Principal Components Representation of the Two-Dimensional Coronal Tongue Surface

Brief Title: Principal Components of 2D Coronal Tongue

Eric Slud[1], Maureen Stone[2], Paul Smith[1], and Moise Goldstein Jr[2]

[1]Mathematics Department, University of Maryland
College Park, MD 20742 USA

Phone: 301-405-5469, Fax: 301-314-0827, Email evs@math.umd.edu

[2] Department of Oral and Craniofacial Biological Sciences
Department of Orthodontics, Univ. of Maryland Dental School
Baltimore, MD USA

December 14, 2001

ABSTRACT: This paper uses principal components (PC) analysis to represent coronal tongue contours for the eleven vowels of English in two consonant contexts (/s/, /l/), based upon five replicated measurements in three sessions for each of six subjects. Curves from multiple sessions and speakers were overlaid before analysis onto a common (x, y) coordinate system by extensive preprocessing of the curves including: extension (padding) or truncation within session, translation, and truncation to a common x-range. Four PC's plus a mean level allow accurate representation of coronal tongue curves, but PC shapes depend strongly on the degree of padding or truncation. The PC's successfully reduced the dimensionality of the curves and reflected vowel height, consonant context, and physiological features.

Acknowledgement: The authors thank Yang Cheng for preliminary analyses, Hsiao-Hui Tsou for analysis of curves measured in reverse orientation, and Bill Levine for useful discussions related to the material of this paper. We also thank the referees and Editor for many detailed suggestions which improved the paper. This research was supported in part by NIH Grant R01 DC 01758.

1 INTRODUCTION

The high dimensionality and lack of fixed landmarks in the human tongue make the parsimonious representation of its deformations a challenging problem. A model based upon linear superposition of a few basis-shapes or factors is very appealing: Principal Components Analysis (PCA) is such an approach (Anderson 1984).

Previously, PCA and factor analysis have been used successfully to reduce the dimensionality of midsagittal tongue contours for vowels based on tongue contour data obtained from images (*cf.* Harshman et al. 1977, Jackson 1988b, Maeda 1990, Hoole 1999). The cross-sectional or coronal tongue profile has not often been represented using PCA. Stone, Goldstein and Zhang (1997) examined 11 vowels in 2 consonant contexts for a single subject and found that two PC's explained 93% of the variance. Moreover, the representations of the tongue shapes as linear combinations of the first two PC's was consistent with traditional phonetically based groupings.

This study extends that work by examining multiple subjects and sessions to determine whether a small number of PC's will still represent vocalic tongue shapes despite this additional "noise."

Two major alternative approaches to tongue surface tracking are provided by fleshpoint measurements, as in X-ray Microbeam or Electromagnetic Midsagittal Articulator (EMMA) data, versus imaging, as in ultrasound, Xray or MRI. With fleshpoint measurements, there is little need to worry about registration across sessions and speakers, but fleshpoints interfere somewhat with natural speech and introduce the methodological problem of extrapolating the tongue surface between and beyond the fleshpoints. In addition, fleshpoints are tracked only at midline, which would provide less complete information about natural speech in 3D and 3D plus time. When the tongue surface is tracked via ultrasound scans, cross-sectional contour measurements can readily be made during natural speech, interpolated reasonably accurately at fine spacings along the contour, and produced along multiple cross-sectional planes with only reangulation of the transducer. However, collection of such measurements across different sessions and speakers leads to important new methodological problems of registration and overlay of the tongue contours for simultaneous analysis. These problems are a primary focus for the present paper.

In ultrasound measurements, the coronal tongue width varies across subjects and sounds. This has particularly important consequences for Principal Components analysis of coronal data, since the lateral range chosen after overlaying of contours necessarily affects the resulting PC's. In particular, it is not at all clear that larger tongues produce uniformly larger lateral coronal measurement ranges. In preliminary examination of curves measured by ultrasound, lateral ranges exhibited a strong but not easily interpreted interaction with respect to speaker and

sound. For this reason, we did not attempt a procrustean width-normalization of the resulting contours by speaker, preferring to emphasize co-registration across speakers in such a way as to maximize the similarity of tongue shapes for the same speech sound and context.

One motivation for this study is to consider PCA as a mechanism for reducing tongue shape dimensionality of 2D contours prior to 3D reconstruction, since there are no instruments that directly collect 3D tongue shape. Two approaches to 3D reconstruction have so far been tried: (i) aligning a series of 2D tongue contours spatially into a surface at each moment in time (Stone and Lundberg, 1996, Lundberg and Stone, 1999), which unfortunately introduces extra degrees of freedom for the independent errors arising from separately measured coronal sections (Stone, 1990); and (ii) modelling the 2D contours parsimoniously using PCA and then reconstructing 3D-surface shapes and motions with the fitted models, as has been done on vocal tract cross-sections by Yehia and Tiede (1997).

Our second motivation is to explore the interaction of subject-specific and phonemic patterns in statistical representation of tongue and vocal tract data. There is a great deal of inter-subject variability in the production of speech. Acoustic and physiological measures show many features that vary across experimental subjects (*cf.* McGowan and Cushing 1999); however, listeners are able to normalize speech across speakers (Johnson and Beckman 1997). Formant frequencies of vowels vary across speakers due to differences in vocal tract morphology (Peterson and Barney 1952, Miller 1989, Hillenbrand et al. 1995). Physiological studies further indicate variant articulations of sounds, which are minimally reflected in the acoustics and not at all perceived. The classic example is /r/, which can be produced with a retroflex or bunched tongue tip (Boyce and Espy-Wilson 1997, and Alwan et al. 1997). More difficult to deal with are

the unsystematic differences found between subjects. Two X-ray Microbeam studies exemplify this difficulty. Johnson et al. (1993) found within-speaker consistency, but between-speaker variability in the production of CVC syllables. Hashi et al. (1998) used speaker normalization (sizing and scaling) in a study of isolated vowels for 20 English and 8 Japanese speakers. After normalization, subject variability decreased in the dorsal-ventral dimension, while in the rostral-caudal dimension it mostly increased, but not consistently. We believe thse authors correctly attributed subject variability in general to anatomical/physiological features represented poorly in the collected data set, features outside the measured range of the collected data set, and idiosyncratic subject speech patterns.

Given such variability in speech production across speakers, one might not expect enough regularity among speakers to allow simple statistical processes, such as PCA, to capture phonetic events. On the other hand, PCA might succeed in reducing dimensionality even without extracting universal behaviors. A technique for extracting 'articulatory prime' shapes from data, allowing non-orthogonal components to scale differently for different speakers, is the PARAFAC model pioneered by Harshman et al. (1977), with further exposition and development by Jackson (1988a,b). The primary concern addressed by PARAFAC is how to modify a small set of prime shapes to account for the variability of sound production by different speakers, without requiring large numbers of parameters to specify tongue shapes for all speaker and sound combinations. Harshman et al. (1977) applied the technique to extract two factors from midsagittal x-ray data on ten vowels spoken by five English speakers, while Jackson (1988b) found three somewhat different non-orthogonal factors in data on 16 Icelandic vowels in two contexts produced by two speakers. Nix et al. (1996) re-analyzed Jackson's data, along with new x-ray tongue shape data

on six English speakers. They reconciled the cross-linguistic results by finding similar sets of two factors which adequately represented the respective data from each dataset. Hoole (1999) applied a hybrid PARAFAC and PCA model to midsagittal tongue pellet data on fifteen German vowels spoken in three consonantal contexts by seven speakers in each of two sessions in which speech rates differed. He found that a PARAFAC model could not be fitted to his full dataset, and that his model led to somewhat different scaling factors across sessions. He also found two PC's in the residuals from a two-factor PARAFAC model, the first PC accounting for 35–49% of the variance among subjects and appearing to represent subject differences in tongue height posterior to the blade. All of these PARAFAC studies modelled only residuals from the mean over all tokens from the same speaker, for tokens with relatively few pieces of articulatory information per token (13 in the Harshman et al. study, and 8 in Hoole's). Our dataset, while cross-classified as extensively as Hoole's, consisted of raw tokens with 120 (x, y) points. While our first objective was accurate reduced-dimensional representation of these tokens, we also discuss under Results below the adequacy of a PARAFAC representation for the PC loadings we obtained.

The present study considers the effect of tongue curve length and subject inhomogeneity on the quality of the PC fits and their representation of phonetic features. We also detail the effectiveness of different data preprocessing approaches in improving the PCA. A further goal is to use PCA to identify subject specific characteristics as well as group task behaviors.

2 Subjects, Speech Materials, and Data

Six normal, adult, native speakers of American English were used as subjects (3 Caucasian females, 2 African-American males, 1 Hispanic male). Subject MS, the same speaker used in Stone et al. (1997), produced new data for this study. Each subject attended three recording sessions at least one week apart and repeated the speech materials while ultrasound and acoustic recordings were made. Methods for the ultrasound recordings of tongue movement are discussed in detail in Stone et al. (1997). The eleven vowels of English (i, ɪ, e, ε, æ, a, ɔ, o, ʊ, u, ʌ), were produced in ∂CVC∂ utterances using two consonant contexts (/s/, /l/).

The methodology of data collection includes a Head and Transducer Support (HATS) system (Stone and Davis 1995) designed to hold the head and transducer steady, in a known relationship to each other. The head is clamped on four sides and the transducer positioned below the chin. The transducer is marked with a line indicating the direction of the beam. The subject's head, with the transducer beneath, is videotaped and inserted into the ultrasound image throughout the recording. Prior to data collection, a video recording is made of the head, and calibrations made of: the occlusal plane, transducer position, and a cm scale. The coronal section was recorded in the region of the palatal vault to encourage the maximal variation of tongue motion and shape. In the vault region there is room for upward tongue motion, and on palatal contact the tongue will reflect its arch-like shape.

The cross-sectional tongue surface for six subjects (MS, MD, SG, CS, GW, and LG) were extracted from digitized ultrasound images recorded on a VCR, using the μ-Tongue (Unser and Stone 1991) software package. Each of the subjects produced a total of 11 vowels $\times 2$ contexts $\times 5$ replications $\times 3$ sessions, for a total of 1980 cross-sectional tongue images. Each image

curve, whatever its length along the x-dimension, is represented in the μ-Tongue output by 120 pairs (x, y), and different curves do not necessarily have the same range of x values. Reasons why this is so are discussed at the beginning of the next Section. After preprocessing, the number of points per curve is chosen to be 101 or more based on the degree of padding chosen. This number is taken to be the same for all curves because comparing different curves as discretized waveforms makes it convenient to reduce them to vectors of the same dimension.

Let (x_{abcdi}, y_{abcdi}), for $a = 1,\ldots,6,\ b = 1,2,3,\ c = 1,2,\ldots,22,\ d = 1,\ldots,5,\ i = 1,\ldots 120,$ be our raw data set, where a indexes subject, b indexes session, c indexes sound/context, d indexes replications within session, and i indexes observations (points) on the image curves.

