with the variation in the acoustic measurements related to the three variation sources (phonemic, anatomical/physiological, and sociolinguistic). In each MANOVA, the talker's gender (''Gender''), the talker's regional background (''Region''), and the vowel token's category (''Vowel'') were used to predict the variation in the transformed acoustic variables. Only the baseline procedure HZ and the three procedures that were most successful at preserving phonemic variation and reducing anatomical/physiological variation, LOBANOV, NEAREY1, and GERSTMAN, were included. The MANOVAs were repeated three times, once with $F0$, $F1$, $F2$, $F3$ as dependent variables, once with $F1$, $F2$, $F3$, and once with only $F1$ and $F2$. This was done to evaluate the effect of eliminating $F0$, and $F0$ as well as $F3$, from the analysis. The multivariate measure of effect size for each set of factors and interaction terms was η^2, which reveals the proportion of the total variation in the dependent variable that is accounted for by the variation in the independent variable. The significance level was estimated using Pillai's trace.[4]

A high value for η^2 in Table V for the factor Vowel indicates that a lot of the phonemic variation in the dependent variables can be predicted by the vowel categories, indicating the preservation of phonemic variation in the acoustic variables. Subsequently, a low value of η^2 for the factor Gender indicates that there is relatively little anatomical/physiological gender-related variation present in the dependent variables. Finally, a high value for η^2 for the interaction between Vowel and Region indicates that sociolinguistic (regional) variation is preserved in the dependent variables. The interaction between Region and Vowel gives a better indication about the presence of regional variation in the data than the factor Region by itself. It seems likely that (large) effects for Region would only be found if the size and shape of the entire vowel systems differ across regions. This does not seem plausible, given the results in Table IV for the cardinal vowels /a/ and /i/, which were relatively stable across regions. Instead, a significant effect of η^2 for Vowel×Region indicates that some vowels show more regional variation than others, which seems plausible, given the relatively high percentages of /ɛ/, /ɪ/, and /ʏ/ in Table IV.

Table V shows that η^2 is highest for the factor Vowel across all procedures. Only for HZ, the largest variation in the dependent variables could be accounted for by the factor Gender (for $F0$ through $F3$, Gender shows a larger effect than Vowel). In contrast, there is no effect for Gender for LOBANOV and NEAREY1, and only a very small effect for GERSTMAN. This corroborates the earlier finding that these three procedures effectively removed all anatomical/physiological variation from the acoustic measurements. No significant effects were found for Region for LOBANOV and NEAREY1, and relatively small effects for HZ and

TABLE V. Results for the four multivariate analyses of variance: η^2 for each significant factor, for each of the four procedures ($p<0.001$). Values of η^2 not significantly different from 0 are not included. For each procedure, the analysis is repeated for three different sets of dependent variables. The number of tokens per analysis is 2880.

	HZ			LOBANOV			NEAREY1			GERSTMAN		
η^2	*F*0 *F*1 *F*2 *F*3	*F*1 *F*2 *F*3	*F*1 *F*2	*F*0 *F*1 *F*2 *F*3	*F*1 *F*2 *F*3	*F*1 *F*2	*F*0 *F*1 *F*2 *F*3	*F*1 *F*2 *F*3	*F*1 *F*2	*F*0 *F*1 *F*2 *F*3	*F*1 *F*2 *F*3	*F*1 *F*2
Vowel	0.527	0.695	0.893	0.579	0.760	0.932	0.556	0.731	0.914	0.568	0.743	0.917
Region	0.075	0.080	0.063	⋯	⋯	⋯	0.041	0.051	0.067	⋯	⋯	⋯
Gender	0.770	0.656	0.537	⋯	⋯	⋯	0.018	0.014	0.014	⋯	⋯	⋯
Vowel×Region	0.120	0.151	0.183	0.150	0.190	0.236	0.126	0.159	0.200	0.139	0.173	0.207
Vowel×Gender	0.064	0.079	0.108	0.014	0.017	0.019	0.011	0.014	0.016	0.019	0.024	0.025
Region×Gender	0.017	0.010	0.011	⋯	⋯	⋯	0.016	0.016	0.019	⋯	⋯	⋯
Vowel×Region×Gender	0.031	0.036	0.036	0.030	0.032	⋯	⋯	⋯	⋯	0.039	0.043	0.039
Vowel×Region×Gender	⋯	⋯	⋯	0.030	0.033	0.033	⋯	⋯	⋯	0.029	0.033	0.032

