Published in final edited form as:

J Acoust Soc Am. 2008 January ; 123(1): 336–346. doi:10.1121/1.2816579.

Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels: Statistical considerations

Michel T.-T. Jackson^{a)} and Richard S. McGowan

CReSS LLC, 1 Seaborn Place, Lexington, Massachusetts 02420

Abstract

In a re-analysis of x rays of speakers producing Swedish vowels, midsagittal pharyngeal dimensions were predicted from anterior tongue positions using procedures based on estimated tongue pellet positions. Principal component analysis was used to reduce the number of pellet degrees of freedom from eight to three prior to applying linear regression from these three independent variables to dependent vocal tract midsagittal cross distances. Except for the regions around the laryngopharynx and uvula, the pharynx dimensions were predictable from linear regressions and were significant at the $p < 0.05$ level. Numerical experiments show that it is crucial to reduce the number of independent variables in tests of statistical significance.

I. INTRODUCTION

One goal of phonetics is to quantitatively understand articulation in human speech production. As part of this goal, the authors are interested in the extent to which some vocal tract dimensions—specifically, midsagittal pharyngeal cross distances—are predictable from more readily observed articulatory measures. The predictability of midsagittal pharyngeal cross distances is of interest for two reasons.

First, a number of modern articulatory research techniques yield high-quality data on the position and movement of the anterior articulators (including the anterior portion of the tongue, the lips, and the upper and lower incisors) during speech, but cannot be extended to the posterior portion of the oral cavity and pharyngeal articulation. These techniques include x-ray microbeam pellet tracking (e.g., Kiritani, 1986; Westbury, 1994), electromagnetic midsagittal articulometry (e.g., Perkell *et al.*, 1992), and optical tracking (e.g., Ramsay *et al.*, 1996). To the extent that pharyngeal dimensions can be predicted or estimated from the positions of the other articulators, including the anterior of the tongue, a “missing” portion of the articulation can be recovered from these data sets. For instance, Kaburagi and Honda (1994) showed that much of the anterior portion of the midsagittal tongue shape, including regions between coils and just posterior to the rearmost coil, can be predicted from electromagnetic midsagittal articulometry data for one speaker of Japanese CVCs, where the Cs had labial, alveolar, and velar places of articulation.

Second, the degree to which pharyngeal dimensions are predictable from other articulatory information is important in attempting to recover articulatory positions from speech acoustics. There are some instances in which the position of a few points on the tongue has been shown

a) Author to whom correspondence should be addressed. Electronic mail: ladmttj@ix.netcom.com.

PACS number(s): 43.70.Bk, 43.70.Kv [BHS]

to be highly correlated with the acoustic output or other articulatory variables. For instance, Nguyen *et al.* (1994) showed that coordinates of three points on the tongue predict the acoustic output of the vocal tract in [s] and [ʃ] in French, and Lindau-Webb and Ladefoged (1989) showed that as few as two fleshpoint positions suffice to determine the entire tongue shape in static American English vowels. Whalen *et al.*, (1999), using the coordinates of four manually estimated fleshpoints from magnetic resonance (MR) images of 11 American English vowels, showed high correlations between these eight variables (x and y coordinates of each fleshpoint) and midsagittal vocal tract cross distances measured along gridlines.

However, there are two reasons to treat these results with caution: First, there are languages in which pharyngeal dimensions—and therefore the acoustic output of the vocal tract—cannot be predicted from anterior tongue position. Tiede (1996) showed that the phonological advanced tongue root (ATR) contrast introduces an articulatory degree of freedom in the pharynx that makes the pharyngeal dimensions vary independently of the anterior portion of the tongue's position. Tiede (1996) also showed that this pattern of variation is different from the one observed, as for example, in American English. Thus, the results presented above have limited interpretive power with regard to phonological considerations. In particular, it would not be expected that pharyngeal cross distance can be predicted from anterior tongue position for languages with ATR contrast.

Second, there are several statistical reasons for treating these results with caution. In these studies, the number of observations on which t tests and F tests are based is small. But most statistical tests, including t and F tests, are based on a theory that requires large numbers of observations. Furthermore, many statisticians (see, for example, Hill and Lewicki, 2006), recommend that multiple regressions have more than ten times as many observations of the dependent variables as there are independent variables. Having a number of independent variables too close to the number of observations in a regression affects the numerical stability of the regression, may lead to overfitting, and increases the likelihood of capitalizing on chance in finding a high correlation in the results.

Similarly, parametric statistical tests, again including t tests and F tests, depend on some assumption of normality in the data sets. By definition, the normal distribution extends to positive and negative infinity. Since the midsagittal cross distance has a minimum possible value of zero and a maximum around 30 or 40 mm, the distribution of the dependent variable in the regressions discussed above *cannot* be normal. Common transformations, such as converting to z scores, do not improve the normality of the distribution of these data as measured by skewness, kurtosis, and higher-moment measures of a distribution's normality (Cohen *et al.*, 2003). Thus, the significance of correlations reported in some speech production studies is open to debate due to the small number of observations and the non-normality of the data. For instance, studies such as Jackson (1988) with 16 observations (vowels) per speaker, and Whalen *et al.* (1999) with 11 observations, typically report high correlations, over 0.95. However, studies using more data, such as Beaudoin and McGowan (2000) with over 3000 observations per speaker, report correlations of around 0.9.

