*Language Variation and Change*, **21** (2009), 413–435.

doi:10.1017/S0954394509990160

# A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics

ANNE H. FABRICIUS
*Roskilde University, Denmark*

DOMINIC WATT AND DANIEL EZRA JOHNSON
*University of York, UK*

ABSTRACT

This article evaluates a speaker-intrinsic vowel formant frequency normalization algorithm initially proposed in Watt & Fabricius (2002). We compare how well this routine, known as the *S*-centroid procedure, performs as a sociophonetic research tool in three ways: reducing variance in area ratios of vowel spaces (by attempting to equalize vowel space areas); improving overlap of vowel polygons; and reproducing relative positions of vowel means within the vowel space, compared with formant data in raw Hertz. The study uses existing data sets of vowel formant data from two varieties of English, Received Pronunciation and Aberdeen English (northeast Scotland). We conclude that, for the data examined here, the *S*-centroid *W&F* procedure performs *at least as well* as the two speaker-intrinsic, vowel-extrinsic, formant-intrinsic normalization methods rated as best performing by Adank (2003): *Lobanov*'s (1971) *z*-score procedure and *Nearey's* (1978) individual log-mean procedure ($CLIH_{i4}$ in Adank [2003], $CLIH_{i2}$ as tested here), and in some test cases better than the latter.

The *S*-centroid vowel normalization procedure, originally presented in detail in Watt & Fabricius (2002), was developed to further research on variation and change in British English vowel systems (e.g., Watt & Tillotson's [2001] discussion of Bradford English). It offered a new normalization method tailored specifically for sociophonetics, optimizing as far as possible comparisons of vowel systems from different speakers, both male and female, without prior

The authors thank Caroline Moreiras and Bronwen Evans for access to Moreiras' (2006) unpublished vowel formant data, Jillian Oddie for assisting with the Aberdeen data collection and analysis, Tyler Kendall for advice on the NORM suite, Victoria Watt with help with processing the Aberdeen data, and Bernhard Fabricius for assistance with exploratory mathematical and programming tasks. Data collection in Aberdeen was supported by the International Association for Forensic Phonetics and Acoustics. Earlier versions of this article were presented at the BAAP (British Association of Academic Phoneticians) Colloquium in Sheffield, UK (with Jillian Oddie), in April 2008 and at Acoustics08, Paris, in July 2008. The authors thank the audiences at both events for comments and suggestions on those occasions.

assumptions as to configurational similarities between the speakers to be compared. The *S*-centroid method has recently been employed in studies of variation in Received Pronunciation (RP) (Fabricius, 2007a, 2007b), London English (Kamata, 2008), South African English (Mesthrie, unpublished manuscript), and, in a modified form inspired by Watt & Fabricius (2002), in a study of the U.S. English of Illinois (Bigham, 2008). The *S*-procedure normalizes a speaker's set of vowel data by expressing each formant value as a proportion of its respective centroid value, which is derived using $F_1$ and $F_2$ maxima and minima for that individual's vowel space (more details follow).

The acoustic phonetics literature already contains a plethora of normalization procedures; see, for example, the 12 different methods investigated and tested in Adank (2003:16–25) (and presented in summary form in Adank, Smits, & van Hout, 2004), as well as other works such as Deterding (1990), Disner (1980), and Hindle (1978), which evaluate a range of algorithms in various ways. The present authors' motivation for bringing yet another normalization algorithm into play was a wish to focus on the particular challenges of the *sociophonetic* methodological practice of visual comparisons of speaker vowel plots. Given that studies of variation and change in U.S. varieties of English (Labov, 1994; Labov, Ash, & Boberg, 2006) have predominantly used visual comparisons of vowel plots to identify the sociophonetic variation patterns under scrutiny (such as the separate stages of the Northern Cities Vowel Shift, for example), we sought to optimize the visual input to sociophonetic studies of variation and change by devising a normalization method that derived normalized plots of vowel spaces for different speakers that were as well matched in size and overlap as possible. Perceptual aspects of the normalization problem are not dealt with at all in this study.[1]

The present article reports on a study we have carried out to further evaluate the performance of the *S*-centroid procedure by testing it more thoroughly than in Watt & Fabricius (2002), using more refined comparison techniques, larger amounts of data, and more than one variety of English. The *S*-procedure has recently been included in the NORM suite of normalization algorithms and plotting functions made available online by Thomas & Kendall (2007), which has made it possible for us to conduct more extensive direct comparisons of the S-procedure with other longer-established normalization algorithms.[2] In this article, then, we explore the merits of the new routine compared with the recommended normalization methods reported in Adank (2003), viz. Lobanov's (1971) *z*-score procedure and Nearey's (1978) individual log-mean procedure (labeled $CLIH_{i4}$ in Adank, 2003), here referred to as *Lobanov* and *Nearey1* (see the note on normalization algorithm details).

In performing the evaluations presented in this study, we employ a series of metrics to gauge an algorithm's performance on the visual parameters we see as crucial for sociolinguistic (more specifically, sociophonetic) purposes. First, we compare the different methods' performance in equalizing vowel space areas, by examining how well they remove variance in vowel space areas between speakers. Second, we measure the degree of intersection of individual vowel

spaces achieved by the algorithms, because an optimal method would achieve the highest possible degree of overlap.[3] Third, we examine two-dimensional geometric relationships between mean vowel points in individual vowel plots and consider how these relationships are represented in normalized data compared with mean raw Hertz data. The study uses two data sets: (1) a corpus of 20 RP speakers compiled from Hawkins & Midgley (2005) and Moreiras (2006), and (2) Scottish English data from Aberdeen in the northeast of Scotland (Watt & Yurkova, 2007).

Our results suggest that the Watt & Fabricius *S*-centroid methods (both the original, hereafter *W&F*, and in a slightly altered version we call "modified *W&F*," hereafter *mW&F*) perform in general at a level somewhere between *Lobanov* and *Nearey1*, and in specific test cases *at least as well as* these two better-known and typologically comparable normalization methods. We also point out some conditions under which all three methods can in fact be too powerful for visual sociophonetic comparisons if the juxtaposition of individual vowel means on the $F_1$-$F_2$ plane is the focus of an investigation. We conclude the article by emphasizing the complex nature of sociophonetic "best practices" choices.

## PREVIOUS LITERATURE

Thomas & Kendall (2007), quoting Disner (1980) and Thomas (2002), enumerated four general goals of vowel normalization procedures:

a. to eliminate variation caused by physiological differences among speakers;
b. to preserve sociolinguistic/dialectal/cross-linguistic differences in vowel quality;
c. to preserve phonological distinctions among vowels;
d. to model the cognitive processes that allow human listeners to normalize vowels uttered by different speakers.

For the present study, we focus on goals (a) and (b); moreover, it is the balance between these two that we see as crucial, and which we want to explore further here. As noted earlier, we do not enter into a discussion of point (d), because it is not relevant for our purposes in this article, although ultimately it has highly important consequences for understanding language and language change. Point (c) comes into play in, for example, Adank's (2003) study of vowel normalization methods, but it is less important for the work presented in this article, because by comparing vowel means, we are not in any sense testing phonological categories, but taking them as given.

Adank's (2003) study provides a useful set of typological classifications for normalization methods that we will briefly present here, using the crucial concepts *intrinsic* and *extrinsic*. Normalization methods can be divided into different types depending on whether the algorithm applies either to individual

vowels or to sets of vowels (according to some researchers, optimally all vowels) of the language variety. The former type are called *vowel-intrinsic*, the latter *vowel-extrinsic*. Moreover, it also possible to classify normalization methods according to whether they are *formant-intrinsic* or *formant-extrinsic*, that is, whether they use information from one formant at a time to normalize a single formant value, or take information from the range of formants of the vowel to normalize a single formant value. Thomas & Kendall (2007) also classified different normalization methods according to whether a method uses information from a single speaker at a time (and thus are *speaker-intrinsic*) or from a population (*speaker-extrinsic*) to normalize vowel data from single speakers. Whereas speaker-extrinsic methods are commonly used in sociolinguistics (especially in North America; see, e.g., Labov et al., 2006; Thomas & Kendall, 2007), they seem to be largely absent from the mainstream acoustic phonetics literature.

