Acoustic-articulatory mapping in vowels by locally weighted regression

Richard S. McGowan[a)]
CReSS LLC, 1 Seaborn Place, Lexington, Massachusetts 02420

Michael A. Berger[b)]
Department of Linguistics, University of Rochester, 503 Lattimore Hall, Rochester, New York 14627

(Received 10 December 2008; revised 5 May 2009; accepted 30 June 2009)

A method for mapping between simultaneously measured articulatory and acoustic data is proposed. The method uses principal components analysis on the articulatory and acoustic variables, and mapping between the domains by locally weighted linear regression, or loess [Cleveland, W. S. (1979). J. Am. Stat. Assoc. **74**, 829–836]. The latter method permits local variation in the slopes of the linear regression, assuming that the function being approximated is smooth. The methodology is applied to vowels of four speakers in the Wisconsin X-ray Microbeam Speech Production Database, with formant analysis. Results are examined in terms of (1) examples of forward (articulation-to-acoustics) mappings and inverse mappings, (2) distributions of local slopes and constants, (3) examples of correlations among slopes and constants, (4) root-mean-square error, and (5) sensitivity of formant frequencies to articulatory change. It is shown that the results are qualitatively correct and that loess performs better than global regression. The forward mappings show different root-mean-square error properties than the inverse mappings indicating that this method is better suited for the forward mappings than the inverse mappings, at least for the data chosen for the current study. Some preliminary results on sensitivity of the first two formant frequencies to the two most important articulatory principal components are presented.
 [DOI: 10.1121/1.3184581]

PACS number(s): 43.70.Bk [DAB] Pages: 2011–2032

I. INTRODUCTION

The acoustic response to articulatory behavior is determined by physical law. Given the physical state of the vocal tract, it is possible to predict the acoustic output using deterministic equations, which may be considered mappings from articulation to acoustic output. Further, these mappings can be inverted with optimization procedures to predict articulatory configuration from acoustics, possibly in a non-unique way.

Articulatory-acoustic models, based on physics, as simple as four-tube models (e.g., Fant, 1960; Badin *et al.*, 1990; Stevens, 1998; McGowan, 2006) and as sophisticated as an articulatory synthesizer (e.g., Mermelstein, 1973; Maeda, 1982, 1990) can be employed to understand the lawful variations between articulation and formant frequencies. As much as such models are useful for conceptual understanding of speech production, they are not direct measures of the articulatory-acoustic relations and are not sufficient for a complete understanding of human articulatory behavior and output acoustics.

An empirical approach is taken here to determine mappings between articulation and output acoustics during vowel production. This approach has the virtues of being based on actual human behavior, of not relying on simplified models of the vocal tract acoustics, and of not relying on published parameters based on measurements of various individuals for which simultaneous articulatory-acoustic data have not been obtained. The empirical approach in this paper relies solely on a corpus of simultaneous acoustic recordings and articulatory measurements.

Several techniques may be used to generate continuous articulatory measurements. These include flesh point measurements such as electromagnetic articulography and X-ray microbeam, and imaging techniques such as magnetic resonance imaging (MRI) and ultrasound. Three-dimensional (3D) MRI (e.g., Engwall and Badin, 1999) can be employed to generate detailed vocal tract shapes that provide much of the information necessary to determine the output acoustics, although additional properties would need to be provided before the output could be accounted for completely, such as vocal tract wall impedance, nasal tract properties, glottal configuration, and sub-glottal properties.

Instead of using three-dimensional imaging by MRI, less comprehensive articulatory data in the form of flesh point data are commonly employed (e.g., Kiritani, 1986; Perkell *et al.*, 1992). One reason for this is the fact that flesh point data are most often recorded simultaneously with acoustic output, whereas MRI data are generally not recorded simultaneously with acoustic output (see, however, Bresch *et al.*, 2006). Further, MRI technology has relatively low temporal resolution, whereas flesh point technologies are faster. This means that the MRI data sets are generally small, but with

[a)]Author to whom correspondence should be addressed. Electronic mail: rsmcgowan@cressllc.net
[b)]Present address: The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom.

point measure technologies the data sets can be larger and contain more natural running speech. In the present work, the point measures in the University of Wisconsin X-ray Microbeam Speech Production Database (XRMB-SPD) (Westbury, 1994) were chosen because of the magnitude of the data set and variety of utterances, and the fact that acoustic signals were recorded simultaneously.

Because the XRMB-SPD only tracks midsagittal flesh points, the articulatory data are impoverished in the sense that acoustic output cannot be determined from these data using the physical theory of acoustics. Midsagittal shape does not determine an area function and output acoustics. Furthermore, even the midsagittal shape of the vocal tract is not completely measured in the XRMB-SPD: the most posterior pellet is located on the tongue dorsum. While it is possible to infer aspects of the midsagittal shape from the points that are measured, such as pharyngeal cross dimension (Whalen *et al.*, 1999; Jackson and McGowan, 2008), other aspects are not determined by the point measures, such as larynx height. On the other hand, it is generally assumed that there is a regular, if non-unique, relationship between midsagittal configuration and area function, and hence formant frequencies.

The intent of the present work is to construct mappings—forward mappings from articulation to acoustics, and inverse mappings from acoustics to articulation—for individual speakers in the XRMB-SPD. The forward mappings are empirically determined analogs to Fant's nomograms, which were derived from mathematical tube models (Fant, 1960, pp. 76–77 and 82–84). Instead of tube lengths and areas, variables derived from flesh point positions will serve as independent variables in the forward mappings in the current work. The nomograms tell us the specific response of formant frequencies to changes in tube parameters, and thus, the variations in sensitivity of the acoustic output to changes in tube geometry. These sensitivities (i.e., magnitudes of slopes in the nomograms) vary across the tube parameter space because the mapping between tube geometry and formant frequency is non-linear. Non-linearity, or changes in sensitivity, can be expected in the mappings between articulatory parameters derived from flesh point data and acoustic output. Thus, one of the goals of this study is to quantify the sensitivity of acoustic parameters to changes in articulatory parameters. This is an important consideration in phonetics, as, for instance, in Stevens' quantal theory (Stevens, 1972, 1989; Wood, 1979).

Another major area of research in the speech sciences is in the speech inverse problem: inferring articulatory information from speech acoustics in an algorithmic manner (e.g., Atal *et al.*, 1978; McGowan and Cushing, 1999; Hogden *et al.*, 2007). The inverse mappings derived in the present work provide a data-driven model to predict acoustics from articulation. They also provide a means of checking the predictions of models that are not derived from simultaneously recorded articulatory and acoustic data.

Our choices in methodology for building empirical mappings are informed by previous related work in speech and mathematical statistics. The two most important parts of the methodology are (1) a method for ordering the importance of independent variables, both acoustic and articulatory, and reducing the number of articulatory degrees-of-freedom from Cartesian pellet coordinates to some smaller, but essential, number; and (2) a least-squares method for approximating a smooth mapping between articulation and acoustics given simultaneously recorded articulatory and acoustic data points. The first element can be a form of function or data decomposition, and the latter element is a form of mathematical regression. The following is a review of previous work on articulatory-acoustic relations that employ one or both of these elements.

Principal components analysis (PCA), factor analysis, or other forms of function or data decomposition in the articulatory domain have become widely used when mapping between articulatory parameters and acoustics (Mermelstein, 1967; Shirai and Honda, 1976; Ladefoged *et al.*, 1978; Maeda, 1990; Yehia and Itakura, 1996; Story and Titze, 1998; Mokhtari *et al.*, 2007; and Story, 2007). These analyses reduce the number of articulatory parameters from an initially large number. This is particularly important when there are a limited number of acoustic parameters, such as three formant frequencies.

Mermelstein (1967) proposed using a Fourier cosine expansion of the log-area function in a largely theoretical study of the relation between area function and formant frequencies. He concluded that if the log-area function is spatially band-limited, then a unique area function can be determined from admittance function poles and zeros. However, while the poles correspond to resonance frequencies when the mouth is open, the zeros correspond to resonance when the vocal tract is closed. The latter are not observable during speech production. Building on Mermelstein's (1967) study, Yehia and Itakura (1996) decomposed the log-area function with a Fourier cosine series. In their mapping from formant frequencies to area functions they employed morphological and "least effort" constraints to alleviate ambiguities in the mapping. They tested their method for inferring area function from formant frequencies on data derived from X-rays of one speaker's 12 French vowels.

Shirai and Honda (1976) measured articulatory parameters taken from X-ray cineradiography of a speaker of Japanese, such as tongue shape, lip position, and jaw angle. The tongue shape was decomposed with PCA, and they related the first two PCA components and other articulatory measures to the first two formant frequencies. They were able to approximately recover the articulation of vowels from the first two formant frequencies using a non-linear regression technique, where the mapping was fitted on a set of 300 simultaneous articulatory and acoustic data frames of the speaker.

Ladefoged *et al.* (1978) used parallel factor analysis (PARAFAC) to decompose two-dimensional tongue shapes measured from X-rays of five speakers' pronunciation of ten American English vowels. They extracted two components for tongue position in the middle of vowels—front raising and back raising—and went on to use multiple regression to specify the two factors in terms of ratios among pairs of the three formant frequencies. Reasonable tongue positions could be inferred from formant frequencies using this

method. The study evaluated the predicted tongue positions in terms of correlations with original midsagittal shapes and in terms of root-mean-square error (RMSE).

Maeda (1990) took the approach of subtracting off important factors in articulatory movement, such as jaw movement, before performing PCA on tongue movement. He termed this "arbitrary factor analysis." In this way, it was easier to assign specific articulatory movements to changes in observed acoustic output than it would have been had PCA been applied to all the data without factoring out certain articulatory movement.

Story and Titze (1998) measured area functions of a single speaker's American English vowels with MRI and decomposed them into principal components, or what they termed "empirical orthogonal modes." They were able to obtain a mapping between the two modes that accounted for most of the variance in the area function and the first two formant frequencies in the form of a two-dimensional grid of iso-coefficient curves (coefficients of the two area function modes) in the formant plane. While ten points in the grid were determined from human empirical data, the remaining data, with 2500 grid intersections, were determined from area function modes by acoustic perturbation theory (Schroeder, 1967).

The recent work of Mokhtari *et al.* (2007) used human subjects for MRI full-volume scans during Japanese vowel production. Because of the noise of the machine, the acoustic recordings were taken separately from the MRI imaging. Also, the linear regressions between formant frequencies and principal components of human area functions were based on less than 40 samples of transitions between the vowels.

Story (2007) used pellet data from four speakers in the XRMB-SPD producing both static vowels and vowel-to-vowel transitions to find two principal components of the cross distances in the front of the mouth of each talker. "Cross distances" are the distances between, say, the tongue and the palate in the midsagittal plane. The amount of data per speaker was greater than the data used by any of Ladefoged *et al.* (1978), Story and Titze (1998), or Mokhtari *et al.* (2007), and it was shown that two PCA components were sufficient to characterize the data set. These components could be mapped to the first two formant frequencies in a largely unambiguous manner. However, the formants were calculated from a normalized area function and the two articulatory PCA components were not directly mapped to the formant data. Further, even this data set is limited and does not account for the sonorant portions of consonant-vowel transitions.

In order to obtain a robust mapping for each individual speaker, between 10 000 and 20 000 data points per speaker for four speakers were taken for the present work. These were simultaneous XRMB-SPD pellet positions and speech acoustic data that included all portions of vowels, including consonant-vowel transitions. PCA was chosen as the method of data decomposition, providing a set of independent variables ordered by the amount of variance accounted for in the data. Thus PCA can reduce the number of independent variables when the higher order components are neglected. For global linear regression, PCA alleviates problems in partial correlation in the independent variables.

Some of the work reviewed above relates articulatory and acoustic parameters using regression (e.g., Shirai and Honda, 1976; Ladefoged *et al.*, 1978; Mokhtari *et al.*, 2007). A form of regression was used in the present work to map both from articulatory coordinates to acoustic coordinates and vice-versa. The regression technique employed here is an adaptation of a method known as locally weighted regression, or "loess" (Cleveland, 1979; Cleveland and Devlin, 1988; Cleveland *et al.*, 1988), which produces a regression that is locally linear but globally non-parametric and non-linear. [Both Mermelstein (1967) and Yehia and Itakua (1996) used the property of local linearity between acoustics and articulation to obtain the inverse mapping.] In standard loess, least-squares regression is performed many times—indeed, for every point in the independent variable space at which we want to evaluate the mapping. For each point being evaluated the data are weighted differently, where the weight assigned to each data point is inversely related to its Euclidean distance to the point under evaluation. This provides a rational method for performing regression when the form of the function relating the independent and dependent variables is not specified, but the variation in local regression parameters can be presumed to be smooth.

The remainder of this paper is as follows. Section II details the procedures of speech segmentation, formant analysis, PCA applied to both pellet and formant data for each speaker, and loess. The loess method is described in some detail because it is novel to studies of speech production. Its particular implementation for constructing forward and inverse mappings, including optimization of loess parameters, is outlined. Section III presents the results of these analyses, starting with the articulatory principal components and their relation to well-known articulatory degrees-of-freedom. The rest of the results pertain to the optimum forward and inverse mappings found for each speaker, and examples of these mappings are shown. The causes of error and the differences in error between the forward and inverse mappings are examined. The sensitivities of formant frequencies to articulatory parameter changes are presented. Section IV provides a discussion focusing on the articulatory PCA, error, and sensitivity before final conclusions are drawn in Sec. V.

II. METHOD

A. Data

Simultaneously recorded acoustic and articulatory data from the XRMB-SPD (Westbury, 1994) were used. The database consists of time-aligned audio and midsagittal pellet position recordings from 57 adult American English speakers, comprising about 15 h of recorded speech. Speech tasks include reading of citation words, sentences, paragraphs, number sequences, and vowel sequences, with some tasks performed at deliberately slower or faster speaking rates.

Pellet coordinates are referenced to a speaker-specific Cartesian coordinate system whose axes are based on ana-

tomical features in the speaker's head. [See Westbury (1994) for the definition of the coordinate system.] Each speaker's time-varying articulatory data contains horizontal and vertical coordinates in the midsagittal plane for each of eight moving pellets. There are four pellets approximately evenly spaced along the tongue centerline; the most anterior tongue pellet is close to the tongue tip and the most posterior is the furthest back the pellet could be placed without inducing a gag reflex. The other four pellets include one pellet on the upper lip and one on the lower lip, and two on the mandible. These pellets give a partial representation of the vocal tract; no information is available about lateral tongue or jaw movement, the velum, or the posterior vocal tract (e.g., pharyngeal dimensions or larynx height), factors that also affect acoustic output. Pellet positions were originally measured at variable rates between 20 and 160 samples/s, but in the database all pellet positions are resampled at an equal rate of 160 samples/s (Westbury, 1994, p. 57).

In the present study only the positions of the six pellets on the tongue and lips were used, resulting in a total of 12 degrees of freedom. We examined four speakers, two males (JW11 and JW18) and two females (JW14 and JW16).

B. Segmentation

All vowel tokens were included in the analysis except those adjacent to nasal, lateral, or rhotic consonants. Also, very short reduced vowels that had little or no formant structure were excluded. Vowels in the context of nasals and laterals were excluded because velar or lateral movement during these vowels would not be captured by any pellet; hence the acoustic changes would not be matched by any articulatory changes. Vowels in rhotic contexts were excluded because the relatively rare but extreme retroflex articulations could potentially complicate reduction in the dimensionality of the articulatory data as well as the acoustic-articulatory regression analysis.

The audio recordings were manually annotated to segment and label the vowels not excluded by the criteria above using PRAAT TextGrids (Boersma and Weenink, 2007). The acoustic and articulatory data within the demarcated vowel intervals formed the basis of the study. The following conventions were used for the placement of interval boundaries. Between the vowel and a preceding obstruent, the boundary was generally placed at the first glottal pulse after the closure release. Similarly, between the vowel and a following obstruent, the boundary was placed at the last glottal pulse before the onset of closure. For the less straightforward case of dividing the vowel from an adjacent oral sonorant, we sought the midpoint of the transition between the two sounds based on auditory judgments and visible formant transitions.

C. Formant tracking, editing, and data pruning

After defining the vowel intervals to be used in the analysis, automated formant tracking was performed on the vowels using PRAAT. The formant analysis was linear predictive coding (LPC)-based using the Burg algorithm. A 25 ms window was used and the centers of neighboring windows were 6.25 ms apart so that the frame rate was the same as the pellet sampling rate of 160 Hz. The maximum number of formants that were identified per frame was 5, with a maximum frequency of 5000 Hz for males and 5500 Hz for females. Pre-emphasis was applied to frequencies above 50 Hz. Only $F1$, $F2$, and $F3$ values were used from the acoustic analysis. The formant tracks were resampled to align acoustic samples with the articulatory samples.

