PARAFAC analysis of the three dimensional tongue shape

Yanli Zheng and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Shamala Pizza

University of California at Los Angeles, LA, CA 90095

(Received

3D tongue shape during vowel production is analyzed using the three-mode PARAFAC model. 3D MRI images of five speakers (9 vowels) are analyzed. 65 virtual fleshpoints (13 segments along the rostral-caudal dimension and 5 segments along the right-left direction) are chosen based on the interpolated tongue shape images. Methods used to adjust the alignment of MRI images, to set up the fleshpoints and to measure the position of the fleshpoints are presented. PARAFAC analysis of this 3D coordinate data results in a stable two-factor solution that explains about 70% of the variance.

PACS Numbers: 43.70.Aj, 43.70.Bk

I. INTRODUCTION

Harshman et al. (1977) demonstrated that a vector of midline-orthogonal vocal tract widths, measured from lateral X-ray cineradiograph tracings, can be represented using a specific three-index factor analysis model called PARAFAC (parallel factors). In their analysis, every tongue shape was represented as a 13-dimensional vector of vocal tract width measurements, arranged in order from the region of the epiglottis to the region of the tongue tip. PARAFAC is similar to standard two-index factor analysis in that it models every vocal tract width measurement as the product of two independent terms: a term that depends on the vocal tract position (commonly called the "factor shape"), and a term that depends only on the vowel identity and speaker identity (commonly called the "factor weight"). PARAFAC differs from standard two-index factor analysis in that the factor weight, in turn, is represented as the product of a vowel-independent speaker weight and a speaker-independent vowel weight. The highly constrained form of the PARAFAC model is both the reason that it is useful, and the reason that it is rarely used in practice. In practice, PARAFAC is rarely used because there are few datasets in nature that satisfy the PARAFAC constraints. When the PARAFAC constraints are satisfied, however, the constrained form of the analysis results in an intuitively appealing simplification of the data. In particular, the vowel weights define a phonemic vowel space that is speaker-independent in the strong sense of that word: if the weights on a particular factor of the vowels /i/ and /u/ are related by a ratio of 2:1 for one speaker, then they are related by a ratio of 2:1 for every speaker.

In practice, PARAFAC has been found to apply to very few types of natural data. Specifically, the literature indicates that the success of PARAFAC analysis depends strongly on the methods used to acquire data, the choice of index variables, and the methods used for statistical pre-processing. Among the many possible measurements of speech articulation, successful PARAFAC analysis has only been reported for radial tongue heights from lateral cineradiograph tracings (Harshman et al., 1977), pseudo-fleshpoint coordinates extracted from lateral cineradiography tracings (Nix et al., 1996) and electromagnetic midsagittal articulometry (EMA; Hoole, 1999). Among the many possible index variables of interest in speech production, PARAFAC has only been reported to explain the interaction among rostral-caudal vocal tract position, speaker identity, and the phonological features of tongue height and tongue fronting. Hoole (1999) found that PARAFAC was unable to model consonant context, and Geng and Moosehammer (2000) found that PARAFAC was unable to model lexical stress. Finally, Harshman and Lundy (1984) report that the success of a PARAFAC analysis depends on the use of adequate statistical pre-processing. The pre-processing methods of Harshman et al. (1977) include implicit vocal tract length normalization, and explicit subtraction of each speaker's mean tongue shape. Hoole (1999), Nix et al. (1996), and Geng and Moosehammer (2000) also report subtracting each speaker's mean tongue shape, but apparently neither study found it necessary to perform any type of vocal tract length normalization.

The purpose of this paper is to demonstrate that PARAFAC successfully represents the three-dimensional shape of the tongue surface extracted from coronal magnetic resonance

(MR) image stacks. Two types of measurement vectors are analyzed: a vector of 3D pseudo-fleshpoint coordinates extracted uniformly from the length and width of the tongue surface, and a vector of 2D pseudo-fleshpoint coordinates extracted from a curve along the tongue surface close to the midsagittal plane. The 2D pseudo-fleshpoint coordinates are structurally similar to the type of data analyzed by Nix et al. (1996). Measurement data are indexed by speaker identity, phonemic vowel identity, and two-dimensional measurement position. A variety of data pre-processing strategies were attempted; the method that yields the best results is similar but not identical to the pre-processing methods of Nix et al. (1996).

II. BACKGROUND: THREE-WAY FACTOR ANALYSIS MODELS

The model used in three-way factor analysis (PARAFAC; Harshman et al., 1977) is represented as follows:

(1) $$d_{ijk} = \sum_{w=1}^{F} c_{iw} v_{jw} s_{kw} + e_{ijk}$$

(2) $$i = 1,2,\cdots,n_c, \qquad j = 1,2,\cdots,n_v, \qquad k = 1,2,\cdots,n_s,$$

where d_{ijk} is the displacement of the i^{th} measurement during utterance of the j^{th} vowel by the k^{th} speaker, c_{iw} is the relative effect of factor w on the displacement of the i^{th} measurement, v_{jw} is the relative contribution of factor w on the j^{th} vowel, s_{kw} is the relative contribution of factor w on the k^{th} speaker, F is the number of factors, n_c is product of the number of fleshpoints (throughout this paper, $n_{fp}=13$ for midsagittal analysis, and $n_{fp}=65$ for three dimension analysis) times number of the dimensions of the

analyzed images, e.g. for mid-sagittal images, $n_c=n_{fp}\times 2=26$; and for 3D images, $n_c=n_{fp}\times 3=195$, n_v is the number of vowels, and n_s is the number of speakers.

PARAFAC analysis is capable of finding only those factors that are used in the same way by all speakers. The PARAFAC model assumes that, for example, if a speaker uses 20% more than the average amount of factor 1 in production of /A/, then he must also use 20% more than the average amount of factor 1 in production of all vowels. The PARAFAC model is too restrictive to accurately represent stress distinctions (Geng and Mooshammer, 2000) or consonant context (Hoole, 1999), but the results of Harshman et al. (1977) indicate that vowel height and fronting may be modeled using PARAFAC.

