How to stretch and shrink vowel systems: Results from a vowel normalization procedure

Christian Geng
Linguistics and English Language, The University of Edinburgh, Midlothian EH8 9AD, United Kingdom

Christine Mooshammer
Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511

(Received 17 May 2007; revised 20 February 2009; accepted 25 February 2009)

One of the goals of phonetic investigations is to find strategies for vowel production independent of speaker-specific vocal-tract anatomies and individual biomechanical properties. In this study techniques for speaker normalization that are derived from Procrustes methods were applied to acoustic and articulatory data. More precisely, data consist of the first two formants and EMMA fleshpoint markers of stressed and unstressed vowels of German from seven speakers in the consonantal context /t/. Main results indicate that (a) for the articulatory data, the normalization can be related to anatomical properties (palate shapes), (b) the recovery of phonemic identity is of comparable quality for acoustic and articulatory data, (c) the procedure outperforms the Lobanov transform in the acoustic domain in terms of phoneme recovery, and (d) this advantage comes at the cost of partly also changing ellipse orientations, which is in accordance with the formulation of the algorithms. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3106130]

PACS number(s): 43.70.Aj, 43.70.Bk, 43.70.Jt [AL] Pages: 3278–3288

I. INTRODUCTION

A. Background

One of the major challenges in experimental phonetics is to overcome the consequences of speaker-specific variability because individual differences obscure the distinction between categories in acoustic and articulatory spaces. Johnson *et al.* (1993), for example, tested the hypothesis that speakers use the same set of articulatory features for the production of the American English vowel system by analyzing tongue contours and jaw movements. They concluded that speakers were very consistent within themselves in the strategies applied for producing different vowels. Between speakers, however, there was a great amount of variability in the way they increased speech tempo, distinguished between tense and lax vowels and also in their overall strategies. This led the authors to the conclusion that the targets of speech production must be specified in terms of the acoustic output. However, as was discussed in Disner (1980), variability in the formant space still reflects speaker-dependent differences due to vocal-tract shapes and sizes, which makes it impossible to compare vowel inventories of different languages by means of formant frequencies taken from natural utterances of human speakers.

The major aim of the current study is to test the usefulness of a normalization procedure heavily inspired by generalized Procrustes analysis (Gower, 1975). Specifically, we applied this normalization procedure to acoustic, articulatory, and anatomic data in order to reveal speaker-independent strategies for the production of German vowels. The motivation for speaker normalization has also been expressed with a stronger conceptual and theoretical bias in the aforementioned paper by Johnson *et al.* (1993). Their discussion of what they term as the "universal articulatory phonetics hypothesis" puts in question the classical tenet that linguistic equivalence classes should be built on the basis of articulatory substance unless it is lawfully possible to relate different speakers' articulatory performance in a systematic way such that the constitution of traditional units such as the phoneme becomes possible. Johnson *et al.* (1993) worked through a whole catalog of influences, which have the potential of putting the universal articulatory phonetics hypothesis in danger: measured fleshpoint, speaking rate, dialect, palate shape, dental occlusion, and articulatory strategy. In fact, in their study, systematic individual differences were retained, making the articulatory definition of equivalence classes a problematic undertaking. There are several ways out of this dilemma: one solution would propose to locate invariance in speech perception (Kingston and Diehl, 1994) or the sensory periphery (Guenther *et al.*, 1998), which assumes that producing phoneme sequences involves the activation of invariant auditory goals. Alternatively, a more practical approach is the use of normalization procedures. Normalization procedures have been successfully applied to articulatory data, e.g., in Harshman *et al.* (1977), Hashi and Westbury (1998), and Beckman *et al.* (1995) or to acoustical data [for reviews see Disner (1980) and Johnson (2005)]. The goals of such procedures can be to abstract from speaker-specific vowel locations and variation patterns in order to compare linguistically different vowel systems in terms of their formant (e.g., Clopper *et al.* 2005) or articulatory spaces (e.g., Jackson, 1988). Within a given language, vowel normalization can be applied to tongue configurations in order to find speaker-independent strategies for speech tempo variation (Hoole, 1999) and for the lexical stress distinction (Geng and Mooshammer, 2000) or to test against predictions made by quantal theory (Beckman *et al.*, 1995).

Traditionally, acoustical vowel normalization methods

 0001-4966/2009/125(5)/3278/11/$25.00

have been divided into "intrinsic" and "extrinsic" methods for vowel normalization, a division which dates back to Joos (1948). Adank *et al.* (2004) argued though that this division might not be fine-grained enough to describe all essential differences between procedures. They extended the usage of the extrinsic vs intrinsic dichotomy to both (a) vowels and (b) formants. This work does not consider vowel intrinsic normalization because it is not clear how it could be related to parallel articulatory data. The same holds for all vowel-extrinsic/formant-extrinsic schemes known to us. The only class containing published normalizations, which correlate well with the aims of the current work, are formant-intrinsic, vowel-extrinsic normalizations. According to Adank *et al.* (2004), this kind of normalization has been the most efficient. Within this class, the Lobanov (1971) transform has been described as one of the most efficient procedures. Still, this kind of normalization is not without drawbacks: As pointed out by Nearey (1989), despite its success for practical reasons, i.e., the most substantial scatter reduction, its cognitive plausibility is questionable because the listener would have to know the formant frequencies of the complete vowel system spoken by this particular speaker in order to recognize a single vowel. This points to the discussion about the psychological reality of entities such as mean formant frequencies and scale factors. In contrast to Nearey, Adank *et al.* (2004, p. 3105) viewed such factors as possibly accounting for listeners' life-long experiences with listening to different types of speakers. Another criticism was noted by Apostol *et al.* (2004), according to which statistical methods cannot be directly related to anatomical differences between speakers, which ultimately underlie the large variation in the formant spaces. The basic hypothesis of their model is that inter-speaker variability in formant spaces "arises from differences among speakers in the respective lengths of their back and front vocal-tract cavities" (p. 337) because of the formant-cavity affiliation.

In a similar vein, namely, to relate articulatory positions to formant frequencies, a number of geometrical normalization procedures has been applied to tongue contours derived during vowel production. These procedures are based on re-expressing fleshpoints on the tongue—usually acquired by the x-ray microbeam system (Westbury, 1994)—as distances to the palate (e.g., Beckman *et al.*, 1995; Hashi and Westbury, 1998; Perkell and Nelson, 1985). Therefore, the pellet positions are translated into a palate-based coordinate system in order to minimize the effects of differences in vocal-tract size and shape on mean articulatory postures. The advantage is that the new palate-based coordinate system is more closely related to the oral part of the area function than the original coordinate space and can therefore more easily be related to spectral properties. Within this framework, as well as for the cavity-affiliation model by Apostol *et al.* (2004), the major aim is to explain and reduce the speaker-dependent acoustical variability by anatomical differences, e.g., due to gender (see Simpson, 2002).