To ensure the best quality formant tracking, the $F1$, $F2$, and $F3$ formant tracks within all of the vowel segments were visually inspected and manually corrected at each sample point where the automatic tracking was deemed faulty. To make this process efficient we developed a graphical formant editor that allows point-and-click corrections to formant values in a dual spectrogram/spectrum display. The spectrum display includes a function for refining an estimated formant frequency by fitting a parabola to the three harmonics closest to the estimate and shifting the formant frequency to the peak of the parabola.

Once the formant analysis of the vowel segments was complete, the samples from all of the segments were pooled for each speaker. Each sample comprised a 3-dimensional acoustic vector and a 12-dimensional articulatory vector. Before continuing in the analysis, each articulatory-acoustic data point was examined automatically for completeness. Any sample missing acoustic data ($F1$, $F2$, or $F3$) or pellet data (any of the two coordinates for the six pellets) was excluded from further analysis. Table I summarizes the data included in the study: the gender of each speaker, the number of vowel segments used, and the number of samples drawn from these segments. The subsequent analyses (PCA and regression) were performed separately on each speaker's data set. Data from different speakers were not pooled. Although multiple data points in a speaker's data set may originate from the same time series, i.e., a vowel trajectory, it is assumed that the error distribution for each measured quantity is statistically independent from others over the same time series.

TABLE I. Gender and the amount of usable data collected for each speaker.

	Gender	No. of segments	No. of samples
JW11	M	763	13 003
JW14	F	828	16 013
JW16	F	974	19 770
JW18	M	666	11 295

D. PCA

For each speaker, PCA was performed twice: once on the speaker's articulatory data points and once on the acoustic data points. The principal components of the articulatory data are denoted $K1, K2, K3, \ldots, K12$, and those of the formant data are denoted $A1$, $A2$ and $A3$. The lower-order principal components were utilized as variables in the subsequent regression analysis: the first four articulatory components $K1-K4$, and all three acoustic components.

E. Locally weighted regression (loess)

1. General method

For each speaker's data set we computed both forward (articulatory to acoustic) and inverse (acoustic to articulatory) mappings using locally weighted regression. Locally weighted regression or loess (Cleveland, 1979; Cleveland and Devlin, 1988; Cleveland *et al.*, 1988) is a form of nonparametric regression by data smoothing. It is computationally intensive but allows one to represent the relationship between one or more independent variables and a dependent variable with few assumptions about the form of that relationship. This makes it suitable for fitting complex regression surfaces for which a suitable parametric function is not known. The general method of loess is now described.

Let $S=\{(\mathbf{x}_i,y_i):i=1\ldots n\}$ be a given set of data points, where y_i is a measurement of the dependent variable and $\mathbf{x}_i$ is a measurement of a *p-tuple* of independent variables. A regression model relating these variables is

$$y_i = g(\mathbf{x}_i) + \varepsilon_i, \tag{1}$$

where ε_i is a zero-mean, normal random error term. In classical regression analysis, g is assumed to belong to some parametric class of functions, such as polynomials, which places practical limits on the variety of surfaces that can be modeled. For example, in the case that a linear relation is assumed to hold between the independent variable $\mathbf{x}$ and dependent variable y, the function g can be estimated in a least-squares sense by $\hat{g}$,

$$y \approx \hat{g}(\mathbf{x}) = \mathbf{x}\cdot\boldsymbol{\beta} + \gamma, \tag{2}$$

where parameters $\boldsymbol{\beta}$ and γ are determined by the data points $(\mathbf{x}_i,y_i)$ and a chosen weighting function of those points, say, w_i (e.g., Chaterjee and Hadi, 1988).

In loess, g is not limited to being a parametric function; it is only assumed to be a smooth function of $\mathbf{x}$. Accordingly, the estimate of g, $\hat{g}$, is computed without fitting a parametric function to the entire data set. Rather the smoothness property of g is exploited to estimate g by locally fitted functions. The smoothness property allows us to assume that for any point $\mathbf{x}$ in the space of the independent variables, $\hat{g}(\mathbf{x}')\approx l_{\mathbf{x}}(\mathbf{x}')$ for $\mathbf{x}'$ near $\mathbf{x}$, where $l_{\mathbf{x}}$ is a locally fitted, low-order polynomial. Strict equality holds for $\mathbf{x}'=\mathbf{x}$; that is, $\hat{g}(\mathbf{x})=l_{\mathbf{x}}(\mathbf{x})$. $l_{\mathbf{x}}$ is obtained by a least-squares fit to the data based on a local weighting function $w_i(\mathbf{x})$ that heavily weights data points $(\mathbf{x}_i,y_i)$ close to $\mathbf{x}$. The locally weighted regression function $l_{\mathbf{x}}$ may be linear or non-linear; in the present study a linear model was used:

$$l_{\mathbf{x}}(\mathbf{x}') = \mathbf{x}'\cdot\boldsymbol{\beta}(\mathbf{x}) + \gamma(\mathbf{x}), \tag{3}$$

where $\boldsymbol{\beta}(\mathbf{x})$ and $\gamma(\mathbf{x})$ are computed just as in standard linear least squares, except that they now depend on $\mathbf{x}$ because the weighting function depends on $\mathbf{x}$. Consequently, the least-squares procedure must be repeated at each value of $\mathbf{x}$ for which we want to solve $\hat{g}(\mathbf{x})$, which makes loess computationally intensive.

To perform loess, one must choose a weight function $w_i(\mathbf{x})$ that assigns weights to data points in S based on distance from $\mathbf{x}$: data points close to $\mathbf{x}$ have large weight, while those far from $\mathbf{x}$ have small weight. A distance metric is also needed. For the distance metric, this study used Euclidean distance in the space of the independent variables, after first scaling the variables by dividing each by its own standard deviation. The weight function was the standard one used by Cleveland (1979), which guarantees a fixed neighborhood size (number of positively weighted data points) regardless of data distribution around $\mathbf{x}$. The weight function has a parameter b between 0 and 1, known as the bandwidth, that expresses neighborhood size as a proportion of the data. Thus, the larger the value of b, the more data points influence the local regression at $\mathbf{x}$. Using a nearest-neighbor algorithm, the neighborhood size is used to determine a neighborhood radius. The radius of the neighborhood of $\mathbf{x}$ for a given bandwidth b, denoted $d(\mathbf{x},b)$, is defined to be the distance from $\mathbf{x}$ to the qth nearest data point, where q is equal to bn rounded to the nearest integer. The weight function is zero beyond this radius. The weight assigned to data point $(\mathbf{x}_i,y_i)$ for the locally weighted regression at $\mathbf{x}$, using bandwidth b is

$$w_i(\mathbf{x},b) = W\left(\frac{\|\mathbf{x}_i-\mathbf{x}\|}{d(\mathbf{x},b)}\right), \tag{4}$$

in which W is the tricube function: $W(u)=(1-u^3)^3$ for $0\leqslant u\leqslant 1$ and $W(u)=0$ otherwise.

2. Computationally efficient loess

In loess, weighted least-squares estimation must be performed for every value of $\mathbf{x}$ at which we want to know $\hat{g}(\mathbf{x})$. This makes the technique computationally expensive for operations requiring many samples, such as plotting. An efficient alternative is to pre-compute $\hat{g}(\mathbf{x})$ at sample values of $\mathbf{x}$, and then interpolate for intermediate values. For the sample set one may choose the original data points, or some strategically selected set of points, such as the vertices of a kd-tree constructed on the data (Cleveland *et al.*, 1988). In the present study the pre-computed sample set was simply the data points. However, rather than directly interpolating the dependent variable at the sample points, better results were obtained by interpolating the local regression parameters computed at those points and generating the dependent variable from the interpolated regression parameters.

Following this approach, a loess model in the current study was fitted to data set S by computing a locally weighted regression at each data point. The resulting model then included one set of local regression parameters for each data point, accompanied by an interpolation scheme, which was as follows. For each data point $(\mathbf{x}_i,y_i\in S)$, let $\bar{\boldsymbol{\beta}}(\mathbf{x}_i)$ and $\bar{\gamma}(\mathbf{x}_i)$ be the pre-computed local regression parameters fitted to the neighborhood of $\mathbf{x}_i$. The regression parameters $\boldsymbol{\beta}(\mathbf{x})$ and $\gamma(\mathbf{x})$ at a novel point $\mathbf{x}$ were estimated by a weighted average of the parameters $\bar{\boldsymbol{\beta}}(\mathbf{x}_i)$ and $\bar{\gamma}(\mathbf{x}_i)$ at the data points, where the data points were weighted using the tricube-based weight function $w_i(\mathbf{x},b)$ defined above. The interpolated regression parameters $\boldsymbol{\beta}(\mathbf{x})$ and $\gamma(\mathbf{x})$ at $\mathbf{x}$ were thus given as

$$\boldsymbol{\beta}(\mathbf{x}) = \frac{\sum_{i=1}^{n} w_i(\mathbf{x},b)\bar{\boldsymbol{\beta}}(\mathbf{x}_i)}{\sum_{i=1}^{n} w_i(\mathbf{x},b)}, \quad \gamma(\mathbf{x}) = \frac{\sum_{i=1}^{n} w_i(\mathbf{x},b)\bar{\gamma}(\mathbf{x}_i)}{\sum_{i=1}^{n} w_i(\mathbf{x},b)}. \tag{5}$$

The predicted value $\hat{g}(\mathbf{x})$ was then calculated using the interpolated parameters $\boldsymbol{\beta}(\mathbf{x})$ and $\gamma(\mathbf{x})$. Note that the bandwidth b used to weight data points for interpolation may differ from that used for the weighted least-squares fit of the local regressions. Thus each loess model constructed in this study had two distinct bandwidth parameters: a *regression bandwidth* b_R, which was used to weight the data points for the local regressions, and an *interpolation bandwidth* b_I, which was used to interpolate the local regression parameters. It will be assumed henceforth that the loess models in this study generate output $\hat{g}(\mathbf{x})$ by interpolating the precomputed regression parameters at the data points in this manner.

3. Grid sampling

Given a loess model $\hat{g}(\mathbf{x})$ with regression parameters $\boldsymbol{\beta}(\mathbf{x})$ and $\gamma(\mathbf{x})$, it is useful to be able to sample these functions in a regular grid over the space of the independent variables. Grid sampling is useful both for generating plots of the regression surfaces and for studying the variation in the regression parameters. The *full grid* associated with a particular data set S has d evenly spaced columns of vertices in each independent variable, generally with $d=20$. For example, a data set with four independent variables will have a $20\times20\times20\times20$ full grid. The span of the grid in each dimension is from -2.5 to 2.5 standard deviations from the mean, using the standard deviation and mean of that dimension in S. A *subgrid* associated with a particular data set is the same as the full grid but missing one or more coordinates of each vertex. Thus it occupies a subspace of the independent variable space. Subgrids are useful for reducing data complexity and for generating 3D plots as a function of the first two independent variables.

Functions are defined for sampling any function of the independent variables at grid vertices. If $f(\mathbf{x})$ is a function of the independent variables, then $[f(\mathbf{x})]=f(\mathbf{x})$ for any $\mathbf{x}$ that is one of the vertices of the full grid. To sample at the vertices of a subgrid, the dimensions that are present are indicated by a superscript; for example, superscript (1,2) indicates that the subgrid occupies only the first two dimensions of the full grid, and values are averaged over the missing dimensions. (The absence of any superscript indicates sampling on the full grid.) Thus, for any vertex $\mathbf{z}$ of the subgrid, $[f(\mathbf{z})]^{(1,2)}$ is equal to the average $[f(\mathbf{x})]$ over all vertices $\mathbf{x}$ of the full grid that agree with $\mathbf{z}$ in the first and second coordinates. The grid sampling functions may be used to sample $\hat{g}(\mathbf{x})$, $\boldsymbol{\beta}(\mathbf{x})$, and $\gamma(\mathbf{x})$ at grid vertices.

4. Model evaluation

Given a loess model $\hat{g}$ fit to data set S, which we may call the fitting set, the performance of the model was evaluated using a completely separate, randomly selected set of data that was held out from the fitting set, called a test set $T=\{(\mathbf{x}_j,y_j):j=1,\ldots,m\}$. The loess model is fitted to the fitting set, without using any data from the test set. The test set is used to evaluate the prediction error of the model using RMSE:

$$\mathrm{RMSE}(\hat{g},T) = \sqrt{\frac{\sum_{j=1}^{m}(\hat{g}(\mathbf{x}_j)-y_j)^2}{m}}. \tag{6}$$

5. Optimum model selection

Of course, prior to testing, the loess model has to be selected by choosing appropriate values for the regression bandwidth b_R and interpolation bandwidth b_I. Too large a value for either bandwidth parameter will generate too smooth a regression surface, while too small a value will lead to overfitting of the data. To optimize model parameters without overfitting to data, a standard technique is to select parameter values that optimize the prediction rate for a separate set of data that was not used during the construction of the model, called a *validation set*. Similar to a test set, a validation set is a randomly selected set external to the fitting set; thus it allows us to see which model best generalizes to new data. However, the validation set may not actually be part of the test set since it is used to select the model parameters and is thus not "unseen" prior to testing of the selected model.

However, when the data pool is limited, it may be difficult to draw an adequately sized validation set separate from both fitting and test data. A common method for validating a model without having a separate validation set is k-fold cross-validation. In this method, the fitting set is randomly partitioned into k equal subsets, and the model is fitted to the data k times. For each fitting iteration i, the ith subset is held out as the validation set, and the remaining $k-1$ subsets are combined to form a reduced fitting set; validation error is equal to RMSE in Eq. (6) calculated over the data in the ith subset. Thus, each data point gets used for validation once and for fitting $k-1$ times. Validation error is averaged over the k trials.

k-fold cross-validation was employed on the fitting set to optimize the two bandwidth parameters b_R and b_I. An exhaustive search of the two-dimensional parameter space was conducted, and for each pair of bandwidth values in the search, validation error was computed using k-fold cross-validation with $k=10$. The regression and interpolation bandwidths yielding the minimum validation error were deemed the optimal bandwidths. The space of bandwidth values to search was determined empirically. If a minimum was not attained in a given search space, then the search space was shifted in the direction of decreasing validation error observed in previous searches. Figure 1 shows the results of a typical search for optimal bandwidths. The loess model being optimized is that of the inverse mapping from the acoustic variables $A1$, $A2$, and $A3$ to the articulatory variable $K1$, for subject JW16. The search space consisted of 15 values of regression bandwidth between 0.002 and 0.03, and 20 values of interpolation bandwidth between 0.0001 and 0.002 for a

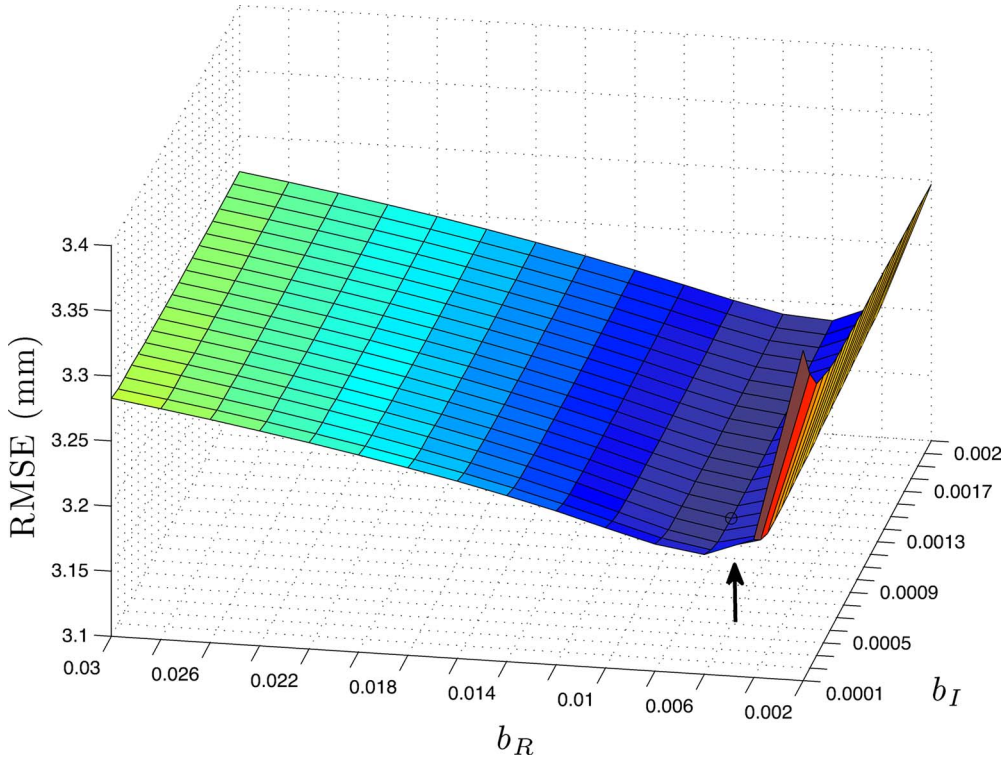

FIG. 1. (Color online) RMSE surface as a function of regression bandwidth b_R and interpolation bandwidth b_I. The loess model being optimized is that of the inverse mapping from the acoustic variables $A1$, $A2$, and $A3$ to the articulatory variable $K1$ for subject JW16. The arrow indicates the location of the minimum RMSE.

total of 300 different trials. The minimum validation RMSE of 3.181 mm was found at b_R=0.006, b_I=0.0005 (indicated by the arrow).

