Analyses of vocal tract cross-distance to area mapping: An investigation of a set of vowel images

Richard S. McGowan^{a)} and Michel T-T. Jackson
CReSS LLC, 1 Seaborn Place, Lexington, Massachusetts 02420

Michael A. Berger
Centre for Speech Technology Research, 10 Crichton Street, University of Edinburgh,
EH8 9AB, United Kingdom

(Received 22 December 2010; revised 2 November 2011; accepted 8 November 2011)

Traditional models of mappings from midsagittal cross-distances to cross-sectional areas use only local cross-distance information. These are not the optimal models on which to base the construction of a mapping between the two domains. This can be understood because phonemic identity can affect the relation between local cross-distance and cross-sectional area. However, phonemic identity is not an appropriate independent variable for the control of an articulatory synthesizer. Two alternative approaches for constructing cross-distance to area mappings that can be used for articulatory synthesis are presented. One is a vowel height-sensitive model and the other is a non-parametric model called loess. These depend on global cross-distance information and generally perform better than the traditional models. © 2012 Acoustical Society of America.
[DOI: 10.1121/1.3665988]

PACS number(s): 43.70.Aj, 43.70.Bk [DAB]

Pages: 424–434

I. INTRODUCTION

Imaging technologies have advanced enough so that it is now possible to measure midsagittal cross-distances simultaneously with cross-sectional areas (Baer *et al.*, 1991; Story *et al.*, 1996; Engwall and Badin, 1999; Ericsson, 2005). The cross-sectional area as a function of position along the vocal tract axis, the *area function*, is essential for any computational model of acoustic wave propagation in the vocal tract, and, hence, for articulatory synthesis (Mermelstein, 1973; Rubin *et al.*, 1981; Maeda, 1982, 1990).

One of the intermediate results for many articulatory synthesizers is a midsagittal outline, or a time series of such outlines. Given a midsagittal outline of the vocal tract, a two-step method is generally used to obtain cross-sectional areas. The first step is to derive the midsagittal cross-distances as a function of the axial position in the vocal tract. This function can be termed the *midsagittal cross-distance function*, or simply, *cross-distances*. To construct this function, a coordinate grid is overlaid onto the midsagittal outline of the vocal tract, so that each gridline can be associated with a unique position on the vocal tract axis (Fig. 1). The lengths of the grid segments within the midsagittal vocal tract are the cross-distances.¹ Then a mapping from the cross-distances to cross-sectional areas is applied to obtain the area function for each desired static midsagittal vocal tract shape.

Traditionally, the mapping from cross-distances to area is assumed to be a local function of the cross-distances: the cross-sectional area at a particular vocal tract axis location is a parametric function of the cross-distance at that location only (Mermelstein, 1973; Sundberg *et al.*, 1987; Perrier

et al., 1992). One of the traditional models has come to be called the α - β model (Heinz and Stevens, 1964; Mermelstein, 1973),

$$A = \alpha d^\beta, \quad (1)$$

where A is cross-sectional area, d is midsagittal cross-distance, or distance between the inner and outer walls of the vocal tract in the same cross-sectional plane, and α and β are parameters. The α and β parameters vary depending on vocal tract axis location.

Most previous studies can be considered to be specializations of the traditional α - β model. Using tomographic images of the pharynx, Sundberg *et al.* (1987) used regression to fit areas as functions of cross-distance squared, thus setting $\beta = 2$ and α is not a function of axial location. From the figures in that paper there can be errors as large as 100% or more made with such fits, even with a limited set of four vowel phonemes. In one of the first vocal tract magnetic resonance imaging (MRI) works, Baer *et al.* (1991) used linear regression to relate cross-distance to areas, as well as to relate the logarithms of these quantities in the pharyngeal region. Their results are reported in their Figs. 15 and 16, correspond to fits with $\beta = 1$ and β variable, respectively. The scatter about the regression lines indicates errors on the order of 100% occur, but they did not allow the parameters to change with position along the vocal tract axis for this analysis. Perrier *et al.* (1992) used a bi-parabolic model for the cross-sectional outlines of the vocal tract, which results in $\beta = 1.5$. The percent error represents standard deviations, so it is not possible to judge maximum error with their model.