A very different approach is the factor analytic treatment of vowel production in which the underlying control mechanisms from the highly correlated coordinates of the tongue-pellets are extracted. In their work, Harshman *et al.* (1977) subjected multi-speaker x-ray tongue contours to the PARAFAC algorithm, yielding two factors consisting of three matrices: (a) the speaker weights, (b) a speaker-independent vowel space, and (c) the so-called articulator weights. A consistent and robust interpretation of the application of PARAFAC arises from these first two factors extracted from flesh-point data as well as from contours and for a number of different languages (see, e.g., Harshman *et al.*, 1977; Hoole, 1999; Jackson, 1988). The first factor, usually dubbed front raising, distinguishes low vowels from high front vowels. In articulatory terms, front raising is a forward movement of the root of the tongue and an upward movement of the front of the tongue. The second factor, back raising, is associated with the formation of back vowels and characterized by an upward and backward movement of the tongue. The current study will make an attempt to evaluate a technique conforming to that of Harshman *et al.* in its applicability to both acoustic and articulatory data but differing in its computational procedure as well as its general orientation. The normalization procedure applied here is based on a method frequently applied in morphometrics/zoology in order to solve the problem of the superimposition of geometrical landmarks. In zoology, this is often helpful for shape comparison between species abstracting from uninformative scaling, translation, and rotations. These methods are often termed "Procrustes" methods (see Gower, 1975, and with a special background in morphometrics Rohlf and Slice, 1990 or Goodall and Green, 1986). Due to its central importance for the current work, this approach will be described in greater detail in the next section.

B. The normalization procedure: A modified generalized Procrustes analysis

These superimposition techniques can be distinguished according to the following aspects: (a) The number of objects to be aligned: Procrustes analysis has originally been designed for the alignment of only two specimens but has later been extended to handle any number of objects in Gower (1975), (b) the nature of the transformation terms to be applied: If only rigid rotations are allowed, orthogonal methods—i.e. Procrustes methods in a more narrow sense—are used, which preserve the angles between data points. If uniform affine deformation is to be applied, the class of methods are called oblique and (c) the optimization strategy applied: If no local shape change is allowed, least-squares fitting methods are appropriate; if one wishes to account for local shape change, more advanced, non-parametric methods based on the median have to be applied. The acoustic and articulatory data corpora in our study contain more than two speakers, such that our algorithm a priori has to refer to the "generalized" case. With respect to the nature of the rotation terms to be applied, a distinction has to be made between articulatory and acoustic data sets. While for the articulatory data, a considerable amount of affine deformation would already be expected due to different vocal-tract morphologies, this is not so clear for the acoustic data set. Goodall and Green (1986) devised a method for checking the amount of

affine deformation necessary to superimpose landmarks in the two-dimensional (2D) case. In cases in which the amount of affine deformation necessary for superimposition is not substantial, superimposition applying orthogonal transformations yields similar results. Concerning the optimization scheme, we used the least-squares technique and not the nonparametric method described in Rohlf and Slice (1990) because we were aiming at uniform rather than local shape change. The approach used here comprises two separate steps: First, across different speakers, a "consensus configuration" representing an average subject is calculated, and, in a second step, this consensus is fitted to the data of individual speakers to result in normalized data.

1. Construction of the consensus object

The consensus configuration is calculated as follows [equivalent to the formulation in Rohlf and Slice (1990)]: First, the data of the n individual speakers are centered and scaled with a standard z-transform. Then, a first version of the consensus object is calculated using

$$A = \frac{1}{n}\sum_{i}^{n} X_i (X_i^T X_i)^{-1} X_i^T, \tag{1}$$

where the X_i are the individual speakers' data matrices after centering and scaling, n is the number of speakers, and superscript T denotes the transpose of a matrix. This version of the consensus object does not resemble the original objects though. In the bivariate case—like in the analysis of formant spaces—the $X(X^TX)^{-1}X^T$ operation transforms each object "so that the variance in the bivariate distribution of landmarks is the same in all directions in the plane for each object" (Rohlf and Slice, 1990, p. 49). The final consensus configuration is calculated as

$$C = A\left(\frac{1}{n}\sum_{i} X_i X_i^T\right)A. \tag{2}$$

Thereafter, C is subjected to a singular-value decomposition, and the final consensus configuration is a matrix of eigenvectors of C subjected to truncation; e.g., for planar configurations, the first two columns are taken (Rohlf and Slice, 1990, p. 49).

2. Calculation of normalized data

The second step consists of calculating reconstructed data for each subject's configuration. These are calculated by post-multiplying the consensus object with a transformation matrix, which in general is calculated as

$$H^* = (X_2'^T X_2')^{-1} X_2'^T X_1' \tag{3}$$

for two objects in the oblique case. X_2 here is the consensus configuration as calculated above, and X_1 is an arbitrary speaker's original configuration. This is equivalent to the equation for the least-squares estimates of partial regression coefficients in multivariate multiple regression. In the case of orthogonal rotation, the rotation matrices are calculated by performing a singular-value decomposition of the product of the object matrices to be superimposed:

$$H = VSU^T, \tag{4}$$

where U and V are such that $X_1'^T X_2' = U\Sigma V^T$ and Σ is a diagonal matrix. S is a diagonal matrix with $s_{ii} = \pm 1$, and the signs of the s_{ii} are taken from the corresponding elements of Σ. Summing up, the main outputs of the procedure are (a) a consensus object, which in our case is the configuration of an average speaker characterized by the statistical properties as described above, (b) the eigenvalues of the transformation matrices, i.e., the diagonal of Σ, as the amounts of uniform deformation in the directions guaranteeing optimal "superimposition" in terms of the least-squares criterion applied, and (c) the normalized data of each single speaker. The data sets described in the following section will be evaluated with respect to these statistics. The results of (b) are often displayed as "a pair of orthogonal axes with lengths proportional to the two eigenvalues and oriented so that the longer axis is parallel to the direction (…) of maximum stretching" (Rohlf and Slice, 1990, p. 48). In the present case, this makes sense for the acoustic data set: A so-called strain cross, defined by the first two eigenvalues and the angle ψ, of a given speaker deviating to a great degree from the unit circle, indicates that this speaker's vowel space needs a higher amount of affine transformation in order to fit to the consensus object as compared to a speaker with a smaller strain cross.

C. Aims of this study

As was pointed out above, the major aim of this study is to evaluate the usefulness of the algorithm we proposed for both acoustic and articulatory data. The rationale of speaker normalization is chosen with regard to the methodological corollary of the universal articulatory phonetics hypothesis, according to which it is not sufficient to report patterns of individual differences among a homogeneous group of speakers, but "we must also consider ways in which such variability is lawful, because this variability must be made to square with the fact that language is a shared system" (Johnson *et al.*, 1993, p. 702). Therefore, apart from demonstrating the success of the normalization suggested, the second aim of the current work is to explore whether the modes of affine deformation are correlated to aspects of vocal-tract morphology.