GERSTMAN. In the light of the discussion of the relevance of the effect for Region versus Vowel×Region, the small effects for HZ and GERSTMAN should not be overrated. Table V shows relatively large effects for all four procedures for Vowel×Region. The effects are largest for LOBANOV and GERSTMAN, indicating that a larger proportion of the sociolinguistic variation in the data can be accounted for after transforming data with these two procedures. Table V shows further that excluding $F0$ from the analysis leads to higher values for η^2 for all four MANOVAs for Vowel and Vowel×Region. Excluding $F0$ as well as $F3$ results in even higher values for η^2 for Vowel and Vowel×Region.[5] In summary, it appears from Table V that, after normalization with LOBANOV and GERSTMAN, the phonemic and the sociolinguistic variation are preserved best of all four procedures in the dependent variables, while the gender-related anatomical/physiological variation appears to be minimized.

IV. DISCUSSION

The aim of this study was to establish to what extent procedures for vowel normalization are suitable for use in language variation research. We carried out three evaluations using eleven normalization procedures that were applied to Dutch vowel data from talkers who were stratified for the factors region and gender.

The procedures were first evaluated on how well they preserved phonemic variation in the transformed vowel data, second on how well they reduced anatomical/physiological variation, and third on how well they preserved sociolinguistic (regional) variation. Given the results for these comparisons, it can be concluded that procedures for vowel normalization can be useful tools in dealing with (unwanted) anatomical/physiological talker-specific variation in studies investigating regional variation in vowel systems. However, this is only valid for a subset of the procedures evaluated: LOBANOV, or Lobanov's (1971) z-score transformation, NEAREY1, or Nearey's (1978) single logmean procedure, and GERSTMAN, or Gerstman's (1968) range transformation. These three procedures were found to preserve phonemic variation best, reduce anatomical/physiological variation most effectively, while at the same time preserving nearly all sociolinguistic variation in the acoustic measurements. After comparing the three sources of variation (vowel, region, and gender) by multivariate analysis, LOBANOV turned out to be the best procedure, although the difference with NEAREY1 is relatively small.

Although this paper does not aim to develop a theory of how listeners normalize vowels, below we discuss the results from a perceptual perspective. Our finding that the three most successful procedures are all vowel-extrinsic procedures and the least successful procedures are all vowel-intrinsic procedures is surprising, because it has been suggested that intrinsic procedures reflect or resemble processes involved in human speech perception better than extrinsic procedures (Syrdal and Gopal, 1986). Vowel-intrinsic models were considered to be more suitable as models for human vowel perception because they, in analogy with human listeners (e.g., Assmann *et al.*, 1982), can normalize a single vowel from a speaker without information about other vowels from that speaker (Nearey, 1989). Vowel-extrinsic procedures, on the other hand, generally require information across multiple vowels (if not all) per speaker to calculate the scale factors necessary for the normalization. Thus, to normalize one vowel from a speaker, the procedure first has to know all other vowel positions of that speaker. Nevertheless, it should not be overlooked that listeners have had years of exposure to different talkers' voices before being able to categorize vowel tokens effectively. Even if listeners are presented with a new speaker, they may use their experience of hearing other, perhaps similar, voices. Given our results for the three vowel-extrinsic procedures, we hypothesize that LOBANOV, NEAREY1, and GERSTMAN can account for the listeners' experience through the use of scaling factors that model the distribution of other vowels produced by the same talker.

But why did some of the vowel-intrinsic procedures perform so poorly? For instance, Syrdal and Gopal's (1986) S & G performed poorer than raw data in Hz at most tasks evaluated. Overall, the poor performance of this procedure can be attributed to the fact that it did not succeed in clustering the transformed vowel data as effectively as most vowel-extrinsic procedures. However, another explanation may be that it incorporates information across different formants (e.g., $F3$-$F2$) for a given vowel token. The overall results

TABLE VI. Classification of the normalization procedures according to whether they use vowel-intrinsic or vowel-extrinsic information, and whether they use formant-intrinsic or formant-extrinsic information.

Information	Vowel-intrinsic	Vowel-extrinsic
Formant-intrinsic	HZ, LOG, BARK, MEL, ERB	GERSTMAN, LOBANOV, NEAREY1
Formant-extrinsic	S & G	NORDSTRÖM, MILLER, NEAREY2

show that vowel-extrinsic procedures that incorporate information across formants (NEAREY2, NORDSTRÖM, and MILLER) perform poorer than those who include only information within formants (LOBANOV, NEAREY1, and GERSTMAN). This pattern is especially clear for NEAREY1 and NEAREY2, which differ only in that NEAREY2 includes information across formants, while NEAREY1 does not. Summarizing, we find that procedures using information across vowels performed better than procedures using only information within vowels and procedures using information within formants performed better than those using information across formants. Given this pattern in the results, we suggest to expand the traditional intrinsic/extrinsic division of procedures to the formants. This way, formant-intrinsic and formant-extrinsic categories are distinguished as well as vowel-intrinsic and vowel-extrinsic categories. The procedures that were evaluated in the present paper are classified according to this extended division in Table VI.