Instead of relying solely on t and F tests, a series of numerical experiments was designed to examine the likelihood of finding our observed results under suitable null hypotheses. These numerical experiments calculate distributions of R in multiple regressions predicting midsagittal pharyngeal cross distance for various numbers of uniformly distributed, randomly generated independent variables. R is the correlation between a dependent variable and an optimally weighted combination of two or more independent variables (Cohen *et al.* 2003). These distributions can be used to determine significance levels for various values of R .

Finally, caution is necessary because the measured or estimated fleshpoint positions, which are the independent variables, are strongly correlated with each other. This produces multicollinearity, which is a well-known problem in multiple regression. Multicollinearity is the situation in which two or more independent variables are strongly correlated with each other. In this situation, regression coefficients are not reliable (Cohen *et al.* 2003).

These statistical concerns were addressed using principal components regression (Cohen *et al.* 2003, p. 428), which uses principal component analysis (PCA) to produce a reduced number of orthogonal (uncorrelated) independent variables from the original independent variables (pellet coordinates). By orthogonalizing the independent variables, multicollinearity is eliminated. Further, reducing the number of independent variables, relative to the number of observations, increases the power of statistical tests. Multiple regression using PCA components as independent variables is known as principal components regression. Previous studies (Shirai and Honda, 1978; Sekimoto *et al.*, 1978; Maeda, 1978, 1990; Beaudoin and McGowan, 2000; Story, 2007) have also used PCA and combinations of PCA with other forms of factor analysis (Hoole, 1999) to study midsagittal tongue shape, particularly in vowel production. Shirai and Honda (1978) used 50 vowel tokens from ten different speakers and Sekimoto *et al.* (1978) used over 1000 vowel tokens from two different speakers. The other studies segregated the observations by speaker, with Maeda (1978) using 77 observations from frames of an x-ray movie, Maeda (1990) using about 1000 observations of vowel production in sentences per speaker, Beaudoin and McGowan (2000) using over 3000 observations per speaker including consonant-vowel transitions, and Story (2007) using data from the centers of 11 vowels and the frames from six vowel pairs per speaker. In his hybrid approach, Hoole (1999) appears to have used 90 observations per speaker. These studies have shown that the number of degrees of freedom for the midsagittal tongue can be reduced to two or three, accounting for between 80% and 97% of the variance in tongue fleshpoint coordinates.

An overview of the current work is as follows. In order to investigate the predictability of midsagittal pharyngeal dimensions, midsagittal vocal tract cross distances were measured along gridlines and fleshpoint positions were estimated on the anterior of the tongue from x-ray tracings. The x-ray tracings come from previous studies (Sundberg 1969, Fant 1965). Principal components regression was used to predict the dependent midsagittal vocal tract cross distances. Numerical experiments were used to produce distributions of R under suitable null hypotheses. The R values from the regressions were compared with these distributions and were examined in order to determine significance levels.

II. PROCEDURE

A. Image acquisition, processing, and coordinate system

The data are tracings of static x-ray images of spoken vowels produced by four speakers of Swedish. The first three speakers are described in Sundberg (1969); photocopies of the original x-ray tracings were re-analyzed in this study. The fourth speaker is described in Fant (1965); photocopies of the x-ray tracing figures as prepared for publication were re-analyzed. The vowel tracings analyzed in this study are summarized in Table I. Each tracing was digitized on a flatbed scanner at a resolution of 300 dpi to produce a digital image that was subsequently contrast enhanced using the MATLAB Image Processing Toolkit.

A notable feature of the tracings of the first three speakers (from Sundberg, 1969) is that they include registration marks that provide accurate image alignment. For each speaker, tracings were translated, rotated, and scaled to bring the registration marks into coincidence using the MATLAB Image Processing Toolkit. The degree of scaling was manually checked by examining the coefficients of the MATLAB image transformation matrix. In all cases, the degree of scaling was less than 5%, indicating that the image acquisition chain, including

photocopying original tracings, scanning and digitization, did not introduce notable image distortions. The accuracy of image alignment was verified by checking that the tracings of the static hard structures of the vocal tract, especially the upper incisors and hard palate, overlapped. The first two rows of Table II summarize the results of comparing the positions of the tips of the upper incisors outline in these images. Since the checkpoint, the tips of the upper incisors, was a considerable distance—approximately 100 mm—from the registration marks that were used for alignment, this was considered a stringent check on the accuracy of the overall image alignment.

After satisfactory image alignment was obtained, an absolute scale in pixels per cm was obtained for the first three speakers from the original tracings' measurements of the length of the midline of the vocal tract. The mean scale of all the images for the first three speakers was 4.6 pixels/mm; the standard deviations were all reported less than 0.6 pixels/mm. The small standard deviations reflect the accuracy of the original tracings and measurements.