Once the optimal bandwidth values were found for a particular mapping, the loess model was constructed on the entire fitting set using the optimal b_R value. Then the optimal b_I value was used in the equations in Eq. (5).

It should be emphasized that in order to maintain an objective evaluation, the test data T may not be used in optimizing the model parameters. Thus k-fold cross-validation was performed only using data from the fitting set S. The validation RMSE values have no relation to the test RMSE mentioned in Sec. II E 4, which is determined only after a model has been selected and fitted.

6. Acoustic-articulatory loess

To construct loess models, each speaker's data set was divided into a fitting set S and a test set T. The test set comprised a random selection of 10% of the data points of the speaker. After discovering the optimal b_R and b_I values for a given model, as described in Sec. II E 5, it was fitted to the fitting set and evaluated against the test set in the manner described above (see Secs. II E 2 and II E 4).

For each speaker, seven loess models were constructed: three forward mappings from the articulatory PCA components $\mathbf{x}=(K1,K2,K3,K4)$ to each of the formants $y=F1$, $y=F2$, or $y=F3$, and four inverse mappings from the acoustic PCA components $\mathbf{x}=(A1,A2,A3)$ to each of the articulatory components $y=K1$, $y=K2$, $y=K3$, or $y=K4$. These totaled seven loess mappings for each of the four speakers: three forward mappings and four inverse mappings. Each model had the optimum bandwidths.

The various loess models generated for each subject are denoted as follows. In the case of a forward mapping from $\mathbf{x}=(K1,K2,K3,K4)$ to say, $y=Fj$, the loess model is denoted $\hat{g}_j^{\rightarrow}$. In the case of an inverse mapping from $\mathbf{x}=(A1,A2,A3)$ to $y=Kj$, the loess model is denoted $\hat{g}_j^{\leftarrow}$.

As shown in Eq. (3), for each locally weighted regression $\hat{g}(\mathbf{x})$ there are functions $\boldsymbol{\beta}(\mathbf{x})$ and $\gamma(\mathbf{x})$, where $\boldsymbol{\beta}(\mathbf{x})$ is a vector of slopes varying as a function of $\mathbf{x}$, and $\gamma(\mathbf{x})$ is the regression constant. $\boldsymbol{\beta}(\mathbf{x})$ will be denoted for each mapping as follows. In the case of a forward mapping, $\hat{g}_j^{\rightarrow}$, the vector of slopes is denoted $\boldsymbol{\beta}_j^{\rightarrow}=(\beta_{1j}^{\rightarrow},\beta_{2j}^{\rightarrow},\beta_{3j}^{\rightarrow},\beta_{4j}^{\rightarrow})$, where $\beta_{ij}^{\rightarrow}$ is the regression slope from Ki to Fj. While all of these quantities depend on $\mathbf{x}$, the argument has been suppressed, as it will be for the rest of the paper. The constant is denoted $\gamma_j^{\rightarrow}$. In the case of an inverse mapping $\hat{g}_j^{\leftarrow}$, $\boldsymbol{\beta}(\mathbf{x})$ is a three-dimensional vector: $\boldsymbol{\beta}_j^{\leftarrow}=(\beta_{1j}^{\leftarrow},\beta_{2j}^{\leftarrow},\beta_{3j}^{\leftarrow})$, where $\beta_{ij}^{\leftarrow}$ is the regression slope from Ai to Kj, which again depends on $\mathbf{x}$. The constant is denoted $\gamma_j^{\leftarrow}$.

The same notation is employed for the pre-computed regression parameters at the data points. Thus $\bar{\boldsymbol{\beta}}(\boldsymbol{x}_i)$ is denoted $\bar{\boldsymbol{\beta}}_j^{\rightarrow}=(\bar{\beta}_{1j}^{\rightarrow},\bar{\beta}_{2j}^{\rightarrow},\bar{\beta}_{3j}^{\rightarrow},\bar{\beta}_{4j}^{\rightarrow})$ in the case of a forward mapping and $\bar{\boldsymbol{\beta}}_j^{\leftarrow}=(\bar{\beta}_{1j}^{\leftarrow},\bar{\beta}_{2j}^{\leftarrow},\bar{\beta}_{3j}^{\leftarrow})$in the case of an inverse mapping. Similarly, $\bar{\gamma}(\mathbf{x}_i)$ is denoted $\bar{\gamma}_j^{\rightarrow}$ for a forward mapping and $\bar{\gamma}_j^{\leftarrow}$ for an inverse mapping.

F. Global linear regression

To provide a baseline of performance, the loess models were compared to standard linear regressions. Thus for each of the seven mappings for each speaker, in addition to constructing a locally weighted regression on the basis of the speaker's fitting data, a single, uniformly weighted linear regression was computed by least-squares fit to the same fitting data. This will be called the "global regression" in contrast to the locally weighted regression. The global regressions were evaluated against the test data in the same manner as the local regression models.

G. Sensitivity

One application of the mappings constructed with loess is to study the sensitivity of variables in one domain to changes in the other domain. In particular, it is of interest to study the variation in the magnitudes of slopes as a function of position in the independent variable space. The magnitudes of the slopes of the forward mappings $\boldsymbol{\beta}_j^{\rightarrow}$ are direct measures of the sensitivity of formant frequencies to changes in the principal components of articulation as a function of the position in articulatory configuration space. The slopes of the forward mappings will be used to define *empirical sensitivity functions*.

To permit systematic examination, the slopes of the forward mappings were sampled in grid format (see Sec. II E 3). However, the resulting data are complex: in the full grid of the forward mapping, there are $20^4=160\,000$ vertices, each of which is evaluated for 12 different slope functions: $\beta_{11}^{\rightarrow}$, $\beta_{12}^{\rightarrow}$, $\beta_{13}^{\rightarrow}$, $\beta_{21}^{\rightarrow}$, $\beta_{22}^{\rightarrow}$, $\beta_{23}^{\rightarrow}$, $\beta_{31}^{\rightarrow}$, $\beta_{32}^{\rightarrow}$, $\beta_{33}^{\rightarrow}$, $\beta_{41}^{\rightarrow}$, $\beta_{42}^{\rightarrow}$, and $\beta_{43}^{\rightarrow}$. To focus on some of the more important aspects of sensitivity, the degrees of freedom of the slope data needed to be reduced. The reduction presented here is simply based on the observation that $K1$ and $K2$, on average, account for 81% of the variation in pellet positions for the four speakers with a range from 77% to 89%. Thus only slopes for which the independent variable was $K1$ or $K2$ were considered, i.e., $\beta_{ij}^{\rightarrow}$ with $i=1,2$, reducing the number of slopes by half. Furthermore, slopes were sampled on the $K1,K2$ subgrid, which has only $20^2=400$ vertices, while averaging over $K3$ and $K4$ coordinates, i.e., using the slope functions $[\beta_{ij}^{\rightarrow}]^{(1,2)}$.

Because we are concerned with magnitudes, the absolute values of slopes are examined here. In fact, for reasons that will become apparent in Sec. IV, it was convenient to consider these values normalized by the predicted formant frequencies. Thus the empirical sensitivity functions, evaluated at each vertex of the $K1,K2$ subgrid, are defined as

$$\nu_{ij}^{\rightarrow}=\frac{[|\beta_{ij}^{\rightarrow}|]^{(1,2)}}{[\hat{g}_j^{\rightarrow}]^{(1,2)}} \quad \text{for } i=1,2 \text{ and } j=1,2,3.$$

For each speaker and each empirical sensitivity function $\nu_{ij}^{\rightarrow}$, two categories of $(K1,K2)$ vertices over the 20×20 subgrid were defined: those with the largest 20% $\nu_{ij}^{\rightarrow}$ and those with the smallest 20% $\nu_{ij}^{\rightarrow}$. These categories were labeled large sensitivity and small sensitivity, respectively. Critical vertices were also saved for later examination. $(K1,K2)$ was a critical vertex for $\nu_{ij}^{\rightarrow}$ if its value of $\nu_{ij}^{\rightarrow}$ is smaller than that of any of the neighboring vertices, where two vertices in a rectangular grid are neighboring if they differ by at most one row or column, or both one row and column. (This criterion was more easily satisfied for the boundary points because there were fewer comparisons to satisfy.)

To provide reference markers for the sensitivity categories and critical points in the $K1,K2$ plane, point vowels were also located in the plane as follows. The formant frequencies and the pellet positions of sustained vowel productions [ɑ], [æ], [u], and [i] of the point vowels from task 14 of the XRMB-SPD were extracted for each of the speakers. (One exception was the [æ] of JW14, which was extracted from task 24.) Using the first two formants, a time sample was chosen for each of the vowels as being the most extreme moment in the articulation of the vowel. For instance, for [ɑ], a time where the first formant frequency was high and the second formant frequency was low, relative to the other values in the file, was chosen as the time that best represented the extreme of the articulation. The pellet positions at that time were projected onto $K1$ and $K2$ to obtain a representation of the vowel articulation in the $(K1,K2)$ coordinates.

TABLE II. Percent of variance accounted for by the first five PCA components of the articulatory data.

	$K1$ (%)	$K2$ (%)	$K3$ (%)	$K4$ (%)	$K5$ (%)
JW11	53.5	23.3	10.4	7.4	1.4
JW14	75.8	13.1	5.3	3.2	1.1
JW16	58.7	21.0	7.9	6.6	1.7
JW18	56.0	24.0	10.1	5.1	1.8

III. RESULTS

A. PCA

The first four components of the PCA of the articulatory pellet data accounted for between 94% and 97% of the variance for each of the four talkers. Table II shows the percentage of variance accounted for by each of the first five components.

Figure 2 illustrates the first four articulatory components for each speaker. As evident in the figure, the first articulatory PCA component, $K1$, corresponds to a low-back to high-front degree-of-freedom in all of the talkers, except for JW16. For JW18, the front-back component is minimal. Also, there appear to be varying amounts of tongue bunching in $K1$ for each of the subjects for the high-front position. $K1$ is associated with varying degrees of lower lip height changes, except in the case of JW14. JW16's $K1$ shows a tongue height degree-of-freedom, and there appears to be a "rocking" motion of the tongue in her $K1$, which has the effect of inducing a tongue blade up-down degree-of-freedom.

For JW11 and JW14 the second articulatory PCA component, $K2$, corresponds to a high-back to a low-front tongue degree-of-freedom, with more height change for JW14 than for JW11 (Fig. 2). There is also some tongue bunching-stretching associated with tongue height changes for JW14. Interestingly, the $K2$ of JW16 shows a low-back to high-front

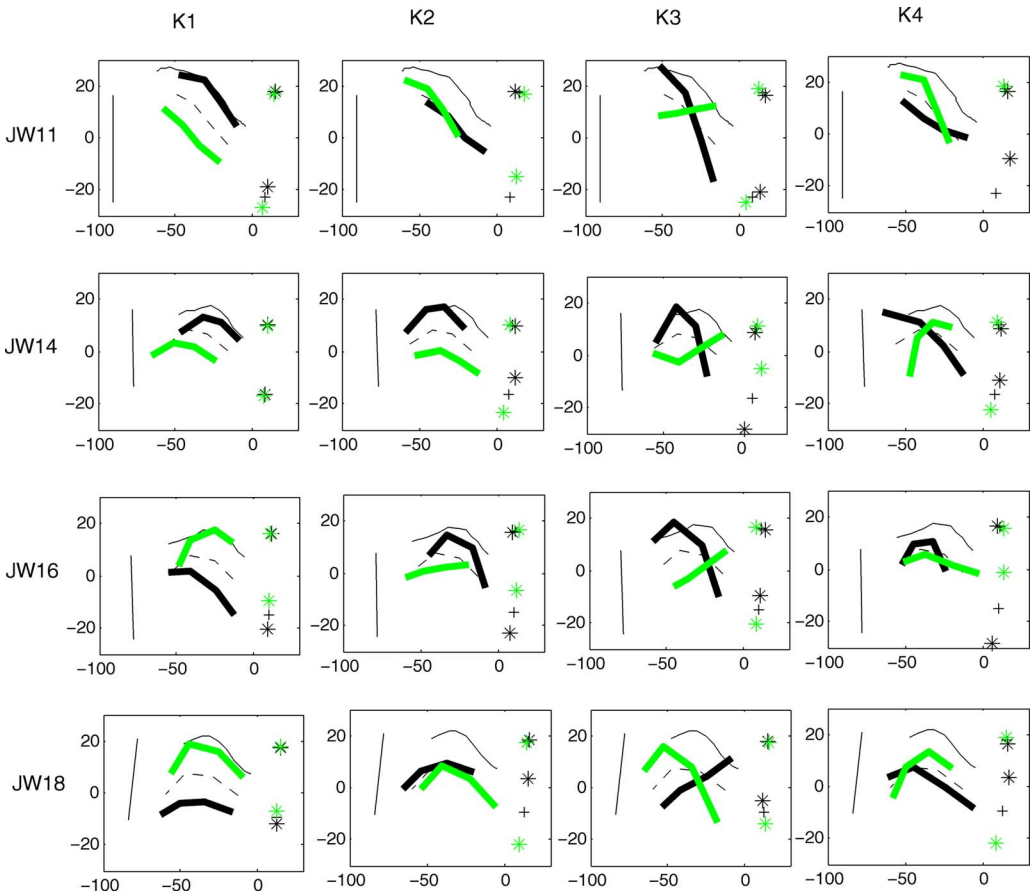

FIG. 2. (Color online) Graphical representation of the first four articulatory principal components. Shown are mean and two extreme values of each component, *K*1, *K*2, *K*3, and *K*4 (by column) for each of the four speakers, JW11, JW14, JW16, and JW18 (by row). The axes are in units of milimeters and correspond to the horizontal and vertical axes as defined in Westbury (1994). The top curve in each plot is the palate trace and the black vertical line is the approximate pharyngeal wall. The dotted curve and "+" marks represent the mean tongue and lip pellet coordinates for the principal component, while the bold curves and " *" marks represent the extreme values.

degree-of-freedom that the others show in their *K*1, with a substantial bunching-stretching degree-of-freedom. JW18's *K*2 exhibits a front-back degree-of-freedom, with some tongue rocking, again inducing a tongue blade up-down degree-of-freedom.

In *K*3, all the subjects show a movement from a high bunched tongue with the tongue blade down to a low stretched tongue with tongue blade up. Thus, *K*3 of all subjects contains simultaneous tongue blade up-down and tongue bunching-stretching degrees-of-freedom. There does not appear to be much consistency among the subjects in their *K*4 components, although bunching-stretching and tongue blade up-down degrees-of-freedom also appear as aspects of the *K*4 components.

There are differences among the subjects in the kinds of tongue shapes/positions and lip positions accounted for in the first four components. A portion of these differences can be accounted for in terms of differences in the placement of the pellets. For instance, many of the differences between JW11 and the other three subjects can be accounted for if the rearmost tongue pellet of JW11 can be considered to be equivalent to the second rearmost in the other subjects.

The relation between familiar articulatory degrees-of-freedom (such as high-low, front-back) and the PCA components can be quantified with an orthonormal set of vectors intended to represent these familiar degrees-of-freedom. These are (1) the tongue front-back degree-of-freedom, with equal weights (0.5) in the four tongue *x*-components and zeros in the four tongue *y*-components and the four lip components; (2) the tongue high-low degree-of-freedom, with equal weights (0.5) in the four tongue *y*-components and zeros in the four tongue *x*-components and the four lip components; (3) the lip-opening degree-of-freedom, with equal and opposite weights (–0.707 and 0.707) in the two lip *y*-components and zeros in the others; and (4) the lip-protrusion degree-of-freedom, with equal weights (0.707) in the two lip *x*-components and zeros elsewhere. The amount that each of the PCA components has of the tongue front-back, tongue high-low, lip-closure, and lip-protrusion degrees-of-freedom can be determined by projecting the PCA component vector onto each of the vectors associated with the familiar degrees-of-freedom, using the dot product. The proportion of the articulatory PCA components that can be attributed to each familiar degree-of-freedom are shown in Fig. 3. The signs within a speaker indicate the relative directions of these projections. For instance, *K*2 of JW11 associates a more low with a more front tongue position.