II. METHOD

A. Data acquisition

Seven native speakers of German (five males, M1–M5, and two females, W1 and W2) were recorded by means of electromagnetic midsagittal articulography (EMMA, AG 100, Carstens Medizinelektronik). All speakers spoke a standard variety of German with at best slight dialectal variations: three speakers (W1, M1, and M4) originally come from South Germany, one speaker (W2) from Saxonia, two speakers (M3 and M5) from Northeast Germany, and one speaker (M2) from Berlin. At the time of recording, the speakers were between 25 and 40 years old and had lived in

Berlin for at least 5 years. The speech material consisted of words containing /tVt/ syllables with the full vowels /iː,ɪ,yː,ʏ,eː,ɛ,ɛː,øː,œ,aː,a,oː,ɔ,uː,ʊ/ in stressed and unstressed positions. Stress alternations were fixed by morphologically conditioned word stress and contrastive stress. Each symmetrical CVC sequence was embedded in the carrier phrase *Ich habe* /'tVtɐ/, *nicht* /tV'taːl/ *gesagt*. (*I said* /'tVtɐ/, *not* /tV'taːl/) with the test syllable /tVt/ in the first word always stressed and in the second word always unstressed. All 15 sentences were repeated six (four speakers) or ten times (three speakers). Four sensors were attached to the tongue, one to the lower incisors, and one to the lower lip. The analyses in this study are limited to the four sensors on the tongue for the remainder of this text and are numbered T1–T4 going from front to back. Two sensors on the nasion and the upper incisors served as reference coils to compensate for head movements relative to the helmet and for the definition of a preliminary coordinate system. This served as the basis for the final reference coordinate system, which was defined by recordings of two sensors on a T-bar, manufactured individually for each subject in order to determine his or her bite plane. Simultaneously, the speech signal was recorded by a digital audio tape recorder. Original sampling frequencies were 400 Hz for the EMMA data and 48 kHz for the acoustical signals. For the analyses, the EMMA signals were low-pass filtered at 30 Hz and downsampled to 200 Hz, while the acoustical signals were downsampled to 16 kHz.

B. Measurements

Formant frequencies of the first and second formant were measured interactively close to the mid of the vowel at the moment of minimal formant movement or, for lax vowels, at a turning point in the F2 trajectory. For estimating the frequencies of the first and second formants, the default settings of the software package SIGNALYZE (⟨http://www.signalyze.com/⟩) were used, i.e., LPC with 15 ms smoothing. The same temporal markers as described for the acoustic analyzes were also used for extracting tongue positions. Both acoustic and articulatory data were then averaged over the six respectively ten repetitions of each vowel. Information about palate shapes was acquired by measuring the artificial EPG palates of all seven speakers by means of a sliding caliper. This procedure gave the 3D coordinates for all EPG electrodes and the 2D coordinates of the palate midline approximately located between the two most central columns. Since the location of the EPG electrodes is adjusted to the speakers' anatomy, e.g., the rear border is aligned with the rear wall of the second molars, the EPG-based palate midline was deemed to be more exact compared to the palate outline, traced by means of EMMA. For addressing the question whether speaker-dependent differences can be explained by their palate shape, two measures were used: the palate length and a doming index (see Johnson *et al.*, 1993), which was calculated as the ratio between the total midsagittal length of individual EPG palates and the vertical distance between the first and the last point on the palate. Higher values indicate a palate with a higher degree of doming. Three sets of data were subjected to the analysis: tongue configurations during the 15 vowels in two stress conditions, measured as X and Y coordinates of the four sensors (30×8 matrix), frequencies of the first and second formant during the vowels (30×2 matrix), and palate outlines specified by 11 x and y coordinates (11×2 matrix). For measuring the palatal outline, we adapted the method described in Fitzpatrick and Ni Chasaide (2002). In short the 3D coordinates of the Reading EPG palates were measured as described above. From this we estimated the midsagittal outline from the positions of the two inner columns of the electrode positions. The outline was then adjusted to the EMMA data by eye.

C. Statistical apparatus

1. Quantification of the relationship pools of quantitative variables

Throughout the current work, methods are used, which require the correlation of data sets with independent as well as independent variables containing more than two variables. The correlations between articulatory configurations and the morphological data set could serve as an example. In this example, the research interest lies in gaining insight about the nature of the main directions of isotropic shape change in these two data modalities, substantial correlations indicating that similar variance components are targeted in both data sets. In order to explore relationships as these, canonical correlation analysis (CCA) was applied. CCA can be seen as the multivariate generalization of product-moment correlation. Considering the two matrices X and Y, the CCA finds a linear combination of the variables of X and a linear combination of the variables of Y of maximal correlation. It has often been claimed that CCA needs many cases compared to the number of variables. But Stevens (1986) discussed sample size in CCA and states that if the canonical correlations are strong (i.e., $R > 0.7$), then even small samples (e.g., $n = 50$) can be sufficient to detect significant correlations most of the time. Another drawback often reported for canonical correlations is that it is reported to be sensitive to multicollinearity among variables. This issue can at least to some extent be settled by reporting redundancy indices, which provide a partitioning of the explained variance into predictors and criteria. The redundancies of individual canonical variates can be summed up to yield an R^2-like measure of the contribution of the predictor-side canonical variates in explaining the criterion-side canonical variates and vice versa.

2. Classification procedures

By means of statistical discrimination, it is possible to check how successful different normalization procedures are at preserving the phonemic identity of vowel tokens. Adank *et al.* (2004) focused on two such classification approaches of vowel tokens. The most basic model, linear discriminant analysis (LDA) assumes that the covariance of each of the classes is identical. A more advanced method, quadratic discriminant analysis (QDA), makes no such assumptions but has the drawback of estimating more parameters, making it more difficult to crossvalidate and more prone to overfitting. These two approaches are just two instances of classification