For the fourth speaker (Fant, 1965) the tracings do not include registration marks. The scanned images of these tracings were translated to bring the tip of the upper incisor on each image into coincidence. Then, between five and ten points on the superior surface of the maxilla on each image were selected, and the best-fit line through these points (equivalent to the major axis of the distribution of the points) was determined. Each image was then rotated to bring this line, which approximated the upper surface of the maxilla, to horizontal. The images were not scaled relative to one another; the scale was stipulated to be 4.5 pixels/mm to make the measurements on these images comparable to the measurements on the images from the first three speakers. Since there were no lengths measured on the original tracings an absolute scale could not be determined, but all the images were scaled so as to make the range of measurements comparable with the range of measurements found among the other speakers.

From each image, more than 300 outline points on the upper and lower lip surfaces, hard palate and upper incisor outline, soft palate outline, tongue, laryngeal structures, and dorsal wall of the pharynx were selected from digitized images. Points on the outlines of the other upper teeth, especially the molars, were also selected where available.

Adopting Westbury's (1994) coordinate system, the maxillary occlusal plane (MaxOP) was estimated as the horizontal axis and the tip of the upper incisors as the origin. For the first three speakers (Sundberg, 1969), the MaxOP was determined from the x-ray tracing that had the most extensive tracing of the upper incisors and other upper teeth. The MaxOP was determined by finding the line through the tip of the upper incisors that was tangent to the outline of the other upper teeth. For the fourth speaker (Fant, 1965), because the outline of the upper teeth was not traced, the MaxOP was estimated by rotating the images to make the upper surface of the maxilla horizontal and passing the horizontal axis through the tip of the upper incisor.

B. Reference points and measurement grids

A measurement grid was constructed for each speaker. Each measurement grid was based on four fixed reference points, P1 through P4, similar to the ones used in Whalen *et al.* (1999). However, Whalen *et al.* (1999) analyzed MR images, whereas this study analyzes tracings of x-ray images, leading to differences in choosing anatomical landmarks.

Teeth do not image well in MR, because conventional MR imaging relies on the magnetic resonance of hydrogen nuclei, and the density of hydrogen atoms in the enamel and dentin of teeth is low. However, bone does image well and thus the bony alveolar ridge can be imaged. In x rays, both teeth and bone project shadows, and the shadow of the roots of the teeth may obscure the shadow of the surrounding bony alveolar socket and vice versa. In order to pick P1 to be on the alveolar ridge, an inflection point on each tracing at the juncture between the

incisor and the alveolar socket in the maxilla was identified visually (see Fig. 1). The standard deviations of the coordinates of this point over all the vowels for each speaker are reported in rows three and four of Table II. The standard deviations are generally less than 1 mm. The mean position of these points for each speaker was used as reference point P1.

The most superior point of the hard palate, P2, is well defined because of the definition of vertical in the coordinate system based on the MaxOP. The highest point on the bicubic spline tracing the outline of the hard palate was determined automatically for each image. Although it is possible that the hard palate could contain a small region that appears to be horizontal, in fact none were found, and the highest point on the hard palate was unique on each tracing. The standard deviations of the coordinates of this point over all the vowels for each speaker are reported in rows five and six of Table II. The mean position of these points for each speaker was used as reference point P2.

The definitions of P3 and P4 used landmarks that could be found in the x rays. P3 is the point on the rear pharyngeal wall that is at the same height as the most anterior point of the anterior tubercle of the atlas. A point corresponding to P3 was identified on each image for each speaker, and the mean position (within speaker) was used as the fixed location of P3 in subsequent analysis. Following Tiede (1996), who described an earlier version of the image-measurement methods used in Whalen *et al.* (1999), the point on the rear pharyngeal wall at the same height as the point at the bottom of the vallecular sinus was adopted to be P4. Table II reports the standard deviations of the x and y positions of P3 in rows seven and eight; P4 in rows nine and ten.

Figure 1 shows the results of applying these procedures to pick reference points for one speaker (BE). The figure shows the joint (x, y)-distribution of these points in all the vowels produced by this speaker by plotting the ellipses where the semimajor and semiminor axes are one standard deviation. The center of the circle passing through P1, P2, and P3 is denoted P0, which is an auxiliary point required to construct the set of gridlines for each speaker.

C. Gridlines

There are three groups of gridlines in this procedure: gridlines in the lower pharynx, which are parallel to each other; gridlines in the upper pharynx and posterior oral cavity, which are radii of the circle centered on P0, and gridlines in the anterior oral cavity, which are also parallel.

1. Lower pharynx—The gridlines of the lower pharynx were specified to be perpendicular to the P3–P4 line, 3 mm apart (Fig. 2). They are constructed so that the topmost gridline passed through P0 (i.e., it is a radius of the circle passing through P1, P2, and P3) and the lowest gridline was specified to be less than 3 mm above the position of the vocal folds in the vowel that had the highest vocal fold position. In this way, there was no missing data on any of the lower gridlines due to the vocal folds having risen above them. These gridlines are approximately perpendicular to the axis of the pharynx and measure the cross distance from the posterior surface of the tongue to the dorsal wall of the pharynx.

Where there appeared to be an acoustically relevant airspace posterior to the opening of the larynx (e.g., posterior to the arytenoid cartilages) in the x-ray tracing, the midsagittal cross distance of that airspace was included in the measured gridline cross distance, but the thickness of the tissue separating the larynx from the posterior airspace was excluded. Similarly, when the vallecular sinus, between the tongue and the epiglottis, was open, the width of the airspace between the tongue and the epiglottis was included but the thickness of the epiglottis itself was excluded.