The abundance of normalization algorithms in the literature has naturally led to a series of studies evaluating them (Adank, 2003; Deterding, 1990; Disner, 1980; Hindle, 1978; Nearey, 1978; Rosner & Pickering, 1994). Our work is of a more limited scope than Adank's (2003) research, because we only consider acoustic comparisons, not perceptual comparisons, as noted previously, but like Adank, we are interested in normalization techniques that reduce variability of physiological origin but preserve sociolinguistically interesting information, though our treatment of these two sources of variation differs slightly from that of Adank.

Adank's (2003) research aimed to find optimally successful vowel normalization algorithms for sociolinguistic purposes within the set she compared (which did not include *W&F*). This was done by evaluating how well procedures eliminate physiological differences (defined as the consequence of vocal tract length, speaker sex, and age) while preserving (regional) sociolinguistic information and preserving or improving the phonemic categorization of the data by seeing how well a normalization algorithm modeled trained speakers' perceptions. Adank employed a range of acoustic and perceptual evaluations of her data, comprising spoken forms from regional varieties of Standard Dutch.

Adank's (2003:99–124) evaluation procedures submitted raw Hertz data and normalized data tokens to a series of linear discriminant analyses (LDAs). The first of the LDAs was used to determine how well the evaluated normalization methods performed in obtaining correct phonemic classifications of the tokens, by determining which method delivered the highest percentage of correctly classified tokens. *Lobanov* and *Nearey* $CLIH_{i4}$ performed best of all procedures on this test (Adank, 2003:91). A second LDA determined how well the different procedures eliminated male-female variation, using a range of combinations of formant values from $F_1$ to $F_3$ plus the fundamental frequency $F_0$. *Lobanov* and *Nearey* $CLIH_{i4}$ again performed best overall at removing this information from the data, such that the resultant capability of the normalized data to be classified as male or female was not better than chance (Adank, 2003:95). Adank's third LDA evaluated how well the procedures classified the tokens as being produced

by younger or older speakers (2003:95). Note that Adank treats the age variable as representing physiological information to be removed, not sociolinguistic information to be retained, as it potentially would be in a study of diachronic change. The acoustic consequences of speaker age in the raw data were much smaller than gender variations in the first place, as the standard Dutch data were diachronically quite stable. Again, *Lobanov* and *Nearey CLIH$_{i4}$* performed best at eliminating this information (Adank, 2003:96). Both age and gender effects were therefore most successfully removed using *Lobanov* and *Nearey CLIH$_{i4}$*. The explicitly sociolinguistic variable in Adank's data was region, because the data consisted of varieties of standard Dutch from across the Netherlands and Flanders, Belgium. In this case, *Lobanov* and *Nearey CLIH$_{i4}$* did remove some of the sociolinguistically interesting regional variation; the two procedures were neither the best nor the worst-performing at this test.

To sum up, *Lobanov* and *Nearey CLIHi4* performed best at removing variation due to speaker sex and speaker age, while performing less than optimally in preserving the sociolinguistically interesting regional variation in the Standard Dutch data. The most successful procedures in the latter test were the vowel- and formant-intrinsic methods, such as Bark and ERB (equivalent rectangular bandwidth), among others (Adank, 2003:98); it should be noted, however, that the latter are not normalization procedures as such, at least in the sense of the term that we adopt here, but rather are psychoperceptual transforms. Bark, ERB, and other vowel-intrinsic, formant-intrinsic algorithms had, however, performed poorly on removing age and gender variation. This led to Adank's overall conclusion that *Lobanov* and *Nearey CLIH$_{i4}$* were the most successful methods, performing best in two out of three tests.

The conclusions of Adank's (2003) acoustic comparisons are interesting for our study, because the procedures she judged as being optimal for sociolinguistic purposes, including *Lobanov* and *Nearey CLIH$_{i4}$*, are typologically *speaker-intrinsic*, *vowel-extrinsic*, and *formant-intrinsic*, and this is precisely the typological profile of the Watt & Fabricius (2002) method. We decided therefore to subject this method, together with *Lobanov* and a method similar to *Nearey CLIH$_{i4}$* (which incorporates $F_0$, $F_1$, $F_2$, $F_3$), which is labeled *Nearey CLIH$_{i2}$* (as it incorporates only $F_1$ and $F_2$) hereafter *Nearey1*, to a series of tests that also aimed at identifying sociophonetically optimal normalization procedures. Our methodology is, however, somewhat different from Adank's, in that we are interested in an evaluation of visual cross-speaker mapping of vowel means, rather than evaluations of the efficiency of normalization methods in sorting vowel tokens into known overarching phonemic categories. The two types of comparisons do complement each other, however.

As noted in the introductory text, our choice of comparison methodology reflects our interest in the widespread practice within sociolinguistics of using visual comparisons of speaker plots to find vowel configurational differences that match speaker age and/or gender (and/or regional) distinctions. Rather than sorting between different possible sources of variability, and seeking to eliminate some and retain others, we simply seek to optimize the process of

visual comparison between vowel plots from any two individuals, regardless of which sociolinguistically relevant factor lies behind the variability. For instance, we test to what extent the set of normalization procedures will preserve known age variation within RP as a sociolinguistically interesting phenomenon, rather than, as Adank (2003) did, treating age as a physiological variable to be eliminated by normalization. Thus, several purely geometric parameters come into focus. One is the ability of the normalization algorithm to "bring vowel spaces together" to enable realistic comparisons that reflect actual, and not artifactual, differences. Optimal cross-speaker mapping enables a comparison of geometric relationships between points in two-dimensional representations of speakers' vowel spaces, whether in informal terms or by means of geometric measures (Fabricius, 2007a, 2007b). As demonstrated with a male/female pairwise comparison in Watt & Fabricius's (2002) paper, the Bark transformation did not facilitate optimal cross-speaker mapping, but the amount of overlap between the two speakers was greatly improved in the *S*-normalized data.

We make no explicit cross-variety comparisons between the RP and Aberdeen English systems that are examined here, because we accept Adank's point that comparisons between different phonological systems can be problematic. Note, however, that Thomas (2002) took a different position on this question; comparisons between phonological systems can be desirable for certain research purposes, and we acknowledge that ideally a normalization method ought to be able to straightforwardly accommodate this need. Nevertheless, here we confine our comparisons to one variety of English (Aberdeen English or RP) at a time; our purpose in including two varieties in the present study is simply to be able to evaluate how the *S*-procedure performs with different vowel systems and data sets.

Figure 1 serves to illustrate the general normalization problem we are evaluating. The figure depicts the different mappings between an older male's and a young female's RP vowel systems. (OM5 corresponds to Hawkins & Midgley's [2005] speakers 1–5); YF5 to Moreiras [2006] YS-5). The non-normalized vowel plot at top left shows the higher formant frequencies of the female speaker to the left and below those of the male speaker, resulting in vowel configurations of very different size and with minimal overlap. The three normalized plots show that greater overlap between the male and female speaker's vowel plots is achieved in each case, and that the vowel polygons have now approached similar sizes. Moreover, comparisons between each individual's internal vowel configurations in the non-normalized plot, and under each normalization condition, show that the three normalization algorithms seem to perform similarly on these two data sets. The relative positions of each individual speaker's TRAP and STRUT vowels, for instance, is well-preserved under normalization. These points are more or less horizontal for the older speaker, whereas for the younger speaker, the two vowel means are almost vertically aligned. These examples illustrate our purpose in the present study: to compare the normalization algorithms on two separate data sets to evaluate performance on a range of tasks especially suited to the typical needs of the discipline of quantitative sociolinguistics.