A more flexible alternative to PARAFAC analysis is provided by the Tucker3 model, represented as follows (Bro, 1998)

(3) $$d_{ijk} = \sum_{w_1=1}^{F_1}\sum_{w_2=1}^{F_2}\sum_{w_3=1}^{F_3} c_{iw_1} v_{jw_2} s_{kw_3} g_{w_1w_2w_3} + e_{ijk}$$

(4) $$i = 1,2,\cdots,n_c, \qquad j = 1,2,\cdots,n_v, \qquad k = 1,2,\cdots,n_s,$$

The definitions of d_{ijk}, c_{iw_1}, v_{jw_2}, s_{kw_3}, n_c, n_v n_s are the same as those in the PARAFAC model, except that a core matrix G $(F_1 \times F_2 \times F_3)$ is introduced, where $g_{w_1w_2w_3}$ is the element of G. The flexibility of the Tucker3 Model allows it to represent regularities that cannot be represented using PARAFAC, but because of this flexibility, it can be difficult to determine when the correct number of factors has been chosen.

Kiers (1991) shows that PARAFAC can be considered a constrained version of the Tucker3 model. Although Tucker3 will always fit better than PARAFAC, Tucker3 tends to use excess parameters when the PARAFAC model is adequate (which is easy to observe from the mathematical model of PARAFAC and Tucker3). Another obvious merit of PARAFAC is its structural uniqueness, whereas the Tucker3 model has rotational freedom and thus gives an ambiguous result.

Based on the considerations above, and based on the result of Harshman et al. (1977), the PARAFAC model was chosen for the analyses reported in this article. Comparison between PARAFAC and Tucker3 analysis results will provide a method for verifying the statistical significance of any given PARAFAC analysis order.

III. METHODS

In this study, magnetic resonance images of five speakers of American English were collected during production of nine English vowels (/A, Θ, E, ɔ, ɛ, Y, ʊ, I, ɪ/). The shape of the tongue was manually transcribed on all coronal images. Reconstructed tongue shapes were analyzed using the PARAFAC three-way factor analysis algorithm.

A. Data Acquisition

The vowel tongue shapes of five speakers (three male: m1, m2 and m3, two female: f1 and f2) were imaged using a fast gradient echo magnetic resonance imaging protocol (GE 1.5T SIGNA scanner, T_e=1.9ms, T_r=150-300ms depending on subject). Four speakers

were natives of Southern California, while one speaker (m2) was a native of northern New York State.

Subjects produced the English vowels /A, Θ, E, o, ɛ, Y, ʊ, I, ɪ/. Subjects were told to imagine the words “father, bat, bet, boat, bait, put, boot, bit, beat,”” respectively, as a guide to pronunciation. Subject training was conducted prior to imaging. During training, subjects lay on a couch in a sound-treated recording room, in a supine position similar to subject position in the MRI scanner. Subjects were then asked to sustain for 12-15 seconds. Pronunciation errors were corrected by the experimenter. When subjects were comfortable with the vowel list, each subject recorded each vowel three times, again sustaining each vowel for 12-15 seconds per repetition (speaker m1 recorded only one sustained example of each vowel). Acoustic recordings were analyzed using the Entropic formant tracker (Talkin, 1987). Vowels produced by subject m3 were too breathy for accurate formant frequency analysis, but were judged to be perceptually valid exemplars of the desired vowels. Formant frequency trajectories of the vowels of every other speaker were measured approximately 1/3, 1/2, and 2/3 of the way through the vowel. All formant frequency measurements were compared to the values reported in (Peterson et al., 1952), and were found to be within the range of variation predicted by that paper.

Coronal and axial image stacks were collected for each vowel, but only coronal images are analyzed in this article. Images are 256x256 pixels, covering a 24cm field of view, with a 3mm inter-slice interval. Each image stack was collected in two breath holds of 15-20 seconds in duration: odd-numbered images were collected in the first breath hold,

and even-numbered images were collected in the second breath hold. Subjects m1 and f1 were instructed to phonate during imaging, but the resulting images suffer from substantial vibration artifact, especially in the region of the vocal folds, ventricular folds, and aryepiglottic ligament. Subjects m2, f2 and m3 were instructed to hold their breath during imaging while maintaining a posture appropriate to the requested vowel.

Images were segmented by hand using custom MRI display and segmentation software (Hasegawa-Johnson et al., 1999a, 1999b). On each coronal image, the location of the entire tongue was outlined, including the dorsal surface of the tongue and all lateral surfaces in contact with the cheek or interdental airway. The remainder of the vocal tract (borders of the palate, teeth, cheek, lips, and pharynx) was outlined in a different color, and the positions of the peak of the palatal dome and of the four gingival margins were marked.

All images and segmentation files are available from the second author, or over the Internet at “http://www.ifp.uiuc.edu/speech/mri”.

B. Image Alignment

Preliminary analysis of our data revealed that different subjects were positioned at different angles within the scanner, and that despite the head restraint, some subjects changed position between one vowel and the next. Before the statistical analysis, it is necessary to correct the misalignment error across the collected data and to measure the 3D shape in a systematic way. The subject head front/back declination (pitch; shown schematically in Fig. 1(a)) was quite obvious in our dataset. Image-alignment is very

important before the fleshpoints-measurements. Tiede (2000) pointed out that measurements according to the tract midline (i.e. method used by Harshman et al. (1977)) are difficult to begin with and sensitive to asymmetrical articulation. In his report, a modified "shape based" semi-polar method was used, which takes into consideration the shape of the hard palate. In our study, we adopted the idea in Tiede's report for our image alignment.

The steps that were taken to align tongue shape data are as follows:

Step 1: Generating pseudo-sagittal images. Since only coronal images were used in our analysis, pseudo-sagittal images were constructed only for the purpose of placing reference points. Pseudo-sagittal images were generated by resampling a sagittal slice through the coronal image stack. Fig 1 (b) shows a pseudo-sagittal image with reference points marked.

Step 2: Marking reference points in the midsagittal plane. As illustrated in Fig. 1 (b), three reference points (R1, R2 and R3) were used, where R1 is the maxillary incisor, R2 is the highest visible point of the palatal vault, and R3 is the point on the rear pharyngeal wall adjoining the anterior apex of the second cervical vertebra.

Step 3: Adjusting the pitch error. R1, R2 and R3 determine a circle, and C is the center of the circle. The sagittal images were rotated by an angle θ, where θ is the angle between the vertical axis and the C-R2 line segment. Fig. 1 (c) shows the midsagittal view before and after the rotation respectively.