An important observation was made by Maeda (1972), as cited in Perrier *et al.* (1992), namely that the mapping parameters from cross-distances to area depends on the phonemic identity of the vowel and tongue shape parameters. This

^{a)}Author to whom correspondence should be addressed. Electronic mail: rsmcgowan@cressllc.net

FIG. 1. (Color online) A midsagittal outline of a model vocal tract for an articulatory synthesizer with the overlaying grid.

finding was confirmed by Sundberg *et al.* (1987) for the pharynx [see especially their Fig. 4(b)], but adequate ways to calculate pharyngeal area functions were not offered. Perrier *et al.* also concluded that one needed to account of “... the position of the section vis-a-vis the constriction...” in order to calculate the cross-sectional area of that section (Perrier *et al.* 1992, p. 61).

Ericsson, in her Ph.D. dissertation (2005), presents a comprehensive set of MRI data on vocal tract cross-distances and cross-sectional areas in vowels produced by an adult male and an adult female, both speakers of Swedish. Further, the work shows that it is necessary to group or categorize articulations according to articulatory properties and/or phonemic identity in order to attain good accuracy in the cross-distance to area function mapping. With these grouping factors the accuracy of the mapping from cross-distances to area functions could be improved by a factor of up to 4 over the traditional α - β model of Heinz and Stevens. The classifications employed by Ericsson are summarized in the first part of chapter seven of her dissertation. The classifications of the vowel tokens from her two subjects in each image plane depend on, for example, the concavity of the tongue surface, and, sometimes phonemic identity. Also, tokens of the same phoneme can be included into different groupings, so that two different models of the cross-distance to area function mappings can apply to different tokens of the same phoneme. These classifications appear to be ad-hoc, and they are not the same for her two subjects.

On the other hand, Ericsson (2005) has come the furthest toward incorporating information other than local cross-distances for quantitative mapping to area functions to obtain more accurate quantitative mappings from cross-distances to area functions. She even makes a qualitative proposal to base the mappings purely on articulatory properties. This is an important step, because one of the uses of an articulatory synthesizer is to explore the acoustic consequences of articulatory configurations without reference to linguistic categories. Thus

there remains a need to find systematic, statistically valid techniques for mapping cross-distances to cross-sectional areas that do not depend on the individual speaker. These new techniques should possess accuracy gains over the traditional, but without using the classifications of Ericsson (2005).

In the present work we use more globally based formulations of the cross-distance to area mappings than the traditional models. Due to the use of global vocal tract information, some of the relations between phonemic identity and articulation found by Ericsson (2005) are implicitly quantified in a form that can be used to provide cross-distance to area mappings for an articulatory synthesizer. The way to generalize the traditional mapping models to use global information is to permit cross-sectional areas at individual vocal tract axis locations to depend quantitatively on geometric features of articulation other than the cross-distance at the same vocal tract axis location, such as the cross-distances at other vocal tract axis locations.

The reanalysis of Ericsson’s dissertation data starts with the traditional α - β models [Eq. (1)], used for comparison purposes here. We go on to confirm that vowel features, if not their phonemic identity, can be a statistically significant factor in the values for the area functions using a covariance analysis. Again, however, these vowel features will need to be related to geometrical properties of vowel articulation in order for them to be useful for articulatory synthesis. Another approach accounts for global midsagittal geometry using a statistical model. A non-parametric weighted regression, or *loess* (Cleveland, 1979; McGowan and Berger, 2009) is employed to construct a mapping from cross-distances to cross-sectional areas. Phonemic identity or feature information is included implicitly in such a regression, because global cross-distance information is included in the independent variables for *loess*.

II. CROSS-DISTANCE TO AREA MAPPINGS

The traditional models for mapping from midsagittal cross-distances to cross-sectional areas that use only local cross-distance information appear to miss information of a more global nature essential for constructing accurate mappings from cross-distances to area functions. In this section two alternative approaches for deriving the cross-distance to area function mapping are introduced and applied to the Ericsson (2005) data: Analysis of covariance to derive different α - β models that depend on vowel features and a non-parametric model called *loess*. These both use global cross-distance information, but with the added virtue of not referring to phonemic identity. Traditional α - β models will also be used for comparison purposes. First, though, a brief description of Ericsson’s data is presented.

A. Description of the Ericsson data

As described in Ericsson (2005), sagittal and coronal/axial (“multi-mode”) MRI images of two tokens each of 11 Swedish vowel allophones produced by one male and one female speaker were acquired using the Philips Medical Systems Gyroscan NT at the Unité de Résonance Magnétique, l’Hôpital Erasme, Brussels. These vowels were [i, y, ʊ, e, œ,

ε , $\text{\text{æ}}$, u , $\text{\text{ɒ}}$, a , o] for the male subject and $[i$, y , $\text{\text{ʉ}}$, e , $\text{\text{ø}}$, ε , $\text{\text{æ}}$, u , $\text{\text{ɒ}}$, a , o] for the female subject, with the vowel length diacritic suppressed. Fifteen multi-mode coronal/axial images were taken approximately perpendicular to the subject's vocal tract airway. Eleven sagittal images in planes approximately parallel to each subject's midsagittal plane were acquired in each token, and the image that was determined to be closest to the subject's actual midsagittal plane was analyzed.