2. Upper pharynx and posterior oral cavity—The second set of gridlines was constructed using the topmost of the previous gridlines as a starting point. From that gridline, which passes through P0, radial gridlines centered at P0 were constructed every 5° until the gridlines were within 5° of being perpendicular to the P1–P2 line segment, i.e., until the gridlines (nearly) coincided with the perpendicular bisector of the P1–P2 line segment. These gridlines cover the region of the oral cavity—pharyngeal cavity junction. As in the lower pharynx, the width of any airspace posterior to the uvula was included in the cross distance measured along the gridline, but the thickness of the uvula itself along the gridline was excluded.

3. Anterior oral cavity—From the last radial gridline, which was within 5° of being perpendicular to the P1–P2 line, gridlines perpendicular to the P1–P2 line were placed every 3 mm as far as P1. These gridlines measure the cross distance from the anterior portion of the tongue to the anterior portion of the hard palate.

Because of variations in overall vocal tract length, a different number of gridlines was produced for each speaker. In addition, varying proportions resulted in a different number of gridlines in the pharynx and oral cavity for each speaker. The gridlines for the same speaker (BE) illustrated in Fig. 1 are shown in Fig. 2.

D. Estimated fleshpoint locations and principal components regression

The procedure in Nix *et al.* (1996) estimates the position (x and y coordinates) of 13 equally spaced points along each tongue curve from traced x-ray images similar to the ones analyzed in this paper. The result is 13 estimated fleshpoint locations, for a total of 26 coordinates. In the present work, their procedure was used to estimate the length of the tongue surface in the midsagittal plane, and then to estimate the location of four fleshpoints.

Numerical integration was used to estimate the length of the tongue surface from the bicubic spline interpolating the selected outline points described in Sec. II A (not points on gridlines) from tongue tip to the point at the bottom of the vallecular sinus. This distance was then divided by 13, effectively dividing the portion of the tongue between the tip and the root of the epiglottis into 13 sections. Estimated fleshpoint locations were then calculated at 1.5, 3.0, 4.5, and 6.0 times this section length. These values were taken in an attempt to roughly simulate typical fleshpoint locations on the anterior of the tongue. The x and y coordinates of each fleshpoint were recorded for each vowel and used as the raw independent variables for the data analysis.

PCA of the fleshpoint coordinates was performed for each subject. The PCA components that accounted for the largest amount of the variance were employed as independent variables in predicting all the midsagittal cross distances in a principal components regression. The principal components regressions reduced the number of independent variables from the eight Cartesian tongue pellet coordinates, as well as alleviating any multicollinearity. These regressions were performed using the SPSS statistical package.

One of the assumptions underlying regression models and significance tests is that the values of the independent variable in the regression are under experimental control and known precisely. However, in this study, the quantities used as independent variables—the x - and y -fleshpoint coordinates—are estimates and not known precisely.

There are two notable sources of errors in the estimation of fleshpoint coordinates. First, of course, there are possible errors in the original x-ray tracing and its duplication (photocopying, scanning, and digitization). These errors are intrinsic to the data and there seems to be no way of estimating their magnitude or effect without independent instrumental control. Another kind of intrinsic uncertainty in the x-ray tracings is overall scaling; however, the presence of

registration marks in some of the original tracings and the degree of overlap of the hard structures of the vocal tract from tracing to tracing provides a relatively strong control on this kind of uncertainty.

A second kind of error would occur if the tongue has substantial nonuniform longitudinal dilation or contraction during vowel production. If, e.g., the surface of the blade of the tongue were to stretch while the surface of the tip of the tongue contracted, the true fleshpoint positions would change in a way that could not be estimated by our algorithm. However, we find no literature suggesting that stretching of the anterior tongue surface occurs in speech. This evidently remains a topic for further investigation.

Despite these possibilities, we believe that the error in the estimates of fleshpoint positions is small. This is mainly because the PCA has residual errors that are comparable to residual errors found in other studies.

E. Numerical experiments

The goal of these numerical experiments was to empirically determine the distribution of R given the measured dependent variables, the midsagittal cross distances along the gridlines, under the hypothesis that the values of the independent variables, the PCA components, are drawn from a uniform random distribution. This distribution of R is, in effect, the expected distribution of R under a null hypothesis of no underlying relationship between the independent and dependent variables. These distributions provide an empirically derived relation between the value of R and the probability of obtaining that value of R by chance, since with random values of the independent variables, there is no relation between the independent variables and the midsagittal cross distances. A particular interest was the effect of the number of independent variables on the statistical significance of the multiple regressions.

A computer program was written to perform this experiment by generating different numbers of uniformly distributed random variates to use as independent variables in a regression model for the midsagittal pharyngeal cross distances for each speaker. The program then calculated the fraction of the total variance accounted for by the regression model (i.e., the R^2), and from that, the R value that would have been obtained in multiple regression. This program performed regressions 10 000 times for each gridline, and sorted the resulting R values into 100 bins, that is, into percentile distributions. The statistical significance (i.e., the probability of incorrectly rejecting the null hypothesis) of R values can be determined from these distributions. For instance, if 5% of the simulated regressions possess an R value greater than 0.92, then the null hypothesis can be rejected when $R > 0.92$ with a 5% probability that, in fact, the null hypothesis is true.