C. Measurement Vector Composition

Harshman et al. (1977) measured the vocal tract width along dorsoventral gridlines. Nix et al. (1996) point out that this method reduces the ability of the model to represent important inter-dimensional correlations; instead, they propose the analysis of "virtual flesh-points" located by segmenting the arc length of the tongue. In our study, spatially unconstrained factors were analyzed with PARAFAC using the (X, Y, Z) coordinates of sixty-five fleshpoints, with thirteen points in the posterior-anterior direction i.e. Y direction, and 5 points in the left-right direction i.e. X direction.

From the coronal tongue outlines, the 3D tongue shape was reconstructed using piece-wise-linear interpolation. In order to determine the location of virtual fleshpoints, the raw data was projected onto the X-Y plane. Projection into the X-Y plane results in many-to-one projection near the lateral margins, where the tongue bulges into the interdental airway. In order to simplify analysis, first, all points inferior to the bulge were discarded, so that tongue height is a single-valued function of X and Y; second, because of the noise at the edge of the X direction in the extracted tongue surface, one twelfth is cut from both edges in the X direction. Then the fleshpoints are placed based on the extracted tongue surface, using the following procedure. First, uniform sampling along the X direction in every coronal slice generates five fleshpoints in each coronal plane. Second, points are connected to form five rostral-caudal piece-wise linear "curves" along the tongue surface. Thirteen fleshpoints are placed on each curve using an equal curve length criterion. Fig. 2 shows the two steps for placing the fleshpoints in the X-Y plane. The Z values (i.e. low/high positions) of the fleshpoints are obtained by piece-wise linear

interpolation between coronal planes. “Equal-curve length” is defined as equal length along each of the five rostral-caudal curves. The main purpose of this step is to distribute the fleshpoints as uniformly as possible over the surface of interest.

Two types of measurement vectors were constructed for subsequent pre-processing and PARAFAC analysis. One of the measurement vectors contained 3D coordinate data from the sixty-five reconstructed fleshpoints described above. The 3D measurement vectors contained the Cartesian (X, Y, Z) coordinates of each fleshpoint, as previously defined. Each 3D measurement vector contained a total of 65×3=195 measurements. In addition to the 3D measurement vectors, 2D measurement vectors were also constructed. Each 2D measurement vector contained a 2D representation of every fleshpoint on the median curve in Fig. 2(b). The 2D Cartesian-coordinate vector contained (Y, Z) measurements of each of these points. Each 2D measurement vector contained a total of 13×2=26 measurements.

The 2D measurement vectors are of interest for two reasons. First, the 2D measurement vectors are similar to the measurement vectors used by Nix et al. (1996). By comparing the 2D measurement results in this paper to the results of Nix et al., it is possible to qualitatively evaluate some of the structural measurement differences that result from the differences in our measurement methodologies (the comparison is not perfect, because of differences in the subject pool; see the discussion section). Second, the 2D measurement vectors are interesting because each 2D measurement dataset is a strict subset of one of the 3D datasets: the data in the 2D Cartesian coordinate vector are a strict subset of the

data in the 3D Cartesian coordinate vector. Harshman and Lundy (1984) proposed the use of split-half analysis in order to determine the validity of a PARAFAC solution. In split-half analysis, the dataset is split into two or more subsets along one of the index variables; if factor weights along the two non-split index variables are unaffected by the split, Harshman and Lundy argue that the PARAFAC analysis should be considered valid. In our case, similarity between the vowel weights and talker weights of 2D and 3D analysis may be considered evidence that the PARAFAC model of the 3D dataset is valid.

D. Data Pre-processing

To make the data compatible with the PARAFAC model, preprocessing is required. In this research, two preprocessing equations were used in order to improve the effectiveness of PARAFAC analysis: removing the mean across the vowels in some dimensions as defined in Eq. (5), and scaling the inter-speaker variation in some dimensions as defined in Eq. (6). In these equations n=1,2,3, is the index of the coordinates (e.g. 1 is the X, 2 is the Y, and 3 is the Z coordinate), and n_s, n_{fp}, n_v are defined as in Eq. (1). The matrix for PARAFAC analysis, d_{ijk} in Eq. (1), is defined by putting $d^{(1)}{}_{ijk}$, $d^{(2)}{}_{ijk}$ and $d^{(3)}{}_{ijk}$ in the same matrix sequentially along the dimension i.

(5)
$$d^{(n)}{}_{ijk} = d^{(n)}{}_{ijk} - \mu_{ik}^{(n)}, \qquad \mu_{ik}^{(n)} \equiv \frac{\sum_{j=1}^{n_v} d^{(n)}{}_{ijk}}{n_v}$$

(6)
$$d_{ijk}{}^{(n)} = \frac{d_{ijk}{}^{(n)}}{s_k^{(n)}}, \qquad s_k^{(n)} = \sqrt{\sum_{i=1}^{n_{fp}} \sum_{j=1}^{n_v} (d_{ijk}^{(n)})^2}$$

Bro (1998) specifies two criteria for the selection of PARAFAC pre-processing algorithms. He demonstrates, both theoretically and empirically, that violation of either criterion may change the statistical structure of the data. For example, data that are properly explained by a two-factor PARAFAC solution may seem to support a three-factor solution if the data are improperly pre-processed. The two criteria are, first, that mean subtraction must be within a single mode of the data matrix, and second, that all scaling operations must follow any mean subtraction operations.

The single-mode mean subtraction rule means that the sum in Eq. (5) must be a sum over only one index variable. The subtraction before scaling rule means that equations like Eq. (6) must follow equations like Eq. (5). In Eq. (5), the mode variable is *j*, the vowel mode. Equation (5) is thus identical to the statistical pre-processing applied by Harshman et al. (1977) and Nix et al. (1996).

In addition to these two essential criteria, Bro (1998) suggests two additional criteria without proving their effect on the statistical results. First, he suggests that scaling should be done "within" a mode, i.e., the sum in an equation like Eq. (6) should be a sum over two of the mode variables (in Eq. (6), the sum is over mode variables *i* and *j*, the fleshpoint and vowel modes). Second, he suggests that one of the modes in the summation in Eq. (6) should always be the mode that was subjected to mean subtraction in Eq. (5).