The tongue front-back movement is the familiar degree-of-freedom that is represented the most often in the PCA components at or above the 20% level. On the other hand,

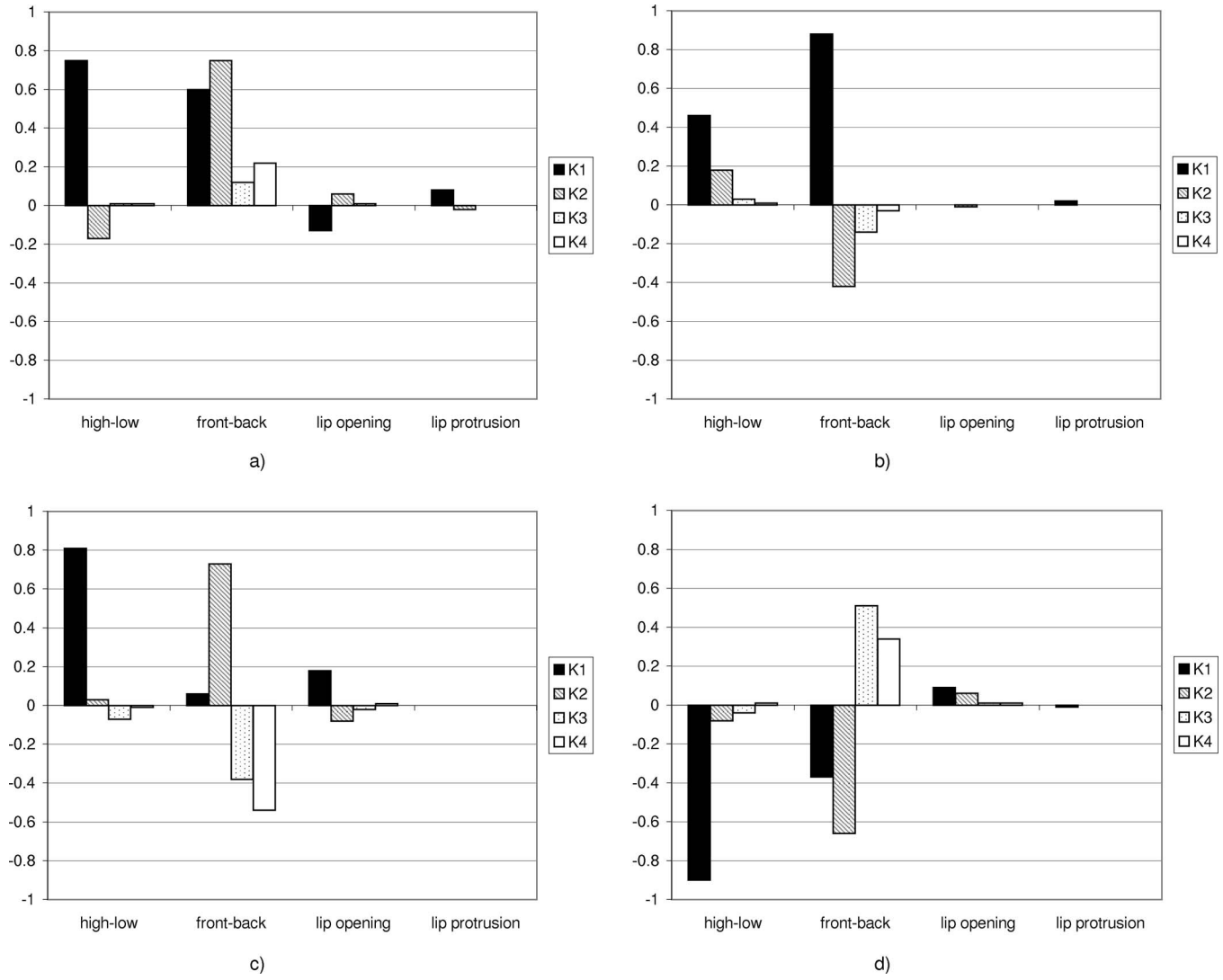

FIG. 3. The normalized projection of each articulatory component—$K1, K2, K3, K4$—onto each familiar degree-of-freedom—high-low, front-back, lip opening, and lip protrusion—for (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

the tongue high-low movement reaches the 20% level only in each speaker's $K1$. Lip protrusion and lip opening are not captured at the 20% level by any of the components for any of the speakers.

PCA analysis was performed on the acoustic data to generate a set of three orthonormal vectors that spanned the formant space. Table III shows the amount of variance accounted for by each of the PCA components, $A1$, $A2$, and $A3$. A detailed examination of the PCA results revealed that for all speakers, $A1$ contained a very large proportion of $F2$, and small amounts of $F1$ and $F3$, because $F2$ was the formant frequency with the largest variance. $F1$ and $F3$ had comparable variances in terms of absolute frequency and these formants were accounted for in $A2$ and $A3$.

The PCA components $A1$, $A2$, and $A3$ spanned the same space as the original formant frequencies $F1$, $F2$, and $F3$ and the two sets of basis vectors can be transformed from one to another with a linear, distance-preserving transformation. However, because the PCA components are orthogonal they can be used as independent variables in global linear regression analysis without the complication of partial correlations between them.

TABLE III. Percent of variance accounted for by the three acoustic PCA components.

	$A1$ (%)	$A2$ (%)	$A3$ (%)
JW11	81.0	13.4	5.6
JW14	72.9	17.0	10.1
JW16	75.8	13.6	10.6
JW18	77.2	13.6	9.2

TABLE IV. Optimum regression bandwidths b_R and the corresponding number of data points in parentheses for the forward mappings. Speakers are arranged by row and dependent variables by column.

	$F1$	$F2$	$F3$
JW11	0.004 (47)	0.004 (47)	0.004 (47)
JW14	0.003 (43)	0.003 (43)	0.003 (43)
JW16	0.003 (53)	0.002 (36)	0.002 (36)
JW18	0.004 (41)	0.003 (30)	0.003 (30)

TABLE V. Optimum interpolation bandwidths b_I and the corresponding number of data points in parentheses for the forward mappings. Speakers are arranged by row and dependent variables by column.

	$F1$	$F2$	$F3$
JW11	0.001 (12)	0.001 (12)	0.001 (12)
JW14	0.0008(12)	0.0008(12)	0.0008(12)
JW16	0.0007(12)	0.0007(12)	0.0007(12)
JW18	0.0009 (9)	0.0009 (9)	0.0012(12)

B. Optimum bandwidths for loess

The regression and interpolation bandwidths that were found optimal for the loess mappings by the cross-validation method are listed in Tables IV and V for forward mappings and in Tables VI and VII for inverse mappings. The neighborhood sizes represented by these bandwidths are found in parentheses.

The interpolation bandwidths are consistently smaller than the regression bandwidths with 7–18 data points included in the neighborhood for interpolation (Table IV versus Table V, and Table VI versus Table VII). The optimum regression bandwidths are consistently two to three times as large for the inverse mappings as for the forward mappings when comparing within subject (Table IV versus Table VI).

C. Forward mappings

It is possible to visualize the forward mappings from $K1$ and $K2$ to the formant frequencies $F1$ and $F2$ by plotting $[\hat{g}_j^{\rightarrow}]^{(1,2)}$for $j=1,2$, in which $F1$ and $F2$ values are averaged over $K3$ and $K4$. Four examples of forward mappings are presented. Figures 4 and 5 show these mappings from $K1$ and $K2$ to $F1$ and $F2$, respectively—that is, $[\hat{g}_1^{\rightarrow}]^{(1,2)}$ and $[\hat{g}_2^{\rightarrow}]^{(1,2)}$—for subject JW11. $K1$ has a relatively large effect on $F1$ with $K2$ having a slight effect. This would be expected from the projections of the $K1$ and $K2$ vectors onto the vector representing tongue height (Fig. 3). Figure 5 shows that $K1$ and $K2$ both have effects on $F2$, and that these parameters interact substantially. This can be expected from the projections of $K1$ and $K2$ onto the vector representing the tongue front-back dimension. Figures 6 and 7 show these averaged mappings from $K1$ and $K2$ to formants $F1$ and $F2$, respectively, for subject JW14. Both $K1$ and $K2$ affect $F1$, but with $K1$ having the largest effect, as can be expected by considering these vectors' projections onto the tongue height dimension. $K1$ and $K2$ also affect $F2$.

TABLE VI. Optimum regression bandwidths b_R and the corresponding number of data points in parentheses for the inverse mappings. Speakers are arranged by row and dependent variables by column.

	$K1$	$K2$	$K3$	$K4$
JW11	0.008 (94)	0.008 (94)	0.008 (94)	0.008 (94)
JW14	0.006 (86)	0.006 (86)	0.010(144)	0.008(115)
JW16	0.006(107)	0.008(142)	0.006(107)	0.006(107)
JW18	0.010(102)	0.010(102)	0.010(102)	0.012(122)

TABLE VII. Optimum interpolation bandwidths b_I and the corresponding number of data points in parentheses for the inverse mappings. Speakers are arranged by row and dependent variables by column.

	$K1$	$K2$	$K3$	$K4$
JW11	0.001 (12)	0.001 (12)	0.0012 (14)	0.0012 (14)
JW14	0.0005 (7)	0.0005 (7)	0.0012 (17)	0.0005 (17)
JW16	0.0005 (9)	0.001 (18)	0.0006 (11)	0.0006 (11)
JW18	0.0007 (7)	0.0011(11)	0.0013 (13)	0.0012 (13)

D. Inverse mappings

The inverse mappings from the formants to each of $K1-K4$ can be viewed as mappings from the $F1,F2$ plane to variables $K1-K4$ when the values are averaged over $F3$—that is, $[\hat{g}_j^{\leftarrow}]^{(1,2)}$ for $j=1,2,3,4$. (The independent variables $A1$, $A2$, and $A3$ have been transformed to $F1$, $F2$, and $F3$ here for purposes of illustration.) Instead of plotting these functions as a set of surfaces, a series of tongue shapes can be drawn in which each tongue shape corresponds to the value of $K1$, $K2$, $K3$, and $K4$ as either $F1$ or $F2$ varies. Examples of such inverse mappings are provided in Figs. 8 and 9. In Fig. 8, $F2$ is held constant, while $F1$ is increased for subjects JW11 [Fig. 8(a)] and JW14 [Fig. 8(b)]. The tongue goes from the thickest lines to the thinnest lines, and the lips from the largest asterisks to the smallest. The expected variation in tongue height and mouth opening is apparent. $F1$ is held constant in Fig. 9, while $F2$ is increased for subjects JW11 [Fig. 9(a)] and JW14 [Fig. 9(b)]. The tongues tend to move forward as $F2$ increases, and as $F2$ reaches the maximum value the mouth must open and the tongue tip drop in order to allow $F1$ to remain constant.

E. Distributions of local regression parameters

Fitting a loess model produces a set of regression parameters $\bar{\boldsymbol{\beta}}(\mathbf{x}_i)$ and $\bar{\gamma}(\mathbf{x}_i)$ at each data point $\mathbf{x}_i$ in the fitting set. We may examine the distributions of those parameters. As an

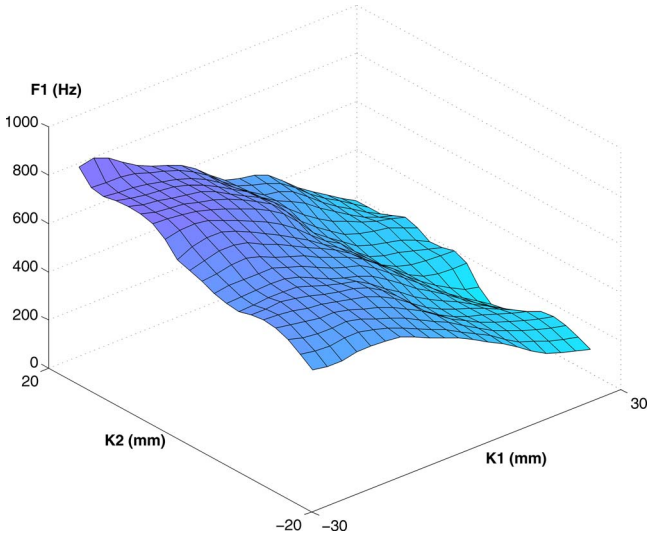

FIG. 4. (Color online) Visualization of the forward map from $K1$ and $K2$ to $F1$ for subject JW11 by a surface plot of $[\hat{g}_1^{\rightarrow}]^{(1,2)}$.

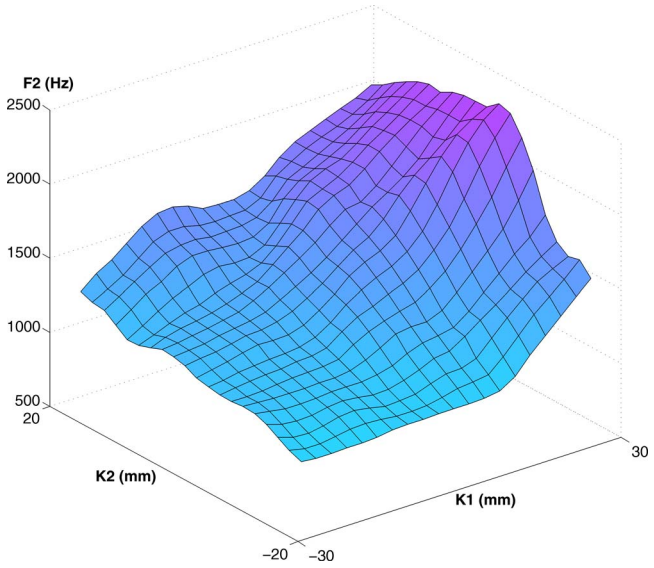

FIG. 5. (Color online) Visualization of the forward map from $K1$ and $K2$ to $F2$ for subject JW11 by a surface plot of $[\hat{g}_2^{\rightarrow}]^{(1,2)}$.

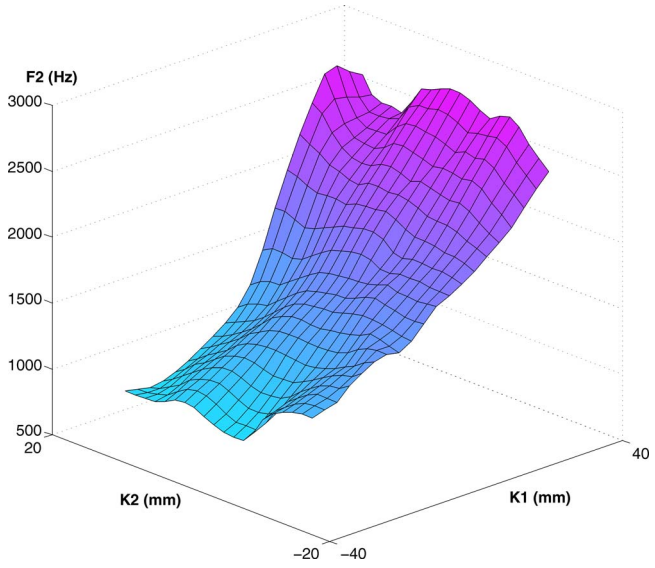

FIG. 7. (Color online) Visualization of the forward map from $K1$ and $K2$ to $F2$ for subject JW14 by a surface plot of $[\hat{g}_2^{\rightarrow}]^{(1,2)}$.

example, Fig. 10 shows the distribution of $\bar{\beta}_{11}^{\rightarrow}$ slope values for JW11. The distribution's tails have been removed in this figure.

The rest of the results are presented in terms of the means and standard deviations of the distributions of the loess regression coefficients. Figure 11 presents the means and standard deviations of the constants $\bar{\gamma}_j^{\rightarrow}$ in the forward mappings [Fig. 11(a)] and $\bar{\gamma}_j^{\leftarrow}$ in the inverse mappings [Fig. 11(b)]. The regression constants resulting from the corresponding global regressions are also indicated. The means of the loess constants and the constants from the global regression are all within one standard deviation of the distribution of constants for both the forward and inverse mappings. Of course the constants of the global regressions representing the forward mappings are simply the average formant frequencies of each speaker's data [Fig. 11(a)]. Because the variables $K1-K4$ for each speaker are the result of a PCA analysis with the mean subtracted out, the constants for the global regression representing the inverse mappings are very close to zero [Fig. 11(b)]. The standard deviations of the constants from the loess regressions for the forward mappings increase with the order of the dependent variable, formant frequency, for all subjects [Fig. 11(a)], where the "order" of formant Fj is j. There is no apparent trend between standard deviation of the constants from the loess regressions for the inverse mappings and the order of the dependent variable, articulatory PCA component [Fig. 11(b)].