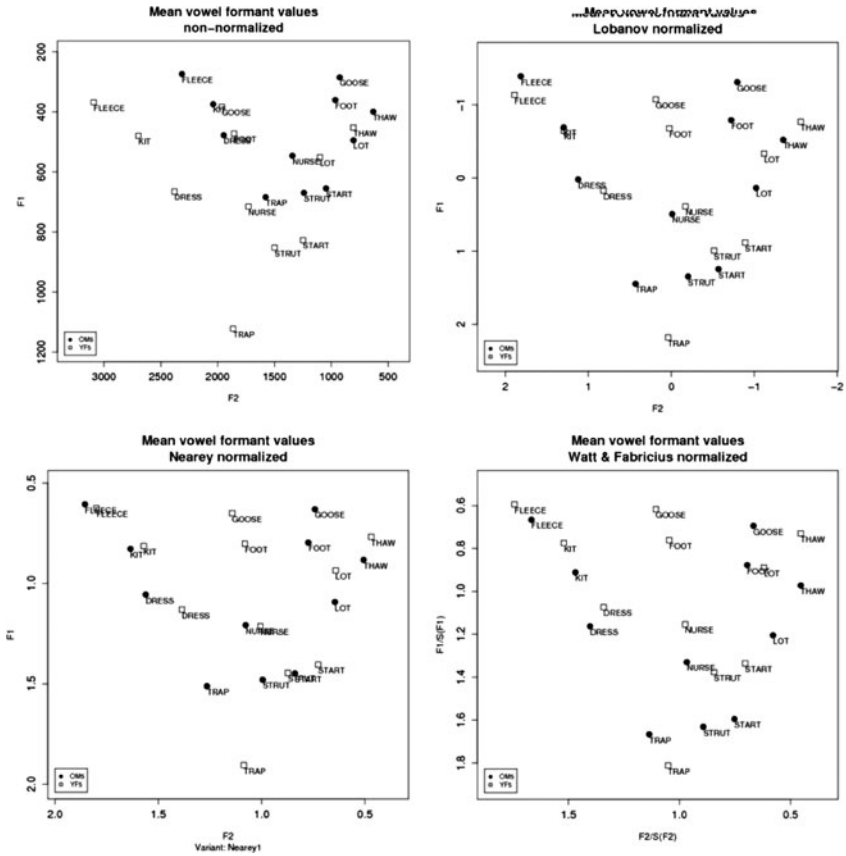


FIGURE 1. Example plots of unnormalized and normalized vowel formant data for two RP speakers.

### *Clarification of versions of Nearey normalization*

Note that the *formant-extrinsic* version of *Nearey*, $CLIH_{s2}$ in Adank's terms, labeled *Nearey2* on the NORM suite Web site, differs from the *formant-intrinsic* version of *Nearey*, also known as $CLIH_{i4}$, and Adank evaluated the latter method as far more successful than the former in her investigation (Adank, 2003). For that reason, in the present article, we examine new comparisons between *Lobanov*, *W&F*, and *Nearey*, now using NORM's version of the formant-intrinsic method, labeled *Nearey1* (in effect $CLIH_{i2}$, because only formants 1 and 2 are employed in the algorithm here). Although this is close to the formant-intrinsic version of *Nearey* evaluated in Adank (2003), it is not identical to it, because Adank's evaluations were performed on $CLIH_{i4}$, which uses $F_0$ and $F_3$ as well as $F_1$ and $F_2$. Moreover, as *Nearey2* is probably the most-commonly used method in North American sociolinguistic studies, we include footnotes with details of our earlier (Fabricius, Watt & Johnson, 2008; Fabricius, Watt, & Yurkova, 2008) comparisons with *Nearey2*, for completeness.

## THE *S*-PROCEDURE

We present here a brief summary of the *S*-centroid normalization method, described more fully in Watt & Fabricius (2002). The *S*-procedure seeks to establish $F_1$ and $F_2$ maxima and minima for each speaker within a sample of vowel measurements, as indicated in Figure 2. This is based on finding and deriving three "point vowels" representing the frontest and lowest and (by derivation) backmost points of the vowel space. Note that the *u′* vowel is not an observed, backmost point, but a derived one, defined as $F_2$ (*u′*) = $F_1$ (*u′*) = $F_1$ (*i*). The procedure then derives a centroid point, *S* (after Koopmans–van Beinum, 1980) from these corner points according to the formula:

$$S(F_n) = \frac{[i]F_n + [a]F_n + [u']F_n}{3}$$

All the observed measurements of $F_n$ are then divided by the *S* value for that formant *n*, and all resulting figures are expressed on the scale of $F_n/S(F_n)$.

Watt & Fabricius (2002) carried out an initial comparison of the *S*-procedure with Bark (Traunmüller, 1990, 1997). Using two RP speakers' data from Deterding (1997), vowel data points from one male and one female speaker were normalized using Bark and the *S*-centroid procedure. The three data sets were then evaluated on the parameters of area agreement and intersection of vowel spaces. This comparison was carried out by means of simple geometrical calculations of triangle area and calculations of ratios of the smaller to the larger vowel triangle.

The numbers in Table 1 reveal that *S*-normalized data showed substantial improvement over both raw Hertz data and Bark-transformed values in the comparisons of the two speakers' area ratio and overlap. The table also includes percentage figures showing the *S*-procedure's improvements in performance over Hertz and Bark (see Watt & Fabricius [2002:165–166] for diagrammatic representations showing area ratios and overlaps under the three conditions).

In this article, we expand upon this earlier evaluation of the *S*-procedure by including larger data sets and two varieties of English—RP and Aberdeen English. Note also that slightly different definitions of the evaluation measures are used in the present article. In Watt & Fabricius (2002), the degree of overlap between the two speakers' vowel triangles was expressed as the percentage of the male speaker's triangle that overlapped with the female speaker's triangle, and vice versa. In the present study, overlap is defined as the intersection of two vowel polygons divided by the union of the same polygons (further details follow).

We also introduce a minor modification of the *S*-procedure, which we here call *mW&F*. We do this in response to the specific critique in Thomas & Kendall (2007) that the original (2002) formula introduced skewing of normalized vowel values in the lower part of the vowel space if the vowel chosen to represent the lowest point of the vowel space, [a], had an $F_2$ value that was significantly higher or lower than that

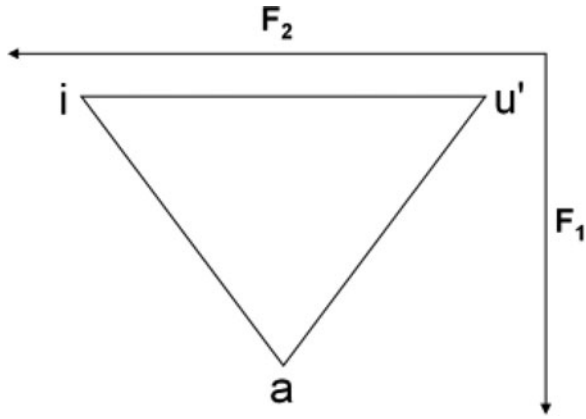


FIGURE 2. Schematized representation of the idealized vowel triangle used for the calculation of *S*. $i$ = min $F_1$, max $F_2$; $a$ = max $F_1$; $u'$ = min $F_1$, min $F_2$, where $F_1$ ($u'$) and $F_2$ ($u'$) = $F_1$ ($i$). Reproduced from Watt & Fabricius (2002:164), with permission.