In conclusion, vowel-extrinsic, formant-intrinsic normalization procedures can be useful and accurate tools for research investigating language variation. Application of these normalization procedures to the measurements of the fundamental frequency. The frequencies of the first three formants produced by different talkers eliminates anatomical/physiological variation. The variation that remains in the data is either phonemic or sociolinguistic in nature. Normalization is especially useful when data from male and female talkers is to be compared, as the successful procedures eliminated all variation related to the talker's gender. An additional benefit for language variation research is that the most successful procedures are also the easiest to implement. Finally, Hindle's (1978) concern, applying normalization procedures may reduce sociolinguistic variation in the acoustic representation along with the anatomical/physiological variation, does not generally hold. Instead, it appears that our results for LOBANOV, NEAREY1, and GERSTMAN confirm results reported in Labov (2001): most sociolinguistic variation was retained in the normalized data.

ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization for Research (NWO) through the Flemish Netherlands Committee (VNC) under project nr. 205-41-069 (PI Roeland van Hout).

[1]We considerer a scale transformation such as a transformation to a bark-scale, or to a mel-scale, to be a normalization procedure as well.

[2]The following procedures were not selected: Wakita (1977), Bladon and Lindblom (1981), Hermansky *et al.* (1985), and Pickering (1986), which were all four evaluated in Deterding (1990), and Harshman (1970), as described in Disner (1980).

[3]Traunmüller (1990) provides a low frequency correction as well as a high frequency correction. We decided not to use either for the following reasons. First, although the low frequency correction ensures that the transformed data resembles the rounded values of Zwicker's (1961) table more closely, Traunmüller (1990) states that the uncorrected form approximates the actual empirical data in Zwicker *et al.* (1957) more closely at low frequencies. Second, the high frequency correction aims to reduce inaccuracies above 20.1 Bark (around 8 kHz), but we were only interested in the frequency regions up to 4 kHz.

[4]One of the appropriate tests available in multivariate analysis of variance, used for reflecting the proportion of the variance in the dependent variable that can be accounted for, given the independent variable(s). See Stevens (1979).

[5]It may suffice to use only $F1$ and $F2$ to describe the data acoustically. To find further evidence for this idea, LDA 1 (cf. Table II) was repeated for the four procedures HZ, LOBANOV, NEAREY1 and GERSTMAN, this time using only (transformed) $F1$ and $F2$ as predictors. The results for HZ showed that 72% of the vowel tokens could be correctly classified when only $F1$ and $F2$ were entered as predictors, as opposed to 79% for LDA 1, for LOBANOV this was 91% as opposed to 92% for LDA 1, for NEAREY1 it was 87% as opposed to 90% for LDA 1, and for GERSTMAN a percentage of 83% was found as opposed to 84% for LDA 1. Overall, the scores for the analysis with $F1$ and $F2$ as predictors are 1–7 percent points lower than the analysis with $F0$, $F1$, $F2$, and $F3$ as predictors. The largest difference (7%) was found for the untransformed data, the percentages for the transformed data decreased only 1%–3%. It thus appears that Dutch may be described relatively effectively using only $F1$ and $F2$, after transformation through LOBANOV, NEAREY1, or GERSTMAN.

Adank, P. (**2003**). "Vowel normalization: a perceptual-acoustic study of Dutch vowels," PhD thesis, University of Nijmegen.

Adank, P., van Hout, R., and Smits, R. (**2004**). "An acoustic description of the vowels of Northern and Southern Standard Dutch," J. Acoust. Soc. Am. **116**, 1729–1738.

Ainsworth, W. A. (**1975**). "Intrinsic and extrinsic factors in vowel judgements," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London).

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustics aspects," J. Acoust. Soc. Am. **71**, 975–989.

Bladon, R. A., and Lindblom, B. (**1981**). "Modeling the judgement of vowel quality differences," J. Acoust. Soc. Am. **69**, 1414–1422.

Boersma, P. (**1993**). "Accurate short-term analysis of fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, **17**, pp. 97–110.

Deterding, D. (**1990**). "Speaker normalization for automatic speech recognition," PhD thesis, University of Cambridge.