3 Pre-Processing Strategy

There are several possible reasons for data preprocessing. First, the ultrasound transducer, although positioned with care using the HATS system (Stone and Davis 1995), may be set differently across subjects and sessions, resulting in arbitrary shifts in x- and y-coordinates. In fact the transducer calibrations indicated that different sessions may have been collected at slightly more anterior or posterior locations. Second and more importantly, even for the same speech-sound, session, and speaker, some tongue contours may have different widths, or more lateral data points. There are two reasons for these width differences. The first is that, in repetitions of the same sound, the volume-preserving nature of the tongue implies that vertical tongue expansion (elevation) must be balanced by anterior-posterior compression and/or lateral narrowing. Another reason is that the edges of the tongue have air beneath them. This air disperses the sound wave before it reaches the tongue, diminishing or eliminating the reflected

surface. Moreover, these two phenomena are related: higher tongue positions are more likely to produce air beneath the edges, and a narrower tongue. Thus, curve extent is not uniform across sounds, and preprocessing decisions must be made concerning the truncation, padding, or extension of the contours. Methodology for registering, smoothing, and interpolating curve data is discussed in a broader statistical context by Ramsay and Silverman (1997).

Three methods (treatments) of respectively aligning curves, equalizing their length, and normalizing their position were studied, and combinations of these resulted in examining a total of seventeen variant preprocessing plans. The first treatment was to define a grid of x-coordinates (x-range) common to all overlaid curves. In some plans, individual curves were extended by padding. The second treatment was to translate or shift the x-coordinates so that the averaged curves within speaker/session/sound could be overlaid with other curves. After shifting, if any is done, all plans truncate the curves to a common x-range. The third treatment was the possible subtraction of a constant mean value from curves to align the curves better in the y direction. After preprocessing, all image curves for all a, b, c share the same equispaced x-value sequence $\{x_i,\ i = 1, \ldots, N\}$, $(N = 101,\ 109$ or $141;$ see Table 3), which are analyzed along with the corresponding smoothing-spline-interpolated y values. All preprocessing and analysis was done using customized **Splus** *version 3.4* functions.

3.1 Length Equalization of Curves

The first preprocessing treatment is to determine a common set of x-coordinates (x-range) to replace the unequal x coordinate values of the image curves. The choices for defining the x-range are to truncate the data beyond the common region, or to extrapolate individual curves

by linear or spline extension, and possibly to pad individual curves with constant values.

The first choice is to truncate the curves to the largest interval common to all images, or:

$$\left[\max_{a,b,c,d} \min_{i} x_{abcdi},\ \min_{a,b,c,d} \max_{i} x_{abcdi}\right] \tag{1}$$

Truncation to a minimal common segment isolates common subject information, but discards interesting and valid data from the longer curves. For the six subjects, the intervals common to all curves were respectively 25.3, 32.0, 25.0, 27.6, 25.3 and 27.7 mm long. The maximal ranges from leftmost to rightmost x-coordinates, $(\min_{bcdi} x_{abcdi},\ \max_{bcdi} x_{abcdi})$, were respectively 58.5, 60.2, 77.0, 61.8, 55.8, and 62.5 mm long. Thus the common interval was 53% as long as the maximal x-range for speaker MD, and 43% for speaker MS. The truncation method provided a very stable PC1 shape across subjects, as much of the variability occurred at the ends; however, it eliminated a large amount of information from the longer curves. For this reason, all truncation analyses are based upon the interval over which at least 3 of the 5 replicate curves for each fixed (a, b, c) were measured. The curve ordinates which were not measured on this interval were interpolated from measurements via smoothing splines. See Figure 1(a) for illustration[1].

A second strategy for equalizing length is to extend the shorter curves to a larger interval by padding, after extrapolating within the maximal interval

$$\mathcal{I}_{abc} = [\min_{d} x_{abcd,1},\ \max_{d} x_{abcd,120}] \tag{2}$$

for each speaker/sound/session combination abc. For shorter curves, coordinate measurements to the left of $x_{abcd,1}$ or to the right of $x_{abcd,120}$ are not available and must be defined either

[1]The data shown in Figure 1 were among the most highly variable replicated curves within a speaker, sound, and session. They are not typical, and were chosen only to illustrate the effect of truncation and padding.

by extrapolation or 'padding' or both. The simplest method is to extrapolate y values linearly from the spline defined between the most extreme observations. However, our early efforts to use these extrapolations made it clear that large gradients among values near the end of the measurement interval lead to unacceptably wild and meaningless swings in the extrapolated y values. While it might still be possible to extrapolate sensibly by controlling or damping the gradients and possibly basing the extrapolation on a larger window of observed y values, we have abandoned the attempt to extrapolate beyond $\mathcal{I}_{abc}$, relying instead on padding by constant values for each measured curve. Curves might be padded with many values: zeros, endpoint averages, or overall averages. After exploring many possibilities, we decided to pad using endpoint averages, i.e., by adding points to the beginning and end of the curve equal to the average $(y_{abcd,1} + y_{abcd,120})/2$ of the first and last y-values. Padded curves can have non-physical discontinuities at the ends of their original x-intervals of measurement if, as often occurs, the values $y_{abcd,1}$ and $y_{abcd,120}$ are much different. But these artifactual discontinuities do not seem to affect much the PCA methods which we adopt to analyze and represent the curve shapes. The discontinuities are much less apparent here (e.g., in Figure 3) than in previous work using zero padding and only one subject (Stone et al. 1997).

Padding is specified with two additional integer parameters illustrated in Figure 1(b): *gap* and *flat*, which have the following meaning. First, for each speaker/session/sound combination *abc*, each replicate curve is first extrapolated linearly beyond its measured range to all of $\mathcal{I}_{abc}$. These extrapolations are very short, typically 1mm or less. The extrapolated curves extended in this way are then interpolated at 101 equispaced x-values spanning $\mathcal{I}_{abc}$, yielding 100 x-increments of size $d_{abc} = \mathit{length}(\mathcal{I}_{abc})/100$. Next we skip over a range of x-coordinates of

length $gap \cdot d_{abc}$ on both sides of $\mathcal{I}_{abc}$ for each of the 5 replicate curves. The omitted x-coordinates — consisting of two segments at either end of $\mathcal{I}_{abc}$, each *gap* percent as long as $\mathcal{I}_{abc}$ — correspond to the *gap* regions in Figure 1(b). The curves are then padded by placing *flat* further points at height $(y_{abcd,1} + y_{abcd,120})/2$ (= 69.25 in Figure 1(b)) spaced d_{abc} apart on each side. These regions of constant padding are the *flat* regions in Figure 1(b). Thus the padded replicate curves for fixed *abc* have a total of $101 + 2 \cdot flat$ points over a common range $2 \cdot (gap + flat - 1)\%$ longer than $\mathcal{I}_{abc}$. For example, consider a specific *abc*, as in Figure 1(b), for which the range $\mathcal{I}_{abc}$ of x-values is $[18.77, 83.493]$, *gap*= 20 and *flat*= 10. Then $d_{abc} = 0.64723$, and the interpolated values for the five *abc* curves would be created at the x-locations 18.77, 19.417, 20.064, ..., 82.198, 82.845, 83.493, and padded ordinate values would be placed at $x =$ 0, 0.647, 1.294, ..., 5.178, 5.825 on the left side and at 96.438, 97.085, ..., 101.615, 102.262 on the right. See Figure 1(b) for illustration of the method of extending and padding. The dots indicate the points actually included on the padded curves: of these, the solid dots represent measured data and the hollow ones linearly extrapolated data. The dashed lines indicate the smoothing-spline interpolated values which would be used to supply equispaced points on a common x-range after the curve has been shifted as described in the following Section. The specific (*gap, flat*) parameter pairs were chosen so that the steep sides of the tongue cross-sections would be smoothed to a constant level at the extremes, over an interval long enough to avoid sharp gradients; but the exact values used in numbered Plans were chosen after some trial and error.

3.2 Laterally Shifted vs Unshifted Curves

Translation of x-coordinates between sessions is an attempt to overlay image curves to reduce session effects and to establish uniform x, y coordinates for each speaker. First, we overlay curves within each speaker, with session $b = 1$ as the standard. We shift the image curves for the same speaker and sound in sessions $b = 2$ and $b = 3$ into the scales for session 1, as follows. For each speaker/session/sound combination abc, a penalized sum of squared differences is used to measure the distance between two overlaid shifted curves, each averaged over replications. That is, denoting by $\mathcal{J}_{abc}$ the interval resulting from the length equalization of Section 3.1, for an x-translation Δx of the curve for session $b\,(= 2, 3)$, we overlay x-intervals $\mathcal{J}_{a1c}$ and $\mathcal{J}_{abc,k} + \Delta x$ and define $\hat{x}_{abc,k}$ for $k = 1, 2, \ldots, 101$ to be an equispaced sequence of 101 points spanning the intersection $\mathcal{J}_{a1c} \cap (\mathcal{J}_{abc,k} + \Delta x)$. We obtain y-coordinate $\hat{y}_{abc,k}$ for the averaged replicate-curves by smoothing-spline-interpolation from the given averaged curves $y_{abc\cdot k}$. With the notation

$$\hat{y}_{abc\cdot} \;=\; \frac{1}{101} \sum_{k=1}^{101} \hat{y}_{abc,k} \tag{3}$$

the translation $\Delta x \,=\, (\Delta x)_{abc}$ is chosen in the first instance to minimize over x in the interval $[-10, 10]$ the sum of squares

$$f(\Delta x) \;=\; \frac{1}{100} \sum_{k=1}^{101} [(\hat{y}_{abc,k} - \hat{y}_{abc\cdot}) \;-\; (\hat{y}_{a1c,k} - \hat{y}_{a1c\cdot})]^2 \tag{4}$$

It turns out that the average cross-sectional tongue shapes for the same speech-sound from one session and speaker to another are occasionally sufficiently different, e.g., with a unimodal average shape in one session and bimodal shape in another, that minimization of $f(\Delta x)$ can lead to large and unreasonable shifts Δx. For this reason, we describe in Appendix 6 a slightly

complicated and artificial test on the function f (for fixed abc) which has the effect in these unusual settings that Δx is chosen as the minimizer not of f but of a *penalized distance*

$$f(\Delta) + dpar \cdot \left[\Delta x - (\frac{1}{2}(x_{a1c,1} + x_{a1c,101}) - \frac{1}{2}(x_{abc,1} + x_{abc,101}))\right]^2 \tag{5}$$

The penalty term with coefficient parameter $dpar$ tends to make the function to be minimized more sharply convex, and to bias the minimizer toward the crude default-value

$$\frac{1}{2}\,(x_{a1c,1} + x_{a1c,101} - x_{abc,1} - x_{abc,101}) \tag{6}$$

The current default for the parameter $dpar$ is 0.2, but the value 0.5 also works well, and in most cases the penalty term is not invoked.