TABLE 1. *Improvements in area ratio and degree of overlap between vowel triangles for speakers A (male) and C (female)*

| | Hz | Bark | *S* |
|---|---|---|---|
| Area ratio (ΔC:ΔA) | 1:3.93 | 1:2.76 | 1:2.16 |
| % improvement over Hz | — | 29.8 | 45 |
| % improvement over Bark | — | — | 21.7 |
| % overlap (ΔC:ΔA) | 46.1 | 49.9 | 99.2 |
| % improvement over Hz | — | 8.2 | 115.2 |
| % improvement over Bark | — | — | 98.8 |
| % overlap (ΔA:ΔC) | 13.7 | 18.1 | 45.8 |
| % improvement over Hz | — | 32.1 | 234.3 |
| % improvement over Bark | — | — | 153 |

*Source:* Data from Deterding (1997). Reproduced from Watt & Fabricius (2002:168), with permission.

of the center point, as defined by the median value of the [i]–[u′] line. To remove this potential skewing, we disregard the $F_2$ value of [a] in the calculation of $S(F_2)$, so that the *S*-value for $F_2$ is equidistant between $F_2$ of [i] and $F_2$ of [u′] and calculated only on the basis of these two values, and not three $F_2$ values as per *W&F* in its original formulation.[4]

## RESEARCH QUESTIONS

Our research questions are therefore formulated as follows. How do the *S*-procedures (*W&F*, and an experimental version of the same procedure, *mW&F*) perform compared with *Lobanov* and *Nearey1* on the following

sociophonetically relevant evaluative parameters, using data sets from two accent varieties of English?

1. Reduction of variance in area ratios of vowel spaces, thus equalizing vowel space areas as far as possible;
2. Improvement of coextensiveness of vowel polygons;
3. Reproduction of two-dimensional vowel configurational relationships within the vowel space, compared with raw Hertz formant data by calculating:

   i. The gradients of lines joining means for DRESS and LOT, both of which are considered to be diachronically stable in RP (Hawkins & Midgley, 2005),
   ii. The gradients of lines joining means for TRAP and STRUT, and, secondly, LOT and FOOT, juxtapositions that have been undergoing diachronic change in RP (Fabricius 2007a, 2007b).[5]

## DATA

The data employed in the comparisons reported here come from three independent sources. The RP vowel formant data are compiled from two sources: 10 male speakers, taken from the sets of 5 oldest and 5 youngest speakers published in Hawkins & Midgley (2005); and 10 female speakers, taken from Moreiras's (2006) data, using the first 5 female speakers in each of two matching age groups (OS1-5, YS1-5). The older speakers were born in the period 1928–1936, and the youngest were born between 1976 and 1981. The study here uses $F_1$ and $F_2$ data for all 11 stressable monophthongs, representing the keyword categories (Wells, 1982) FLEECE, KIT, DRESS, TRAP, START, LOT, THOUGHT, FOOT, STRUT, NURSE, and GOOSE. The measurements are all taken from tokens of stressed monophthongal citation forms in a /hVd/ frame within wordlists, recorded in sound-treated conditions.

The Aberdeen data in this study (documented in part in Watt & Yurkova, 2007) consist of measurements of vowel tokens from six speakers, male and female, aged between 21 and 62 years. The study uses $F_1$ and $F_2$ data for eight vowels in Aberdeen English: FLEECE, FACE, DRESS, TRAP, START, THOUGHT, GOAT, and GOOSE. This system differs from the RP set for various reasons. FACE and GOAT are included because they are monophthongal, not diphthongal, in Aberdeen English. As Aberdeen English lacks a LOT/THOUGHT contrast, LOT is not included separately. KIT is a centralized vowel in Aberdeen English and so was not suitable for use as a point on the perimeter of the vowel space; because of the frequency of tapped [ɾ] in this variety, START was included as it is not heavily rhotacized. Data tokens in the Aberdeen English data are taken from word lists, with target vowels in a variety of phonetic contexts, recorded in a quiet sound-treated room at Aberdeen University. Although arising from different recording settings, the data in all cases are measurements of careful speech forms that have been straightforward for acoustic measurement by the respective authors. For

that reason, we consider the data sets equally valid as test cases for normalization comparisons, as well as being sociolinguistically realistic.

## METHODS

Normalization procedures were performed using the NORM suite,[6] using the *Watt & Fabricius*, *Lobanov* (*speaker intrinsic*), and *Nearey1* (*speaker-intrinsic*) routines available on that Web site. Normalizations were carried out without using Thomas & Kendall's (2007) "scaling factor."[7] *mW&F* was programmed by the third author and performed in the same manner as the other NORM suite normalizations. For the subsequent testing procedures, the areas of individual vowel spaces and their unions and intersections were calculated using the R package gpclib by Roger D. Peng, adapted from the GPC (General Polygon Clipper) library.[8]

First, Test 1 involved comparisons of performance in equalizing vowel space areas. The better the normalization procedure at equalizing vowel space areas, the less variation there should be among speakers in the sample. Because the area calculation results were not immediately comparable across methods because the units of each algorithm were different, we used the squared coefficient of variation (SCV) to quantify the reduction of variance across different methods:

$$\mathrm{SCV} = (\mathrm{SD}/\mathrm{Mean})^2$$

To compare how well the four normalization methods perform against the raw Hertz data, we then divided each method's SCV by the Hertz SCV, which gave the proportion of variance remaining after normalization. This latter value subtracted from 1 gave the proportional reduction in variance for each normalization procedure. These results were then compared statistically using Pitman-Morgan's test of homogeneity of variance between correlated samples (Cohen, 1990).[9]

Second, for Test 2, a measure of the intersection of any individual vowel polygon with all other speakers' vowel polygons in the data set was obtained. This was done by calculating the intersection of two vowel polygons divided by the union of the same polygons to obtain an overlap value. The comparison for any individual speaker was with all other speakers in the (RP or Aberdeen English) data set. Sets of overlap values were compared statistically using paired *t*-tests.

Third, because studies of sociolinguistic variation in vowels often involve observations of variation in juxtapositions of vowel tokens (e.g., Labov, 1994; Thomas, 2001), we considered it important to investigate how the three normalization algorithms performed in showing visual relationships between vowel tokens on the $F_1$-$F_2$ plane. If a change is visualized as a shifting vowel positions, which means that vowels of the older generation are in a visually different configuration from those of the younger generation, it is important that the normalization procedure employed does not seriously distort these

geometrical relations. In Test 3, therefore, our third comparative parameter was the extent to which the normalization algorithms preserved two-dimensional geometric relationships within individual speakers' vowel spaces as compared with the raw Hertz data. This particular set of comparisons was carried out on the 20 RP speakers only. This part of the study is intended to be exploratory and not comprehensive. As already discussed, the small set of comparisons we examine here are:

a. The gradient of a line between the points DRESS and LOT relative to the horizontal axis, as an example of a stable configuration for older and younger speakers of RP;
b. The gradients of the lines between (i) the points TRAP and STRUT (relative to the horizontal) and (ii) the points LOT and FOOT (relative to the vertical) as examples of changing configurations with known differences between older and younger speakers of RP.

These calculations were carried out using the geometrical function of arctangent for gradient calculations (Fabricius, 2007a).

## RESULTS

### *Test 1: Equalizing vowel space areas*

Test 1 thus determined which normalization method performed best at reducing variance in vowel space areas, as described previously. Table 2 presents the proportional reduction of area variance obtained for each normalization procedure. According to the measures obtained for the RP data in Table 2, *Nearey1* performs worse than the three other normalization methods,[10] because this procedure removed only 7% of the variance within in the raw Hertz area ratios for the RP data. *W&F* and *mW&F* perform similarly and considerably better, removing 35% and 39% of variance, respectively. *Lobanov* is highly efficient, in that 92% less variance is found in the RP data following normalization. In the case of the Aberdeen data, the various normalization methods perform more similarly with respect to each other, but again with approximately the same rankings, with *Lobanov* achieving the best removal of variance. *Nearey1* performs least optimally, and the two *W&F* algorithms are again ranked between the two other methods, achieving better removal of variance in the Aberdeen data than in the RP material. *Nearey1, W&F*, and *mW&F* thus perform with the same ranking but from a better starting point with the Aberdeen data. The lowest in the rank, *Nearey1*, achieves a 60% reduction in variance.[11] Statistical testing of these comparisons was then carried out by means of a Pitman-Morgan test of homogeneity of variance between correlated samples. We took $p < .05$ as the significance threshold. The $p$ values for each individual comparison are given in Table 3.

TABLE 2. *Performance of normalization algorithms in removing variance between vowel space areas*

| Improvement over Hertz | *Nearey1* | *W&F* | *mW&F* | *Lobanov* |
|---|---|---|---|---|
| RP | 0.070 | 0.350 | 0.389 | 0.923 |
| Aberdeen | 0.601 | 0.877 | 0.865 | 0.974 |

*Note:* Higher values indicate a better result.