For each speaker, a range of numbers of independent variables was simulated. Simulating the x and y coordinates of the four estimated fleshpoints would yield eight independent variables. However, there cannot be more independent variables than one less the number of observations. Since some of the speakers in this study produced fewer than nine vowels (see Table I), eight independent variables could not be simulated for those speakers (see row two of Table III). Further, numerical degeneracy also arose in some cases when the number of independent variables was two or three fewer than number of vowels, further reducing the maximum number of independent variables that could be simulated. Finally, it was stipulated that the a significance level of $0.01 < p$ should be attained with $R < 0.99$. With these constraints, the potential maximum number of independent variables simulated for each speaker is shown in the third row of Table III. The minimum number of independent variables simulated for each speaker was three.

III. RESULTS

A. PCA components

The cumulative variance accounted for by the principal components is summarized in Table IV for each speaker. Table IV shows that the first three PCA components account for more than 99% of the variance in the estimated fleshpoint positions. The remaining components clearly account for negligible amounts of the variance. Therefore only the first three components derived by PCA were retained for the remaining analysis. With the small number of observations (vowels) per subject it would have been possible to proceed with two principal components accounting for at least 97% of the variance. However, in other circumstances with small numbers of observations, three components would need to be retained. This could occur if a set of observations contains consonant-vowel transitions and rhoticized articulations (cf. Hoole, 1999). It will be shown that even with a small number of observations (between 7 and 13), a reduction to three PCA components is sufficient for statistically meaningful results.

The “shapes” of the first three PCA components, in terms of the pellet displacements from mean position associated with variation in each component, are shown in Fig. 3 for the same speaker (BE) as illustrated in Fig. 1. The component shapes are plotted together with the mean position averaged across all vowel tracings. Although we are not focusing on the interpretations of these components in this study, the first component resembles the “front raising” component of Harshman *et al.* (1977). This has been found to be true in a number of other studies, e.g., Shirai and Honda (1978) Sekimoto *et al.* (1978), Maeda (1978), Zerling (1979), Jackson (1988), and Beaudoin and McGowan (2000). The second and third components are more variable. However, there is evidence of a “back raising” component and either a “bunching” or a “tongue blade raising” component.

B. Principal component regressions predicting midsagittal vocal tract cross distances

Multiple regression models for each speaker were constructed with the first three PCA components as independent variables and midsagittal vocal tract cross distances measured along the gridlines as dependent variables using SPSS. The model sum of squares, error sum of squares, R values, and the R threshold corresponding to the $p < 0.05$ significance level according to the F test for each gridline are plotted in Fig. 4–Fig. 7 for the four subjects. In general the R values are over 0.9 except in the regions of the laryngopharynx and uvula. In particular, the midsagittal pharyngeal cross distances that are the focus of this study generally exhibit high values of R that are significant according to an F test.

The gridline construction procedure distributes different numbers of gridlines in different parts of the vocal tract for each speaker (see Fig. 4–Fig. 7). For simplicity, Table V reports the mean R for the gridlines listed as being in the pharyngeal region of each speaker in Fig. 4–Fig. 7. All of these values are over 0.9. The predictability of the vocal tract cross distances measured along the gridlines in the oral cavity is extremely high, which is not surprising since the estimated fleshpoints are also in the oral cavity.

Closer examination of the results reveals other trends. First, the variance of the midsagittal cross distance, as measured along the gridlines, is lowest in the laryngopharyngeal and uvular regions for all four speakers. These are also areas in which the F test suggests the regression model to be nonsignificant. Another trend worth noting is that even when R is fairly high, the regression model can be nonsignificant according to the relevant F test. For example, Fig. 5 shows that gridline 9 for speaker JS has an R over 0.92 but nonetheless, fails to achieve significance at the $p < 0.05$ level.

C. Numerical experiments

Figure 8 plots the distribution of R resulting from the numerical experiments for each gridline together with the $p < 0.05$ (95th percentile) and $p < 0.01$ (99th percentile) levels, and the thresholds for the same significance levels according to F tests (calculated according to Hays, 1981). For speaker BE (Fig. 8(a)) with six independent variables, the $p < 0.05$ significance level according to the F test is at $R \approx 0.97$. But the empirically derived 95th percentile of the R distribution—the dark gray band in the figure—requires a larger $R (\approx 0.98)$. Even for the speaker with the most vowels (observations), F, the 95th and 99th percentile levels derived from the numerical experiments require a larger R than the $p < 0.05$ and $p < 0.01$ significance levels according to the F test. Figure 8(d) shows the results with eight independent variables. The $p < 0.05$ significance level according to the F test is at $R \approx 0.96$; the empirically derived 95th percentile of the R distribution is ≈ 0.98 . Similarly, the $p < 0.01$ significance level according to the F test is $R \approx 0.985$, but the empirically derived 99th percentile level is $R \approx 0.99$. Similar results obtain for speakers JS (Fig. 8(b)) and RL (Fig. 8(c)).