Deterding, D. (**1997**). "The formants of monophthong vowels in Standard Southern British English Pronunciation," J. Int. Phon. Assoc. **27**, 47–55.

Disner, S. (**1980**). "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am. **67**, 253–261.

Gerstman, L. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 78–80.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched noise data," Hear. Res. **47**, 103–138.

Hagiwara, R. (**1997**). "Dialect variation and formant frequency: The American English vowels revisited," J. Acoust. Soc. Am. **102**, 655–658.

Harshman, T. (**1970**). "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-model factor analysis," In Working Papers in Phonetics, **16**, Phonetics Lab UCLA.

Hermansky, H., Hanson, B. A., and Wakita, H. (**1985**). "Low-dimensional representation of vowels based on all-pole modeling in the physiological domain," Speech Commun. **10**, 509–512.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acous-

tic analysis of American English vowels,'' J. Acoust. Soc. Am. **97**, 3099–3111.
Hindle, D. (**1978**). ''Approaches to formant normalization in the study of natural speech,'' in *Linguistic Variation, Models and Methods*, edited by D. Sankoff (Academic, New York).
Labov, W. (**2001**). *Principles of Linguistic Change: Vol. II: Social factors* (Blackwell, Oxford).
Ladefoged, P., and Broadbent, D. E. (**1957**). ''Information conveyed by vowels,'' J. Acoust. Soc. Am. **29**, 88–104.
Lobanov, B. M. (**1971**). ''Classification of Russian vowels spoken by different speakers,'' J. Acoust. Soc. Am. **49**, 606–608.
Miller, J. D. (**1989**). ''Auditory-perceptual interpretation of the vowel,'' J. Acoust. Soc. Am. **85**, 2114–2134.
Most, T., Amir, O., and Tobin, Y. (**2000**). ''The Hebrew Vowel System: raw and normalized acoustic Data,'' Lang Speech **43**, 295–308.
Nearey, T. M. (**1978**). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Indiana).
Nearey, T. M. (**1989**). ''Static, dynamic, and relational properties in speech perception,'' J. Acoust. Soc. Am. **85**, 2088–2113.
Nearey, T. M. (**1992**). ''Applications of generalized linear modeling to vowel data,'' in Proceedings of the 1992 International Conference on Spoken Language Processing, 583–587.
Nearey, T. M., Assmann, P., and Hillenbrand, J. (**2002**). ''Evaluation of a strategy for automatic formant tracking,'' J. Acoust. Soc. Am. **112**, 2323.
Nordström, P. E. (**1976**). ''Female and infant vocal tracts simulated from male area functions,'' J. Phonetics **5**, 81–92.
Peterson, G. E., and Barney, H. L. (**1952**). ''Control methods used in the study of the vowels,'' J. Acoust. Soc. Am. **24**, 175–184.
Pickering, J. B. (**1986**). ''Auditory vowel formant variation,'' PhD thesis, Oxford University.
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.
Stevens, J. P. (**1979**). ''Comments on Olson: Choosing a test statistic in multivariate analysis of variance,'' Psychol. Bull. **86**, 355–360.
Stevens, S. S., and Volkmann, J. (**1940**). ''The relation of pitch to frequency: A revised scale,'' Am. J. Psychol. **53**, 329–353.
Syrdal, A. K. (**1984**). ''Aspects of a model for the auditory representation of American English vowels,'' Speech Commun. **4**, 121–135.
Syrdal, A. K., and Gopal, H. S. (**1986**). ''A perceptual model of vowel recognition based on the auditory representation of American English vowels,'' J. Acoust. Soc. Am. **79**, 1086–1100.
Traunmüller, H. (**1990**). ''Analytical expressions for the tonotopic sensory scale,'' J. Acoust. Soc. Am. **88**, 97–100.
Van de Velde, H., van Hout, R., and Gerritsen, M. (**1997**). ''Watching Dutch change,'' J. Sociolinguistics **1**, 361–391.
Watson, C. I., Maclagan, M., and Harrington, J. (**2000**). ''Acoustic evidence for vowel change in New Zealand English,'' Language Variation and Change **12**, 51–68.
Wakita, H. (**1977**). ''Normalization of vowels by vocal tract length and its application to vowel identification,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-25**, 183–192.
Zwicker, E. (**1961**). ''Subdivision of the audible frequency range into critical bands (Frequenzgruppen),'' J. Acoust. Soc. Am. **33**, 248.
Zwicker, E., Flottorp, G., and Stevens, S. S. (**1957**). ''Critical bandwidth in loudness summation,'' J. Acoust. Soc. Am. **29**, 548–557.
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.