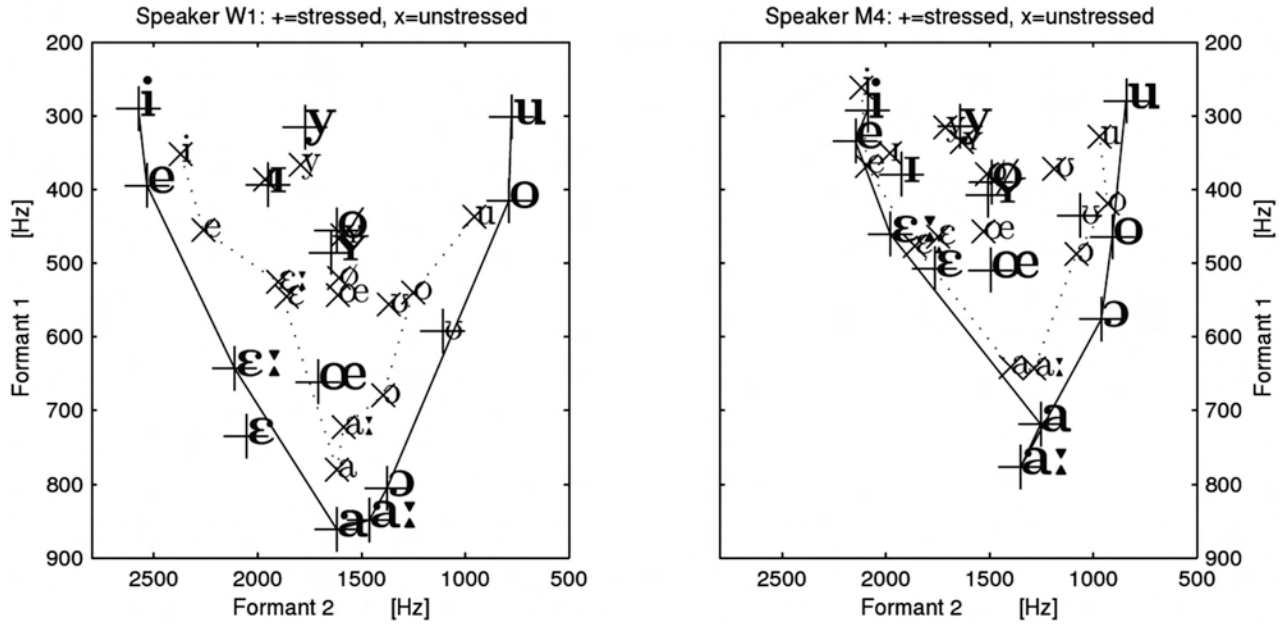

FIG. 1. Means of formant frequencies for stressed (bold and Large symbols with "+") and unstressed vowels [light and smaller symbols with "×") for a female (left) and a male speaker (right)].

procedures, with the LDA being one standard choice.[1] Having said that, it is obvious that it is necessary to be selective with regard to the classification procedures chosen. We compared two classification procedures, including (i) LDA—for comparability purposes with Adank *et al.* (2004) and (ii) an additional logistic discriminator (LOGDA) (Ripley, 1994). This logistic classifier has the advantage of making less assumptions than both QDA as well as LDA.

III. RESULTS

This section is concerned with the evaluation of the normalization procedure applied in this study. It will roughly be organized in analyses of the two modalities: acoustic and articulatory. The acoustic description of the data set involves comparing raw and normalized F1/F2 spaces. The capability of the approach to preserve phonemic identity, in particular with respect to the stress condition manipulated in the data set comprises both acoustic and articulatory modalities. Of special interest for the articulatory data is the question of how the deformations of the articulatory spaces are to be related to measurements of anatomical characteristics of the palate, as captured in terms of (a) results of the normalization procedure of palate outlines and (b) scalar doming indices (Johnson *et al.*, 1993). This also justifies the parsimonious selection of acoustic features: As we were primarily interested in vowel quality as manifested in tongue shapes, we have limited ourselves to the analysis of the first two formants in the acoustic domain. In other words, we were not interested in acoustic features such as F0 as it has stronger prosodic correlations—at least in our corpus—with stressed vowels carrying pitch accents. Similar arguments hold against F3, which mostly is associated with rounding information.

A. Formant spaces

In order to give a general impression of speaker-dependent differences for the production of German vowels, formant spaces of two speakers from Bavaria, one female (W1) and one male (M4), are presented in Fig. 1. The means of the frequencies of the first and second formant are indicated as bold symbols with a "+" sign for the stressed vowels. Unstressed vowels are indicated by "×." For reasons of clarity, the marginal vowels are connected by lines.

As can be seen, the speakers differ not so much in the relative location of vowels in the vowel system, with the exception of some minor relative changes in the location of the low vowels. However, the two speakers do differ in the way they realized the stressed-unstressed distinction: On the one hand, speaker W1 reduces the unstressed vowels consistently toward the center of the formant space. On the other hand, speaker M4's formant frequencies of unstressed vowels differ to a greater degree for the back and low vowels from their stressed counterparts, whereas the front vowels are only slightly affected. The direction of change for the back and low vowels suggests a more fronted and closer constriction for the unstressed vowels. Therefore, the acoustical results suggest that this speaker produces the stressed vowels with a greater contrast between the neighboring consonants and the vowels.

The consensus object, shown on the left in Fig. 2, gives more evidence for the latter strategy for reducing vowels; i.e., the back and the lower vowels are centralized when unstressed but not the high and mid palatal vowels /i, y, e, ø, ɪ, ʏ/, which change only very little. This finding correlates well with the observation that their constriction location is already quite close to the constriction location of the neighboring apical stops. The formant values of the remaining unstressed vowels change in the direction of front high vowels.

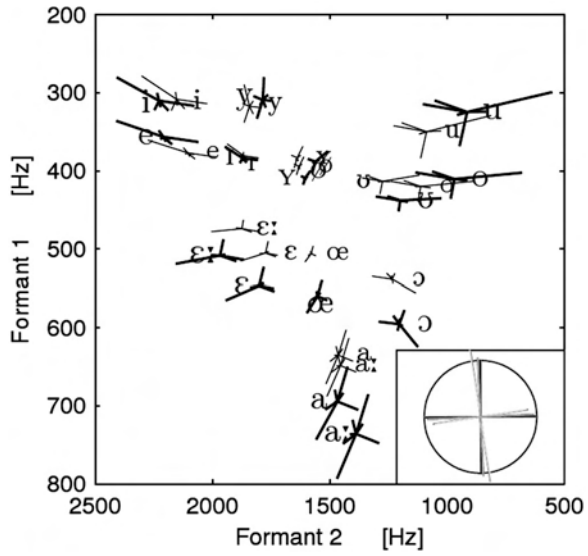

FIG. 2. Formant consensus object of stressed vowels (large symbols) and unstressed vowels (small symbols). Lines correspond to the distance between the speaker-dependent models for each of the speakers and the formant consensus object. Inset: amount of necessary speaker-dependent affine transformation compared to the consensus object, all speakers, displayed as strain cross.

As was explained in Sec. I B 2, strain crosses display the amount of necessary affine transformation of individual speakers in order to fit to the consensus object. The more the crosses deviate from the cross in the unit circle in length and orientation, the more affine transformation was necessary for the corresponding speaker. Only for two speakers (M1 and W2) do the strain crosses deviate clearly from the unit circle. This implies that not much affine deformation is applied to single speakers' vowel spaces in order to achieve the best fit.

1. Cross-method comparisons

One way of evaluating the quality of a normalization procedure is to measure the amount of scatter remaining after normalization, which has been considered good practice as a tool for the evaluation of speaker normalizations since Disner (1980). Scatter remaining is computed as the percentage of ellipse areas of transformed data relative to the ellipse areas of the raw data. Table I summarized the results for our method and the Lobanov transform. In terms of scatter reduction, our method outperforms the Lobanov transform, which was confirmed by the results of a t-test ($t=3.67$, $p<0.01$, df=29). Another pattern which becomes evident from an inspection of Table I is that not all vowels are normalized to the same degree, an effect which appears to be common to both procedures applied. While there is a considerable amount of scatter reduction for most vowels, there are instances where there is even more variability after normalization. While technically possible, such a result runs counter the objective of normalization. This pattern is particularly prominent for stressed /uː/. Further, from visual inspection of the dispersion ellipses, it seemed evident that our procedure affects the orientations of the ellipses to a larger degree than the Lobanov transform, in particular for back vowels. This was evaluated by calculating the ellipse orientations of the raw data and the normalizations and then by bootstrapping the correlation coefficients of these angles. While the ellipse orientations of the Lobanov correlated well with the orientations of the original data ($r=0.94^{**}$, 95% CI [0.86, 0.97]), this was not the case for the Procrustes-influenced approach ($r=0.29$, $p=0.11$, 95% CI [−0.08, 0.58]). The high CI range (0.67) for our method further evidences that only a subset of ellipse orientations is affected (see also Fig. 3).