After overlaying all fifteen curves (3 sessions by 5 replications) for each speaker and sound, the x-coordinates for the same sound and different subjects are standardized: with subject MS ($a = 1$) fixed as standard, we shift the (set of fifteen) curves for each other subject ($a = 2, \ldots, 6$), by a minimum-distance method exactly analogous to that given above for a single speaker[2].

We chose to allow the shift Δx to vary with speaker and session only, but not sound, since the HATS instrument adjustments vary with speaker and session but not with speech-sound. Therefore, we actually translate curves only by the average across speech-sounds of the shifts obtained as above for each speaker and session. We did check to see whether shifting different amounts for different sounds would give more closely overlaid curves, despite the fact that the

[2]Another order in which shifting and overlaying of curves could have been done would be to divide the speakers into more homogeneous groups and first shift within groups. This was tried, in connection with the 2- and 4-speaker groups of Figure 7 below. The analytic results for the pooled 6-speaker dataset created in this way were virtually indistinguishable from those using a pooled dataset created by the original method.

instrumentation settings were held constant across sound. But in fact, maintaining separate shifts for different sounds just introduced additional noise.

After performing the operations summarized in Section 3.1 and this Section, we truncate all curves to a common x-range, and interpolate them once more via smoothing splines (with very little smoothing) to a common set of $N = 101 + 2 \cdot \mathit{flat}$ equispaced x-values x'_i, with y-coordinates denoted $y'_{abcd,i}$.

3.3 Subtracting Constants and Norm-Standardizing

To align the curves better in the y direction and make their shapes more comparable, a prerequisite for a meaningful Principal Components Analysis, a constant was subtracted from each image curve. The first choice of constant to subtract, and the choice adopted in all further discussion, was the mean curve-ordinate, $\sum_{i=1}^{N} y'_{abcdi}/N$, which is usual in PCA. Subtraction of other choices of constants was tried, but led to much worse representation of curves as a constant plus a linear combination of fitted PC's.

Beyond subtracting a constant y-value, normalization of curve ordinates by multiplicative scaling was considered as a method to reduce the effects of size differences in oral morphology. Norm-standardization was tried as one computational strategy, as in Abeles and Goldstein (1977) and Stone et al. (1997), and we compared the fits of PCA models under various combinations of equalization and shifting both with and without norm-standardization. However, the resulting PC fits (Table 2, Unit= Y cases), were slightly worse than for the unstandardized data. Moreover, norm-standardization in this dataset would have diminished the inter-sound and inter-subject differences. Therefore, the strategy was not used in the analyses below.

Another strategy for standardization would have been to re-scale curves, in either the x- or y-coordinates or both, for individual speakers. In fact, different speakers were observed to have quite different amplitudes of vertical tongue-displacements, but the corresponding curve lengths were not very different. The average curve-length by speaker, in mm, before truncation to a common range; variance in mm^2 by speaker of the mean curve-ordinate over all sounds, sessions and replications; and average of per-curve y variance by speaker, in mm^2, were as follows:

[Table 1 goes about here.]

Thus, although different speakers had quite different behavior with respect to vertical displacements, their pattern of variability in average level was somewhat different from the pattern of average variability of their individual curves, and these two quantities would have been expected to track together if the differences between speakers were due to a simple scaling effect. Also, there was not much difference by speaker in the average curve lengths. Nevertheless, since this approach seems physiologically reasonable, we investigated the effect of re-scaling the data by speaker, so as to equalize y- and/or x-scales. The results, as measured by (appropriately scaled) mean-squared errors of representation of the resulting curves by 2 or 4 PC's, indicated that re-scaling did not help, so we did not pursue it further in this study.

3.4 Conclusions from Pre-Processing Comparisons

Many combinations are available for the preprocessing methods and parameter choices described above. These choices relate closely to model *parsimony*: that is, whether data standardizations which require retaining constant values for x- and y-translations and possibly curve norms for certain combinations of speaker, sound, session, and replication have sufficient impact upon the

PCA model fits to be worth the cost in parametrization. That there is such a cost can be seen by asking whether the fit would be hurt more by deleting some of these shift and scaling constants rather than by reducing the number q of PC's. There is one loading constant per curve per PC which must be retained in the fitted models. Because of the inherent noise in a data set that includes multiple speakers and sessions, the preprocessing was considered important to reduce the variability of the curves, so that fewer PC's would be needed to represent the data.

[Table 2 goes about here.]

The seventeen preprocessing plans studied were compared using data from only the first two speakers (MS and MD, the first two to supply data), with parameters defined and results displayed in Table 2. The different model fits were judged by per-observation Mean-Squared Fitting Error (MSE) averaged over session, sound, and replication.

Our analyses of PCA and MSE were performed for many different preprocessing plans, which we now summarize before displaying the results. We numbered the plans, and grouped them according to their equalization method: truncation (Trnc) or extension with padding (Ext). All of the plans using the truncation method truncated all curves within each session, sound, and speaker to the interval over which at least three of the five curves were measured. The 'Param' column of the Table has no entry for the Truncation-method plans, and contains the pair *gap, flat* for methods with Extension. The notation that x-shifting is *abc* or *ab* means respectively that there is one (Δx) shift for each speaker/session/sound combination or one for each speaker/session combination. The column 'Unit' has entry Y if unit-norm standardization was used, N otherwise. The 'x-range' column denotes the length of the common x-coordinate range after preprocessing including any shifting. Finally, the last 3 columns of the Table are

the MSE's averaged over *bcdi* and summed over $a = 1, 2$ when the indicated number q of principal components were fitted to each set of 5 replicated curves.

Of the plans summarized in the Table, all those with Truncation (*Trnc*) give per-observation MSE results for spline-interpolated curves with 101 points; those with Extension (*Ext*) are based on curves with either 109 points (Plans 10, 11, 12, 16), 121 points (Plans 14, 15, 17), or 141 points (Plan 13). All plans with Extension use the displayed parameters in solving for optimal shifts *except* Plans 12 and 13, which shift by exactly the same amounts as Plan 5.

Our conclusions from evaluating the seventeen plans[3] can be summarized as follows:

(1.) Preprocessing plans which equalize length by truncating curves to a common range achieve small MSE's for PCA model fits, because the curves encompass only part (the middle) of the tongue, which naturally has less variety and extent of deformation than the whole. However, even the best of these plans (Plan 5) loses much useful information about the lateral tongue.

(2.) Equalizing length by extension and padding of shorter curves promotes accurate x-shifting and overlay by removing excessive, non-physical, edge effects. By contrast, linear extrapolation causes worse MSE's due to erratic positions of the lateral endpoints of curves.

(3.) When overlaying the curves, x-shifting the data sets reduces MSE. Shifting works better for speaker/session than for speaker/session/sound, reflecting the experimental constancy of tranducer registration within session.

(4.) The shifting penalty (*dpar*) is needed in a relatively small number of cases, primarily where bimodal and unimodal curves must be overlaid. The penalty terms prevent bad shifts

[3]Actually, we tried many more plans and combinations of parameters than are included here. The conclusions are representative also of the plans shown in Table 2 and of others not shown.

that cause the common range to be unreasonably short or bias the curve shapes.

(5.) Norm-standardization does not do much in this setting. The performance of models was actually a little worse after standardization. Scaling in either the x- or y-coordinate (or both) before preprocessing also did not improve the fit of the PC representations.

Therefore, we chose two following preprocessing combinations to pursue, one using truncated and the other padded curves. Curves were either truncated to the subinterval over which 3 of the 5 replications were measured (Plan 5), or were padded (Plan 12). In either case, curves were shifted using $dpar = 0.2$; the mean y-value of each curve was subtracted (after shifting), retaining the average $y'_{abcd\cdot}$ as PC0; and curves were *not* normalized multiplicatively.

4 Results of Data Application

The first subject (MS) is identical to the one studied in Stone et al. (1997). The PC's calculated on her data alone are comparable to the PC's found earlier (Stone et al. 1997, Fig. 3), which is reassuring since the speech materials are the same. The differences are that the newer method of inserting a gap and padding with the endpoint average value makes negligible the discontinuities seen at the edges of the earlier PC's, which were zero padded with no gap. A second difference is that the earlier PCA had norm-standardized the y-direction, resulting in less influence for peaked curves and more influence for flat ones.

We consider first the effect of curve length on the analysis. Table 3 indicates the amount of per-curve average mean square error (MSE) expressed as a percentage of the sum of squares (%SSQ). The MSE's are based on the full six-speaker data set, for each of three preprocessing plans (plans 5, 12, 13 of Table 2). The plans correspond to truncation and two successively

more aggressive methods of extension (padding) of tongue contours, so that the common x-ranges of the contours have respective lengths of 23 mm, 33 mm, and 43 mm. The %SSQ's are incrementally attributable to the following model fitting steps: the initial SSQ based on first subtracting a single mean y-value for the whole dataset; next subtracting a single mean for each speaker/sound/session combination (the abc row); subtracting a mean y-value for each individual curve (PC0); and fitting each of four successive PC's. The small size, under all Plans, of the %SSQ for PC0 indicates that the differences between replicate-curve means within session are pure noise. At the bottom of Table 3, the *Curve mean* values are the percentages due to subtracting the mean ordinate for each curve, which are equal to the sum of the PC0 and *abc* percentages. The *Sounds* row shows that at the same location on each curve, a large %SSQ is due to differences between sounds. This value is calculated as usual for Analysis of Variance (Scheffé 1959), without regard to the PC model fitting. The %SSQ due to session, following PC0, is negligible (not shown).

[Table 3 goes about here.]

Figure 2 shows the shape of the first 4 (non-constant) PC's based on Plan 5. With this degree of truncation, leaving a common x-range of only 23.2 *mm*, the striking finding is that the PC's are virtually mirror-image symmetric in pairs (after a multiplication by -1 of PC4 as displayed). The interpretation is that tongue profiles averaged over speakers, sounds, and sessions are strongly symmetric on a narrow middle x-range, while individual curves continue to exhibit asymmetry. The dominant component, PC1, is relatively stable in shape across speakers in the sense that its shape changes little as one re-calculates it for different subsets of 3 or 4 out of 6 speakers. The second through fourth PC's were also reasonably stable among the first 5

speakers, but, even on the narrow x-range, were erratic when re-calculated on subsets including speaker LG.

Although the PC's based on the severe truncation of Plan 5 are stable and interpretable, our objective is effective representation of tongue curves over a sufficiently broad range to be acoustically and physiologically informative. Therefore, the rest of our analyses use Plan 12 of Table 2, which was based on moderate extension and padding of individual curves. This was the best of the extended plans (see Tables 2 and 3) and preserved 10 mm more of the lateral tongue than Plan 5. Figure 3 shows how different the 4 PC's are over the broader range. The PC's in Figures 2, 3, and 7 should not be interpreted as typical tongue shapes: it is their linear superpositions which effectively represent behavior of session-average curves (and, less effectively, of individual replicate curves) both in the middle and at the edges of the x-range. The mirror-image symmetry apparent in the PC's of Figure 2 is absent from Figure 3, largely because of the broader x-range; fitting longer tongue contours causes very different features to emerge in the dominant PC's. Nevertheless, there is a strong overall similarity of shape between PC's 1, 3, and 4 in Figures 2 and 3. PC2 adds a strong asymmetric component to the shapes.