Provided that all four criteria are satisfied, Bro (1998) demonstrates that scaling does not change the number of factors in the PARAFAC solution. Instead, scaling as in Eq. (6) is equivalent to the use of a weighted least-squared-error metric in the PARAFAC model fitting process. The effect of Eq. (6), in particular, is to normalize each talker's data to unit sum-squared variation, so that talkers with greater variability and/or larger vocal tracts do not dominate the PARAFAC fitting process.

Subject to Bro's ordering constraints, experiments were performed to determine the combination of mean-subtraction and scaling operations with the highest percentage of variance explained by both 2D and 3D PARAFAC models. Operations similar to Eq. (5) were applied to the inter-talker mean, the inter-vowel mean, or the inter-fleshpoint mean (three possibilities). The resulting data were either scaled or not scaled; if scaled, all possible combinations of index variables were considered. Similar operations were repeated for all possible combinations of X, Y, and Z coordinate data, and the resulting processed data were in each case fitted using a PARAFAC model. The highest percentage of variance explained by PARAFAC modeling of any candidate dataset was 70.6% for 3D analysis, 76.2% for 2D analysis. The optimum pre-processing strategy was not unique: more than a dozen different candidate pre-processing strategies resulted in the same percentage of variance explained. Among the set of optimum pre-processing strategies, the strategy with the simplest algorithm is as follows:

Step 1: Remove the mean across the vowels using Eq. (5) in all the three coordinates that are used in the analysis;

Step 2: Correct for inter-speakerscale differences in the low-high and left-right coordinates (X and Z Coordinates) using Eq. (6).

E. Choosing the Number of Factors

The core consistency diagnostic (Bro, 1998) is a new approach for determining the appropriate number of components for a PARAFAC model. By observing the mathematical models of PARAFAC and Tucker3, one discovers that PARAFAC is a restricted Tucker3 model in which the core matrix *G* is defined to equal the super-diagonal matrix *I*. The definition of a super-diagonal matrix *I* is as follows:

(7)
$$i_{w_1 w_2 w_3} = \begin{cases} 1, & where\, w_1 = w_2 = w_3 \\ 0, & otherwise \end{cases}$$

The "goodness of fit" of a PARAFAC model can be verified by first computing t_{iw}, v_{jw} and s_{kw} using PARAFAC, and then calculating the least squares Tucker3 core *G* given t_{iw}, v_{jw} and s_{kw}. If the PARAFAC model is valid, the core matrix *G* should be super-diagonal. By simply monitoring the distribution of super-diagonal and off-super-diagonal elements, one can assess whether the model structure is reasonable or not. There are two simple criteria: 1) the super-diagonal elements should be all close to one, and 2) off-super-diagonal elements should be close to zero. If these criteria are met, the model is not over fitting. If these criteria are not met, then either too many components have been extracted, the model is mis-specified, or gross outliers disturb the model. It is possible to calculate the super-identity of G to obtain a single parameter for the model quality. The super-identity or core consistency is defined here as:

(8)
$$Core\ Consistency = 100 \times \left(\frac{1 - \sum_{w_1=1}^{F} \sum_{w_2=1}^{F} \sum_{w_3=1}^{F} (g_{w_1 w_2 w_3} - i_{w_1 w_2 w_3})^2}{\sum_{w_1=1}^{F} \sum_{w_2=1}^{F} \sum_{w_3=1}^{F} (i_{w_1 w_2 w_3})^2} \right)$$

Where $g_{w_1 w_2 w_3}$ are elements of the core matrix *G*. The probability distribution of Core Consistency has not been analyzed, but Bro et al. (2002) suggests that a core consistency of at least 90% is a good indicator of a valid model.

F. Checking the Reliability of the Solutions

Generally there are two steps to check the reliability of the solutions (Bro, 1998). First, one should check for the convergence of the solution, i.e., the algorithm should reach the same solution when initiated from several different random start points. Second, one should check for the signs of a degenerate solution. A typical sign of degeneracy is that two of the components become almost identical, but with opposite contributions to the model.

In the PARAFAC model, each rank-one component Z_w can be expressed as the vectorized rank-one array obtained as follows, where $\otimes$ denotes the Kronecker tensor:

(9)
$$Z_w = C_w \otimes V_w \otimes S_w,\ w=1,2,\ldots,F,$$

(10)
$$C_w = [c_{1w}, c_{2w}, \cdots, c_{n_c w}]',\ V_w = [v_{1w}, v_{2w}, \cdots, v_{n_v w}]',\ S_w = [s_{1w}, s_{2w}, \cdots, s_{n_s w}]'$$

For the degenerate model, the loading vectors in component *f* and component *g* will be almost equal in shape, but negatively correlated. The similarity between component *f* and

component g can be measured using the congruence coefficient $\cos(Z_f,Z_g)$ (Tucker, 1951), defined as:

$$\cos(Z_f, Z_g) = \cos(C_f, C_g)\cos(V_f, V_g)\cos(S_f, S_g) \tag{11}$$

Bro (1998) suggests that a degenerate model is indicated whenever the congruence coefficient between any components f and g is less than or equal to -0.85.

IV. RESULTS

The percentage of variance explained for 2D mid-sagittal analysis and 3D analysis are shown in Table I. For purposes of comparison, Table I also shows the results reported by Harshman et al. (1977) and Nix et al. (1996). Section A describes the results of core consistency tests and congruence coefficient tests conducted in order to verify that a two-factor PARAFAC analysis is appropriate for these data. Factor shapes, vowel loadings and speaker loadings of PARAFAC analysis from both 2D and 3D analysis are described in sections B and C.

A. Core Consistency Diagnostic and Congruence Coefficients

As mentioned in the last section, the *core consistency diagnostic* is used to decide how many factors underlie variations in the data. Fig. 3 shows core consistency results for one, two, three, and four-factor PARAFAC models. The tests of both 2D and 3D Cartesian models are shown in Fig. 3. The two-factor PARAFAC model has a Core Consistency of nearly 100%. The three-factor model has a Core Consistency of about 65% (2D) to 82% (3D). Core Consistency measurements in this range are described as

"problematic" by Bro et al. (2002); they report on some valid solutions with Core Consistency measurements in this range, but they do not claim that Core Consistency measurements in this range always indicate a valid model. In order to ensure that the results are meaningful, the rest of this paper will assume a two-factor PARAFAC model.