For each subject, planes were designated, orthogonal to the midsagittal plane, and approximately orthogonal to the vocal tract axis. Planes 1–5 proceeded from the glottis up to the mid pharynx, and planes 6–10 from the top of the pharynx through the oro-pharyngeal region to the front of the soft palate or rear of the hard palate (Figs. 2 and 3). For the female subject, the planes intersecting the hard palate or incisors were in order from posterior to anterior 11, 12, 15, and 14 (Fig. 2). For the male these planes were 11, 14, and 13 (Fig. 3). (These planes are from posterior to anterior despite the fact that they are not numbered consecutively.) The region that included the lips was imaged in plane 13 for the female and 12 for the male. Midsagittal cross-distances were measured from the multi-mode images at each image plane for both tokens of each of the 11 vowels. The cross-sectional area of the vocal tract airway was also measured from the multi-mode images. The midsagittal image was used for comparison with x-ray images collected from the same subjects in earlier work and to provide global orientation of the multi-modal images.

B. Nonlinear and linear α – β models

At each image plane it is assumed that the cross-sectional area can be expressed as a power function of the midsagittal

FIG. 2. Sample midsagittal vocal tract outline and image plane position for the female speaker, traced from [Ericsson \(2005\)](#), Figs. 2–11. Note that the first image plane falls at or below the level of the glottis, and therefore is not considered here. Note also the numbering of image planes 1–12 from glottis to oral cavity, but 13–15 (as in the original) from lips back. The last image plane (i.e., image plane 13) falls beyond the end of the oral cavity for some vowels, and therefore also is not considered.

FIG. 3. Sample midsagittal vocal tract outline and image plane position for the male speaker, traced from [Ericsson \(2005\)](#), Figs. 2–11. Note that the first image plane falls at or below the level of the glottis, and therefore is not considered here. Note also the numbering of image planes 1–11 from glottis to oral cavity, but 12–14 (as in the original) from lips back. The last image plane (i.e., image plane 12) falls beyond the end of the oral cavity for some vowels, and therefore also is not considered here.

cross-distance at that image plane, as shown in Eq. (1). Two methods are used to fit the α – β model to the Ericsson data here. One method uses non-linear regression with d as the independent variable and A as the dependent variable. This is termed the non-linear α – β model here. The other method uses linear regression with $\log(d)$ as the independent variable and $\log(A)$ as the dependent variable. This is termed the linear α – β model here, and is the one employed by [Heinz and Stevens \(1964\)](#). The mappings from cross-distances to area that result from such analyses are traditional because a cross-sectional area at an axial location depends only on the cross-distance at the same axial location.

The results of the reanalysis of the data in each image plane with the traditional nonlinear and linear α – β models are shown in Fig. 4. The scatter plots for both α and β as a function of normalized distance from the glottis are fit with fourth degree polynomials. The polynomial fits show a consistent shape for the non-linear and linear models and between the male and female [Figs. 4(a) and 4(b)]. The same consistency holds for the female subject's β , but not for the male's β [Figs. 4(c) and 4(d)].

C. Analysis of covariance and the height-sensitive α – β model

An intermediate step is sought for including phoneme identity information in the mappings from cross-distances to cross-area in an implicit manner. Phonological features are explored as independent variables in estimating the mapping from cross-sectional distances to cross-sectional areas. The phonological features used to describe the vowels are

FIG. 4. Scatter plots of the α and β parameters as a function of normalized distance from the glottis in the linear ($\circ$) and nonlinear ($*$) α - β reanalyses of the male and female subjects in [Ericsson \(2005\)](#). The lines are fourth degree polynomial fits to the data for the linear (—) and nonlinear (---) models. (a) Coefficient α for female subject, (b) coefficient α for male subject, (c) power β for female subject, and (d) power β for male subject.

backness (categorized as front or back), height (high [i, y, u], mid [e, æ, ε, o] or [e, ø, ε, o], or low [æ, a, ɒ]), and rounding (rounded or unrounded). It should be noted that rounding in Swedish vowels is not redundant with backness as there are both unrounded and rounded front vowels /i/ versus /y/, as well as the rounded back vowels /u, o, ɒ/. Exploratory analysis of covariance (ANCOVA) is carried out using the SPSS Generalized Linear Models procedure as a method of screening the independent variables in order to determine which phonological features are likely to contribute to an area function model. This analysis is not directly applicable to the generation of area functions for articulatory synthesis without further geometric information. However, it does provide insight into how vowel identity affects the cross-distance to area mappings, as well as a statistical model that can be compared quantitatively with other models that are appropriate for articulatory synthesis.