TABLE 3. *Pitman-Morgan p values on test of homogeneity of variance between correlated samples, RP, and Aberdeen English data*

| Procedure pairs | RP ($N = 20$) | Aberdeen ($N = 6$) |
|---|---|---|
| Hertz/*Nearey1* | .804 | **(0.00159)*** |
| Hertz/*W&F* | .213 | **(0.00073)*** |
| Hertz/*mW&F* | .143 | **(.00315)*** |
| Hertz/*Lobanov* | **(1.17e–07)*** | **(.00098)*** |
| *Nearey1/W&F* | **(.04)**$^{\dagger}$ | **(.0101)*** |
| *Nearey 1/mW&F* | **(.00591)*** | **(.0446)**$^{\dagger}$ |
| *Nearey1/Lobanov* | **(8.18e–08)*** | **(.0122)**$^{\dagger}$ |
| *W&F/mW&F* | .405 | .664 |
| *W&F/Lobanov* | **(5.58e–7)*** | .126 |
| *mW&F/Lobanov* | **(8.23e–07)*** | .126 |

*$p < .001$, $^{\dagger}p < .05$; bold values are significant.

In the case of the RP data, *Lobanov* performs significantly differently from all other normalization procedures in the test, and from raw Hertz. Somewhat counterintuitively, *W&F* and *mW&F* perform significantly better than *Nearey1*, but not significantly better than Hertz. This may be due to a small sample size; *W&F* does almost always equalize or centralize the area better than *Nearey1*, but not in all individual cases. In the case of the Aberdeen data, the results show that all four methods perform significantly differently from (and, again, comparing Table 2, better than) Hertz, and, in addition, here the original *W&F* procedure and *Lobanov* perform significantly better than *Nearey1*. The two Watt & Fabricius *S*-centroid procedures, *W&F* and *mW&F*, perform similarly, and not significantly differently from *Lobanov*. Across the two data sets, then, the two *W&F* procedures and *Lobanov* consistently outperform *Nearey1* on Test 1. In summary then, a performance scale for Test 1 (equalizing vowel space areas) can be posited thus:

$$Lobanov \geq W\&F,\ mW\&F > Nearey1 > \text{Hertz}$$

### *Test 2: Improving vowel space overlap*

Test 2 concerned measures of vowel space overlap. To reiterate the methodology discussion previously, the procedure employed for investigating vowel space

overlap was to compare the intersection divided by the union of a speaker's space with the combined union of the spaces of all other speakers in the data set (importantly, treating the RP and Aberdeen data as separate sets due to their different underlying phonological systems). This comparison determined how well each speaker's space overlaps with those of the rest of the group, so that any normalization procedure's performance can be compared with the overlaps found in the raw Hertz data. Table 4 shows how well each normalization procedure, and raw Hertz, performed on average in obtaining overlap between any single speaker's vowel space and the union of all other speakers' vowel spaces in the data set.[12]

Paired *t*-tests were then run on these overlap ratios, with the results shown in Table 5. The results show that most normalization methods did indeed perform significantly differently from each other. The only pair that did not perform significantly differently on either data set was *W&F*/*Nearey1*. Note that *mW&F* performed significantly better than *W&F* on both sets of data and also proved to perform significantly better than *Nearey1* on the RP data set. For Test 2, then, a scale representing the different algorithms' performance can be posited thus:

$$Lobanov > mW\&F \geq W\&F, Nearey1 > \text{Hertz}$$

We can now summarize the results of Tests 1 and 2 in an overview. Note first that these two related tasks compare the different normalization algorithms' ability to achieve improved area ratios between speakers and better overlaps/coextensiveness of vowel spaces—in other words, in different ways, to "bring vowel spaces together" on both parameters. Comparing the results of the tests, we can conclude that all four vowel-extrinsic formant-intrinsic normalization methods show improvement over raw Hertz on the two parameters we have tested so far. For the data sets tested in this study, these performances can be ranked in the following order from greatest improvement to least improvement:

$$Lobanov > mW\&F \geq W\&F \geq Nearey1$$

*Test 3: Preserving vowel mean juxtapositions*

The discussion turns now to Test 3. These comparisons sought to examine the extent to which normalization procedures could preserve geometric relationships between the mean points representing the average placements of vowels within individual speakers' acoustic vowel spaces, using the raw Hertz values as a starting point for the comparison.[13] The needs of sociophonetic studies to examine changes in relative placements of vowel data on the $F_1$-$F_2$ plane are foregrounded through this test. Thus, an optimal angle-preserving normalization algorithm would be one that did not significantly distort the relative positions

TABLE 4. *Average vowel space overlaps between any single speaker's vowel space and all other speakers' vowel spaces, RP, and Aberdeen data sets*

| Vowel space overlaps | Hertz | *Nearey1* | *W&F* | *mW&F* | *Lobanov* |
|---|---|---|---|---|---|
| RP average | .380 | .444 | .452 | .500 | .564 |
| Aberdeen average | .444 | .571 | .598 | .618 | .688 |

TABLE 5. *Paired t-test p values for overlap comparisons, RP and Aberdeen English data*

| Procedure pairs | RP ($N = 20$) | Aberdeen ($N = 6$) |
|---|---|---|
| Hertz/*Nearey1* | **(.0076)*** | **(9.160e–05)*** |
| Hertz/*W&F* | **(.0073)*** | **(.0065)**† |
| Hertz/*mW&F* | **(6.302e–05)*** | **(.005)**† |
| Hertz/*Lobanov* | **(1.709e–05)*** | **(.0034)**† |
| *Nearey1/W&F* | .4911 | .4085 |
| *Nearey1/mW&F* | **(2.525e–05)*** | .1979 |
| *Nearey1/Lobanov* | **(.0014)*** | **(.0455)**† |
| *W&F/mW&F* | **(2.567e–08)*** | **(.0141)**† |
| *W&F/Lobanov* | **(.0002)*** | **(.0096)**† |
| *mW&F/Lobanov* | **(.0246)**† | **(.0334)**† |

*$p < .001$, †$p < .05$; bold values are significant.

found in the raw Hertz data, or, for example, best retained differences between generations as seen in Hertz data.

It must be noted at the outset of this particular discussion that there is an inverse relationship between area optimizations of the kind investigated in Tests 1 and 2 and angle preservation in a vowel space under normalization as investigated in Test 3. It would be geometrically impossible to devise a normalization method that optimized vowel space areas and at the same time entirely preserved relational, configurational angles. The principal aim of Test 3, then, is not to define an absolute standard for angle preservation, because that would preclude gains in the other areas we are interested in for this study, but to explore how the different normalization methods might affect angle configurations at different positions with the vowel space.

The following examples illustrate some differences in performance on the RP data for the three normalization methods. We begin by reviewing one vowel configuration that is relatively stable over the two generations, the geometric relationship between the DRESS and LOT vowels. We then look at two RP vowel juxtapositions that have been found to be undergoing quite dramatic diachronic change: TRAP-STRUT and LOT-FOOT (Fabricius, 2007a, 2007b). In RP over the course of the 20th century, TRAP has lowered and backed, STRUT has shifted toward the center of the vowel space, and FOOT has fronted (and incidentally, unrounded). By examining these movements as changing relative juxtapositions

between vowel means, expressed as angles, we can quantify vowel changes that shift vowel locations around the vowel space. The illustrative vowel juxtapositions examined here will show that different normalization procedures do affect geometric angles somewhat differently. Moreover, the results of Test 3 show that the separate normalization algorithms performed differently depending on whether the vowel-space internal angle is calculated relative to the horizontal or the vertical (for more details on the angle method in sociophonetics, see Fabricius, 2007a).