Congruence coefficients for two-factor PARAFAC models are listed in Table I. Table I, supporting the conclusions of Fig. 3, shows that two-factor PARAFAC is appropriate for these data.

B. Analysis of Semi-midsagittal data

Results of the PARAFAC analyses of the semi-midsagittal data using Cartesian coordinates are shown in Figs. 4. The semi-midsagittal data consist of the (Y, Z) coordinates of the 13 fleshpoints located on the middle "curve" in Fig. 2 (b).

Fig. 4 shows the shape factors, vowel weights, and speaker weights resulting from the results of analyzing Cartesian coordinate data. In these figures, the "mean shape" c_{i0} is defined as the average shape across all the speakers and all the vowels, c_{iw}^{*} is the c_{iw} after the scaling adjustment (meaning that the factors corresponding to the X-coordinates in the c_{iw} is back-scaled using $s^{(1)}$, and the factors corresponding to the Z-coordinates in the c_{iw} is back-scaled using $s^{(3)}$), while the "plus shape" c_{iw+}, "minus shape" c_{iw-} and s(n) are defined as:

(13) $$c_{iw+} = c_{i0} + c^{*}_{iw} \times \max_{j}(v_{jw}) \times \operatorname{mean}_{k}(s_{kw})$$

(14) $$c_{iw-} = c_{i0} + c^{*}_{iw} \times \min_{j}(v_{jw}) \times \operatorname{mean}_{k}(s_{kw})$$

(15) $$s^{(n)} = \frac{1}{n_s} \sum_{k=1}^{n_s} s_k^{(n)}$$

In the figure, shape factor 1 (upper plot in Fig. 4(a)) appears to represent raising or lowering of the entire tongue. The raised tongue has a peak about 2/5 of the way from the back, while the lowered tongue is flat over the anterior 2/3. Shape factor 2 (lower plot in Fig. 4(a)) appears to represent anterior-posterior movement of the entire tongue.

Vowel weights are shown in Fig. 4(b). In the analysis by Harshman et al. (1977), the vowel space formed a rough quadrilateral, with vowels arranged around the quadrilateral in the order /ι,ε,I,E,Θ,A,ο,Y, υ/. The arrangement of vowels in Fig. 4(b) is similar, except that /o/ has a higher factor 2 weighting than /υ, ο/.

In the results of Harshman et al. (1977), the vowel quadrilateral was tilted approximately 30 degrees relative to the factor axes; Harshman et al. described their factors as a "front raising" factor and a "back raising factor." Factor 2 seems to represent the distinction between back and front vowels. Factor 1 could be said to represent the distinction between high and low vowels. The vowels /E, ε, ι, Y,υ/ are produced with negative factor 1 weights, apparently indicating raising of the whole tongue. The vowels /A, Θ/ have positive factor 1 weights, apparently indicating lowering of the whole tongue. The vowels /Y, ο, υ/ are produced with positive factor 2 weights. apparently indicating

posterior tongue raising. The vowels /E, ɛ, I, ɪ, Θ/ have negative factor 2 weights, apparently indicating lowering of the posterior tongue.

Speaker loadings are shown in Fig. 4(c) Speaker f2 and f1 weights factor 1 most heavily, and speaker m3 weights factor 1 least heavily. Factor 2 is most heavily utilized by speaker m2. Speaker f2 has a negative factor 2 weight, meaning that the tongue shape measurements acquired from MRI of this subject may be troublesome.

C. Analysis of 3D Data

Results of the PARAFAC analyses of 3D Cartesian coordinate data are shown in Fig. 5 and Fig. 6. The “plus”, “minus” and “mean” shapes are defined same as in the analysis of semi-midsagittal data.

Midsagittal views of the shape factor are similar to the 2D case. The correlation coefficient between the 2D factors and the corresponding 26 dimensions of the 3D factor is 0.99. And vowel and speaker loadings are almost the same as the 2D case with correlation coefficients 0.98 and 0.99 respectively. This is an interesting result because the results of the 2D analysis is exactly split half analysis (Harshman et al. (1984)) of the 3D analysis. The result also confirms that two-factor PARAFAC is suitable for our data.

V. DISCUSSION AND CONLUSION

The two-factor PARAFAC model is capable of extracting linguistically meaningful and statistically valid 2D and 3D factors from our MRI-derived tongue data. The 2D analysis

works as the split-half analysis of the 3D analysis. Splitting the measurement vector into a pseudo-midsagittal part and a non-midsagittal part does not change analysis results in the nonsplitted modes (i.e. vowel and speaker modes).

The 2D MRI measurement vector analyzed in this paper is similar in form to the X-ray measurement vector analyzed by Nix et al. (1996), but the results of PARAFAC analysis are not identical. The percentage of variance explained using MRI data is lower than the percentage of variance explained using X-ray data (Table I), and the vowel space is somewhat distorted. In order to better understand the difference in results between PARAFAC of X-ray and PARAFAC of MRI, it is worth considering the similarities and differences between our methodology and the methodology of Nix et al.

First, like the X-ray measurement vector of Nix et al., the 2D MRI measurement vector consists of the Y and Z coordinates of 13 fleshpoints uniformly distributed between the tongue tip and the base of the tongue. Nix et al. spaced their fleshpoints uniformly in a sagittal (Y-Z) projection, while ours were spaced uniformly in an axial (X-Y) projection; as a result, our data cover the dorsum of the tongue more densely and the tongue root less densely than the data of Nix et al. We have compared pseudo-sagittal and coronal segmentations of the same data, and the comparison indicates that coronal segmentation is an accurate representation of tongue position as far back as the tip of the epiglottis. Position of the tongue root between tip of the epiglottis and base of the epiglottis was part of Nix et al.'s measurement vector, but not ours, and this difference may explain some of the differences in the resulting vowel spaces.

Second, Nix et al. pre-processed their data by removing the inter-vowel average value of each measurement. Our pre-processing methods included the same inter-vowel mean subtraction, as well as inter-talker scaling of the X and Z coordinates. Scaling after mean subtraction is guaranteed not to change the statistical structure of the data (Bro, 1998), but will change the relative weight assigned by the PARAFAC algorithm to fitting errors incurred by different talkers. We found that scaling using Equation 6 improved the percentage of variance explained by the model, but did not change the resulting vowel space. Other types of pre-processing sometimes changed the vowel space substantially, and are therefore not reported in this paper.