TABLE I. Scatter remaining after normalization in percent of the original data.

	Model		Lobanov	
	Stressed	Unstressed	Stressed	Unstressed
i	34.63	29.56	61.73	26.92
ɪ	11.20	5.60	74.06	68.58
y	26.09	19.03	182.59	75.24
ʏ	1.47	3.43	17.88	44.16
e	19.92	8.98	23.91	12.58
ɛː	4.34	3.16	23.80	60.34
ɛ	3.00	1.67	33.41	32.97
ø	4.77	1.61	29.25	21.23
œ	5.17	1.97	48.68	59.73
aː	75.77	27.88	92.50	66.84
a	24.98	21.42	48.43	10.53
o	43.46	10.38	209.51	42.11
ɔ	19.36	11.94	30.73	64.01
u	151.72	40.43	206.62	37.73
ʊ	6.09	5.05	39.35	75.18
$\bar{x}$	28.80	12.81	74.83	46.54
sd	38.06	11.73	65.54	21.56

Apart from scatter reduction, normalization procedures

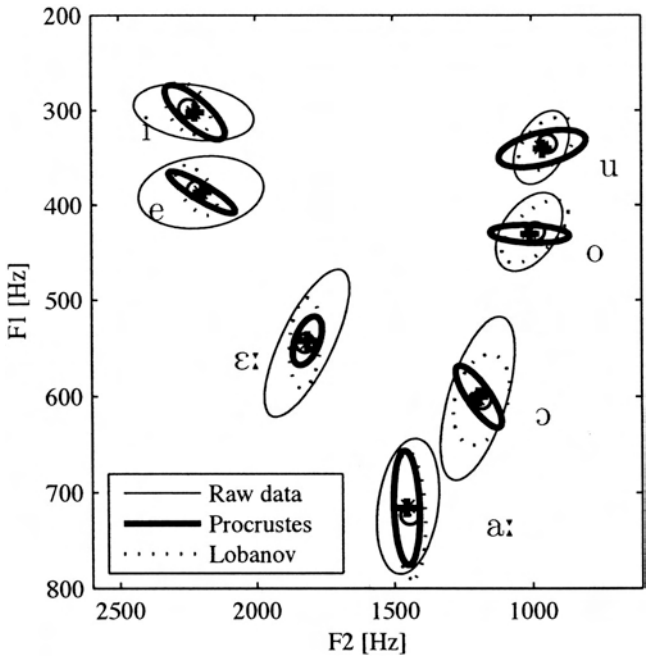

FIG. 3. 1−σ dispersion ellipses for raw formant data (thin lines), the speaker-dependent Procrustes data (bold lines), and the Lobanov transformed data (dashed line). Centroids are indicated by ○ (Lobanov), + (Procrustes), and × (raw data), respectively. For reasons of clarity, only the cardinal vowels are plotted here.

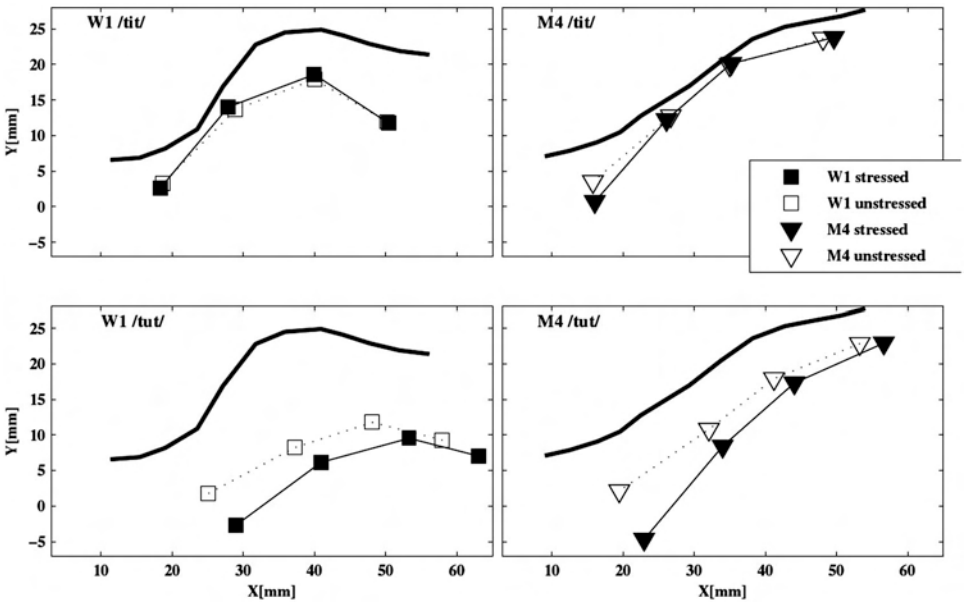

FIG. 4. Comparison between averaged articulatory configurations of the two speakers W1 (thin lines in left panels) and M4 (right panels). The upper two panels show tongue configurations for /tit/ (left) and the lower panels for /tut/. The upper curved lines show the outlines derived from EPG palates. Tongue configurations marked with filled markers were measured at the midpoints of stressed vowels and the empty markers for the corresponding unstressed vowels.

should also maintain the relative positions of items, e.g., in order to compare different languages or dialects. In order to ensure that this is not the case, we conducted a MANOVA to compare the formant means of the raw data, the Lobanov transformed data and the modeled Procrustes data. The differences were insignificant [Wilk's lambda $F(4,34)=0.00006$, $p=1$].