As Table 6 shows, *no* normalization procedure produces a perfect fit with raw Hertz values for each individual speaker. *Lobanov*, *W&F* and *mW&F*, and *Nearey1* all diverge from the raw Hertz values by a factor that seems to increase as the Hertz angle increases. Thus, at 1 degree (see OF1, YM2, YF1, and YF2), the deviation from Hertz under the *Nearey1*, *W&F*, *mW&F*, and *Lobanov* algorithm conditions is minimal. At 7, 9, and 11 degrees, however, the deviation from Hertz is larger (see OM4, YM3, YM5). Thus, what looks like an essentially stable configuration between older and younger speakers in the Hertz conditions (an average difference of 6 degrees only, between −2 and 4) might under the *W&F*, *mW&F*, and *Lobanov* normalizations be interpreted as a small difference (of 17, 18, 17, and 20 degrees, respectively). Note for example the *Lobanov* result for OM1, where a Hertz value of −4 becomes −24 under normalization, an average small angle difference (6 degrees on average) in the raw Hertz values (as shown in the last line of Table 6) becomes a larger angle difference under normalization conditions, as much as 20 degrees under Lobanov normalization. Whereas the raw Hertz data essentially shows diachronic stability, the normalized data could be interpreted as showing a slight generational difference as a consequence of what is essentially a systematic artifact of normalization in these data.

Bearing these findings in mind, we then also examined two generationally different and changing configurations in RP to see how well the different normalization methods reproduced these juxtapositions and, importantly, reproduced the generational differences between them. All three normalization algorithms performed differently, depending in part on whether angles/slopes were calculated against the horizontal or the vertical dimension using the arctangent method (Fabricius, 2007a). For the TRAP-STRUT juxtaposition, the angle is calculated relative to horizontal with TRAP as the apex of the angle; for LOT-FOOT, the calculation is carried out with LOT as the apex and relative to the vertical. Tables 7 and 8 show mean values for vowel data from groupings within the same 20 RP speakers as used earlier; individual values for each speaker are given in tables in Appendix 1.

As Table 7 reveals, all methods overestimate the angle relative to the horizontal and underestimate it relative to the vertical compared with the Hertz values, by factors which, as we saw in Table 6, seem to be related to the size of the angle to begin with. Whereas a Hertz value of 2 degrees is only affected minimally (becoming 5 and 4 degrees under *Nearey1*, *W&F*, and *Lobanov*), a larger value will be affected quite dramatically; see, for example, the TRAP-STRUT younger speakers' average angle configurations, from 41 to 66 and 67 degrees).

TABLE 6. *RP DRESS-LOT juxtaposition angles by speaker and normalization method*

| Angle of DRESS-LOT by RP speaker | Hertz | *Nearey1* | *W&F* | *mW&F* | *Lobanov* |
|---|---|---|---|---|---|
| OM1 | −4 | −11 | −11 | −12 | −24 |
| OM2 | −3 | −9 | −8 | −8 | −9 |
| OM3 | −3 | −7 | −8 | −7 | −10 |
| OM4 | −7 | −20 | −21 | −22 | −23 |
| OM5 | −1 | −2 | −3 | −3 | −3 |
| OF1 | 1 | 2 | 2 | 2 | 3 |
| OF2 | −2 | −6 | −7 | −6 | −8 |
| OF3 | 0 | 0 | 0 | 0 | −1 |
| OF4 | −2 | −5 | −5 | −5 | −5 |
| OF5 | 2 | 4 | 5 | 5 | 5 |
| YM1 | 6 | 16 | 17 | 16 | 19 |
| YM2 | 1 | 3 | 3 | 3 | 3 |
| YM3 | 9 | 24 | 24 | 22 | 22 |
| YM4 | 3 | 9 | 8 | 7 | 7 |
| YM5 | 11 | 29 | 32 | 32 | 31 |
| YF1 | 1 | 2 | 2 | 2 | 3 |
| YF2 | 1 | 3 | 3 | 3 | 3 |
| YF3 | 3 | 9 | 8 | 8 | 9 |
| YF4 | 3 | 8 | 9 | 8 | 10 |
| YF5 | 5 | 15 | 14 | 14 | 15 |
| Average OS | −2 | −5 | −6 | −6 | −8 |
| Average YS | 4 | 12 | 12 | 11 | 12 |
| Difference (OS average – YS average) | 6 | 17 | 18 | 17 | 20 |

OF = older female; OM = older male; OS = older speaker; YF = younger female; YM = younger male; YS = younger speaker.

TABLE 7. *Angle values of TRAP-STRUT and LOT-FOOT across age groups in raw Hertz and under three normalization conditions*

| Average angle values | Hz | *Nearey1* | *W&F* | *Lobanov* |
|---|---|---|---|---|
| TRAP-STRUT relative to horizontal, older speakers | 2 | 5 | 5 | 4 |
| TRAP-STRUT relative to horizontal, younger speakers | 41 | 66 | 66 | 67 |
| LOT-FOOT relative to vertical, older speakers | 32 | 15 | 14 | 11 |
| LOT-FOOT relative to vertical, younger speakers | 81 | 65 | 66 | 65 |

TABLE 8. *Mean differences in angles, older speakers compared with younger speakers*

| | Hz | *Nearey1* | *W&F* | *Lobanov* |
|---|---|---|---|---|
| TRAP-STRUT angle relative to horizontal | | | | |
| Mean differences between older and younger groups | 38 | 61 | 61 | 64 |
| Standard deviation | 12 | 25 | 26 | 30 |
| LOT-FOOT angle relative to vertical | | | | |
| Mean differences between older and younger groups | 49 | 51 | 52 | 54 |
| Standard deviation | 24 | 13 | 12 | 9 |

As well as examining absolute average values of angles for each age group, we also explored how well each normalization algorithm preserved the differences in angle juxtaposition across generations observed in the raw Hertz data. Table 8 presents the values for the *mean difference* between any one older speaker's angle configuration and the average of all younger speakers in the sample, for both TRAP-STRUT and LOT-FOOT, as well as the standard deviation around the mean, in raw Hertz and under three normalization conditions.[14] When the data are examined from this perspective, we also find a difference in performance for the normalization algorithms between the configurations. *Nearey1*, *W&F*, and *Lobanov* perform much better at replicating the LOT-FOOT configuration, which is based on angles calculated against the vertical. In this case, a mean difference between older and younger speakers' average values of 49 degrees is matched very well by the *Nearey1* result of 51, *W&F* result of 52, and the *Lobanov* result of 54 degrees. Note also that the standard deviations around these averages are smaller than the corresponding raw Hertz values. All three methods perform less well with the TRAP-STRUT configuration, because a Hertz average difference of 38 degrees becomes 61 and 64 under normalization, with a greater standard deviation than the raw Hertz value.

Although the results presented here are only exploratory, they suggest that this parameter of vowel means juxtapositions could be worth exploring in further studies. Our observations of comparisons between *W&F*, *Lobanov*, and *Nearey1* suggest that all three algorithms perform approximately equally well at preserving the angle relationships seen in the raw Hertz data configurations if these values are calculated relative to the vertical. No method stands out among them in the case examined here: that of the LOT-FOOT juxtaposition. However, in the case of angles calculated relative to the horizontal, as seen with DRESS-LOT and TRAP-STRUT, all methods reconfigure the mean points such that angles under normalization diverge to a greater or lesser extent from the raw Hertz angles. To summarize the results of Test 3, we conclude that even though we see differences between the raw Hertz data and under the four normalized conditions, all four normalization algorithms either preserve relationships equally well or change angle configurations in similar ways. On that basis, we see no reason to scale the performances under Test 3 and thus regard all four algorithms as performing equally well. Note that earlier studies (Fabricius, Watt, & Johnson, 2008; Fabricius, Watt, & Yurkova, 2008) comparing *Nearey2* (called *Nearey* in those papers) with *W&F* and *Lobanov* showed that *Nearey2* achieved very high rates of replication of Hertz angle configurations; however, importantly, this angle match-up was not accompanied by better equalization of vowel space areas or improved overlap/coextensiveness of the vowel polygons, leading to an overall downgrading of *Nearey2*'s performance in earlier work. *Nearey1*, applied to the RP and Aberdeen English data of this study, performs similarly to *Lobanov* and *W&F*/*mW&F* on Test 3. Moreover, optimal replication of angle configurations seems to be obtained when angles are calculated relative to the vertical.