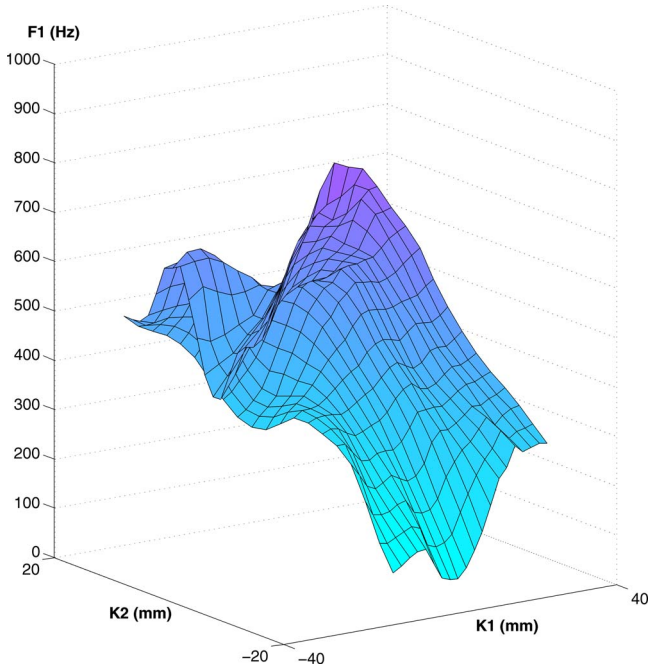

FIG. 6. (Color online) Visualization of the forward map from $K1$ and $K2$ to $F1$ for subject JW14 by a surface plot of $[\hat{g}_1^{\rightarrow}]^{(1,2)}$.

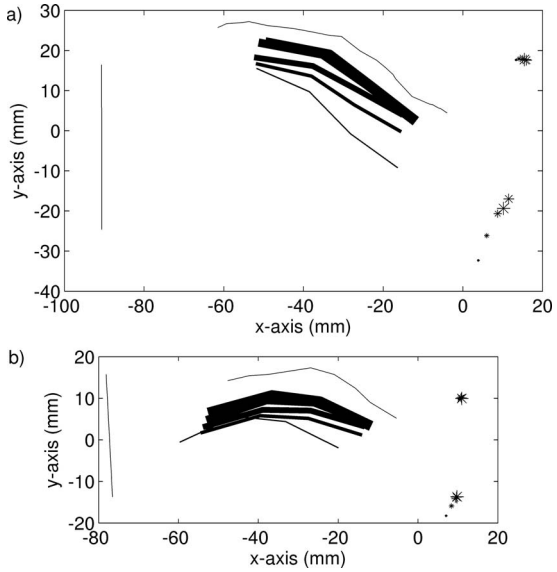

FIG. 8. Visualization of the inverse maps from $F1$ and $F2$ to $K1$, $K2$, $K3$, and $K4$. Tongue and lip positions predicted by $[\hat{g}_j^{\leftarrow}]^{(1,2)}$, $j=1,2,3,4$, are shown for a series of $F1$ increases for fixed $F2$. Tongue moves from thick line to thin. (a) JW11 with $F2=1556$ Hz and $F1$ from 145 to 877 Hz and (b) JW14 with $F2=1664$ Hz and $F1$ from 53 to 1019 Hz.

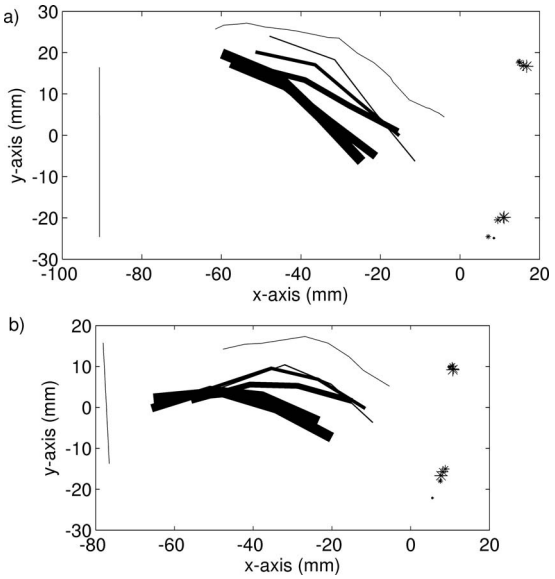

FIG. 9. Visualization of the inverse maps from $F1$ and $F2$ to $K1$, $K2$, $K3$, and $K4$. Tongue and lip positions predicted by $[\hat{s}_j^{\leftarrow}]^{(1,2)}$, $j=1,2,3,4$, are shown for a series of $F2$ increases for fixed $F1$. Tongue moves from thick line to thin. (a) JW11 with $F1=530$ Hz and $F2$ from 644 to 2377 Hz and (b) JW14 with $F1=613$ Hz and $F2$ from 522 to 2910 Hz.

Figure 12 shows the means and standard deviations of the slopes $\bar{\boldsymbol{\beta}}_j^{\rightarrow}=(\bar{\beta}_{1j}^{\rightarrow},\bar{\beta}_{2j}^{\rightarrow},\bar{\beta}_{3j}^{\rightarrow},\bar{\beta}_{4j}^{\rightarrow})$, $j=1,2,3$, of the forward mappings for each speaker. The corresponding quantities from the global regression are also indicated. The means of the loess slopes are very close to the corresponding global regression slopes when compared to the size of the standard deviations. There is a consistent trend of increasing standard deviation with the order of either the independent variable (i.e., the i in Ki and $\bar{\beta}_{ij}^{\rightarrow}$) and the order of the dependent variable (i.e., the j in Fj and $\bar{\beta}_{ij}^{\rightarrow}$). Figure 13 shows the mean

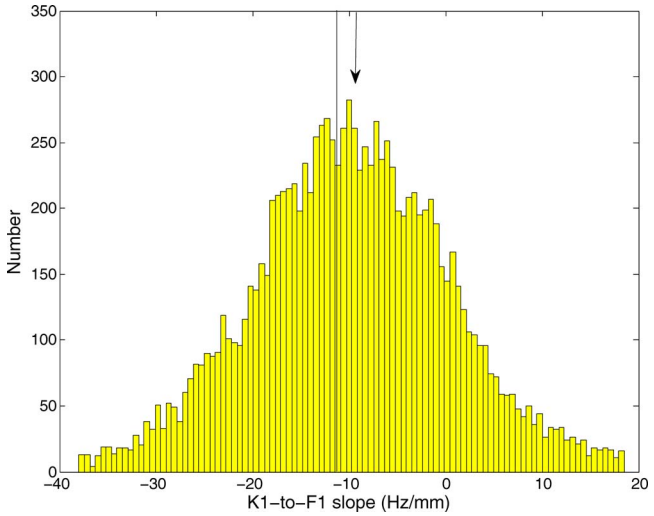

FIG. 10. (Color online) Histogram of local $K1$-to-$F1$ slope, or $\bar{\beta}_{11}^{\rightarrow}$, at the data points in the forward mapping for JW11. The arrow denotes the mean of the distribution, and the vertical line indicates the value of the $K1$-to-$F1$ slope obtained in global regression. The slope data have been trimmed to cut the tails from the distributions. The tails for a distribution X are defined as follows. Let x_q=value of X that marks the qth percentile of the X distribution, and let x_{mean}=mean of the X distribution. Then y is in one of the tails if $x_{25}-y>1.5|x_{25}-x_{\text{mean}}|$ or $y-x_{75}>1.5|x_{75}-x_{\text{mean}}|$.

and standard deviations of the loess slopes $\bar{\boldsymbol{\beta}}_j^{\leftarrow}=(\bar{\beta}_{1j}^{\leftarrow},\bar{\beta}_{2j}^{\leftarrow},\bar{\beta}_{3j}^{\leftarrow})$, $j=1,2,3,4$, of the inverse mappings for each speaker, along with the corresponding slopes from the global regressions. Again, the slopes from the global regressions are very close to the means of the corresponding slopes from the loess regressions. While there is a general trend for increasing standard deviation with the order of the independent variable (i.e., the i in Ai and $\bar{\beta}_{ij}^{\leftarrow}$), there is no apparent trend with the order of the dependent variable (i.e., the j in Kj and $\bar{\beta}_{ij}^{\leftarrow}$).

A sample of correlation coefficients among the slopes and constant was calculated. This was done for JW11 and JW14 in their loess forward mappings to $F1$ and $F2$. The independent variables $K1$ and $K2$ were the focus. Thus, the three correlation coefficients among constants, $\bar{\gamma}_j^{\rightarrow}$, and slopes, $\bar{\beta}_{1j}^{\rightarrow}$ and $\bar{\beta}_{2j}^{\rightarrow}$, for fixed $j=1$ or 2 were calculated for each JW11 and JW14. The three correlation coefficients among the absolute values of these quantities were also calculated for both speakers. The correlation coefficients in Tables VIII and IX indicate varying degrees of co-variation among loess parameters and their absolute values. Given the large number of data points, all of these correlation coefficients are significant at the $p<0.001$ level (Bickel and Docksum, 1977, pp. 221 and 472).

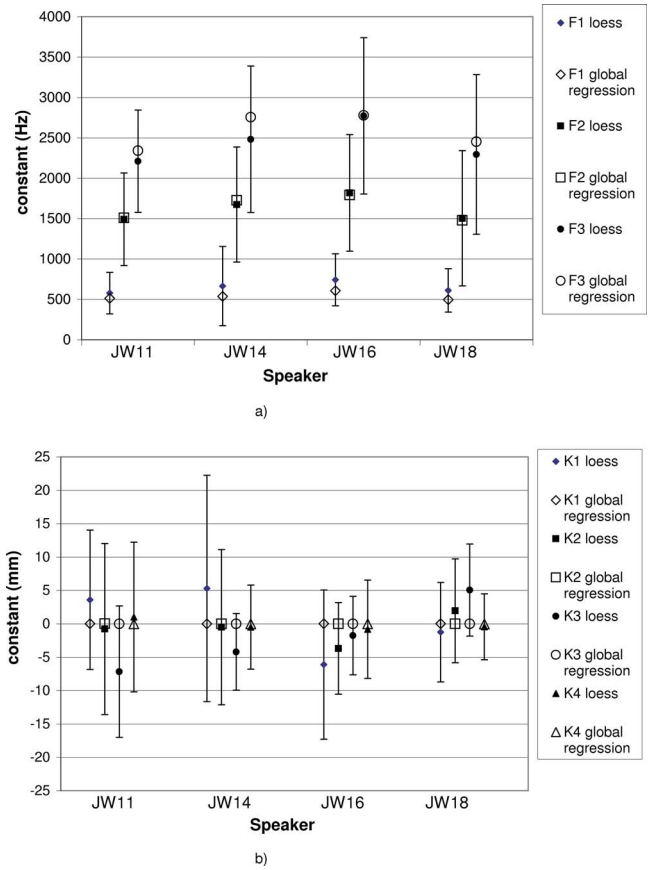

FIG. 11. (Color online) Means and standard deviations of the local regression constants ($\bar{\gamma}_j^{\rightarrow}$ and $\bar{\gamma}_j^{\leftarrow}$) at the data points. The constants obtained in global regression are also shown. Labels refer to the dependent variable. (a) $\bar{\gamma}_j^{\rightarrow}$ (forward mappings) and (b) $\bar{\gamma}_j^{\leftarrow}$ (inverse mappings).

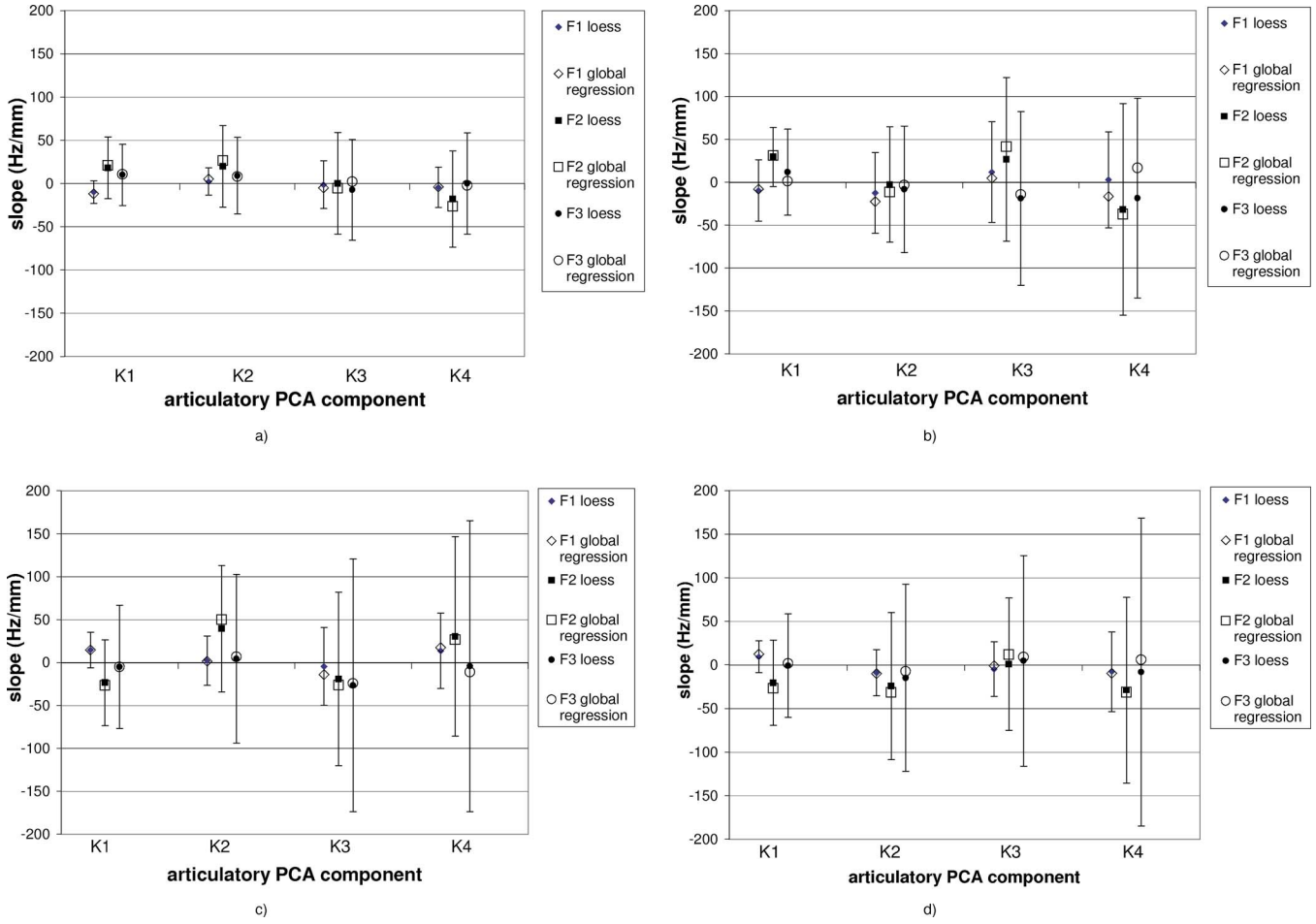

FIG. 12. (Color online) Means and standard deviations of the local regression slopes of the forward mappings at the data points (i.e., distributions of $\bar{\beta}_{ij}^{\rightarrow}$ for $i=1,2,3,4$ and $j=1,2,3$). Corresponding global regression slopes are also plotted. (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

F. Test RMSE

Figure 14 shows the RMSE results from applying the forward loess mappings to each speaker's test data set. If it is assumed that the formant frequency values are distributed in a normal distribution about a mean, then the RMSE is one standard deviation of the distribution. There is a general trend for increasing RMSE with the order of the dependent variable, or formant frequency number, for each speaker. Figure 15 shows the percent decrease in the RMSE of the forward loess compared to the corresponding global regression applied to the same test data. The improvement in fit from the global regression to the loess regression is consistently greatest for $F2$ in terms of percent reduction in RMSE for each subject.

Figure 16 exhibits the test RMSE values for the inverse loess mappings. There was a general trend, though not completely consistent, of decreasing RMSE error from $K1$ and $K2$ to $K3$ and $K4$ for each speaker. This can be contrasted with increases in RMSE from $F1$ to $F3$ in the forward mappings for each speaker. Figure 17 shows the percent decrease in RMSE for the loess from the global regression for the inverse mapping. The amount of decrease for the inverse mappings is generally less than for the forward mappings (Fig. 15).