Third, the difference between tongue surface measurements acquired using MRI and those acquired using X-ray may be substantial, for several reasons. In an X-ray projection, there is no way to tell whether the observed tongue surface is midsagittal, right of midsagittal, or left of midsagittal. MRI produces true 3D images, so there is no lateral-midsagittal ambiguity. X-ray also has important advantages over MRI. First, bony anatomical landmarks such as the tip of the incisor and the second cervical vertebra are visible in X-ray, but are invisible or ambiguous in MRI. Inter-talker differences in head position can be easily corrected when bony fixed landmarks are visible. In MRI, inter-talker differences in head position must be corrected using less precise soft tissue landmarks, such as the pharynx wall and the peak of the palate. It is possible that ambiguity in the position of soft tissue landmarks may result in some imprecision in the measurement data analyzed using PARAFAC. The second important difference between

X-ray and MRI is in the type of speech being performed by subjects. The MRI data acquired in this paper were acquired during sustained production of the target vowels, while the X-ray data used by Nix et al. were frames selected from a cineradiograph of continuous fluent speech. Engwall (2000) compared the static vowel positions observed during MRI to the dynamic vowel extrema observed using electropalatography and electromagnetic articulometry. His data suggest that the vowels observed in MRI experiments are hyperarticulated, hypercanonical productions, related but not identical to the productions observed during fluent speech. Engwall's results suggest that the vowel space derived from an MRI study should be slightly more extreme than the vowel space derived from an X-ray study. It also seems reasonable to speculate that sustained phonation is more difficult for subjects than fluent articulation, and that the increased difficulty of the task may cause extra measurement variability that can not be explained using a two-factor PARAFAC model.

Fourth, the five talkers analyzed by Nix et al. (1996) were imaged prior to 1972 (Harshman et al., 1972). All of the five talkers that we imaged were born after 1972. It is possible that dialect changes since 1972 may have changed the tongue positions used to produce the vowels of English. All five talkers analyzed by Nix et al. (1996) were male; in our study, three talkers were male, and two were female. Male-female differences in vocal tract size may have caused extra variability in our data that was not well explained using a two-factor PARAFAC model.

VI. CONCLUSION

In summary, the interaction between speaker identity and vowel identity in MRI-derived tongue surface shape measurements can be modeled using a two-factor PARAFAC statistical model. The two-factor model is statistically valid and the resulting vowel space represents most of the information in the traditional tongue height and tongue fronting relationships.

ACKNOWLEDGMENTS

The work reported in this paper was supported by NIH Fellowship F32 DC 00323-01, and by a grant from the University of Illinois Research Board.

REFERENCES

Bro, R. (**1998**), "Multi-way analysis in the Food Industry, Models, Algorithms, and Application," PhD Thesis, Royal Veterinary and Agriculture University, Denmark

Bro, R and Kiers, H (2002), "A new efficient methods for determining the number of components in PARAFAC modesl," J. Chemom. (in press)

Engwall, O. (**2000**), "Are statical MRI data representative of dynamic speech?Results from a comparative study using MRI, EPG and EMA," *Proc ICSLP 2000 , I:17-20.*

Geng, C and Mooshammer, C. (**2000**), "Modeling the German stress distinction," Proc. 5th Speech Production Seminar, Kloster Seeon, Germany.

Peterson, G. and Barney, H. (1952), "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am., **24**(2):175-184.

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

Harshman, R. A., and Lundy, M. E. (1984), " The PARAFAC model for three way factor analysis and multidimensional scaling," in Research Methods of Multimode Data Analysis, edited by H. G. Law, C.W. Snyder, J.A. Hattie, and R.P. MacDonald (Praeger, New York, pp122-215.

Hasegawa-Johnson, M.A., Cha, J.S., Pizza, S., and Haker, K. (**1999a**), "MRCAT: A Case Study in Human-Computer Interface Design," in Proc. Int. Conf. on Public Participation and Information Technology, Lisbon.

Hasegawa-Johnson, M.A., Cha, J.S., and Haker, K. (**1999b**), "CTMRedit: A Matlab-based Tool for Viewing, Editing, and 3D Reconstruction of MR and CT Images," in Proc. Meeting BMES/EMBS, Atlanta, 1170.

Hasegawa-Johnson, M.A., Cha, J.S., and Haker, K. (**in review**), "Tongue Height and Formants Show Talker-Independent Vowel Categories, but Oral Area Does Not."

Heinz, J.M. and Stevens, K.N. (**1965**), "On the relation between lateral cineradiographs, area functions, and acoustic spectra of speech," Proc. Fifth Internat. Congress of Acoustics, Paper A44.

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

Kiers, H.A.L. (**1991**), "An efficient algorithm for PARAFAC of three-way data with large numbers of observation units," Psychometrika, **56**, 147.

Narayanan, S. S., Alwan, A.A., and Haker, K. (**1997**), "Towards articulatory-acoustic models of liquid approximants based on MRI and EPG data. Part I: The Laterals," J. Acoust. Soc. Am **101**, 1064-1077.

Nix, D. A., Papcun, G., Hogden, J. and Zlokarnik, I. (**1996**), "Two cross-linguistic factors underlying tongue shapes for vowels," J. Acoust. Soc. Am. **99**, 3707-3717.

Talkin, D. (1987), "Speech Formant Trajectory Estimation Using Dynamic Programming with Modulated Transition Costs," J. Acoust. Soc. Am. 82.

Tiede, M. K. (**2000**), "An MRI-based morphological approach to vocal tract area function estimation," ATR Technical Report TR-H-XXX.

Tucker, L. R. (**1951**), "A method for synthesis of factor analysis studies," Personal Research Section Report No. 984, Dept. of the Army, Washington D. C.