On the other hand, when the number of independent variables is limited to only three, the empirically derived 95th and 99th percentile levels differ from the thresholds according to F test by less than 0.001. As the number of independent variables decreases, the empirical thresholds and the F -test thresholds converge. Thus, in Fig. 4–Fig. 7, the dashed lines denoting the R level for $p < 0.05$ significance remain virtually unchanged as a result of the numerical experiments.

IV. DISCUSSION

This study confirms that three, or even two, degrees of freedom can describe anterior tongue position during vowel production in Swedish. Three principal components suffice to describe more than 99%, and two principal components more than 97%, of the variance in the anterior tongue position (as measured by pellet positions) in these vowels for each speaker. In addition, it has been shown that these measures of the anterior tongue position can be sufficient to describe tongue position in the pharynx, excluding the regions of the laryngopharynx and uvula. At least for the Swedish vowels examined here, restricting the data to anterior tongue pellet positions would not lose any essential part of the pattern of variation of tongue position in the mid-pharynx; there is no component or “gesture” that affects the posterior of the tongue to the exclusion of the anterior. This is shown by the fact that the cross distance of the posterior portion of the tongue from the rear pharyngeal wall can be predicted from the anterior pellet positions, via principal components analysis and regression equations.

This study does not attempt to formulate the components of tongue position exactly in Swedish vowel production. We do not know, in fact, whether or not different speakers exploit similar or different articulatory possibilities to produce roughly similar acoustic output. Furthermore, in languages which have pharyngealized or ATR vowels (cf. Jacobsen 1978; Lindau 1979; Tiede 1996), there will be different degrees of freedom or components, but there seem likely to be at least three. In consonants there may also be additional degrees of freedom (cf. Zerling 1979; Hoole 1999).

The results of the numerical experiments show that it is somewhat more difficult to demonstrate significance with this data than it might appear. First of all, it is not possible to construct numerically valid regression models with eight independent variables for all speakers. Second, the numerical experiments with the maximum number of independent variables result in thresholds for the $p < 0.05$ significance level ranging from $R > 0.96$ to $R > 0.98$. But the threshold R value according to the F test can be calculated, and in each case, the F test gives a more liberal, and potentially erroneous, threshold for significance with these data than the numerical experiment. Multiple regression models relating four pellet positions with eight

independent variables to midsagittal pharyngeal cross distances may be valid, but simple F tests are not appropriate for determining significance levels because the data do not obey the assumption of multivariate normal distribution, and the number of observations is small.

For multiple regression models with only three independent variables derived via PCA, the values of R accorded significance by numerical experiment are virtually the same as the values of R judged significant by the F tests. Thus, the approach to determining significance levels by numerical experiment is more conservative than the F test, but they converge as the ratio of number of observations-to-independent variables increases.

This study shows that it is methodologically difficult to justify using raw x and y coordinates from fleshpoint positions (x-ray microbeam pellets, electromagnetic midsagittal articulometry sensor coils) as independent variables in statistical analysis of static vowels. The problems induced by the high correlations among the independent variables and the small number of observations relative to the number of independent variables make the true threshold for significant results substantially higher than routine application of statistical methods suggests. Therefore it is imperative to reduce the number of independent variables in this kind of multiple regression, and to use numerical experiments to test statistical significance.

This work shows that a large portion of the posterior position of the tongue can be deduced—quantitatively—from the anterior position in these vowels with a high degree of statistical confidence. This in turn implies that if the position of the anterior portion of the tongue can be deduced from its acoustic output, then much of the entire tongue position is, in turn, recoverable from the acoustic description of a vowel, via pellet positions on the anterior portion of the tongue. In fact, the recent work of Story (2007) has shown that two PCA factors of cross distance in the oral cavity can be predicted from the first two formant frequencies. Because the formant frequencies are shaped by the pharyngeal, as well as oral, cavity, Story's work in English indirectly corroborates the findings here. This opens the way to both a compact representation of the vowel (anterior tongue position only) as a basis for articulatory speech synthesis, and to the inverse task, namely reconstruction of the articulatory configuration in a given vowel from its acoustic representation.

For speech production research in general it should be borne in mind that statistical significance cannot always be quantified using parametric statistics. There can be a number of reasons for this, including a small number of observations with respect to number of independent variables (e.g., Whalen *et al.*, 1999), high variability in the data, as is often the case with children (e.g., McGowan *et al.*, 2004), the complexity of algorithms, such as PARAFAC (parallel factors, e.g., Jackson, 1988), or because the data are not normally distributed. In these cases numerical experiments and/or nonparametric statistics provide a means for enabling us to make statistically meaningful statements.

ACKNOWLEDGMENTS

This work was supported by Grant No. NIDCD-001247 to CReSS LLC. The authors would also like to acknowledge the personal help of Dr. Sundberg and Dr. Fant, who supplied the x-ray tracings analyzed in this article.

References

- Beaudoin, RE.; McGowan, RS. Proceedings of the Fifth Seminar on Speech Production: Models and Data, Kloster Seeon, Bavaria, Germany. Ludwig-Maximilians-Universität; 2000. Principal components analysis of x-ray microbeam data for articulatory recovery; p. 225-228.
- Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Vol. 3rd ed.. Mahwah, NJ: Erlbaum; 2003. p. 26, 69, 390, 391-428.