Figure 4 shows the difference between the mean tongue curve (solid line) and the curves deviating from the mean in the direction of each PC with loading ± 1 standard deviation. PC0 represents vertical translation. PC1 reflects arching/grooving at midline. PC2 indicates left/right asymmetry, and PC3 represents a bimodal shape with a narrow groove. To show that these PC's do a very effective job in representing replication- , session-, and speaker- averaged tongue profiles, Figure 5 exhibits *by sound* averaged tongue shapes (dashed lines) and model-fits (solid lines) derived from PC0, PC1, and PC2, including the worst such fits. It can be seen that

the PC0–PC2 model has some difficulty representing higher-order shapes such as /se/. With the inclusion of PC's 3 and 4, however, the model-fits are indistinguishable visually from the averaged curves. To put these small model-fitting errors in context, the root-mean-square or *rms* per-observation errors (the square roots of the per-observation MSE's after preprocessing by Plan 12, some of which are displayed in Table 4 below) range from 0.6 to about 1.4 mm. By contrast, the typical measurement error in tongue ordinates is 0.5–1 mm, and individual measured curves occasionally contain errors of up to 2 mm near the middle of the x-range due to edge-detection anomalies of the μ-Tongue software. The raw or preprocessed data are unsmoothed and reflect the noise inherent in the ultrasound images. Smoothing or averaging allows better fits via PC's.

Examination of the PC loadings in Figure 6(a) indicates some phonetic grouping of the data into high vowels — those shaped by palatal contact — and the other, non-high vowels. These categories reflect physiological features more than phonetic or acoustic ones. Grooving reflects the muscle activity required to pull the midline tongue inwards to form a groove. Arching reflects the shape of the palate during palatal contact as well as the muscle activity required to elevate the tongue. Within Figure 6(a) are three clusters of PC-derived tongue shapes typical of the extreme regions of vowel concentration. A large negative loading on both PC's 1,2 denotes the high midsagittal tongue arch shaped by palatal contact and seen for the high vowels (Figure 6(b)). A positive PC2 and neutral PC1 results in a shallow but well-defined central groove (upper-middle region of Figure 6(a), tongue shape in Figure 6(c)). Many of the non-high vowels had a positive PC1 and negative PC2, with contour as in Figure 6(d). The differences between the PC1 and PC2 loadings reflect shapes with left-to-right asymmetry in these data.

The two together produce a more symmetrical gesture if loadings are in the same direction, and increased asymmetry if not. Negative PC1 loadings tend to be symmetrical because the palate influences the arched shape. Positive PC1 loadings give less symmetrical shapes because the grooving is entirely due to muscle activity, which may be unequal bilaterally.

PC's 3 and (inverted) 4 add a shallower, narrower midsagittal groove and left or right asymmetry to curves that are not well fitted by PC1 and PC2 alone. The PC4 loadings are quite a bit more positive for /s/ context than for /l/, with max at 0.65 for /s æ s/ and /s ʊ s/. PC3 loadings also tend to be more positive for /s/ context than for /l/ (with the notable exceptions of vowels ʊ and ʌ). By contrast, the PC1 by PC2 loadings of vowels in /l/ context very closely track the pairs of loadings for the corresponding vowels in /s/ context. Thus the PC's indicate that coronal tongue shapes have few degrees of freedom (grooved/arched and asymmetry), as was found previously by Stone and Vatikiotis-Bateson (1995) with the additional observation of minor context effects in the higher order shape components.

Our second interest in this work is the effect of subject variability. While it seems necessary to retain a sufficiently broad common x-range for overlaid curves, a weakness of preprocessing methods with extension and padding of curves is that the resulting PC's show less consistency over different subsets of speakers. This means that additional subjects may change the shape of the PC's. We can see this effect in the present data set. Two of our speakers (SG, LG) had particularly long curves, so that with or without padding, the rightmost 10mm of their curves after shift-overlaying were not included in the common x-range. We illustrate in Figure 7 the differences between the first 2 PC's calculated independently for this group of 2 speakers (top) and for the other 4 speakers (bottom). Subject SG was female and LG male, but SG was large

boned and had a fairly large head and jaw, while LG was bi-lingual Hispanic. Inspection of SG's data indicates that asymmetrical tongue shapes created grooves and arches that were not centered at transducer midline. This was not due to the contour alignment strategy, as we first thought, but rather to the particular subject's unusual productions.

Inspection of both subjects' data indicated bright well-defined tongue surfaces. Ultrasound imaging does not capture the extreme edges of the tongue because there is air beneath them. In these subjects, it is possible that we captured more tongue tissue, rather than larger tongues. In either case the salient features of the curves were different from those produced by speakers with narrower tongues, resulting in a wider depression for PC1 and a different shape for PC2 than in the PCA of the 4 other subjects (Figure 7). The tongue arch largely represented by PC1 is considerably more spread out, even bi-modal, in the 2-speaker group than in the 4-speaker group. This reflects the variety of lateral positions for arching in the broader tongues. As indicated in Table 4, the curves for the 2 speakers SG and LG were not fitted materially worse by the final PC's than were the curves for other speakers. The only anomaly, perhaps, is the very large MSE for fitting /lɪ/ in the 2-speaker group using a 2-PC model.

Large subject- and sound-specific differences in the quality of model-fitting can be seen in Table 4 below. The sounds for which speaker-specific root-mean-square fitting errors are displayed are the same as in Figure 5. The PC0 table shows that LG (the speaker with the broadest tongue) has a consistently large range in tongue ordinates, while the curves from CS are particularly flat. For CS, there is relatively little vertical shape to fit, so all fits were good, even with only 2 PC's. In the 2 PC models, the six speakers showed their worst fits for different sounds. There is a clear interaction between subject and sound in Table 4. This lack of phonetic

consistency may be partially explained by a recent factor analysis study of midsagittal pellet positions during vowel production in German (Hoole, 1999). That study found considerable intersubject variability in the region of the tongue 2.5–3 cm back from the tip, especially for the coronal sound /t/. The ultrasound sections used in the present study were taken from roughly the same region of the tongue, which could explain some of the subject variability. Moreover, the present data had only coronal contexts. In Table 4.B, the relative variability of PC loadings shows a very different pattern for the different speakers. LG is unique in the large amount of variability explained by PC2 and is at the high end of variance due to PC0–PC1 as well. These differences capture LG's uniqueness and show why he has such influence on the PC shapes. Here, as always, users of PCA should be careful when including data from a highly anomalous subject into a general PC analysis.

[Table 4 goes about here.]

[Table 5 goes about here.]

Table 5 displays the interaction between sound and speaker in a different way, through the pattern of sound-by-speaker average loadings for PC1. Of the 22.92% of the total sum of squares not already accounted for by subtracting a constant from each (*abcd*-indexed) curve, all but 8.62% of the original sum of squares is accounted for by the PC1 loadings. However, the PC1 loadings averaged over *acd* (i.e., distinct for each sound) or over *bcd* (i.e., distinct for each speaker) respectively explain only 3.07% and 4.62% of the original sum-of-squares. Regardless of whether PC1 effects for sounds or speakers are entered first, the bulk of the sum of squares explained by PC1 derives from the sound-by-speaker interactions of PC1 loadings.

In the present study, unlike most other PCA or factor analysis studies of the tongue, the data are coronal. The cross-sectional tongue at this location has shapes that consist primarily of midline arches and grooves. The truncated PC shapes reflect this limited repertoire quite clearly, the extended curves less so because they add subject inhomogeneity at the edges as a feature of the PC shape.

4.1 *PCA versus PARAFAC Modelling*

To conclude our data analysis, we investigated the extent to which our fitted PC models are compatible with a PARAFAC model. In the present context and notation, the latter model says

$$y_{abcd,i} \; - \; \mu_{abcd} \; = \; \sum_{j=1}^{4} \, w_{aj} \, l_{bc,j} \, f_i^{(j)} \; + \; \epsilon_{abcd,i} \tag{7}$$

where the means μ can be identified with our 'PC0'; the new principal factors $\mathbf{f}^{(j)}$, $j = 1, \ldots, 4$, are no longer orthogonal but span (approximately) the same space of residual tokens as the previous PC1–4; the loadings on the PC's or principal factors do not depend upon replication d; and the model-errors ϵ are assumed independent and normally distributed with means of 0 and variances either all identical or possibly depending on speaker (subscript a). Since 4 PC's so effectively captured the variation in the coronal tokens (after pre-processing via Plan 12), we restricted our attention to a hybrid PARAFAC/PCA analysis in which the principal factors $\mathbf{f}^{(j)}$ are assumed to span *exactly* the same subspace of 109-dimensional vectors as the orthogonal PC's $\mathbf{P}^{(j)}$, $j = 1, \ldots, 4$. (Throughout the following, $A^{(j)}$ denotes the j'th column of a matrix A.) In that case, equation (7) indicates a highly desirable reduction of dimension

from the fully general set of $6 \times 22 \times 3 \times 4 = 1584$ loadings for PC0–4, down to

$$\dim(W) + \dim(l) + \dim(B) \;=\; 6 \times 4 \;+\; 22 \times 3 \times 4 \;+\; 4 \times 4 \;=\; 304 \tag{8}$$

where B is a 4×4 matrix transforming the old PC's to the new factors. Replacing the indices bc by the single 66-level index k, and letting Y_{ad} and E_{ad} for each a, d respectively denote the 109×66 matrices of values $y_{abcd,i} - \mu_{abcd}$ and $\epsilon_{abcd,i}$ indexed by (i, k), we can re-write the equation (7) as

$$Y_{ad} \;=\; \mathbf{P}\, B\, W_a\, V^t\, \Lambda^t \;+\; E_{ad} \quad , \qquad \mathbf{P}\, B = \mathbf{F} \quad , \qquad L \;=\; \mathbf{\Lambda} V = L \tag{9}$$

Here $\mathbf{P}$ denotes the 109×4 matrix whose columns are the (first) 4 orthonormal PC's; $\mathbf{F}$ is the 109×4 matrix whose columns are the principal factors $\mathbf{f}^{(j)}$; the nonsingular 4×4 matrix B transforms PC's to factors; the 66×4 matrix L has elements $L_{kj} \equiv l_{bc,j}$ and is represented as a coordinate-change given by the 4×4 matrix V applied to the 66×4 matrix Λ with orthonormal columns spanning the same space as those of L ; and the matrices W_a are 4×4 diagonal, with $(W_a)_{jj} = w_{aj}$.