Table I. Variance explained and Congruence coefficients by two-factor PARAFAC models

Variance Explained			Congruence coefficients
Our Analysis	**2D**	**76.2%**	**0.08**
	3D	**70.7%**	**0.06**
Harshman et al. (1977)		**92%**	**0.07**
Nix et al. (1996)		**88%**	**N/A**

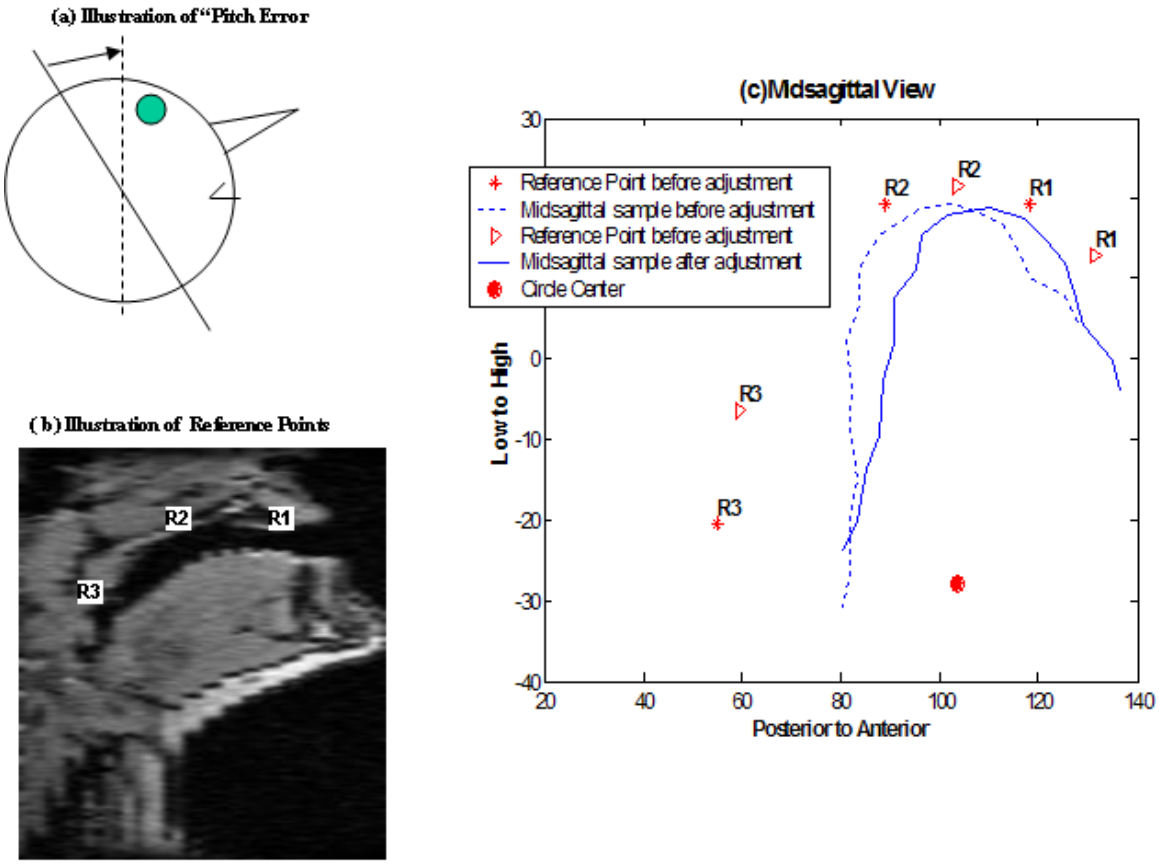

Figure 1. Image alignment methods. (a) Illustration of the "Pitch Error". (b) Illustration of the positions of reference points R1, R2 and R3. (c) Midsagittal view of tongue surface before and after alignment adjustment. Dotted line represents the tongue surface before the rotation adjustment, solid line represents the tongue surface after the rotation adjustment, star points represent the reference landmarks before the adjustment, left triangles represent the reference landmarks before the adjustment, and solid circle represents the center of the reference circle.

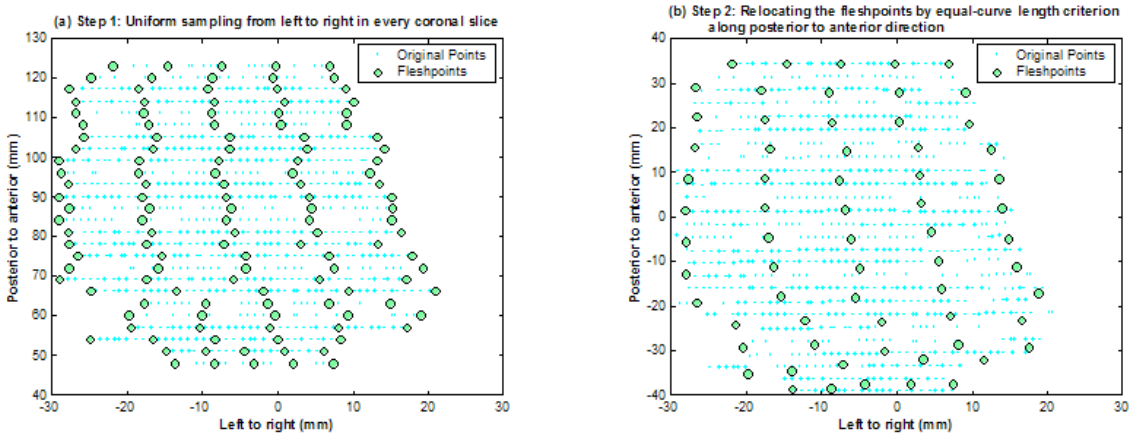

Figure 2. Two steps for placing the fleshpoints in the X-Y Plane. (a) Step 1, the fleshpoints are uniformly spaced from left to right in every coronal slice. (b) Step 2, the fleshpoints are repositioned according to an "equal-curve length" criterion along each of the five rostral-caudal tongue surface curves.