G. Sensitivity

The results of the sensitivity analysis are presented by sectioning a schematic two-dimensional articulatory vowel space of each speaker into regions of large and small sensitivity of acoustic variables to articulatory variables. For each speaker, schematic articulatory vowel spaces were constructed by connecting the four point-vowel articulations projected onto their $(K1,K2)$ coordinates to form a quadrilateral. Adjoined to this quadrilateral was another quadrilateral representing the region between the projected [i] and [u] articulations and the palate. The resulting polygon formed a schematic of the articulatory vowel space in $K1-K2$ space.

The schematic articulatory vowel space of each speaker was divided into regions of large and small sensitivity according to each empirical sensitivity function $\nu_{ij}^{\rightarrow}$, $i=1,2$, $j=1,2$. Boundaries of the regions were defined by visual inspection. The results were grouped into four figures with Fig. 18 corresponding to $F1$ sensitivity to changes in $K1$ (magnitudes of $\nu_{11}^{\rightarrow}$), Fig. 19 corresponding to $F1$ sensitivity to changes in $K2$ (magnitudes of $\nu_{21}^{\rightarrow}$), Fig. 20 corresponding to $F2$ sensitivity to changes in $K1$ (magnitudes of $\nu_{12}^{\rightarrow}$), and Fig. 21 corresponding to $F2$ sensitivity to changes in $K2$ (magnitudes of $\nu_{22}^{\rightarrow}$).

Figure 18 shows a general tendency for the regions of

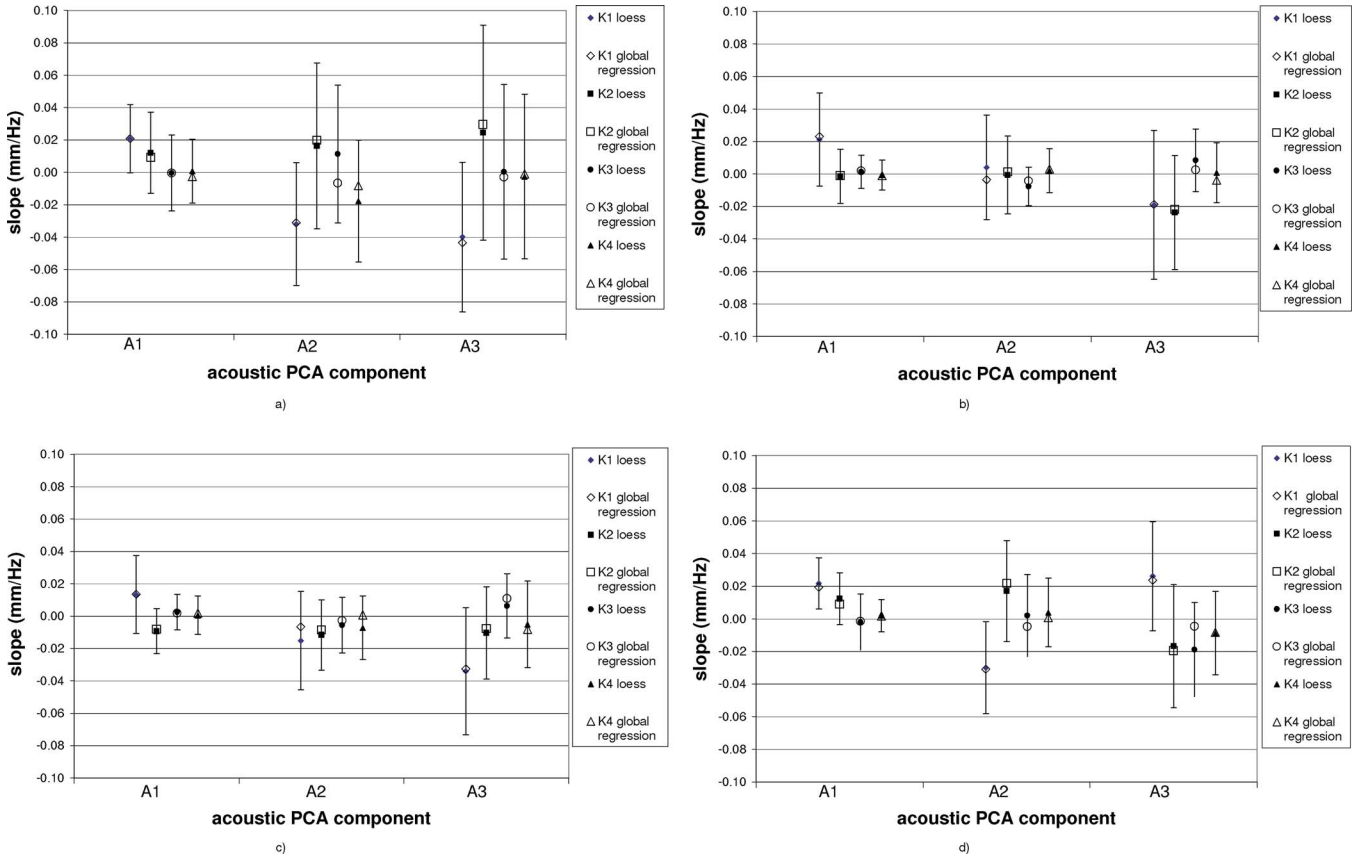

FIG. 13. (Color online) Means and standard deviations of the local regression slopes of the inverse mappings at the data points (i.e., distributions of $\bar{\beta}_{ij}^{\leftarrow}$ for $i=1,2,3$ and $j=1,2,3,4$). Corresponding global regression slopes are also plotted. (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

largest sensitivity of $F1$ to $K1$, or largest $\nu_{11}^{\rightarrow}$, to be associated with high vowels and the least sensitive with low vowels. There are differences among the speakers, with JW11 [Fig. 18(a)] and JW18 [Fig. 18(d)] exhibiting similar patterns. (These are sketches where differences in the amount of area covered by the shadings are not a reliable indication of differences in sensitivity.) JW14 [Fig. 18(b)] appears to have the pattern of JW11 and JW18 rotated counterclockwise through her vowel space. JW16 has neither large nor small sensitivity in the front part of her vowel space. Figure 19 for the sensitivity of $F1$ to $K2$ change again shows that regions of large sensitivity, or large $\nu_{21}^{\rightarrow}$, are in the high part of the vowel space, and regions of small sensitivity are in the low part of the vowel space. In fact, for JW11 the sensitivity regions are quite similar between $\nu_{11}^{\rightarrow}$ and $\nu_{21}^{\rightarrow}$. In contrast, JW14's pattern for $\nu_{21}^{\rightarrow}$ appears to be a clockwise rotation of her pattern for $\nu_{11}^{\rightarrow}$ [Figs. 18(b) and 19(b)]. JW18 shows a similar small rotation [Figs. 18(d) and 19(d)]. JW16 exhibits a sensitivity pattern in $\nu_{21}^{\rightarrow}$ similar to that of JW11, except her region of small sensitivity appears to extend higher into her vowel space [Figs. 19(c) and 19(a)].

The sensitivities of $F2$ to $K1$ change and $K2$ change (Figs. 20 and 21) show more regions of large and small sensitivity than the sensitivity of $F1$ to $K1$ change and $K2$ change (Figs. 18 and 19). This is to be expected because the higher frequency formant depends on finer-scaled details of articulation than the lower frequency formant does. All of the speakers appear to have large $F2$ sensitivity to $K1$ change, or large $\nu_{12}^{\rightarrow}$, in the high-back region of the vowel space and a region of small sensitivity somewhere in the low region of the vowel space with, often, an accompanying region of large sensitivity (Fig. 20). On the other hand, the regions of large sensitivity of $F2$ to $K2$ change, or large $\nu_{22}^{\rightarrow}$, are in the back part of the vowel spaces, except for JW16 (Fig. 21). The regions of small sensitivity of $F2$ to $K2$ change are most notably in the front of the vowel spaces, except for JW11, who shows small sensitivity around the low-back vowel (Fig. 21).

The critical articulation points in the $K1,K2$ subgrid—i.e., those subgrid vertices with $\nu_{ij}^{\rightarrow}$ values smaller than that of any of the neighboring vertices—were examined in the context of the point vowels projected onto the $K1,K2$ plane. The critical articulations that corresponded closely to point vowels for each speaker are noted in Table X.

TABLE VIII. Correlation coefficients for local regression parameters at the data points with means subtracted in the forward mappings. The subscript j denotes the dependent variable (formant) number.

	$\bar{\gamma}_j^{\rightarrow}$ by $\bar{\beta}_{1j}^{\rightarrow}$	$\bar{\gamma}_j^{\rightarrow}$ by $\bar{\beta}_{2j}^{\rightarrow}$	$\bar{\beta}_{1j}^{\rightarrow}$ by $\bar{\beta}_{2j}^{\rightarrow}$
JW11, $j=1$	−0.174	−0.188	0.133
JW11, $j=2$	−0.471	−0.271	0.303
JW14, $j=1$	0.232	0.181	−0.032
JW14, $j=2$	−0.217	−0.075	0.048

TABLE IX. Correlation coefficients for absolute values of the local regression parameters at the data points with means subtracted in the forward mappings. The subscript j denotes the dependent variable formant number.

	$\lvert\vec{\gamma}_j\rvert$ by $\lvert\vec{\beta}_{1j}\rvert$	$\lvert\vec{\gamma}_j\rvert$ by $\lvert\vec{\beta}_{2j}\rvert$	$\lvert\vec{\beta}_{1j}\rvert$ by $\lvert\vec{\beta}_{2j}\rvert$
JW11, $j=1$	0.683	0.365	0.362
JW11, $j=2$	0.719	0.618	0.599
JW14, $j=1$	0.545	0.500	0.272
JW14, $j=2$	0.529	0.424	0.301

The results show that there is little consistency among the speakers as to the point vowels that could be considered to have critical $(K1, K2)$ coordinates.

IV. DISCUSSION

A. Articulatory PCA

The lower-order articulatory PCA components in the present study accounted for less of the variance than is typical for PCA analyses of a small number of vowels. Typically two principal components will account for at least 95% of the variance for a small number of static vowel images (on the order of 10) (e.g., Jackson, 1988; Mokhtari *et al.*, 2007; Jackson and McGowan, 2008). On the other hand, for subject JW16, more of the tongue position variance was accounted for in the present study than was accounted for in the Beaudoin and McGowan (2000) study involving the same subject, also with thousands of data points. That is, the sum of the variances accounted for in the first one to four components is always greater in the present work than in the Beaudoin and McGowan (2000) study, where tokens were chosen automatically based on a criterion of minimum tongue-to-palate distance. It appears that the present data set based on vowels alone and consonant-vowel transitions that excluded nasal, lateral, and rhotic contexts was more restrictive than the earlier study.

The degrees-of-freedom that were dominant in the $K1$ and $K2$ components were familiar high-low and front-back degrees-of-freedom, except that the $K1$ and $K2$ components of JW16 additionally contained a large amount of tongue blade up-down and tongue bunching-stretching degrees-of-freedom. $K3$ and $K4$ contain varying amounts of tongue bunching-stretching as well as tongue blade up-down degrees-of-freedom. These degrees-of-freedom would be expected to appear because transitions to and from consonants have been included in the data set. None of the articulatory PCA components have substantial amounts of lip opening or lip protrusion. This could be due to the fact that there is only one phonologically rounded monophthong in American English, and that the vowels in the XRMB-SPD appeared with both unrounded and rounded consonants. The latter fact may mean that lip position was not correlated with tongue position.

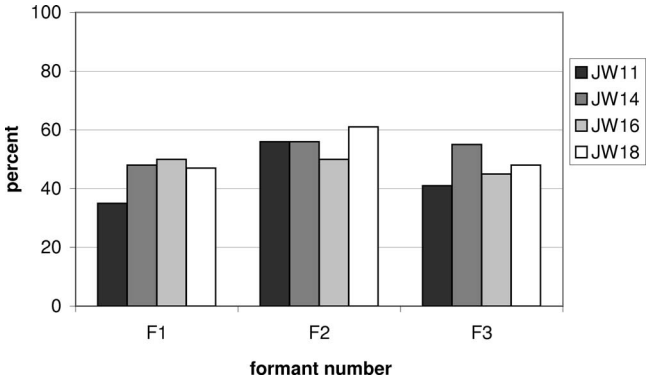

FIG. 15. Percent improvement in test RMSE of loess over global regression for the forward mappings.

B. Distributions of local regression parameters

In Sec. III E, the wide distributions of constants and slopes of loess were noted. It is of some interest to investigate whether there is co-variation among the constant and slope values in order to know whether the values in the tails of their distributions could be accounted for in terms of compensation. Evidence for at least a small amount of such covariation was found.

However, a comparison between corresponding cells in Tables VIII and IX show that the correlation coefficients for the absolute values are larger, by a factor of 1.5–8.5, than their corresponding coefficients of the signed parameters. This indicates that while there may be a negative (or positive) relation between two parameters, there are substantial numbers of data points where a positive (or negative) relation holds between them. Further, there is a strong correlation in magnitudes no matter the sign. The covariance among loess parameters cannot be neglected, yet the sign of the covariance is only weakly determined compared to the covariance in magnitude.

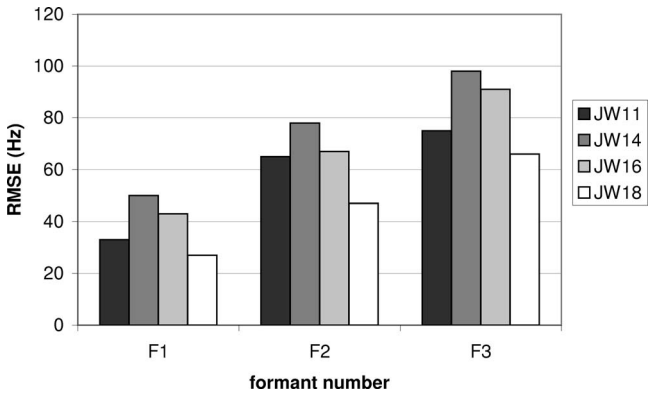

FIG. 14. Test RMSE of the forward loess mappings.

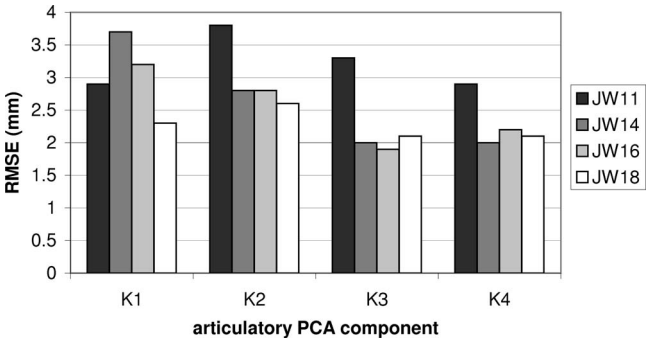

FIG. 16. Test RMSE of the inverse loess mappings.

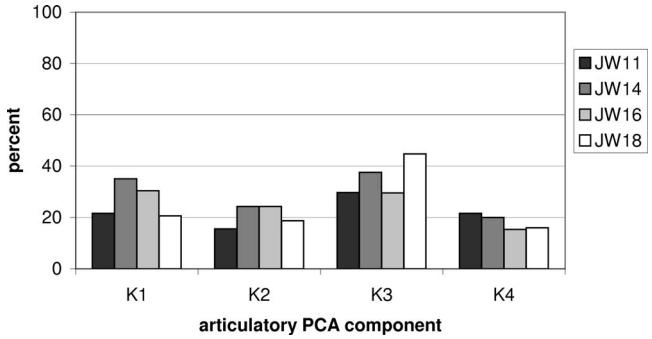

FIG. 17. Percent improvement in test RMSE of loess over global regression for the inverse mappings.

C. Measurement error and test RMSE

The different contributions to error—in measuring articulatory coordinates, in measuring formant frequencies, and in relating the two domains—are now considered. Much of the discussion will focus on how measurement error relates to the test RMSE of the loess models. It will be seen that the RMSEs for the forward loess mappings are better than expected based on measurement error considerations, but the RMSEs for the inverse loess maps are not as good as would be expected based on estimated measurement errors. Explanations for the latter phenomenon will be offered.