## CONCLUSIONS

As stated earlier, the principal aim of this study is to evaluate the Watt & Fabricius (2002) *S*-centroid procedure in relation to two other typologically similar normalization algorithms, *Lobanov* and *Nearey*, and to illustrate its performance over different vowel systems and data sets. The comparisons contained in the three tests presented suggest that all of the three speaker-intrinsic, vowel-extrinsic, and formant-intrinsic procedures we have examined here have strengths and weaknesses in relation to the three individual tasks, all of which evaluate performance in obtaining optimal conditions for visual comparisons of vowel plots. Our results show that whereas *Lobanov* is the most successful technique with regard to improving overlap and optimizing area ratios between pairs of speakers, Watt & Fabricius's method performs nearly as well, and in some cases better than Nearey's $CLIH_{i2}$, here labeled *Nearey1*. Our results also show that the Watt & Fabricius *S*-centroid method performs similarly to *Lobanov* and *Nearey1* in the angle preservation comparisons. Moreover, all three algorithms perform well at preserving angles calculated against the vertical dimension. On the basis of the tests we have carried out in the present article on two data sets, then, this study concludes that the *S*-centroid *W&F* procedures perform *at least as well* as the two most-recognized speaker-intrinsic, vowel-extrinsic, formant-intrinsic normalization methods, *Lobanov*'s (1971) *z*-score procedure and *Nearey's* (1978) individual log-mean procedure ($CLIH_{i4}$ in Adank [2003], $CLIH_{i2}$ as tested here), and in some test cases consistently better than *Nearey1*.

*W&F*'s potential advantage over other methods could lie in the economy of data needed to perform the algorithm, because over and above the measurements being examined, it requires data only from the vertices of a triangular vowel space,[15] not the entire vowel space (however one might define that; see Watt & Fabricius 2002 and previous discussion for details, and for further discussion, Fox & Jacewicz 2008). This economy may have a high priority for an empirical study if large numbers of speakers are used (e.g., Bigham, 2008; Kamata, 2008). Second, the *S*-centroid procedure is a flexible speaker-intrinsic method that can be adapted for the language variety concerned; see, e.g., Bigham (2008), where Illinois English data were normalized using a centroid derived from four and not three vertices, because of the typical vowel configurations of that accent variety. In this way, the normalization procedure can be adapted to varieties of a language with a more quadrilateral than triangular vowel space.

Our message here, echoing Thomas & Kendall (2007), is also that the choice of a normalization algorithm for any one data analysis scenario should remain more a matter of inductive testing of individual data sets than deductive decree and as such, will be dependent on a range of factors. As Thomas & Kendall (2007) point out, all vowel-extrinsic normalization methods operate optimally when all the vowels of speakers' vowel systems are included.[16] A researcher must also consider the purpose of the analysis itself in any particular empirical case. If the purpose of the study is to examine the placement of vowel means or clouds of tokens across

generations because of suspected chain shifts or mergers in progress, the choice of normalization algorithm can have repercussions for the configurational relationships seen in the data. Researchers thus need careful guidelines that indicate what the potential consequences of the choice between various algorithms might be for their particular data set, and need to test this thoroughly for themselves. A guide to best practice for normalization in sociophonetics (see Watt, Fabricius, & Kendall, forthcoming) must point the way toward empowering researchers to resolve these types of complex questions.

## NOTES

**1.** This is not to deny the importance of perceptual factors, but our principal interest here is not to try to simulate in any direct way how listeners process phonetic input.

**2.** The NORM suite can be accessed at: http://ncslaap.lib.ncsu.edu/tools/norm/.

**3.** Comparisons of removal of variance in vowel space areas and comparisons of overlap can be viewed as mutually reinforcing, in that they test related goals. Achieving greater size agreement between vowel polygons and achieving more extensive overlap of vowel spaces is not necessarily mathematically the same process, but in the context of examining vowel formants in the $F_1$-$F_2$ plane, the two processes are closely related and reinforce each other.

**4.** One anonymous reviewer points out to us that Thomas & Kendall's (2007) critique of vowel-extrinsic normalization procedures has two nuances. First, that the different weightings of different vowel systems between front and back vowels can skew comparability between normalized vowel sets (a critique that applies to the Nearey and Lobanov normalization procedures), and second, that the Watt & Fabricius method does not have a way to deal with the fact that some vowel systems are more quadrilateral in shape than triangular, and that the $F_1$ value of /a/ in that procedure risks being misleading as a corner point that "imposes" a triangular shape on a otherwise quadrilateral system. Bigham's (2008) study of Illinois English deals with this latter problem by using a quadrilateral frame for the calculation of a (Watt & Fabricius-inspired) *S*-centroid normalization procedure, and this type of solution would be one we recommend as being true to the spirit of the *S*-centroid method.

**5.** TRAP has lowered and retracted, STRUT is now often located immediately above TRAP instead of behind it, and FOOT has fronted (and unrounded) in the course of the 20th century (Fabricius, 2007a, 2007b).

**6.** Normalization procedures for the NORM suite can be found at: http://ncslaap.lib.ncsu.edu/tools/norm/norm.php.

**7.** For details of the scaling factor, see: http://ncslaap.lib.ncsu.edu/tools/norm/about_norm1.php.

**8.** See further the General Polygon Clipper library at: http://www.cs.man.ac.uk/~toby/alan/software/.

**9.** The Pitman-Morgan test operates on a slightly different version of these numbers, as it first adjusts each area by dividing it by the mean of the areas being compared. The set of adjusted area figures for each normalization method (now centered around 1) are compared with those for each of the other normalization methods in turn. For each comparison, the Pitman-Morgan test returns a *p* value for the null hypothesis that the two sets of figures have equal variances.

**10.** Earlier tests of *Nearey2* normalization with the same data (Fabricius, Watt and Yurkova., 2008) gave values of .071 for RP and .670 for Aberdeen English, respectively, so the two Nearey algorithms performed similarly for this task.

**11.** Because Nearey's normalization algorithms operate with a uniform scaling factor, they perform better on data sets where all values for Speaker X are proportionally related to all values for any Speaker Y. This is especially notably characteristic of the Aberdeen English data (6 speakers) to a greater extent than the RP data (20 speakers), where there is greater variation in the overall shapes of vowel polygons (because vowel shifts over generations are evident in this data).

**12.** Corresponding values for *Nearey2* were .445 and .583 for RP and Aberdeen English, respectively (Fabricius, Watt and Yurkova, 2008).

**13.** It could be argued that it would be desirable for a sociophonetic study to derive "normalized" vowel configurations that eliminate individual variation, in order to obtain an overall picture of the direction of community change, and thus that altering vowel configurations through normalization was in fact as unobjectionable as altering vowel space areas. This could indeed be the case, but it is also useful to know to in what manner such configurations in Hertz data are altered by normalization algorithms in the first place.

**14.** These difference values were calculated by comparing each individual older speaker to the group of younger speakers as a whole, that is, finding the individual differences between an older speaker and the average of the younger speakers' angle values, and then deriving the average difference between younger and older speakers for all of the older speaker group. Thus Table 8 shows how big the diachronic difference in angle juxtapositions was between the two age groups (on average 38 degrees for the TRAP/STRUT configuration, 49 degrees for the LOT/FOOT configuration).

**15.** Or quadrilateral vowel space, as employed in Bigham (2008).

**16.** For further discussion, see Thomas & Kendall (2007; URL: http://ncslaap.lib.ncsu.edu/tools/norm/about_norm.php) and the links to discussions of individual algorithms contained therein.

## REFERENCES


Adank, Patti. (2003). *Vowel normalization: A perceptual-acoustic study of Dutch vowels.* Ph.D. thesis, Katholieke Universiteit Nijmegen.