The extreme reduction of loading-dimension implied by (7) can now be assessed in two stages. First, regarding $M_a = BW_aV^t$ for $a = 1, \ldots, 6$ as a general set of 4×4 matrices, (9) says

$$\mathbf{P}^t\, Y_{ad} \;=\; M_a\, \Lambda^t \;+\; \mathbf{P}^t\, E_{ad} \tag{10}$$

This equation decomposes the observed set of $120 = 6 \cdot 5 \cdot 4$ vectors $Y_{ad}^t\, \mathbf{P}^{(j)}$ (of dimension 66, one vector for each (a, d, j)) into 4 orthonormal PC's given by the columns of Λ, and the adequacy of such a representation can be assessed by a PCA, via the ordered decreasing set of eigenvalues of $\frac{1}{120} \sum_{(a,d,j)} Y_{ad}^t\, \mathbf{P}^{(j)}\, (\mathbf{P}^{(j)})^t\, Y_{ad}$. The first 9 of these eigenvalues in the

present setting are 3408.8, 1817.2, 358.8, 281.6, 224.4, 152.8, 122.1, 74.9, 72.9, and the first four account for only 81.9% of the sum of all 66 of the eigenvalues. As we will confirm below, this suggests that PARAFAC reduces dimension too much to represent the data.

A second implication of (9), even if the representation (10) could be accepted with Λ defined as the matrix of first 4 orthonormal PC's associated with the vectors $Y_{ad}^t \mathbf{P}^{(j)}$, is that

$$\mathbf{P}^t Y_{ad} \Lambda \;=\; B\, W_a\, V^t \;+\; \mathbf{P}^t E_{ad} \Lambda \quad , \quad 1 \le a \le 5, \quad 1 \le d \le 6 \tag{11}$$

Now let the 4×4 matrix M_a and the other matrices on both sides of (11) be understood as 16-dimensional vectors. Equation (11) says that the $30 = 6 \cdot 5$ vectors on the left side are represented as linear combinations (with coefficients w_{aj} which can depend upon a), of the four undetermined 16-dimensional vectors corresponding to the matrices $B^{(j)}\,(V^{(j)})^t$. This again is a decomposition of a given matrix into principal factors, and its adequacy can be assessed by the extent to which the largest four eigenvalues of the matrix $\frac{1}{30}\sum_{a,d} (\mathbf{P}^{(j)})^t Y_{ad} \Lambda \Lambda^t Y_{ad}^t \mathbf{P}^{(j)}$ dominate the 12 others. In fact, the sum of the four largest of these eigenvalues is 97.3% of the sum of all of them. So this second stage of PARAFAC reduction does not distort the data.

Finally, there is a third stage of reduction implicit in (7). The vectors constructed from the elements of the matrices $U_j \equiv B^{(j)}\,(V^{(j)})^t$, which were sought in the previous stage as principal factors, cannot in fact be arbitrary 16-dimensional vectors: instead, these matrices are constrained by definition to be of rank one. However, we do not assess in any way the validity of this last property.

We check directly whether the net effect of the first two reductions leads to a model which represents the data adequately. The model we are testing— which is *less* special than PARAFAC,

as shown above – has the form

$$\frac{1}{5}\sum_{d=1}^{5} Y_{ad} \;\approx\; \mathbf{P}\sum_{j=1}^{4} w_{aj}\, U_j\, \Lambda^t \tag{12}$$

where $\mathbf{P}$ is the matrix of (known) orthonormal PC's; Λ is the (initially unknown) 66×4 matrix of PC's from the PCA of (10); and $M_a \equiv \sum_{j=1}^{4} w_{aj}\, U_j$. With all of the matrices Λ, W_a, U_j estimated via PCA as above, (12) implies for all a, b, c, i, (again with the notation $(b,c) \equiv k$ and with $\cdot$ denoting averaging over the replication-index d)

$$y_{abc\cdot,i} \;\approx\; \mu_{abc\cdot} + \sum_{r=1}^{4}\sum_{j=1}^{4}\sum_{s=1}^{4} w_{aj}\,(U_j)_{rs}\,\Lambda_{ks}\,P_i^{(r)} \tag{13}$$

The approximation given in equation (13) can be assessed visually in Figure 8, where the right-hand sides of equation (13) are plotted as functions of i for selected (a,b,c), and compared with the left-hand sides of (13) and with the approximation to the left-hand sides using the representation in terms of (constant means and) PC1–PC4. The PARAFAC model is seen to be an unacceptable simplification of the loading arrays for these data. We do not attempt to create and interpret PARAFAC factors for our data because the model's misspecification would rob them of value. An analysis similar to the one just presented, but assuming that neither replication nor session affects the measured curves (except through independent noise), gave very similar results in terms both of PCA eigenvalues and fitted models.

The dimension-counting argument presented above to show just how great a dimensional reduction is implicit in the PARAFAC model, strongly suggests that the PARAFAC model may be least adequate in datasets which are highly cross-classified. The present data *are* highly cross-classified, like those of Hoole (1999), who also found a PARAFAC representation inadequate.

5 Conclusions of Data Application

From the preprocessing and data analysis described above, we arrive at several general conclusions. The degree of truncation used in preprocessing the curves can be used to reveal different features of the curve in coronal data sets: the length of the x-coordinate range interacts with PC shape. The truncated curves provided stable PC's (especially PC1). Entering the 6 subjects in any order resulted in PC's essentially the same as the final set by about the third subject. The extended curves were very different for different subjects, so that the PC's changed shape as more subjects were added. The PC's generally represented 2 degrees of freedom, arching/grooving and asymmetry. In addition, as more subjects were added to the data pool, the interval of x-values common to all overlaid curves decreased considerably, necessitating padding to maintain a reasonable tongue width. In future, a scaling parameter for each speaker might well be needed to deal with large numbers of speakers. The 2PC models plus a constant level represented the simpler curves quite well, but not the higher-frequency oscillations in the shapes (Figure 5). Model-fitting with 4 PC's plus a constant did a generally excellent job of reproducing session- or speaker-averaged curves and an adequate job of reproducing individual curves. A PARAFAC modelling approach with up to four principal factors did not adequately represent the data.

More generally, our experience with this dataset convinces us that the success of PCA or any other method of statistical analysis of tongue images during speech depends critically on the method of preprocessing used, especially the overlaying, trimming, or padding of curves used at the ends of the token images. While PCA is not guaranteed to be successful in every setting, it is a particularly attractive format for reducing the dimensionality of measured curves without unnecessary modeling assumptions. Moreover, the arrays of loadings retained from highly cross-

classified tokens invite the investigators to distinguish those classifying factors which, singly or through interactions, strongly influence token shapes.

The results of this work can be further explored and exploited in several ways. First, the stability of the truncated PC's means they may be used to detail features of vocal tract constriction across subjects. For example, most fricative constrictions would fall within the truncated length of the tongue. Thus, constriction shape, duration, etc., could be studied across subjects using truncated curves. Moreover, the sensitivity of the extended PC's to subject variation could be used as a starting point for classification of speakers.

6 Appendix: Procedure used to Invoke *dpar* Penalty

The test on f which is used to decide whether the penalty term is added, can be summarized as follows. First, a quadratic function

$$y \;=\; a(\Delta x)^2 + b(\Delta x) + c \tag{1}$$

is fitted by least squares to f evaluated at $0, \pm 1, \pm 2, \ldots, \pm 10$. If any of three conditions holds, the penalty-term is added: (i) $\hat{a} \leq 0$, (ii) the minimizer $-\hat{b}/(2\hat{a})$ differs from the minimizer of f by more than 2.5, or (iii) the function f somewhere on $[-10, 10]$ falls below the quadratic

$$\hat{c} - \frac{\hat{b}^2}{4\hat{a}} + \frac{\hat{a}}{2}\,(\Delta x + \hat{b}/(2\hat{a}))^2 \tag{2}$$

by at least 0.5. Of course, the default parameters 2.5 and 0.5 respectively entering (ii), (iii) are arbitrary and could be adjusted.

References

[1] Abeles, M. and Goldstein, M. Multispike train analysis. *Proc. IEEE*, 65: 762-73 (1977)

[2] Alwan, A., Narayanan, S., and Haker, K. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *J. Acoust. Soc. Am.* 101(2): 1078-1089 (1997).

[3] Anderson, T.W. **An Introduction to Multivariate Analysis** (Wiley, New York 1984).

[4] Boyce, S. and Espy-Wilson, C.Y., Coarticulatory stability in American English /r/. *J. Acoust. Soc. Am.* 101(6): 3741-3753 (1997).

[5] Harshman, R., Ladefoged, P., and Goldstein, L. Factor analysis of tongue shapes. *J. Acoust. Soc. Am.* 62: 693-707 (1977).

[6] Hashi, M., Westbury, J., and Honda, K. Vowel posture normalization. *J. Acoust. Soc. Am.* 104: 2426-2437 (1998).

[7] Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. Acoustic characteristics of American English vowels, *J. Acoust. Soc. Am.* 97(5): 3099-3111 (1995).

[8] Hoole, P. On the lingual organization of the German vowel system. *J. Acoust. Soc. Am.* 106(2): 1020-1032 (1999).

[9] Jackson, M., Phonetic Theory and Cross-Linguistic Variation in Vowel Articulation. *UCLA Working Papers in Phonetics* 71, (1988a).

[10] Jackson, M., Analysis of tongue positions: Language-specific and cross-linguistic models. *J. Acoust. Soc. Am.* 84: 124-143 (1988b).

[11] Johnson, K., and Beckman, M. Production and perception of individual speaking styles. *Ohio State University Working Papers in Linguistics* No. 50: 115-12 (1997).

[12] Johnson, K., Ladefoged, P, and Lindau, M., Individual differences in vowel production. *J. Acoust. Soc. Am.* 94: 701-714 (1993).

[13] Lundberg, A. and Stone, M., Three-dimensional tongue surface reconstruction: practical considerations for ultrasound data. *J. Acoust. Soc. Am.* 106: 2858-2867 (1999).

[14] Maeda, S., Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. pp. 131-150 in: **Speech production and speech modeling**. Ed. Hardcastle, W. and Marchal, A. (Dordrecht: Kluwer Acad. Publ., 1990).

[15] McGowan, R. and Cushing, S., Vocal tract normalization for midsagittal articulatory recovery with analysis by synthesis. *J. Acoust. Soc. Am.* 106(2): 1090-1105 (1999).

[16] Miller, J. D., Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.* 85: 2114-2134 (1989).

[17] Nix, D., Papcun, G., Hogden, J. and Zlokarnik, I., Two cross-linguistic factors underlying tongue shapes in vowels. *J. Acoust. Soc. Am.* 99: 3707-3717 (1996).

[18] Peterson, G. and Barney, H., Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24: 175-184 (1952).

[19] Ramsay, J. and Silverman, B. **Functional Data Analysis** (Springer, New York 1997).

[20] Scheffé, H. (1959) **The Analysis of Variance** (John Wiley, New York 1959).

[21] Stone, M., A three-dimensional model of tongue movement based on ultrasound and x–ray microbeam data. *J. Acoust. Soc. Amer.* 87: 2207-2217 (1990).