B. Articulatory spaces

In the second part of this section, lingual configurations for the 15 vowels of German in stressed and unstressed positions are used as input to the normalization procedure. In order to exemplify the consequences of speaker-dependent morphology and inconsistent sensor placements, we turn to a description of speaker-dependent strategies for the stress distinction. Figure 4 shows tongue configurations (X-and Y-coordinates) for the four tongue sensors *before* normalization. The upper two panels of this figure show the tongue configurations during the vowel /i/ in stressed (filled symbols) and unstressed positions (unfilled symbols). On the left side, data of the female speaker W1 are displayed together with her palate outline, and on the right side, those for the male speaker M4. Typically, for /i/ the tongue is braced against the sides of the palate. In the lower two panels of Fig. 4, the tongue configurations are presented for the back vowel /u/. The speakers W1 and M4 not only differ to a great degree in the way they produce the vowel /i/ but also in their palate shapes. Speaker W1 has an extremely steep and domed palate, whereas the palate of speaker M4 is rather flat. Accordingly, the tongue configuration for /i/ of speaker W1 is bunched toward the palate and speaker M4's is shaped straighter and oriented in parallel with the palate outline. Despite these immediately obvious differences between these two speakers, both show little differences between stressed and unstressed tongue configurations for /i/. For /u/ the tongue shapes again show a pronounced difference between the speakers. In contrast to /i/, however, stress affects the tongue configurations for both speakers, with the unstressed /u/ being produced with an elevated tongue tip and a generally more fronted tongue body. Based on these observations, it seems reasonable to conjecture (i) that interindividual differences in anatomy can to a large extent be made responsible for the patterns just described, (ii) that the speakers' vowel gestures are still functionally equivalent, i.e., forming a close palatal constriction with the front part of the tongue for /i/ and a uvular constriction with the rear part of the tongue for /u/, and (iii) that speakers apply a general strategy for producing the stress distinction.[2]

In order to test these assumptions, the normalization procedure was separately applied to both the articulatory data and the palate outlines. Speaker-dependent data were calculated according to Eq. (3). The results are shown in Fig. 5. Plotted in bold lines are the X-and Y-coordinates of palate and tongue configurations of the consensus objects, the thinner lines representing the speaker-dependent normalized data of speaker W1 (left) and speaker M4 (right). These two speakers are extreme in their configurations; the remaining speakers compromise between these two participants and therefore are closer to the consensus object. These speakers were displayed to illustrate the problem to be tackled in the articulatory domain. As can be seen in the upper two panels, speaker-dependent tongue configurations for the high front vowel /i/ deviate very little from the consensus object. This is mainly due to the fact that the applied transformations rigorously twist the tongue configuration of speaker M4 in order to fit in the consensus object. For the back vowel /u/ there is more speaker-dependent deviation from the consensus object. Concerning the stressed-unstressed distinction, tongue configurations of /i/ differ only very slightly with a somewhat lower tongue tip position for stressed /i/. For /u/, however, the speaker-dependent and the consensus configurations of stressed /u/ are clearly more retracted, and the tongue tip points downward. Even though the speakers vary in the extent of the difference, the direction is similar for the two speakers and the consensus object.

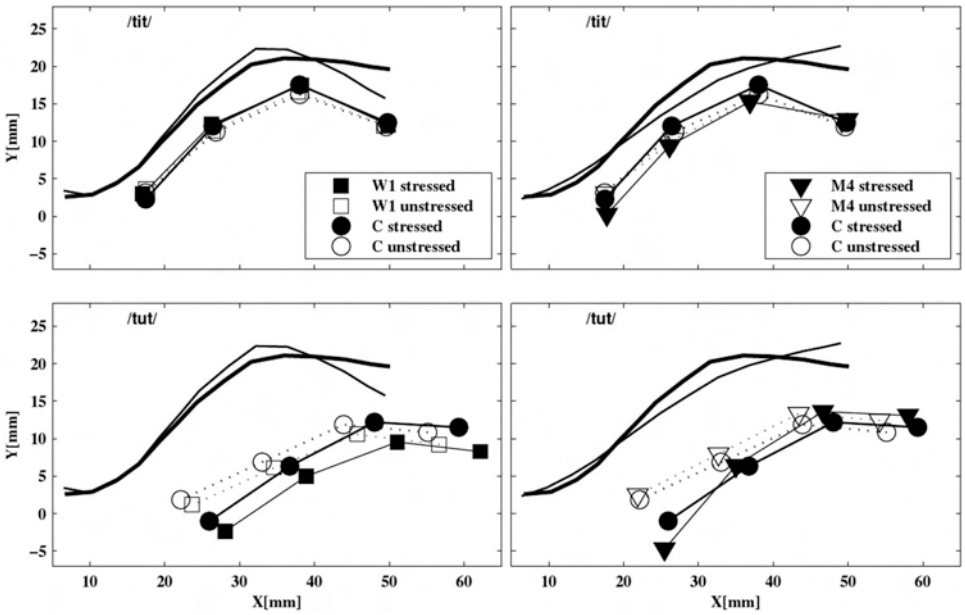

FIG. 5. Speaker-dependent modeled articulatory configurations (tongue contours with filled square symbols) and the consensus objects (contours with circles) for /tit/ (upper panels) and /tut/ (lower panels). Palate contours of the consensus object are printed as thin lines, the contours of speaker-dependent modeled data as bold lines. Modeled data of speaker W1 are presented on the left side and that of speaker M4 on the right side.

C. Relationship between tract morphology and speaker-dependent modeled data

By means of CCA, we aimed at summarizing the relations between the eigenvalues—which measure the amount of affine deformation—in the tongue and the palate data sets as analyzed by the normalization. Recall that the eigenvalues of the transformation matrix relating the consensus object to the individual articulatory spaces contain the amount of uniform affine deformation relating the individual speaker to the consensus. In the morphological data set, these directions have visually interpretable meanings in 2D-Euclidean space, which is not the case for the articulatory vowel spaces. Therefore, we truncated the matrix of eigenvalues in the articulatory data set by visual inspection of the scree plot. Its inspection suggested the use of four eigenvalues, and these were entered in the canonical correlation as predictors (7 speakers $\times 4$ matrix) and both eigenvalues of the palate outline analysis served as criteria (7×2 matrix). Given the low explanatory power of the data set, the results of the following analyses are to be considered as exploratory descriptions. Further note that the division in predictors and criteria is meaningless in CCA; the number of extracted canonical correlations is equal to the minimum number of variables in either set. The expected behavior of the analysis is as follows: Apart from finding substantial canonical correlations between predictor and criterion variables, we expected higher redundancies for the morphological variates, given the tongue configurations. The canonical correlations in this analysis were substantial with values of 0.94 and 0.78, and the summed redundancies over both covariates amounted to 71% of the variance on the criterion side (palate outline data) but only to 52% on the predictor side. In other words, the correlation appears to be substantial and, more importantly, the normalization procedure captures similar directions in the tongue and palate data. Table II summarizes these results.

As a next step, selected intermediate results as obtained from the analysis of tongue shapes were related to the palate doming variable. Here, standard multiple regression was used in order to roughly describe the correlational structure between eigenvalue predictors as derived from palate outline and tongue configuration data sets. To our surprise, the multiple regression for the palate outline eigenvalues on the doming index was not substantial with only 11% of explained variance. In contrast, the regression of the first four eigenvalues of the tongue analysis was substantial, explaining 95% of the variance. Furthermore, it was again (as in the canonical analysis described) necessary to include more than two predictors to capture the anatomical variance. In summary—given the exploratory character of these analyses—a substantial proportion of the variance captured by the normalization procedure when applied to articulatory configurations seems to be shared with what is extracted from the palate outlines.

D. Prediction of phonemic identity

A further check of the success of normalization procedures is to make attempts at measuring the increment of predictability of phonemic identity caused by normalization procedures [see Adank *et al.* (2004) for a more detailed adoption of such a rationale]. This is possible by means of statistical

TABLE II. Summary of analyses relating tract morphology to tongue analyzes.