Adank, Patti, Smits, Roel, & van Hout, Roeland. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America* 116:3099–3107.

Bigham, Douglas. (2008). *Dialect contact and accommodation among emerging adults in a university setting*. Ph.D. thesis, The University of Texas at Austin.

Cohen, Ayala. (1990). Graphical methods for testing the equality of several correlated variances. *The Statistician* 39(1):43–52.

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

_____. (1997). The formants of monophthong vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association* 27:47–55.

Disner, Sandra. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America* 67:253–261.

Fabricius, Anne. (2007a). Variation and change in the TRAP and STRUT vowels of RP: A real time comparison of five acoustic data sets. *Journal of the International Phonetic Association* 37 (3):293–320.

_____. (2007b). Vowel formants and angle measurements in diachronic sociophonetic studies: FOOT-fronting in RP. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken. 1477–1480. Available at: http://www.icphs2007.de/.

Fabricius, Anne, Watt, Dominic, & Johnson, Daniel E. (2008). A new speaker-intrinsic vowel formant frequency normalization algorithm for sociophonetics. Paper presented at Acoustics08, Paris.

Fabricius, Anne, Watt, Dominic, & Yurkova, Jillian. (2008). A new speaker-intrinsic vowel normalisation algorithm for sociophonetics. Paper presented at BAAP 2008, University of Sheffield.

Fox, Robert A., & Jacewicz, Eva. (2008). Analysis of total vowel space areas in three regional dialects of American English. *Journal of the Acoustical Society of America* 123(5):3068.

Hawkins, Sarah, & Midgley, Jonathan. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association* 30:63–78.

Hindle, Donald. (1978). Approaches to vowel normalization in the study of natural speech. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 161–171.

Kamata, Miho. (2008). *An acoustic sociophonetic study of three London vowels*. Ph.D. thesis, University of Leeds.

Koopmans–van Beinum, Florien J. (1980). *Vowel contrast reduction: An acoustical and perceptual study of Dutch vowels in various speech conditions*. Ph.D. thesis, University of Amsterdam.

Labov, William. (1994). *Principles of linguistic change, volume 1: Internal factors*. Oxford: Blackwell.

Labov, William, Ash, Sharon, & Boberg, Charles. (2006). *The atlas of North American English: Phonology, phonetics, and sound change*. Berlin: Mouton de Gruyter.

Lobanov, Boris M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* 49(2B):606–608.

Mesthrie, Rajend. (unpublished manuscript). Socio-phonetics and social change: Deracialisation of the GOOSE vowel in South African English. Submitted.

Moreiras, Caroline. (2006). *An acoustic study of vowel change in female adult speakers of RP*. B.A. thesis, University College London.

Nearey, Terry. (1978). *Phonetic feature systems for vowels*. Ph.D. dissertation, University of Alberta. (published 1978, Indiana University Linguistics Club).

Rosner, Burton S., & Pickering, John B. (1994). *Vowel perception and production*. Oxford: Oxford University Press.

Thomas, Erik R. (2001). *An acoustic analysis of vowel variation in New World English*. Publication of the American Dialect Society 85. Durham, NC: Duke University Press.

_____. (2002). Instrumental phonetics. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *The handbook of language variation and change*. Oxford: Blackwell. 168–200.

Thomas, Erik R., & Kendall, Tyler. (2007). NORM: The vowel normalization and plotting suite. Available at: http://ncslaap.lib.ncsu.edu/tools/norm/index.php.

Traunmüller, Hartmut. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*. 88(1):97–100.

_____. (1997). Auditory scales of frequency representation. Available at: http://www.ling.su.se/staff/hartmut/bark.htm.

Watt, Dominic, & Fabricius, Anne. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1-F2 plane. *Leeds Working Papers in Linguistics and Phonetics* 9:159–173. Available at: http://www.leeds.ac.uk/linguistics/WPL/WP2002/Watt_Fab.pdf.

Watt, Dominic, Fabricius, Anne, & Kendall, Tyler. (forthcoming). More on vowels: Plotting and normalization. In M. di Paolo & M. Yaeger-Dror (eds.), *Sociophonetics: A student's guide*. London: Routledge.

Watt, Dominic, & Tillotson, Jenny. (2001). A spectrographic analysis of vowel fronting in Bradford English. *English World-Wide* 22(2):269–302.

Watt, Dominic, & Yurkova, Jillian. (2007). Voice onset time and the Scottish Vowel Length Rule in Aberdeen English. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany. 1521–124. Available at: http://www.icphs2007.de/.

Wells, John C. (1982). *Accents of English*. 3 vols. Cambridge: Cambridge University Press.

## APPENDIX

*Individual speakers' data,* TRAP-STRUT *and* LOT-FOOT *juxtapositions*

| Speakers | Hz | *Nearey1* | *W&F* | *Lobanov* |
|---|---|---|---|---|
| OM1 TRAP STRUT | −9 | −24 | −24 | −45 |
| OM2 TRAP STRUT | 7 | 20 | 19 | 20 |
| OM3 TRAP STRUT | 4 | 12 | 14 | 17 |
| OM4 TRAP STRUT | 3 | 9 | 9 | 10 |
| OM5 TRAP STRUT | 2 | 7 | 8 | 9 |
| OF1 TRAP STRUT | −3 | −8 | −9 | −13 |
| OF2 TRAP STRUT | 1 | 3 | 4 | 4 |
| OF3 TRAP STRUT | 4 | 12 | 13 | 17 |
| OF4 TRAP STRUT | −16 | −38 | −38 | −41 |
| OF5 TRAP STRUT | 30 | 56 | 57 | 58 |
| Average | 2 | 5 | 5 | 4 |
| YM1 TRAP STRUT | 32 | 61 | 61 | 65 |
| YM2 TRAP STRUT | 74 | 85 | 84 | 85 |
| YM3 TRAP STRUT | 28 | 56 | 56 | 53 |
| YM4 TRAP STRUT | 66 | 81 | 79 | 79 |
| YM5 TRAP STRUT | 27 | 55 | 58 | 57 |
| YF1 TRAP STRUT | 35 | 65 | 67 | 70 |
| YF2 TRAP STRUT | 42 | 69 | 68 | 68 |
| YF3 TRAP STRUT | 43 | 71 | 70 | 71 |
| YF4 TRAP STRUT | 25 | 56 | 57 | 60 |
| YF5 TRAP STRUT | 36 | 65 | 65 | 65 |
| Average | 41 | 66 | 66 | 67 |
| OM1 LOT FOOT | 53 | 25 | 25 | 12 |
| OM2 LOT FOOT | 14 | 5 | 5 | 5 |
| OM3 LOT FOOT | −5 | −2 | −2 | −1 |
| OM4 LOT FOOT | 43 | 17 | 16 | 15 |
| OM5 LOT FOOT | 50 | 23 | 20 | 18 |
| OF1 LOT FOOT | 19 | 7 | 6 | 4 |
| OF2 LOT FOOT | 46 | 20 | 17 | 14 |
| OF3 LOT FOOT | 68 | 39 | 37 | 29 |
| OF4 LOT FOOT | 5 | 2 | 2 | 2 |
| OF5 LOT FOOT | 29 | 12 | 11 | 11 |
| Average | 32 | 15 | 14 | 11 |
| YM1 LOT FOOT | 83 | 70 | 69 | 66 |
| YM2 LOT FOOT | 80 | 63 | 64 | 62 |
| YM3 LOT FOOT | 81 | 67 | 68 | 69 |
| YM4 LOT FOOT | 75 | 53 | 58 | 60 |
| YM5 LOT FOOT | 80 | 65 | 62 | 62 |
| YF1 LOT FOOT | 80 | 62 | 59 | 55 |
| YF2 LOT FOOT | 86 | 78 | 79 | 78 |
| YF3 LOT FOOT | 85 | 75 | 76 | 75 |
| YF4 LOT FOOT | 74 | 49 | 47 | 44 |
| YF5 LOT FOOT | 84 | 73 | 73 | 73 |
| Averages | 81 | 65 | 66 | 65 |