[22] Stone, M. and Davis, E.P., A head and transducer support system for making ultrasound images of tongue/jaw movement. *J. Acoust. Soc. Amer.* 98(6): 3107-3112 (1995).

[23] Stone, M., Goldstein, M., & Zhang, Y., Principal component analysis of cross sections of tongue shapes in vowel production. *Speech Communication* 22: 173-184 (1997).

[24] Stone, M. and Lundberg, A., Three-dimensional tongue surface shapes of English consonants and vowels. *Jour. Acoust. Soc. Amer.* 99: 3728-3737 (1996).

[25] Stone, M. and Vatikiotis-Bateson,E., Trade-offs in tongue, jaw and palate contributions to speech production. *J. Phonetics* 23: 81-100 (1995).

[26] Unser, M. and Stone, M., Automated detection of the tongue surface in sequences of ultrasound images, *Jour. Acoust. Soc. Amer.*, 94: 3001-3007 (1991).

[27] Yehia, H. and Tiede,M., A parametric three-dimensional model of the vocal tract based on MRI data. *Proc. ICASSP-97*, Apr. 20-24, Munich (1997)

TABLE 1. Summary data on curves by speaker.

	MS	MD	CS	GW	SG	LG
Mean Curve Length	44.86	46.83	44.86	39.29	48.83	46.07
Var of Avg. Ordinate	9.43	11.52	8.09	11.76	16.02	26.68
Avg. Curve Var.	605.79	418.24	290.34	414.09	271.74	736.92

Table 2. Comparison of Preprocessing Plans Using 2 Subjects' Data.

Parameters defining 17 pre-processing plans, plus the plans' per-curve residual MSE summed over 2 speakers, after removing from each curve a constant level (PC0) and q additional PC terms.

Plan	Ext	Param	Shift	dpar	Unit	x-range	q=0	q=2	q=4
1	Trnc	*	None	*	N	25.8	7.1	0.91	0.064
2	Trnc	*	abc	0.0	N	19.2	7.6	0.29	0.006
3	Trnc	*	ab	0.0	N	29.1	8.0	0.63	0.062
4	Trnc	*	abc	0.2	N	27.8	7.6	0.65	0.060
5	Trnc	*	ab	0.2	N	29.1	8.0	0.63	0.057
6	Trnc	*	abc	0.5	N	27.8	7.6	0.65	0.059
7	Trnc	*	abc	0.5	Y	27.8	7.6	0.67	0.062
8	Trnc	*	ab	0.5	N	29.1	8.0	0.63	0.057
9	Trnc	*	ab	0.5	Y	29.1	8.0	0.66	0.060
10	Ext	10, 4	abc	0.5	N	37.3	10.3	1.49	0.380
11	Ext	10, 4	ab	0.5	N	37.9	10.2	1.32	0.362
12	Ext	10, 4	ab	0.2*	N	37.9	9.4	1.34	0.335
13	Ext	10,20	ab	0.2*	N	47.5	10.0	1.51	0.528
14	Ext	1,10	abc	0.5	N	35.4	11.1	1.56	0.388
15	Ext	1,10	ab	0.5	N	36.1	11.2	1.41	0.358
16	Ext	10, 4	abc	0.5	Y	37.3	10.3	1.51	0.384
17	Ext	1,10	abc	0.5	Y	35.4	11.1	1.59	0.393

TABLE 3. Average per-curve MSE Attributed to Stages of Fitting Constant Levels and PC's, in 6-Speaker Dataset.

	Trunc. (Plan 5)	Ext. (Plan 12)	Ext. (Plan 13)
# points	101	109	141
x-range, mm	23.2	33.0	42.7
Initial SSQ	2230.2	1990.7	2306.4
% SSQ due to			
abc	85.01	74.48	62.00
PC0	2.60	2.61	2.78
PC1	8.59	14.29	20.58
PC2	2.60	3.93	6.76
PC3	0.77	2.81	4.04
PC4	0.28	0.79	1.69
Residual	0.14	1.10	2.16
% SSQ due to			
Curve mean	87.6	77.1	64.8
Sounds	69.2	61.3	43.2

TABLE 4. ROOT-MEAN-SQUARED FITTING ERRORS PER OBSERVATION FOR MODELS: 4 PC'S PLUS PC0; 2 PC'S PLUS PC0; AND PC0 ALONE.

For each speaker-sound pair, the table entries for each model are the square-root averaged squared residuals (in *mm.*) over observation, session, and replication. Y-coordinate values were interpolated to the same set of x-coordinate values on the common x-range for all shift-overlaid curves for all 6 speakers.

4-PC Model	/li/	/lo/	/lu/	/sa/	/se/	/sʊ/
MS	0.82	0.80	0.72	0.96	1.12	0.85
MD	0.73	0.50	1.06	1.01	0.95	0.93
CS	0.57	0.61	0.74	0.55	0.54	0.60
GW	0.73	0.51	1.07	1.02	0.97	0.84
SG	1.43	1.42	0.61	0.64	0.66	0.64
LG	1.29	0.82	0.99	0.67	0.85	0.66

2-PC Model	/li/	/lo/	/lu/	/sa/	/se/	/sʊ/
MS	1.46	1.23	0.86	1.23	1.38	1.21
MD	1.10	0.61	1.09	1.09	1.17	0.98
CS	0.91	0.73	0.80	0.69	0.67	0.78
GW	1.10	0.70	1.10	1.16	1.24	1.10
SG	2.27	1.50	1.37	0.71	1.07	0.83
LG	2.26	1.20	1.42	1.52	1.74	1.51

PC0 Model	/li/	/lo/	/lu/	/sa/	/se/	/sʊ/
MS	3.63	2.56	2.16	2.56	1.90	2.43
MD	3.67	1.40	1.53	2.29	1.67	2.07
CS	1.28	1.13	1.19	2.42	1.16	2.34
GW	3.79	1.45	1.56	2.31	1.72	2.00
SG	3.32	1.58	2.04	1.29	1.70	1.62
LG	4.73	1.56	2.29	2.59	4.26	2.90

TABLE 4.B. VARIANCE OF PC LOADINGS, BY SPEAKER.

	MS	MD	CS	GW	SG	LG
PC0	8.93	10.97	7.77	11.23	15.54	26.19
PC1	354.84	201.83	99.39	178.74	120.68	389.48
PC2	15.01	42.67	15.13	47.42	44.83	112.27

TABLE 5. Percent SSQ under Plan 12 Due to Successive Stages of Model Fitting in 6-Speaker Dataset.

Fitting Stage	% SSQ Due to Fitting	Residual % SSQ
Overall mean	*	100.00
abcd mean = PC0	77.08	22.92
PC1 × speaker	3.07	19.85
PC1 × spkr × sound	9.41	10.44
Higher PC1 × abcd	1.82	8.62
Overall mean	*	100.00
abcd mean = PC0	77.08	22.92
PC1 × sound	4.62	18.29
Higher PC1 × abcd	9.67	8.62

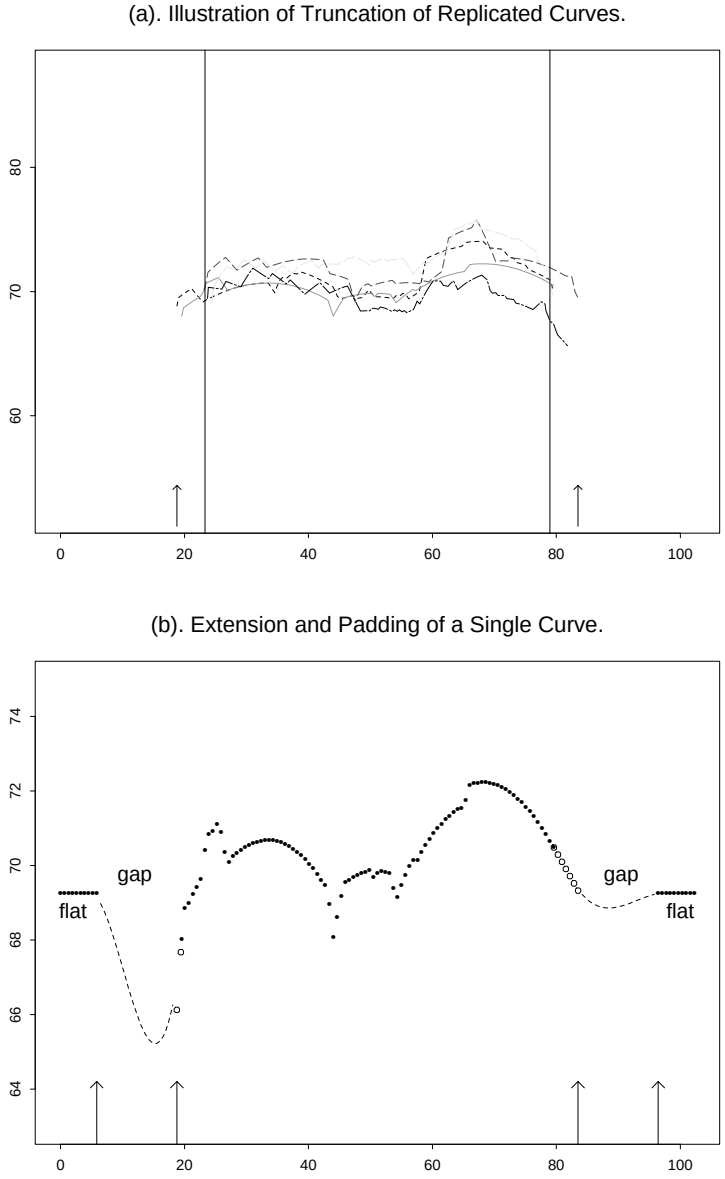

Figure 1: Alternative preprocessing steps for speaker SG, session 3, sound */læ/*. The y-coordinates are distances in mm from transducer to tongue surface. (Note the different vertical scales in the two panels.) (a) Vertical lines define the interval where 3 or more curves are measured, to which all 5 curves are truncated. Arrows bound the observed x-range. (b) Extended padded curve created using *gap*= 20, *flat*= 10, on the solid-line curve in (a). Measured points on curve are solid dots, and linearly extrapolated points are hollow. Vertical arrows bound the *gap* region. Dashed lines are smoothing-spline interpolated curves on the *gap*.