For each image plane within the pharyngeal and oral cavities, ANCOVA is performed on the data with the cross-sectional area treated as the dependent variable; the independent variables are the midsagittal cross-distance (the covariate), the categorical factors backness, height, and rounding, and the interactions cross-distance by backness, cross-distance by height, and cross-distance by rounding. There are 22 observations (2 tokens of each of 11 vowels) for each ANCOVA. For each speaker, 13 ANCOVAs are carried out, as 13 image planes fell entirely within the pharyngeal and oral cavities of the speaker. For the female speaker, image planes 2–12, 15, and 14 are analyzed (Fig. 2); for the male, planes 1–11, 14, and 13 (Fig. 3). Some images from the other planes either fell

below the pharynx, through the laryngeal structures, or intersected the lip opening, beyond the oral cavity.

The statistical significance of each of the independent variables is used to select which independent variables to retain in a final model. Since the vowel inventory is not balanced (e.g., there are more high vowels than low vowels), type III sums of squares are used. Type III sums of squares give the best unbiased estimates of the marginal means ([Hill and Lewicki, 2006](#); see also [Cohen et al., 2003](#)); e.g., the mean cross-sectional area in a particular image plane of high vowels, averaged across all combinations of backness and rounding. The Wald χ^2 test is used to determine the significance level of each independent variable's contribution to the ANCOVA. The Wald test is a test of the significance of a regression slope, B . When used as a test versus the null hypothesis that $B = 0$, the ratio of the square of the regression coefficient to the square of the regression coefficient's standard error has a χ^2 distribution ([Cohen et al., 2003](#)).

Because there are seven independent variables (cross-distance, backness, height, rounding, and the interactions cross-distance by backness, cross-distance by height, and cross-distance by rounding), and 13 image planes for each speaker, there are a total of 91 χ^2 tests for each speaker. Since this is comparable with the size of the data set (13 planes $\times$ 22 vowel tokens = 286 observations), a moderately stringent significance level of $p < 0.01$ seemed appropriate in order to reduce the likelihood of including an independent variable that only achieved significance by chance in the final model—a type I error.

The results of the 13 ANCOVAs for each speaker are summarized in Table I. For each of the independent variables,

TABLE I. Summary results of ANCOVAs.

Speaker	Independent variable	$p < 0.01$ in image plane													Number of image planes with $p < 0.01$
		2	3	4	5	6	7	8	9	10	11	12	15	14	
Female	Cross-distance	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	12
	Backness						✓	✓							2
	Height								✓		✓				2
	Rounding							✓	✓						2
	Cross-distance: Backness														0
	Cross-distance: Height								✓	✓	✓	✓	✓		5
	Cross-distance: Rounding							✓	✓						2
Speaker	Independent variable	$p < 0.01$ in image plane													Number of image planes with $p < 0.01$
		1	2	3	4	5	6	7	8	9	10	11	14	13	
Male	Cross-distance			✓	✓		✓		✓	✓	✓	✓	✓		8
	Backness									✓	✓	✓			3
	Height														0
	Rounding									✓					1
	Cross-distance: Backness								✓	✓	✓				3
	Cross-distance: Height						✓	✓	✓		✓	✓			5
	Cross-distance: Rounding									✓			✓	✓	3

the planes in which the given independent variable attained a significance level of at least $p < 0.01$ are indicated. These results suggest that the midsagittal cross-distance and the cross-distance by height interaction are the most important effects in a multiple regression model for estimating the cross-sectional area measured in each image plane. The cross-distance is significant at the $p < 0.01$ level in 12 of the 13 planes and the cross-distance by height interaction is significant in 5 of the 13 planes for the female speaker. For the male speaker, the cross-distance is significant at the $p < 0.01$ level in 8 of the 13 planes, and the cross-distance by height interaction is also significant in 5 of the 13 planes. For both speakers, the cross-distance is significantly related to the cross-sectional area through most of the vocal tract. However, the cross-distance by height interaction is only significant in the anterior portion of the vocal tract. None of the other independent variables is significant at the $p < 0.01$ level for more than 3 of the 13 image planes. Therefore, the midsagittal cross-distance and the cross-distance by height interaction are retained in the final model, here known as the *height-sensitive α - β model*.