- Fant, G. Formants and cavities. In: Zwirner, E.; Bethge, W., editors. Proceedings of the Fifth International Congress of Phonetic Sciences. 1965. p. 120-141.
- Harshman R, Ladefoged P, Goldstein L. Factor analysis of tongue shapes. *J. Acoust. Soc. Am* 1977;62:693–707. [PubMed: 903511]
- Hays, WL. Statistics. Vol. 3rd Ed.. New York: CBS; 1982. p. 483-486.
- Hill, T.; Lewicki, P. Statistics: Methods and Applications. Tulsa, OK: Stat-Soft; 2006. p. 346
- Hoole P. On the lingual organization of the German vowel system. *J. Acoust. Soc. Am* 1999;106:1020–1032. [PubMed: 10462807]
- Jackson MTT. Analysis of tongue positions: Language-specific and cross-linguistic models. *J. Acoust. Soc. Am* 1988;84:124–143. [PubMed: 3411040]
- Jacobsen, L. Ph.D. dissertation. Los Angeles: University of California; 1978. DhoLuo Vowel Harmony: A Phonetic Investigation.
- Kaburagi T, Honda M. Determination of sagittal tongue shape from the positions of points on the tongue surface. *J. Acoust. Soc. Am* 1994;96:1356–1366. [PubMed: 7963000]
- Kiritani S. X-ray microbeam method for the measurement of articulatory dynamics: Techniques and results. *Speech Commun* 1986;5:119–140.
- Lindau M. The feature expanded. *J. Phonetics* 1979;7:163–176.
- Lindau-Webb M, Ladefoged P. Methodological studies using an x-ray microbeam system. *UCLA Working Papers in Phonetics* 1989;72:82–90.
- Maeda, S. Une analyse statistique sur les positions de la langue: Etude préliminaire sur les voyelles françaises (“A statistical analysis of tongue positions: Preliminary study of French vowels”). *Actes des 9èmes Journées d’Études sur la Parole* (Acts of the Ninth Speech Study Session); France; 31 May — 2 June; Lannion. Lannion: Groupement d’Acousticiens de la Langue Française; 1978. p. 191-200.
- Maeda, S. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, WJ.; Marchal, A., editors. *Speech Production and Speech Modeling*. Dordrecht: Kluwer; 1990. p. 131-150.
- McGowan RS, Nitttrouer S, Manning CJ. Development of [ɹ] in young, Midwestern, American children. *J. Acoust. Soc. Am* 2004;115:871–884. [PubMed: 15000198]
- Nguyen N, Hoole P, Marchal A. Regenerating the spectral shapes of [s] and [ʃ] from a limited set of articulatory parameters. *J. Acoust. Soc. Am* 1994;96:33–39.
- Nix DA, Papçun G, Hogden J, Zlokarnik I. Two cross-linguistic factors underlying tongue shape for vowels. *J. Acoust. Soc. Am* 1996;99:3707–3717. [PubMed: 8655802]
- Perkell JS, Cohen M, Svirsky M, Matthies M, Garabieta I, Jackson M. Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *J. Acoust. Soc. Am* 1992;92:3078–3096. [PubMed: 1474223]
- Ramsay JO, Munhall KG, Gracco VL, Ostry DJ. Functional data analyses of lip motion. *J. Acoust. Soc. Am* 1996;99:3718–3727. [PubMed: 8655803]
- Sekimoto, S.; Imagawa, H.; Kiritani, S. *Ann. Bull., Res. Inst. Logop. Phoniatr.* Vol. 12. Univ. Tokyo: 1978. Dynamic characteristics of tongue movement in the production of connected vowels; p. 11-20.
- Shirai, K.; Honda, M. Estimation of articulatory motion. In: Sawashima, M.; Cooper, FS., editors. *Dynamic Aspects of Speech Production*. Tokyo: University of Tokyo Press; 1978.
- Story BH. Time-dependence of vocal tract modes during production of vowels and vowel sequences. *J. Acoust. Soc. Am* 2007;121:3770–3789. [PubMed: 17552726]
- Sundberg, J. Stockholm Royal Institute of Technology, Speech Transmission Laboratory/Quarterly Progress and Status Reports. Vol. 1. 1969. Articulatory differences between spoken and sung vowels in singers; p. 33-46.
- Tiede MK. An MRI-based study of pharyngeal volume contrasts in Akan and English. *J. Phonetics* 1996;24:399–421.
- Westbury JR. On coordinate systems and the representation of articulatory movements. *J. Acoust. Soc. Am* 1994;95:2271–2273. [PubMed: 8201122]
- Whalen DH, Kang AM, Magen HS, Fulbright RK, Gore JC. Predicting midsagittal pharynx shape from tongue position during vowel production. *J. Speech Lang. Hear. Res* 1999;42:592–603. [PubMed: 10391625]

Zerling, J-P. Thèse de Troisième Cycle. France: Université de Nancy; 1979. Articulation et Coarticulation dans les Groupes Occlusive-Voyelle en Français (Articulation and Coarticulation in Occlusive-Vowel Groups in French).

FIG. 1.