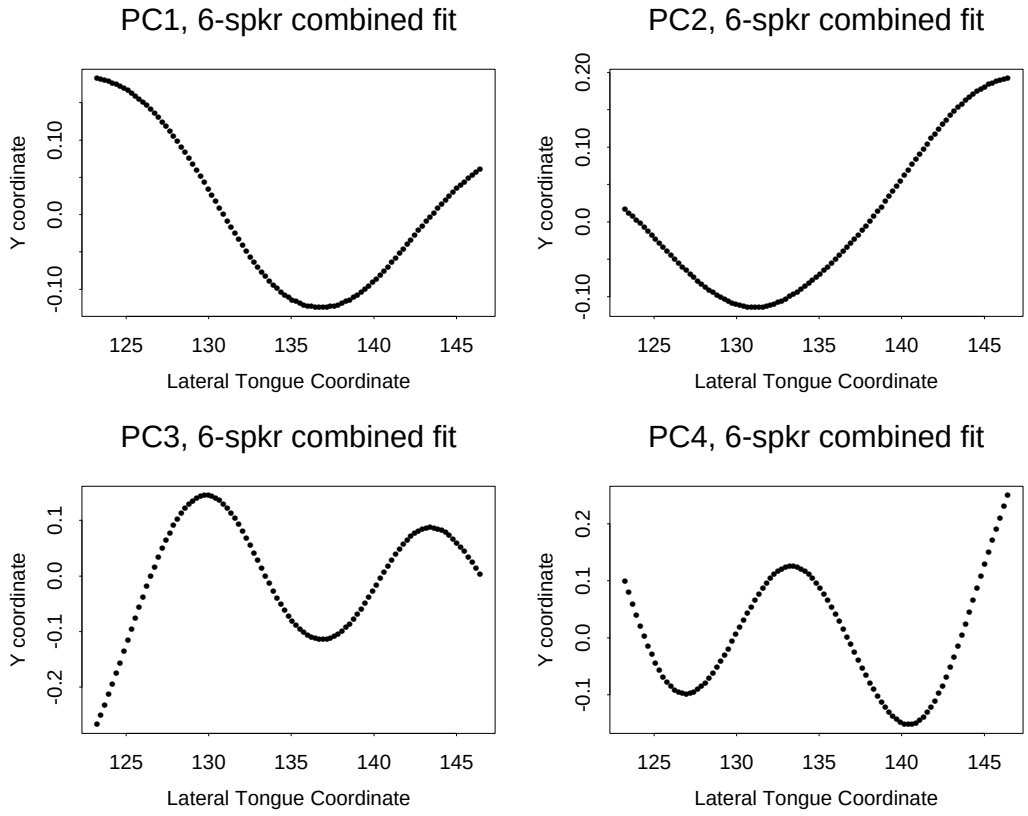

Figure 2: The first 4 non-constant PC's based on superposition over a common x-range of overlaid tongue curves for the full 6-speaker dataset. Data were preprocessed by shifting with $dpar = 0.2$ and truncating to the common ranges over which 3 of each set of replicates were measured. Preprocessing parameters are as in Plan 5, Table 2. Percent of variance (after subtraction of curve mean) accounted for by the four PC's was respectively 69.4, 21.0, 6.2, and 2.2.

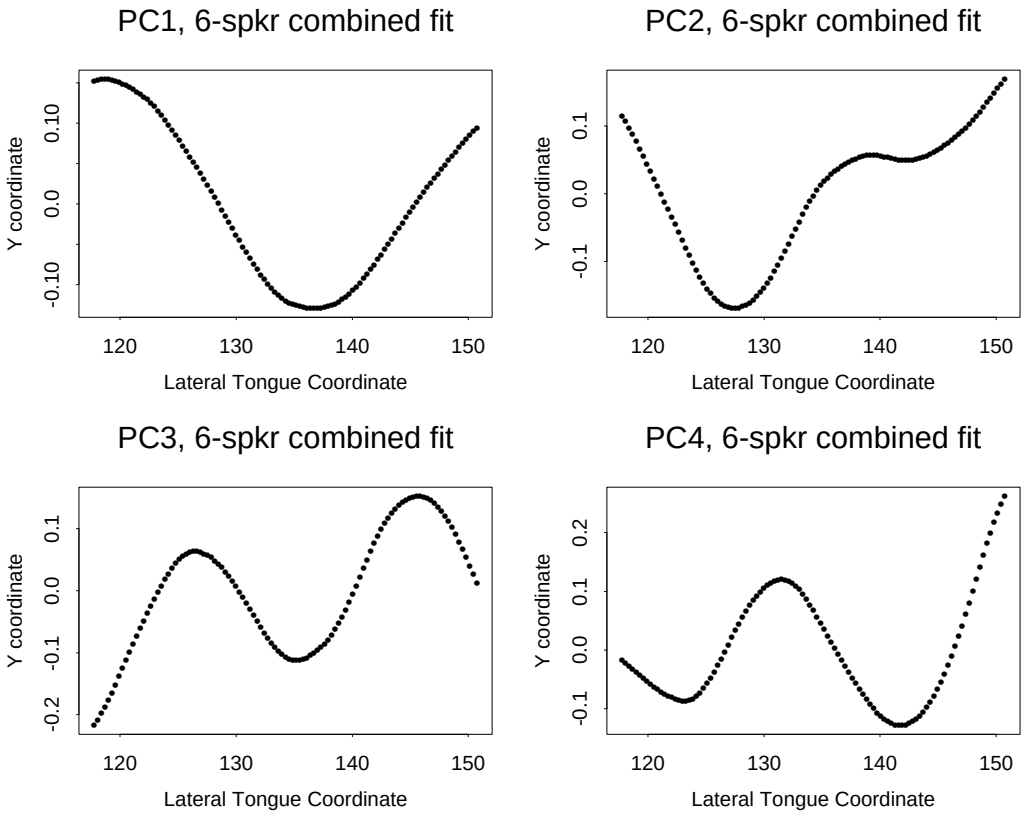

Figure 3: The first 4 non-constant PC's based on superposition over a common x-range of overlaid tongue curves with endpoint-average padding, for the full 6-speaker dataset. Data were preprocessed by shifting curves exactly as for Figure 2, with common x-range found after extending and padding shifted curves using parameters *gap*= 10 and *flat*= 4: preprocessing was done according to Plan 12, Table 2. Percent of variance (after subtraction of curve means) accounted for by the four PC's was respectively 62.4, 17.1, 12.3, and 3.4.

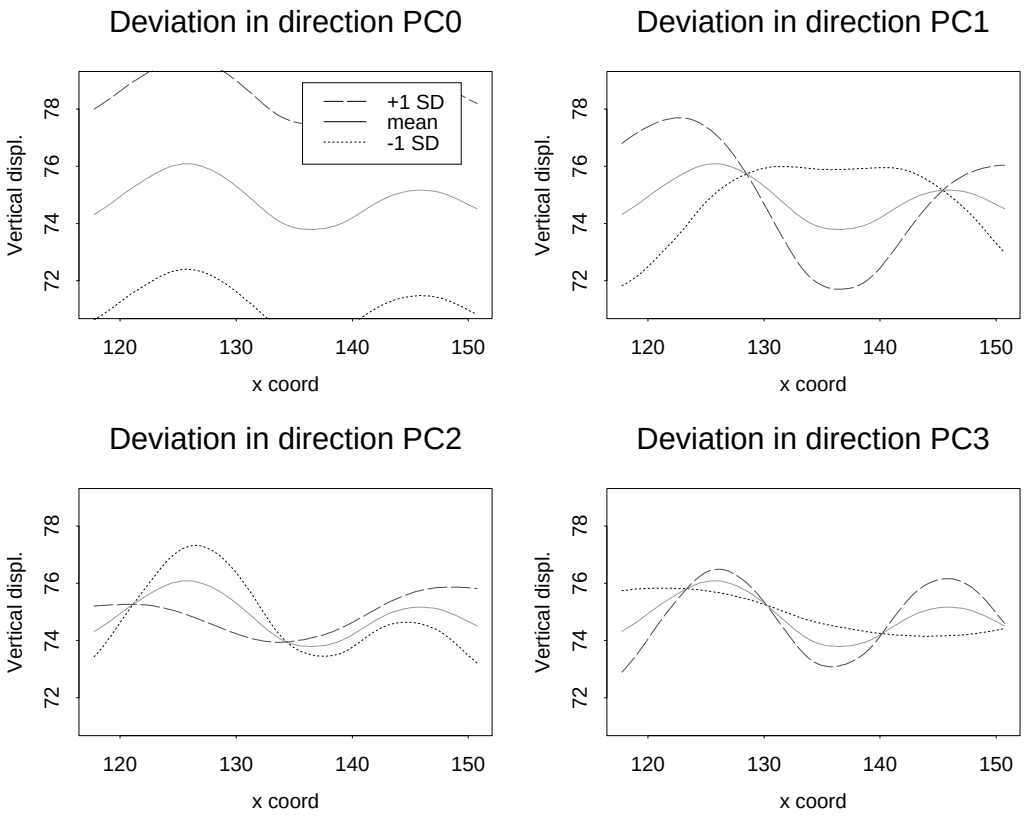

Figure 4: Deviations from mean tongue-curve in the 6-speaker dataset, by ±1 standard deviation of loading in direction of each of the first four PC's.

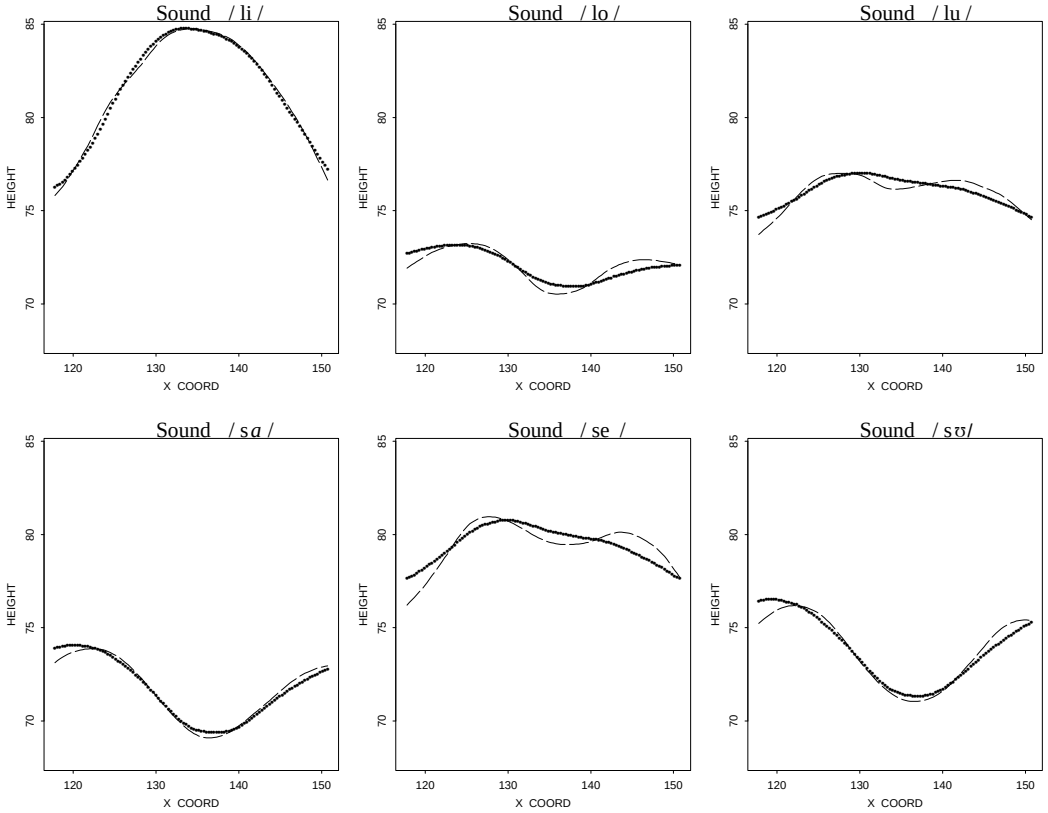

Figure 5: Display of 2-PC models for 6 sounds, after preprocessing by Plan 12 exactly as in Figure 3, including 4 fits with larger errors: tongue curves averaged over replication, session, and speaker (dashed line), and model based upon fitting a constant level plus linear combination of 2 PC's (solid line). The 4-PC model fits are virtually identical to the dashed lines.