There are two sources of formant variability that cannot be attributed to measured articulatory variability. One source is the movement of the speech articulators whose positions were not measured. An example of an articulator coordinate that was not measured for the XRMB-SPD is larynx height. Westbury (1983), using X-ray cineradiography, measured changes in larynx height of 17 mm during speech. Wood

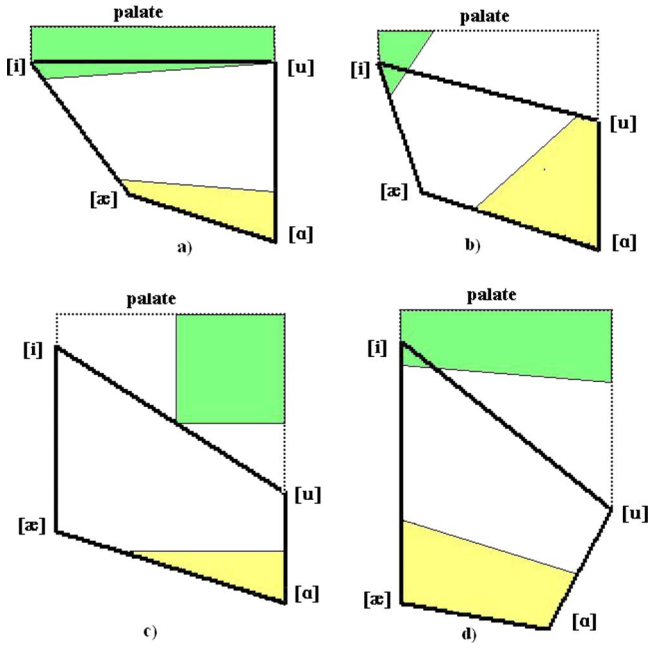

FIG. 18. (Color online) Schematic articulatory vowel spaces with regions of large (dark shading) and small (light shading) empirical sensitivity of $F1$ to $K1$ ($\vec{\nu_{11}}$). The vowel spaces are defined by the articulation of the point vowels projected onto ($K1$,$K2$) coordinates. (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

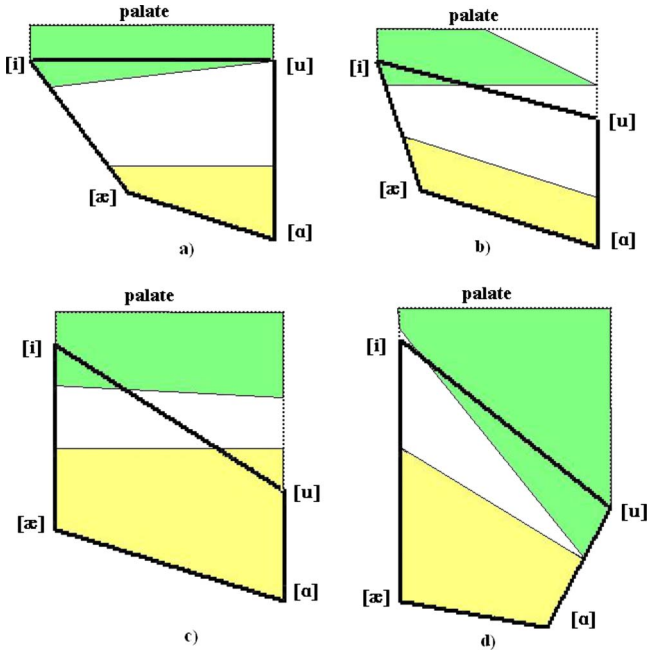

FIG. 19. (Color online) Schematic articulatory vowel spaces with regions of large (dark shading) and small (light shading) empirical sensitivity of $F1$ to $K2$ ($\vec{\nu_{21}}$). The vowel spaces are defined by the articulation of the point vowels projected onto ($K1$,$K2$) coordinates. (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

(1979) measured a maximum range in larynx height of only 5 mm for all the speakers he studied in 15 different languages and dialects. Perkell (1969, pp. 38–42) measured larynx height changes of nearly 15 mm in isolated utterances of a single speaker. However, he did state that these variations were diminished in running speech. Changes in larynx position can have a substantial effect on the formant frequencies. For example, Lindblom and Sundberg (1971) showed in experiments with a synthesizer that $F1$ and $F2$ in vowel production could change by up to 8% with a 10 mm change in larynx height. There are other articulatory dimensions that are not measured, such as pharyngeal dimensions, velum position, and all lateral dimensions. But these effects have been minimized here. It has been shown that in English and Swedish, pharyngeal dimensions are largely predictable from the positions of the front of the tongue during static vowel production (Whalen *et al.*, 1999; Jackson and McGowan, 2008). Furthermore, restricting vowel contexts to exclude nasal and lateral consonants has minimized the amount of velar and lateral variation for any given set of tongue pellet coordinates. Thus, it would seem to be reasonable to assume that the unseen articulatory coordinates should cause a variation of at most 10%–15% in formant frequencies for a given set of tongue and lip pellet coordinates.

Another source of variability for formant frequency values is measurement error incurred in using LPC analysis with human correction. The errors in the formant frequency measurements themselves are estimated to be less than about 5%. In total, one can expect a "cloud" of $F1$, $F2$, and $F3$ measures of radius of about 15%–20% of the mean to be associated with each data point in the articulatory space. In terms of formant frequency values, this would correspond to

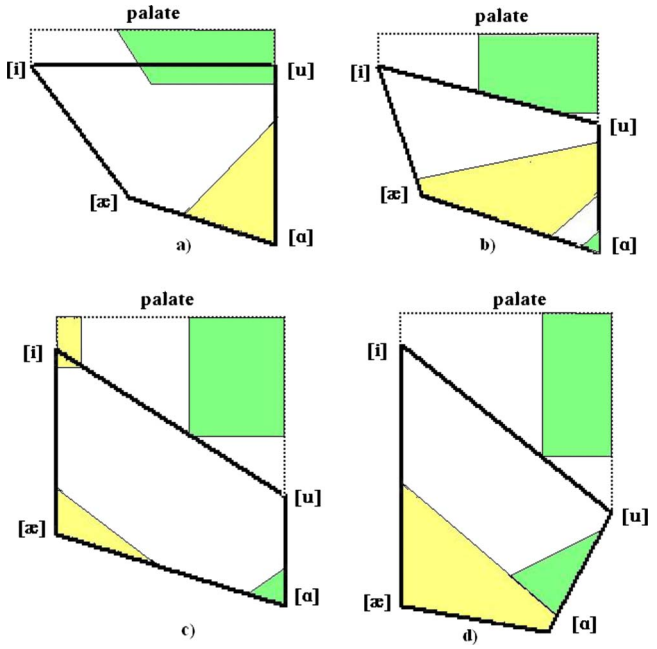

FIG. 20. (Color online) Schematic articulatory vowel spaces with regions of large (dark shading) and small (light shading) empirical sensitivity of $F2$ to $K1$ ($\nu_{12}^{\rightarrow}$). The vowel spaces are defined by the articulation of the point vowels projected onto ($K1$,$K2$) coordinates. (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

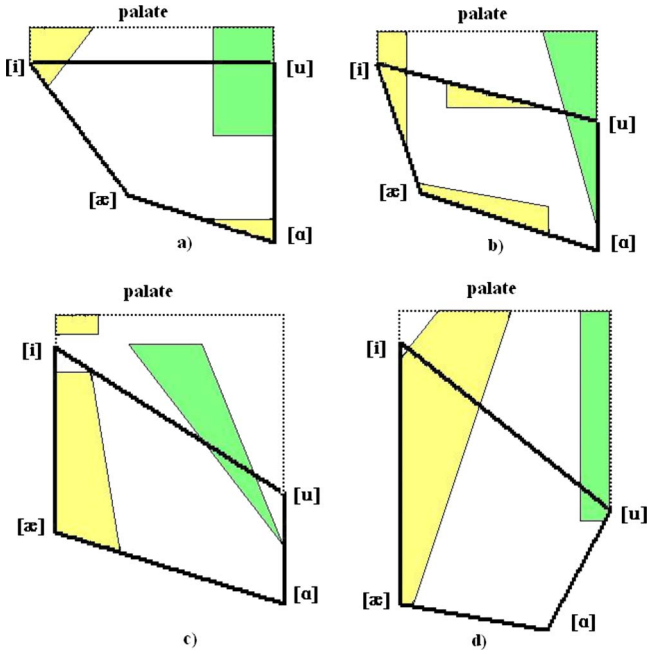

FIG. 21. (Color online) Schematic articulatory vowel spaces with regions of large (dark shading) and small (light shading) empirical sensitivity of $F2$ to $K2$ ($\nu_{22}^{\rightarrow}$). The vowel spaces are defined by the articulation of the point vowels projected onto ($K1$,$K2$) coordinates. (a) JW11, (b) JW14, (c) JW16, and (d) JW18.

radii of 75–100 Hz in $F1$, 225–300 Hz in $F2$, and 375–500 Hz in $F3$ based on means of 500, 1500, and 2500 Hz for $F1$, $F2$, and $F3$ respectively.

We may compare the above estimates of the radii of the formant clouds with the RMSEs of the forward loess mappings. If the error estimates can be understood to be on the order of two standard deviations of a normal formant frequency error distribution (accounting for 84% of the error), and the RMSEs are expected to be one standard deviation of the error distribution, then the error estimate should be twice the RMSE. In the loess mappings, doubling the RMSE for $F1$ gave a value in the range from 54 to 100 Hz, for $F2$ from 94 to 156 Hz, and for $F3$ from 132 to 196 Hz (Fig. 14). These data indicate that twice the standard deviation is on the order of 10%–20% of the mean for $F1$ and 6%–10% of the mean for $F2$ and $F3$. Thus, forward loess mappings for $F2$ and $F3$ perform better than would be expected based on error estimates. Thus it is reasonable to reduce the estimated standard deviation in the formant error distribution to coincide with the forward mappings' RMSEs.

We have estimated the RMSE of $F1$ to be 5%–10% of the mean value of $F1$ and the RMSEs of $F2$ and $F3$ to be 3%–5% of the mean values of $F2$ and $F3$, respectively. These values compare favorably with the error obtained by Mermelstein (1967) when he replaced an area function for six vowels with the first six cosine components of that area function. This replacement produced maximum errors of about 3% in the formant frequencies (Mermelstein, 1967). In contrast to Mermelstein's (1967) investigation from a single set of six model area functions, the present investigation is based on large numbers of articulatory configurations from human talkers with four orthogonal articulatory components spanning the articulatory space.

The estimated positional error for stationary pellets is 0.15 mm (Westbury, 1994, p. 69). Further, it can be expected that fast moving articulators (400 mm/s) will move 0.6 mm during the time it takes for the tracking raster to be generated, thus creating an error due to recording time delay. Another possible source of error is the time delay between the articulatory position and the recorded acoustics. This is error caused by the fact that the speed of sound is finite and the distance of the microphone to the mouth was about 110 mm. It is estimated that the acoustic signal is recorded 0.3 ms after it is generated by the vocal tract, which corresponds to about 5% of the time between sampled frames of data (at 160 Hz). For a fast moving articulator this would correspond to only 0.03 mm of movement, which is small compared to the error caused by the recording time delay. With these considerations, a reasonable estimate of maximum position measurement error is 0.6 mm for each pellet coordinate. If the position measurement errors are independent across the pellet coordinates, then the maximum total magnitude of error for the six pellets would be about 2 mm. In particular, this can be considered to be the measurement error in each of the articulatory PCA components, $K1$, $K2$, $K3$, and $K4$.

TABLE X. The point vowels projected onto $K1$,$K2$ coordinates that closely matched critical values of ($K1$, $K2$) for each empirical sensitivity function $\nu_{ij}^{\rightarrow}$.

	JW11	JW14	JW16	JW18
$\nu_{11}^{\rightarrow}$	[i]		[ɑ]	[ɑ]
$\nu_{12}^{\rightarrow}$	[æ]	[i]	[æ]	[u]
$\nu_{12}^{\rightarrow}$	[ɑ], [æ], [i]		[æ]	[ɑ]
$\nu_{22}^{\rightarrow}$	[æ], [i]		[ɑ]	[u]

If, as in the estimate of the acoustic error, the estimated articulatory measurement error is about two standard deviations of the articulatory error distribution, then it is possible to compare this estimate with twice the test RMSE in inverse loess mappings. Doubling the RMSEs of the inverse mappings gives values between 3.8 and 7.6 mm. The estimated corresponding quantity for the error distribution of 2 mm is well below this range. Causes for this discrepancy and the differences between the forward and inverse mappings in terms of the relation between error distributions and RMSE are proposed below.

One factor contributing to the RMSE of the inverse mappings being two to four times larger than the estimated standard deviation of the error distribution of the articulatory dependent variables is the compensatory ability of the vocal tract. For a given triple of formant frequencies ($F1$, $F2$, and $F3$), there are many vocal tract shapes able to produce it (Atal *et al.*, 1978). Not only can $K1$, $K2$, $K3$, and $K4$ compensate for one another but the unseen articulators, such as the larynx, can do the same. Thus, a point in the acoustic space can map into a region of the articulatory space. This is without any consideration of measurement error. It can be speculated that this is the reason why the optimum regression bandwidths are at least twice as large for the inverse mappings as for the forward mappings of each speaker (Tables IV and VI). It is a reasonable possibility that more acoustic variables need to be included in the inverse mapping to ensure that the inverse mappings are as unambiguous as possible. It can be proven that the RMSE is a positively biased estimate of the error standard deviation if independent variables with significant effect have not been included in the analysis (Chatterjee and Hadi, 1988, pp. 40–42).

Another contribution to the larger RMSE of the inverse mappings might involve the assumptions made in regression analysis. When regression is performed it is assumed that the independent variables are known exactly, or at least the independent variables should be known more precisely than the dependent variables. Therefore it of interest to know how the magnitudes of the errors in articulatory variables compare to those of the acoustic variables, or how the errors in the articulatory PCA components compare to errors in the acoustic PCA components. The standard deviations of the error distribution in the formant frequencies were estimated by the RMSEs of the optimum loess forward mappings, because the original estimates of those standard deviations were found to be too large. On the other hand the 1 mm estimate will be used for the standard deviation in the error distribution of each of the articulatory PCA components. In order to compare magnitudes of errors for physical variables (articulatory and acoustic) with different dimensionalities, it is necessary to normalize both sets of variables. They are normalized by the standard deviations of the coordinate data in the data set for each speaker. For instance, the estimated standard deviation in the error distribution of $K2$ was normalized by the total standard deviation in the $K2$ data.

The normalized estimated error standard deviations averaged over the four speakers for $K1$, $K2$, $K3$, and $K4$ are 0.10, 0.17, 0.27, and 0.33, respectively. The same quantities for $A1$, $A2$, and $A3$ are 0.18, 0.27, and 0.44. The ordering of these variables by increasing magnitude of the normalized estimated error standard deviations is $K1$, $K2$, $A1$, $(A2, K3)$, $K4$, $A3$, which indicates that the lower-order articulatory quantities are determined with less error than the lower-order acoustic quantities. The average normalized estimated error standard deviations are 0.22 ± 0.10 for the articulatory variables and 0.29 ± 0.13 for the acoustic variables. While these averages are not significantly different, the lower-order articulatory variables do appear to be known more accurately than the acoustic variables. These considerations call into question the assumption that the independent variables are known relatively precisely for the loess regressions for the inverse mappings.

To compare these results with previous work in the inverse mapping, reference can be made to Ladefoged *et al.* (1978) and Mokhtari *et al.* (2007). The magnitudes of the RMSEs for $K1-K4$ shown in Fig. 16 are close to, or less than, those of the inverse mappings constructed by Ladefoged *et al.* (1978) for five subjects each producing five vowels (Ladefoged *et al.*, 1978, Table II). Mokhtari *et al.* (2007) investigated one subject speaking a series of vowels for which there were 35 frames of data. They performed regression using three formants as the independent variables and articulatory PCA components of the area function as the dependent variables. For each PCA component they reported the RMSE normalized by the standard deviation of that component: 0.22 and 0.45 for the first and second PCA components, respectively (Mokhtari *et al.*, 2007, Fig. 4). By performing the same normalization using the standard deviations of $K1$ and $K2$, the results of the inverse loess may be compared with these values. The normalized RMSE averaged across the four subjects was 0.29 ± 0.05 for $K1$ and 0.50 ± 0.05 for $K2$.

We have provided only rough estimates of the error expected in both the acoustic and articulatory domains. However, to provide quantitative confidence intervals for both the mappings and the parameters in the mappings, such as $\boldsymbol{\beta}(\mathbf{x})$, the error in both domains should be estimated using the RMSEs and the properties of the loess mappings. There are well-known, established methods to do this with parametric linear regression (e.g., Bickel and Docksum, 1977, pp. 267–268). These methods can be extended to loess, but with some complications. In both parametric linear regression and loess, the estimates of the dependent variable may be viewed as a linear transformation of the data in the dependent variable domain. For parametric linear regression this transformation is an idempotent projection, while for loess the transformation is not idempotent, and it is not even symmetric (Cleveland and Devlin, 1988). This makes computing the estimated variances much more computationally intensive: it involves computations of the traces of products of $N \times N$ matrices, where N is the number of data points. These computations were prohibitive for the current data sets of 10 000–20 000 data points. In the future we will research methods for finding approximations to these traces. Another procedure for finding confidence intervals would be to use Monte Carlo simulation. These issues merit further research.