	CCA	
Can correlations		
	CC I	0.94
	CC II	0.78
Redundancies		
	R^2_{yx}	0.71
	R^2_{xy}	0.52
	Multiple regressions on doming index	
Predictor		
	Tongue eigenvalues	95%
	Palate eigenvalues	11%

TABLE III. Percentages of correctly classified vowels by LDA and LOGDA. The predictors were first and second formant values (acoustic data set) or EMMA coil positions of the four tongue sensors (articulatory data set). Vowel quality with 30 levels (15 German monophthongs is stressed and unstressed positions) served as dependent variables for both data sets. Percentages are given for the whole data sets and separately for stressed and unstressed subsets. In brackets: benefit from normalization (in %).

Data set	Method		Raw	Normalized	Lobanov
Acoustic					
	LDA	Whole	47	80(33)	63(16)
		Stressed	54	82(26)	67(13)
		Unstressed	39	77(38)	60(21)
	LOGDA	Whole	51	83(32)	68(17)
		Stressed	57	86(29)	73(16)
		Unstressed	44	80(36)	63(19)
Articulatory					
	LDA	Whole	41	74(33)	···
		Stressed	44	70(25)	···
		Unstressed	38	77(39)	···
	LOGDA	Whole	48	87(39)	···
		Stressed	59	90(31)	···
		Unstressed	37	83(46)	···

discrimination. The predictors used in the current context were the first and second formant frequency values in the acoustic data set and EMMA coil positions for four tongue sensors in the articulatory data set. The data were pooled over speaker and vowel identity in stressed and unstressed positions, yielding 210 (7 speakers $\times$ 15 vowels $\times$ 2 stress conditions) cases altogether. Vowel identity (15 German monophthongs in stressed and unstressed positions) was the dependent variable. In order to compare the normalization we used with an alternative normalization scheme, the acoustic data set was also Lobanov transformed and subjected to the same discrimination analysis. The calculation of an analog to the Lobanov-normalization in the articulatory case was not considered as meaningful and therefore not undertaken. Results were evaluated in terms of percentage correctly classified tokens, which can be calculated from confusion matrices between vowel quality as intended by the speaker and vowel quality as predicted by the classification procedures. These percentages were calculated both on the whole data set and separately for vowels in stressed and unstressed conditions in order to reveal potentially different effects of the normalization procedures on stressed and unstressed tokens. Classification results were compared by means of McNemar χ^2-tests. In a first step, LOGDA was compared with LDA. LOGDA had the tendency to perform better than LDA, although this effect depended on whether the classification procedures were applied to the raw data or to one of the normalized data sets. For example, LOGDA achieved significantly higher amounts of correctly classified items than LDA for the Lobanov data ($\chi^2=2.7$, $p=0.049$), but this effect did not reach the level of significance neither for the procedure proposed here ($\chi^2=1.33$, $p=0.12$) nor for the raw data ($\chi^2=2.23$, $p=0.067$). In the following we report the results of our analyses with both classification procedures. In contrast to the comparison of classification methods, the comparison of normalization procedures yielded the by far more substantial results (summarized in Table III): Regardless of the classification method used, the normalization procedure proposed performed substantially better than the Lobanov procedure (LDA: $\chi^2=12.38$, $p=0.0002$, LOGDA: $\chi^2=11.69$, $p=0.0003$). This result is in accordance with the amount of scatter reduction reported above.

IV. SUMMARY AND DISCUSSION

In this paper, we have described and applied a normalization procedure applicable to articulatory and acoustic vowel spaces. The procedure consists of constructing a so-called consensus object with the property that the normalized acoustic or articulatory spaces have equal variances in the main directions of affine deformation and performing multiple multivariate regression analysis of this consensus object on the raw configurations to yield speaker-specific normalized data. We delivered a qualitative description of raw and normalized configurations as the first empirical step. For the transformation from the consensus to the speaker-specific formant spaces, only little affine transformation appeared to be necessary (see Fig. 1), which is equivalent to the orthogonal and affine versions of the algorithm yielding very similar results. Still, a higher degree of scatter reduction of speaker-specific variation was achieved by our procedure as compared to Lobanov speaker normalization (see Table I). This benefit in terms of scatter reduction comes at the cost of partially more aggressively transforming the orientation of the data for some vowel categories. For the articulatory data, substantial affine transformation was necessary in order to map the speaker-specific data onto the consensus object. As a result, affine transformations yielded tongue configurations for which a large part of speaker-specific pellet placement and shape differences were removed (see Fig. 5).

We also tested the assumption that the normalization procedure for the articulatory spaces captures directions in the data which correspond to primary dimensions of uniform shape change. For this purpose we applied the normalization

procedure to a control data set of anatomical characteristics, i.e., palate outlines. The substantial correlations obtained suggest that the procedure's success is at least partly related to the removal of uniform shape change differences between individual speakers.

Finally, in order to measure the procedure's capability to recover phonemic identity, we performed separate discrimination analyzes of original and normalized vowel spaces in both articulatory and acoustic domains on phonemic identity. Phoneme recovery probabilities for both articulatory and acoustic data increased substantially in terms of percentages of correctly classified tokens (see Table III). In particular, the classification rates for the unnormalized acoustic data are relatively low in comparison to results published in the literature. For example, Adank *et al.* (2004) reported correct classification rates for unnormalized data (i.e., their "Hz" condition) of about 80% and of more than 90% after Lobanov transformation. However, we do not consider this as alarming for the following reasons: First, they entered additional predictors ($F0$ and $F3$) in their discriminators; second, their data set contained more speakers than ours (160 vs 7), and, presumably most important, ours contain more than three times as many categories in the criterion (30 vs 9) due to the more crowded German vowel system and the additional word stress condition. A further point to consider for the articulatory data set is that the discrimination procedure had no access to lip rounding information, which presumably provides the most important information for the distinction between rounded and non-rounded front vowels than the tongue data presented to the discrimination procedure here.

Another striking aspect of this analysis is that unstressed vowels benefit more from the result of the normalization. This pattern in principle holds for both the articulatory and the acoustic data sets, although there is a large difference in the performance of the discrimination procedure already in the unnormalized baseline for the articulatory data. This observation is compatible with a scenario according to which unstressed vowels are more prone to coarticulatory influences of the consonantal environment, which is sensitive to the normalization (Mooshammer and Geng, 2008).

Still, as already mentioned in the Introduction, the present study was designed not only to propose an alternative normalization scheme but also to attempt to quantitatively relate aspects of vocal-tract anatomy to the functioning of the normalization. This second aim of the study resembles classical tenets of the universal articulatory phonetics hypothesis (Johnson *et al.*, 1993) according to which interindividual differences should be lawfully related to factors such as others vocal-tract geometry. This was achieved by validating the directions extracted by our normalization method against independently extracted models of palate shapes, which resulted in high correlations of these independently extracted directions. This suggests that the normalization procedure in the articulatory data set partially operates on interindividual differences related to aspects of tract morphology.