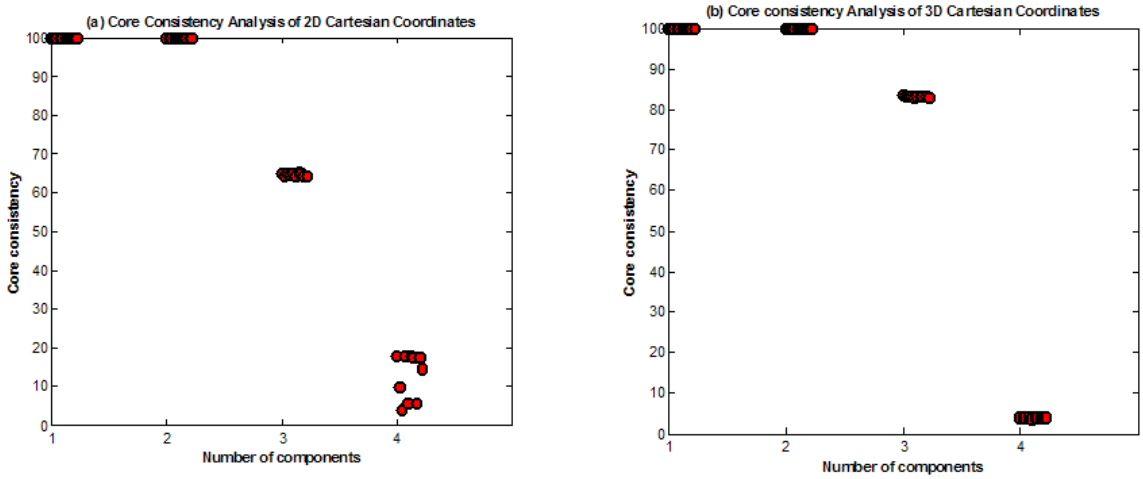

Figure 3. Core consistency diagnostic. (a) Test results of 2D Cartesian-coordinate analysis with type A normalization. (b) Test results of 3D Cartesian-coordinate analysis with type A normalization. (Different circles represent the results of PARAFAC analysis beginning with different start points.)

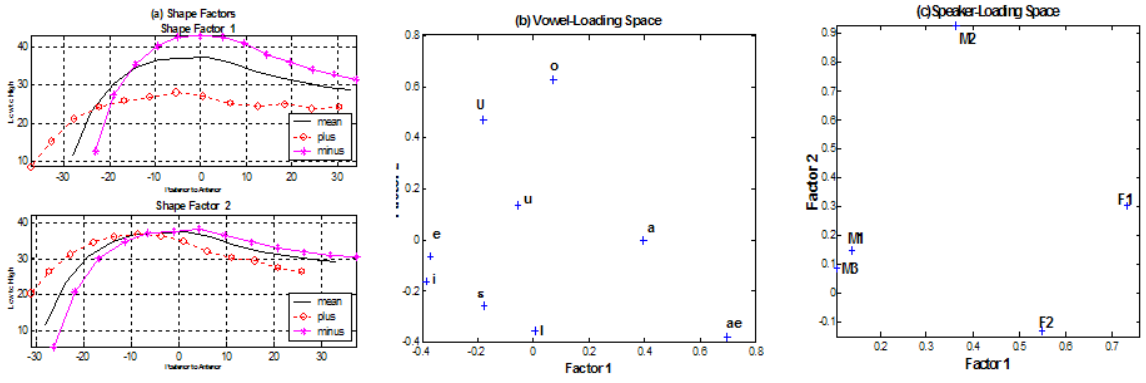

Figure 4. Factors of 2D Cartesian-Coordinate Analysis. (a) Shape factors. (b) Vowel loading space. (c) Speaker loading space.

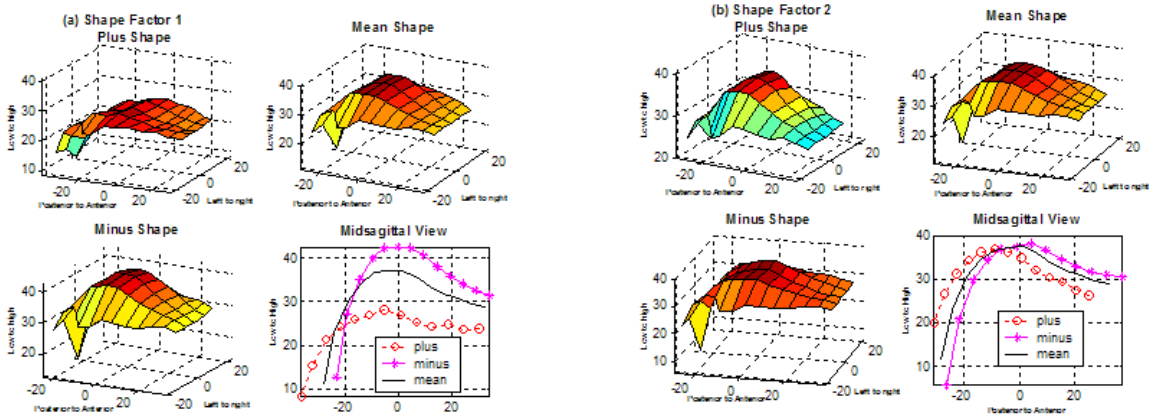

Figure 5. Shape Factors of 3D Analysis. (a) Shape factor 1. (b) Shape factor 2.

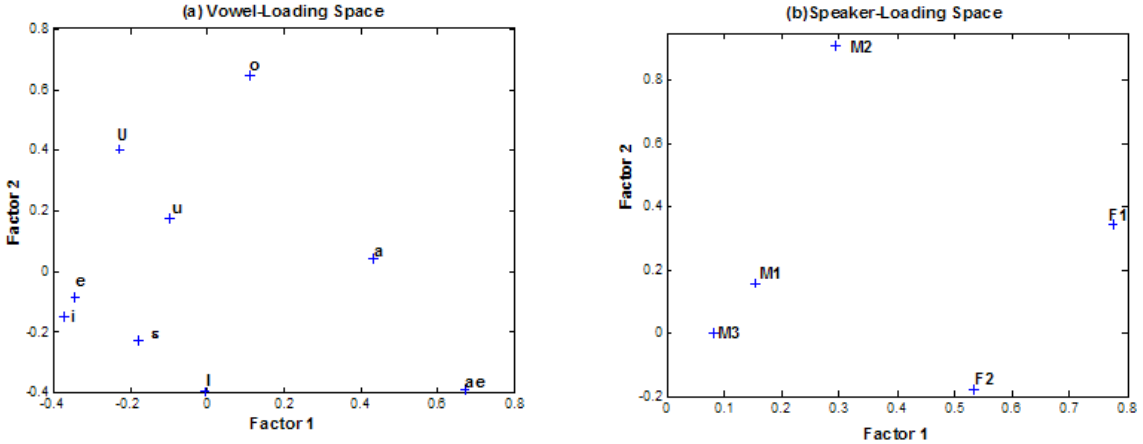

Figure 6. Vowel and Speaker Loading Spaces of 3D Analysis. (a) Vowel Loading Space. (b) Speaker Loading Space.