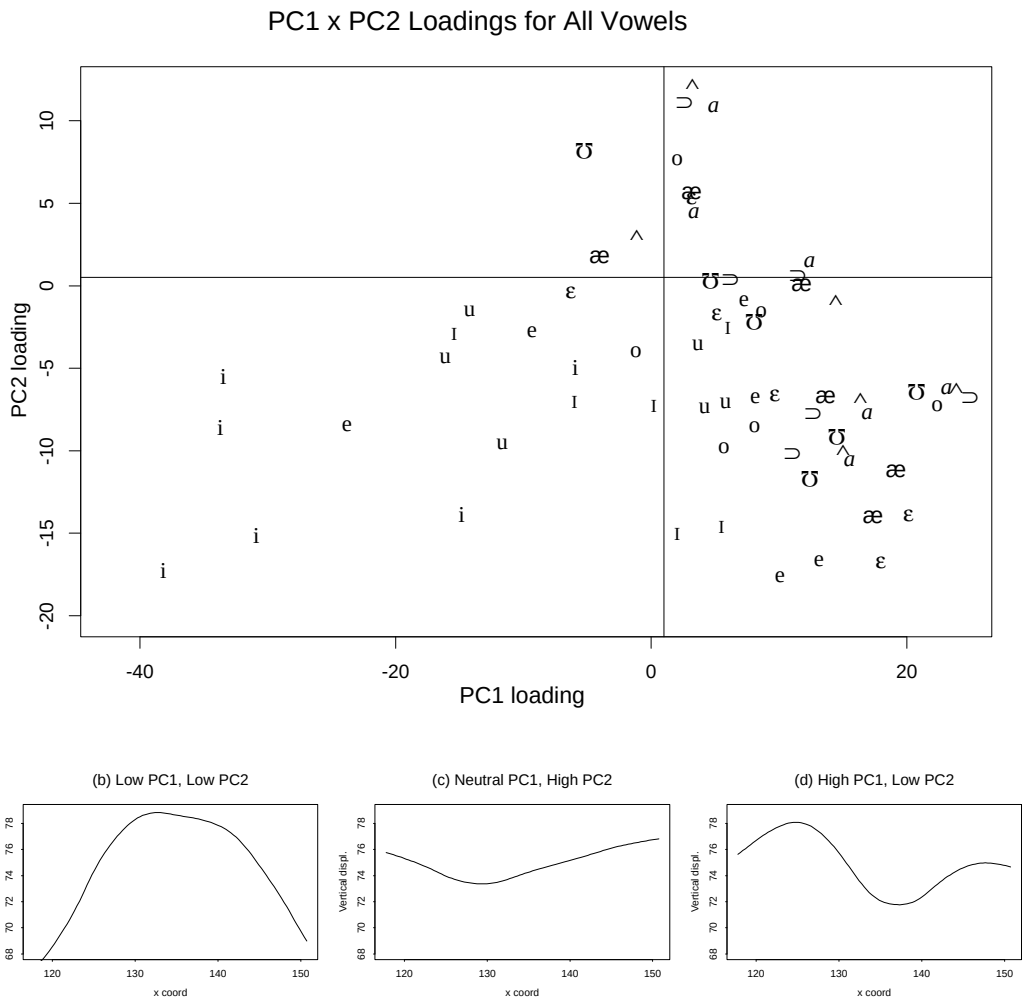

Figure 6: (a) Scatterplot over PC1 versus PC2 loadings, averaged over session and replication, for /l/-context vowels of 6 speakers. (b) Tongue contour corresponding to the PC0–4 model with PC0, PC3, and PC4 loadings taking values equal to their average over the dataset, and loadings for PC1 and PC2 respectively set to the values -35 for P1 and -15 for PC2. (c) Tongue contour for PC0–4 model with average loadings for PC0, PC3, and PC4, and with PC1 loading 3, PC2 loading 10. (d) Tongue contour for PC0–4 model with average loadings for PC0, PC3, and PC4, and with PC1 loading 18, PC2 loading -11.

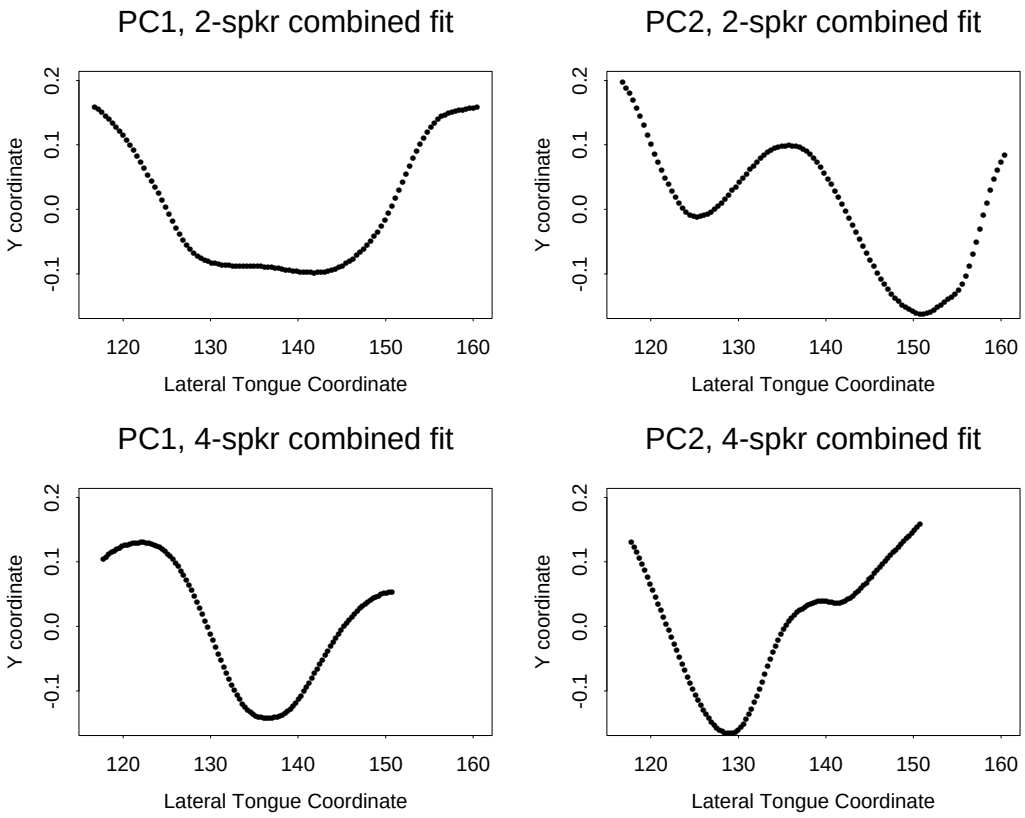

Figure 7: First two (non-constant) PC's for dataset on two speakers with broad tongues (SG, LG — upper pictures) and for dataset on four other speakers (MS, MD, CS, GW — lower pictures). Note that the range (117.8, 150.7) of x-coordinates for the 4-speaker PC's was considerably shorter than the range (116.8, 160.4) for the 2-speaker PC's. Percentages of variance (after subtraction of constant mean level for each curve) accounted for by PC1 and PC2 were respectively 72.8, 11.8 in the upper pictures and 69.7, 15.5 in the lower.

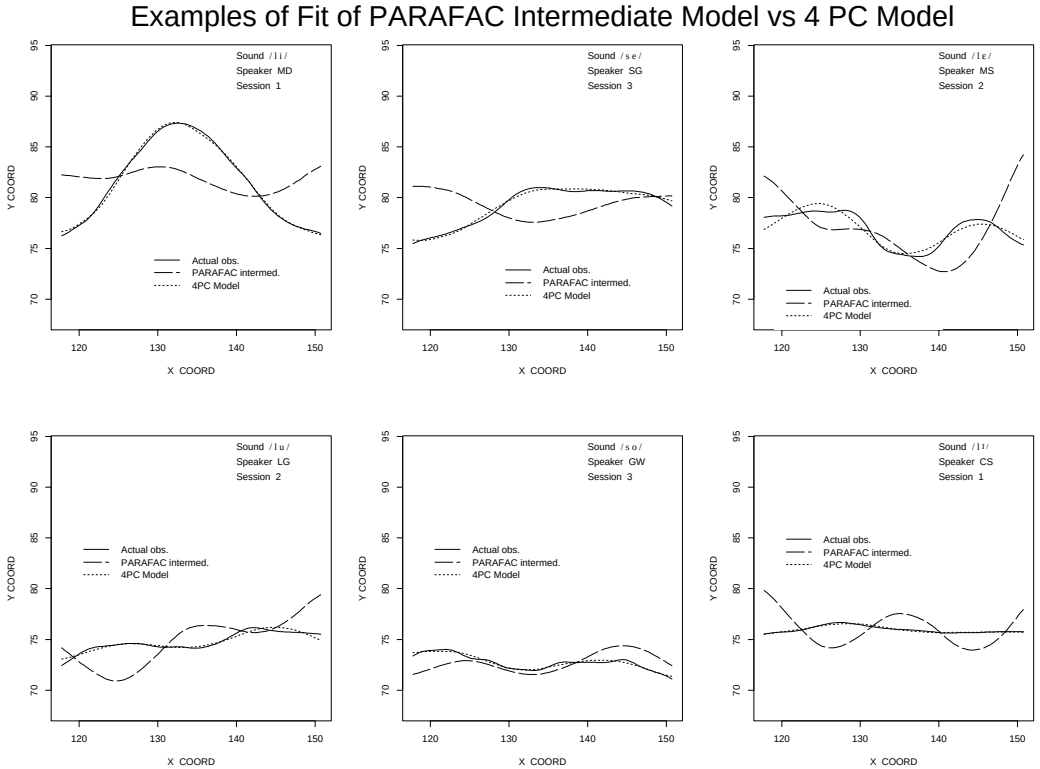

Figure 8: Comparison of fit, for 6 selected combinations of speaker, sound, and session, of the intermediate PARAFAC model (13) [long-dashed line] versus the actual data averaged over replications [solid line] and the model based on constant level plus 4 PC's [short-dashed line], after preprocessing data via Plan 12.