Another issue confronting estimation of error is covariation in the independent variables. Since the independent

variables are PCA components, they are orthogonal in the global sense, but they are not necessarily so for each locally weighted regression, performed here at each data point. It is entirely possible that the independent variables are nearly linearly dependent in certain of the local regressions. Averaging over the higher-order independent variables was used here for visualization and data-reduction purposes. However, this averaging lessened the effects of co-variation that could cause the loess mappings and the measures of formant sensitivity to be noisier. To provide quantitative confidence intervals for both the mappings and the parameters in the mappings, future work should address the issue of co-linearity of independent variables in local regions and co-variation of loess regression coefficients between regions.

D. Sensitivity

The results on sensitivity of $F1$ and $F2$ to $K1$ and $K2$ changes near the point vowels [ɑ], [æ], and [i] can be viewed in terms of the geometric properties of $K1$ and $K2$ shown in Figs. 2 and 3, and the effects of those properties in simple acoustic tube models. Sensitivities of formants to articulatory change for the vowels [ɑ] and [i] can be understood using two-tube models, while [æ] can be viewed as a slightly flared tube (Perkell, 1969, p. 55; McGowan, 2006). The tube properties of [u] are more complex and discussion of this vowel would be more speculative. Further, the discussion will focus on subjects JW11, JW14, and JW18, with only brief mentions of JW16 because of the difficulty in characterizing the $K1$ component of JW16 due to its tongue rocking motion.

Ehrenfest's theorem applied to acoustic tubes was introduced to the speech production community by Schroeder (1967). In this community it is commonly known as acoustic perturbation theory and it is useful for the discussion of sensitivity in conjunction with the two-tube models. Briefly, if, for a given resonant frequency f, cross-sectional tube area is decreased a small amount in a region where acoustic potential energy density is greater than acoustic kinetic energy density, then the resonant frequency increases, that is, $df/f > 0$. The opposite occurs if the kinetic energy density is greater than the potential energy density. To apply this theorem, one needs to compute resonant frequencies and their corresponding energy density distributions through the tube, which is done numerically. However, for a sufficiently low resonant frequency f (e.g., $F1$) and a two-tube model, there is a simplification. For a two-tube model, let the tube closest to the glottis have length L_1 and area A_1, and the tube closest to the lips have length L_2 and area A_2, and let $\alpha = A_2/A_1$. Assuming that the total tube length remains constant, it can be shown that

$$\frac{df}{f} \approx \frac{d\alpha}{\alpha} - \frac{dL_2}{L_2}. \tag{7}$$

In other words, the fractional change in frequency equals the fractional change in area ratio minus the fractional change in length of the front tube, to first order. This is a theoretically derived sensitivity function that relates sensitivity to vocal tract geometry.

JW11, JW14, and JW18 all show large $F1$ sensitivity to $K1$ changes ($\nu_{11}^{\rightarrow}$) in a region around [i] (Fig. 18). The $K1$ of each of these speakers has the effect of moving the tongue toward or away from the palate (Fig. 2). With the tongue moving toward the palate it is plausible that both the effective α decreases rapidly and L_2 increases. According to Eq. (7) these factors would provide a rapid decrease in $F1$, which would explain the large $F1$ sensitivity near [i]. Near [ɑ], on the other hand, all three speakers show small sensitivity of $F1$ to $K1$. It is plausible that near [ɑ], as $K1$ moves the tongue away from the pharynx, α still decreases, as does L_2, thus creating conditions for a small relative change in $F1$ according to Eq. (7). The fact that [æ] is also in a region of small $\nu_{11}^{\rightarrow}$ for JW11 and JW18 suggests that near [æ], the $K1$ of these two speakers tends to balance the changes in α and L_2 in a way similar to the way it balances them near [ɑ].

It was noted in Sec. III G that JW11 and JW18 had similar regions of large and small $F1$ sensitivities to $K1$ [Figs. 18(a) and 18(d)]. This agrees with geometric similarities in their $K1$ components. For both talkers, $K1$ possesses a high-low degree-of-freedom that is proportionately greater than the front-back degree-of-freedom, and the two degrees-of-freedom are of the same polarity, in that "high" belongs with "front" (Fig. 3). The rotated shading pattern for JW14 noted in Sec. III G [Fig. 18(b)] may be due to the fact that her $K1$ has much more of a front-back degree-of-freedom compared with the high-low degree-of-freedom.

Considering the $\nu_{21}^{\rightarrow}$ sensitivity patterns in Fig. 19, JW11 and JW14 both show large $F1$ sensitivity to $K2$ changes ($\nu_{21}^{\rightarrow}$) in a region around [i] [Figs. 19(a) and 19(b)], and JW18 just fails to include [i] in the region of large sensitivity [Fig. 19(d)]. All the speakers include [ɑ] and [æ] in their region of small sensitivity. The same perturbation-theory pictures explaining the sensitivity of $F1$ to $K1$ also seem to apply to the sensitivity of $F1$ to $K2$. JW11 and JW14 share the property that high corresponds to back in their $K2$ [Figs. 3(a) and 3(b)]. On the other hand, JW18 possesses the opposite correspondence between the high-low and front-back degrees-of-freedom (i.e., high corresponds to front) [Fig. 3(d)], and this could be the reason for the clockwise rotation of his regions of sensitivity that excludes [i] from his region of large sensitivity.

The sensitivities of $F2$ to both $K1$ and $K2$ ($\nu_{21}^{\rightarrow}$ and $\nu_{22}^{\rightarrow}$) have more fragmented sensitivity regions as well as more proximate regions of small and large sensitivity (Figs. 20 and 21). Some of the observed trends in $F2$ sensitivity can be explained in terms of changes in constriction size. Changes in constriction size produce relatively large changes in formant frequencies, as can be seen, for example, in perturbation theory. The high-back (palato-velar) regions have large $F2$ sensitivity to both $K1$ and $K2$, with the exception of JW16's $F2$ sensitivity to $K2$ [Fig. 21(c)]. The explanation for this consistently large sensitivity is that both high-low and front-back degrees-of-freedom will change constriction size in this region, thus substantially affecting $F2$ (see also Stevens, 1998, pp. 366–367). Whether or not "front" is coincident with "high" does not seem to matter in the palato-velar region. On the other hand, in the region of [ɑ], the two degrees-of-freedom do not have the same effect on constric-

tion degree. Thus, the reason that $F2$ has large sensitivity to $K1$ in this region for JW14 [Fig. 20(b)] but small sensitivity for JW11 and JW18 [Figs. 20(a) and 20(d)] has to do with the different proportions of high-low and front-back degrees-of-freedom in the different speakers' $K1$s. JW14's $K1$ has a larger proportion of front-back than the other two speakers, and thus her $K1$ affects constriction degree more than the $K1$s of JW11 or JW18. In the region of [i], the reason that high-front vowels have small $F2$ sensitivity to $K2$ in the cases of JW11 and JW14 [Figs. 21(a) and 21(b)] may be due to the $K2$ components of these speakers possessing a relatively large front-back degree-of-freedom with a small high-low degree-of-freedom polarized so that high corresponds to back. This means that $K2$ has the tongue make a tighter constriction while increasing the back cavity length; these two actions have opposite effects on $F2$ and so largely cancel each other out. Further, with high corresponding to back the constriction degree changes at the palate are minimal, and they are dominated by changes in the place of constriction as the tongue moves along the hard palate.

Concerning the "critical" $(K1,K2)$ coordinates defined in Sec. III G, the implications for quantal theory are not conclusive. We see only inconsistent inclusion of projected point vowels at critical coordinates (Table X). However, there are two weaknesses to our analysis of critical points: (1) defining the critical points in $K1,K2$ space and projecting the point vowels to that space may be less informative than an analysis with the critical points and vowels in full $(K1,K2,K3,K4)$ coordinates and (2) the regular 20×20 grid may not be fine enough for a good classification of critical points. Further, quantal theory applied to vowels involves consideration of place-of-constriction, whereas the current analysis does not distinguish between place- and degree-of-constriction.

V. CONCLUSION

A method for obtaining mappings from measured articulatory coordinates to measured acoustic coordinates and the inverses of these mappings has been proposed and examined in this paper. The method involves PCA and a local linear regression procedure named loess. Four speakers from the XRMB-SPD were analyzed using these procedures. PCA was performed on both their pellet coordinate data and their formant frequency data, and these PCA components were used as independent variables in the various forward and inverse mappings constructed using loess. Loess models were made more computationally efficient by pre-computing weighted least squares at the data points and interpolating regression parameters between data points. The parameters of the loess models, regression bandwidth and interpolation bandwidth, were optimized by k-fold cross-validation. By example, it was seen that the forward and inverse mappings were reasonable when viewed as mappings from the two independent variables with the largest PCA variance and averaged over the remaining one or two independent variables. There were wide distributions in the local regression parameters at the data points. Further, there is some evidence of co-variation among these parameters. A discussion of the relation between test RMSE of the loess models and the standard deviation of the error distributions revealed that the RMSEs of the forward mappings are probably a good estimate of the error standard deviations for formant data, but that the RMSEs of the inverse mappings are not a good estimate of error standard deviations for the articulatory data. There may be two fundamental problems with the inverse mappings: (1) there are not enough acoustic parameters to construct a good mapping and (2) the assumption that the independent variables are known with relatively little error is not a good assumption. One important application of loess is in the study of the sensitivity of acoustic parameters to changes in articulatory parameters. It was possible to find regions of large and small sensitivity as a function of the two lowest-order articulatory components when the slopes and formant values are averaged over the remaining two articulatory components.

There are future directions for research into this particular method of constructing empirically determined mappings between articulation and acoustics. The acoustic data were found to be lacking in the number of degrees-of-freedom and the accuracy with which they could be measured. Further, these data require some time consuming hand editing. Work in speech technology may provide some clues on where to look for improved acoustic variables, such as line spectral densities (Qin and Carreira-Perpiñán, 2007). An avenue that would make the results more easily interpretable would be to use the method of arbitrary factor analysis proposed by Maeda (1990). The high-low and front-back dimensions could be factored out of the articulatory data before PCA is performed on the residual. Another significant improvement in this method would be to be able to provide quantitative confidence intervals for both the mappings and the parameters in the mappings. This would require development of more efficient computational methods and merits further research. Finally, an analysis of local co-linearity among independent variables should be performed, and co-linearity should be removed, perhaps by allowing a reduction in the number of independent variables locally.

ACKNOWLEDGMENTS

The authors would like to thank Rebekka Puderbaugh for her help in analyzing the acoustic data. Thanks to Professor Joyce McDonough for making sure that this research proceeded and for her encouragement to the authors. This work was supported by NIDCD-001247 to CReSS LLC.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract," J. Acoust. Soc. Am. **63**, 1535–1555.

Badin, P., Perrier, P., Boë, L.-J., and Abry, C. (**1990**). "Vocalic nomograms: Acoustic and articulatory considerations upon formant convergence," J. Acoust. Soc. Am. **87**, 1290–1300.

Beaudoin, R. E., and McGowan, R. S. (**2000**). "Principal component analysis of x-ray microbeam data for articulatory recovery," in *Proceedings of the Fifth Seminar on Speech Production* (Institute for Phonetics and Speech Communication, Munich, Germany), pp. 225–228.

Bickel, P. J., and Docksum, K. A. (**1977**). *Mathematical Statistics: Basic Ideas and Selected Topics* (Holden-Day, San Francisco, CA).

Boersma, P., and Weenik, D. (**2007**). PRAAT: Doing phonetics by computer (Version 4.5.16) (computer program) (Last vewed February, 2007), from http://www.praat.org.

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. (**2006**). "Synchronized

and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794.
Chatterjee, S., and Hadi, A. S. (**1988**). *Sensitivity Analysis in Linear Regression* (Wiley, New York).
Cleveland, W. S. (**1979**). "Robust locally weighted regression and smoothing scatter plots," J. Am. Stat. Assoc. **74**, 829–836.
Cleveland, W. S., and Devlin, S. J. (**1988**). "Locally weighted regression: An approach to regression analysis by local fitting," J. Am. Stat. Assoc. **83**, 596–610.
Cleveland, W. S., Devlin, S. J., and Grosse, E. (**1988**). "Regression by local fitting: Methods, properties, and computational algorithms," J. Econometr. **37**, 87–114.
Engwall, O., and Badin, P. (**1999**). "Collecting and analyzing two- and three dimensional MRI data for Swedish," TMH-QPSR Report No. 40, KTH, Stockholm, Sweden.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).
Hogden, J., Rubin, P., McDermott, E., Katagiri, S., and Goldstein, L. (**2007**). "Inverting mappings from smooth paths through R^n to paths through R^m: A technique applied to recovering articulatory information from acoustics," Speech Commun. **49**, 361–383.
Jackson, M. T.-T. (**1988**). "Analysis of tongue positions: Language-specific and cross-linguistic models," J. Acoust. Soc. Am. **84**, 124–143.
Jackson, M. T.-T., and McGowan, R. S. (**2008**). "Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels: Statistical considerations," J. Acoust. Soc. Am. **123**, 336–346.
Kiritani, S. (**1986**). "X-ray microbeam method for the measurement of articulatory dynamics: Techniques and results," Speech Commun. **5**, 119–140.
Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (**1978**). "Generating vocal tract shapes from formant frequencies," J. Acoust. Soc. Am. **64**, 1027–1035.
Lindblom, B. E. F., and Sundberg, J. E. F. (**1971**). "Acoustical consequences of lip, tongue, jaw, and larynx movement," J. Acoust. Soc. Am. **50**, 1166–1179.
McGowan, R. S. (**2006**). "Perception of synthetic vowel exemplars of four year-old children and estimation of their corresponding vocal tract shapes," J. Acoust. Soc. Am. **120**, 2850–2858.
McGowan, R. S., and Cushing, S. (**1999**). "Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis," J. Acoust. Soc. Am. **106**, 1090–1105.
Maeda, S. (**1982**). "A digital simulation method of vocal-tract system," Speech Commun. **1**, 199–229.
Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modeling*, edited by J. Hardcastle and A. Marchal (Kluwer Academic, Dordrecht), pp. 131–149.
Mermelstein, P. (**1967**). "Determination of the vocal-tract shape from measured formant frequencies," J. Acoust. Soc. Am. **41**, 1283–1294.
Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. **53**, 1070–1082.
Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. (**2007**). "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," J. Phonetics **35**, 20–39.
Perkell, J. S. (**1969**). *Physiology of Speech Production* (MIT Press, Cambridge, MA).
Perkell, J. S., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M. (**1992**). "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096.
Qin, C., and Carreira-Perpiñán, M. A. (**2007**). "A comparison of acoustic features of for articulatory inversion," in *Interspeech 2007*, pp. 2469–2472.
Schroeder, M. R. (**1967**). "Determination of the geometry of the human vocal tract by acoustic measurement," J. Acoust. Soc. Am. **41**, 1002–1010.
Shirai, K., and Honda, M. (**1976**). "An articulatory model and the estimation of articulatory parameters by nonlinear regression method," Electron. Commun. Jpn. **59**, 35–43.
Stevens, K. N. (**1972**). "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, edited by E. E. David, Jr. and P. B. Denes (McGraw-Hill, New York), pp. 51–66.
Stevens, K. N. (**1989**). "On the quantal nature of speech," J. Phonetics **17**, 3–45.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).
Story, B. H. (**2007**). "Time-dependence of vocal tract modes during production of vowels and vowel sequences," J. Acoust. Soc. Am. **121**, 3770–3789.
Story, B. H., and Titze, I. R. (**1998**). "Parameterization of vocal tract area functions by empirical orthogonal modes," J. Phonetics **26**, 223–260.
Westbury, J. R. (**1983**). "Enlargement of supraglottal cavity and its relation to stop consonant voicing," J. Acoust. Soc. Am. **73**, 1322–1336.
Westbury, J. R. (**1994**). *X-Ray Microbeam Speech Production Database User's Handbook* (University of Wisconsin, Madison, WI).
Whalen, D. H., Kang, A. M., Magen, H. S., Fulbright, R. K., and Gore, J. C. (**1999**). "Predicting midsagittal pharynx shape from tongue position during vowel production," J. Speech Lang. Hear. Res. **42**, 592–603.
Wood, S. (**1979**). "A radiological analysis of constriction locations for vowels," J. Phonetics **7**, 25–43.
Yehia, H., and Itakura, F. (**1996**). "A method to combine acoustic and morphological constraints in the speech production inverse problem," Speech Commun. **18**, 151–174.