This virtue to some extent might at the same time be the largest drawback: Allowing orthogonal rotational or even affine transformations in addition to the Lobanov scalings clearly has a strong desirable effect on phoneme classification rates, but in some instances such transformations might be too aggressive and distort the categorical structure. In particular, ellipse orientations are affected more strongly by our procedure than it is the case for the more conservative Lobanov transform. This particular observation on ellipse orientation and the likewise less conservative nature of formalism and transformation applied might lead to speculations about detrimental effects of the procedure also in other situations. Still, as demonstrated, the affine version of the procedure produces output showing how the fit was achieved. However, affine transformation might be undesirable in some settings. Then, it might be indicated to revert to the orthogonal version—the generalized Procrustes analysis—or to the Lobanov transform. Still, even in such cases, the analysis provides relevant diagnostic information on the data set and therefore presents a useful additional tool for the phonetic practitioner.

ACKNOWLEDGMENTS

Special thanks to Dirk Fischer, Suse Fuchs, and Daniel Pape for their help with analyzing the data and to Jörg Dreyer for assisting in the EMMA recordings. We are also very grateful to Jonathan Harrington, Phil Hoole, and Michel Scheffers for invaluable comments on earlier drafts of this paper. Any remaining errors are our responsibility. This work was supported by the German Research Council (DFG) Grant No. GWZ 4/5-2(A1).

[1]But, as Hastie *et al.* (1995, p. 2)—quoting Michie *et al.* (1994)—note, often the "LDA produces the best classification results because of its simplicity […]. LDA was among the top 3 classifiers for 11 of 22 data sets studied in the STATLOG project." Further, they mention that the cases of shortcomings of LDA can be paraphrased as "saying that a single prototype per class is insufficient." In this line of reasoning, the use of arbitrary decision regions is prohibitive because more than one prototype per category runs against the concept of vowel normalization, which aims at augmenting the coherence between individual tokens and the phonetic class.

[2]As outlined in greater detail in Mooshammer and Geng (2008) and Mooshammer and Fuchs (2002), there is an interesting asymmetry in German: Tense vowels are substantially shortened in unstressed syllables, whereas lax vowels tend to maintain their duration in stressed and unstressed positions. Nevertheless, for both vowel series the tongue tip is elevated and fronted in unstressed position because of the adjacent alveolar stops. Therefore, we concluded that German unstressed vowels are generally produced with a greater degree of coarticulation, irrespectively of durational shortening.

Adank, P., Smits, R., and van Hout, R. (**2004**). "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. **116**, 3099–3107.

Apostol, L., Perrier, P., and Bailly, G. (**2004**). "A model of acoustic interspeaker variability based on the concept of formant-cavity affiliation," J. Acoust. Soc. Am. **115**, 337–351.

Beckman, M. E., Jung, T., Lee, S., de Jong, K., Krishnamurthy, A., Ahalt, S., Cohen, B., and Collins, M. (**1995**). "Variability in the production of quantal vowels revisited," J. Acoust. Soc. Am. **97**, 471–490.

Clopper, C., Pisoni, D., and de Jong, K. (**2005**). "Acoustic characteristics of the vowel systems of six regional varieties of American English," J. Acoust. Soc. Am. **118**, 1661–1676.

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

Fitzpatrick, L., and Ni Chasaide, A. (**2002**). "Estimating lingual constriction location in high vowels: A comparison of EMA- and EPG-based measures," J. Phonetics **30**, 397–415.

Geng, C., and Mooshammer, C. (**2000**). "Modeling the German stress dis-

tinction," in Proceedings of the Fifth Speech Production Seminar, pp. 161–164.
Goodall, C. R., and Green, P. (**1986**). "Quantitative analysis of surface growth," Botanical Gazette **147**, 1–15.
Gower, J. C. (**1975**). "Generalized Procrustes analysis," Psychometrika **40**, 33–51.
Guenther, F. H., Hampson, M., and Johnson, D. (**1998**). "A theoretical investigation of reference frames for the planning of speech movements," Psychol. Rev. **105**, 611–633.
Harshman, R. A., Ladefoged, P., and Goldstein, L. (**1977**). "Factor analysis of tongue shapes," J. Acoust. Soc. Am. **62**, 693–707.
Hashi, M., and Westbury, J. (**1998**). "Vowel posture normalization," J. Acoust. Soc. Am. **104**, 2426–2437.
Hastie, T., Tibshirani, R., and Buja, A. (**1995**). "Flexible discriminant and mixture models," in *Neural Networks and Statistics*, edited by J. Kay and D. Titterington (Oxford University Press, New York).
Hoole, P. (**1999**). "On the lingual organization of the German vowel system," J. Acoust. Soc. Am. **106**, 1020–1032.
Jackson, M. T. T. (**1988**). "Analysis of tongue positions: Language-specific and cross-linguistic models," J. Acoust. Soc. Am. **84**, 124–143.
Johnson, K. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. Pisoni and R. Remez (Blackwell, Oxford), pp. 363–389.
Johnson, K., Ladefoged, P., and Lindau, M. (**1993**). "Individual differences in vowel production," J. Acoust. Soc. Am. **94**, 701–714.
Joos, M. (**1948**). "Acoustic phonetics," Language Monograph 23, Supplement to Language **24**, 1–136.
Kingston, J., and Diehl, R. (**1994**). "Phonetic knowledge," Language **70**, 419–454.
Lobanov, B. M. (**1971**). "Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am. **49**, 606–608.
Michie, D., Spiegelhalter, D., and Taylor, C. (**1994**). *Machine Learning, Neural and Statistical Classiffication* (Ellis Horwood, New York).
Mooshammer, C., and Fuchs, S. (**2002**). "Stress distinction in German: Simulating kinematic parameters of tongue tip gestures," J. Phonetics **30**, 337–355.
Mooshammer, C., and Geng, C. (**2008**). "Acoustic and articulatory manifestations of vowel reduction in German," J. Int. Phonetic Assoc. **38**, 117–136.
Nearey, T. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.
Perkell, J., and Nelson, W., (**1985**). "Variability in production of the vowels /i/ and /a/," J. Acoust. Soc. Am. **77**, 1889–1895.
Ripley, B. (**1994**). "Neural networks and related methods for classification," J. R. Stat. Soc. Ser. B (Methodol.) **56**, 409–456.
Rohlf, F. J., and Slice, D. (**1990**). "Extensions of the Procrustes method for the optimal superimposition of landmarks," Syst. Zool. **39**, 40–59.
Simpson, A. (**2002**). "Gender-specific articulatory-acoustic relations in vowel sequences," J. Phonetics **30**, 417–436.
Stevens, J. (**1986**). *Applied Multivariate Statistics for the Social Sciences* (Erlbaum, Hillsdale, NJ).
Westbury, J. R. (**1994**). *X-ray Microbeam Speech Production Database User's Handbook, Version 1.0* (Waisman Center on Mental Retardation & Human Development, Madison, WI).