Reference points for speaker BE. P1—mean position of the alveolar ridge reference point on all nine digitized vowel tracings; P2—mean position of the highest point of the palate; P3—mean position of the superior point on the rear pharyngeal wall; P4—mean position of the inferior point on the rear pharyngeal wall; P0—center of circular arc through P1, P2, and P3.

FIG. 2.
Gridlines for speaker BE constructed based on the reference points P1 through P4.

FIG. 3.

PCA component shapes for speaker BE. Pellet displacements from mean position associated with variation in each component: (a) first principal component; (b) second principal component; (c) third principal component.

FIG. 4.

Regression results for speaker BE. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R ; dashed line shows the level at which R is significant at the $p < 0.05$ level according to the F test.

FIG. 5.

Regression results for speaker JS. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R ; dashed line shows the level at which R is significant at the $p < 0.05$ level according to the F test.

FIG. 6.

Regression results for speaker RL. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R ; dashed line shows the level at which R is significant at the $p < 0.05$ level according to the F test.

FIG. 7.

Regression results for speaker F. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R ; dashed line shows the level at which R is significant at the $p < 0.05$ level according to the F test.

FIG. 8.

The distribution of R from multiple regressions with the final maximum number random, uniformly distributed independent variables and midsagittal cross distances at each gridline as dependent variables. The light gray region of each distribution shows the fraction of regressions below the 95th percentile of each gridline's R distribution, the dark gray shows fraction between the 99th and 99th percentiles, and the white shows the regressions above the 99th percentile. The solid horizontal line denotes the minimum R value that would need to be attained for $p < 0.05$ according to an F test, and the dashed line the minimum R values that would need to be attained for $p < 0.01$ according to an F test. Subject BE is shown in (a); subject JS in (b); subject RL in (c); and subject F in (d).

TABLE I

Vowels analyzed in this study. “✓” indicate that the vowel was produced by the given speaker; “n/a” indicates that it was not; a phonetic transcription indicates that a variant with different phonemic length (e.g., [ø] rather than [ø:]) was produced.

	[i:]	[e:]	[æ:]	[y:]	[ø:]	[œ:]	[ɥ:]	[u:]	[o:]	[ɑ:]	[θ]	[a]	[ɔ]
BE	✓	✓	✓	✓		n/a	✓	✓	✓	✓	n/a	n/a	n/a
JS	✓	✓	n/a	✓	✓	n/a	✓	✓	n/a	✓	n/a	n/a	n/a
RL	✓	✓	✓	✓	✓	n/a	✓	✓	✓	✓	n/a	n/a	n/a
F	✓	✓	✓	✓	✓	✓	[ɥ]	✓	✓	✓	✓	✓	✓

TABLE II

Standard deviations (mm) in the x and y coordinates of reference points across images. P1 is on the alveolar ridge; P2 is the most superior point of the hard palate; P3 is on the dorsal wall of the pharynx at the level of the anterior tubercle of the atlas; P4 is on the dorsal wall of the pharynx at the level of the bottom of the vallecular sinus. Coordinates are relative to the tip of the upper incisor.

Speaker:	BE	JS	RL	F
Upper incisor- x	0.3	0.5	0.3	N/A
Upper incisor- y	0.3	0.5	0.3	N/A
P1- x	0.7	0.9	0.3	0.9
P1- y	0.4	0.5	0.1	0.8
P2- x	1.4	0.8	1.3	5.2
P2- y	0.7	1.0	0.7	0.8
P3- x	1.2	0.6	1.2	1.8
P3- y	1.5	2.8	1.3	1.2
P4- x	1.8	1.0	1.9	2.8
P4- y	3.3	4.3	1.6	3.5

TABLE III

Numerical experiment parameters. Although a potential maximum of eight independent variables (an x - and a- and a y -coordinate for each of four estimated fleshpoints) could be simulated, the small number of observations (vowels) and / or numerical degeneracy limits the actual maximum number of independent variables that can be used in each numerical experiment.

Speaker:	BE	JS	RL	F
Number of vowels	10	7	9	13
Potential maximum number of independent variables	8	6	8	8
Final maximum number of independent variables	6	3	5	8

TABLE IVCumulative variance accounted for by PCA components of eight pseudopellet variables (mm²; %)

Speaker:	BE	JS	RL	F
Total variance	402.56 (100)	218.40 (100)	143.91 (100)	214.51 (100)
Component 1	365.17 (90.71)	186.96 (85.60)	104.78 (72.81)	162.58 (75.79)
Components 1–2	395.93 (98.35)	213.46 (97.73)	139.60 (97.01)	209.38 (97.61)
Components 1–3	400.6 (99.51)	216.99 (99.35)	143.17 (99.49)	213.19 (99.39)
Residual components 4–8	1.96 (0.49)	1.41 (0.65)	0.74 (0.51)	1.32 (0.61)

TABLE V

Mean R for pharyngeal region gridlines. The pharyngeal regions include gridlines 11 through 30 for speaker BE; 11–24 for speaker JS; 11–28 for RL; and 12–24 for F. (See Fig. 4–Fig. 7 for comparison.)

Speaker:	BE	JS	RL	F
R in pharyngeal region	0.96	0.93	0.96	0.97