The height-sensitive α - β model consists of separate linear α - β models for each of the three vowel height categories. The model accounts for the variation of cross-sectional area with a different mapping for different vowel height groups, which is represented by different values of α and β for each of the three vowel height groups. The values of α and β estimated for each group of vowels are plotted as a function of the normalized distance along the vocal tract from the glottis in Fig. 5. Figure 5 also shows the fourth-order polynomial fits.

In Figs. 5(a) and 5(b), it can be seen that most of the values of α are in the range of 1.0–2.0, if the low pharyngeal region is excluded. However, the polynomial fits show relatively large values of α in the anterior of the vocal tract for the male [Fig. 5(b)]. These larger values of α are found mainly among the low

vowels, and to a lesser extent, among the mid vowels. The variation in α according to vowel height in the height-sensitive regression model accounts for the significant cross-distance by height effect found in the ANCOVAs summarized in Table I. The increase in the values of α toward the anterior of the male vocal tract seen in the height-sensitive models for low vowels [Fig. 5(b)] is largely responsible for the larger values of α in the nonlinear α - β model in the male subject [Fig. 4(b)].

On the other hand, in Figs. 5(c) and 5(d), it can be seen that the values of β do not vary strongly over most of the length of the vocal tract, with most estimated values of β in the range of 1.0–1.5, again, excluding the low pharyngeal region. This range is consistent with the values of β estimated by both linear and non-linear regression as described earlier [Figs. 4(c) and 4(d)].

D. Locally weighted regression, loess

Locally weighted regression, or loess, is a method used to estimate a non-parametric mapping from independent variables to a dependent variable (Cleveland, 1979; Cleveland and Devlin, 1988; Cleveland *et al.*, 1988; McGowan and Berger, 2009). Loess assumes that the mapping is a smooth function of the independent variables, and that, locally, the surface representing the mapping can be approximated by a low order polynomial of the independent variables. Here, the local approximation is linear and the linear regression parameters, a slope for each independent variable and an intercept, are estimated at each point in the space of the independent variables using nearby data points. The nearby data points are weighted by a function that diminishes monotonically with distance from the data point for which the mapping is being approximated.

Prior to the loess analysis, the data must be uniformly resampled. For each vowel token of a speaker, a vector of cross-distances (independent variables) and a corresponding

FIG. 5. Scatter plots of the α and β parameters as a function of normalized distance from the glottis in the height-sensitive α - β reanalysis of the male and female subject in [Ericsson \(2005\)](#): High (*), mid (□), and low (○). The lines are fourth degree polynomial fits to the data for the high (—), mid (---), and low (···). (a) Coefficient α for female subject, (b) coefficient α for male subject, (c) power β for female subject, and (d) power β for male subject.

vector of cross-sectional areas (dependent variables) are obtained from the 13 MRI image planes for which the analyses described above are performed. However, the image planes are not in the same location along the vocal tract from token to token. To obtain uniform axial positions, cross-distances and cross-sectional areas from the image planes are interpolated and uniformly resampled at a spacing of 1% of the normalized vocal tract length. This results in 101 vocal tract positions per vowel token, z_n , $1 < n < N = 101$, with a corresponding vector of values for cross-distances $\mathbf{d} = (d_1, d_2, \dots, d_N)$, where d_n is the cross-distance at axial location z_n . A vector for cross-sectional areas can be written $\mathbf{A} = (A_1, A_2, \dots, A_N)$, where A_n is the cross-sectional area at axial location z_n .

Our particular implementation of loess can be described symbolically as follows. For each location on the vocal tract axis, z_n , the cross-sectional area, A_n , is assumed to be a smooth function of the variable vector $\mathbf{d}$,

$$A_n = A_n(\mathbf{d}), \quad (2)$$

where derivatives of A_n with respect to the components of $\mathbf{d}$ are assumed to exist and to be continuous. With these assumptions, A_n can be approximated at $\mathbf{d}'$ near any physically realizable $\mathbf{d}$ by a linear function, say $\ell_{n,\mathbf{d}}(\mathbf{d}')$, so that $\ell_{n,\mathbf{d}}(\mathbf{d}')$ is as close to $A_n(\mathbf{d}')$ as desired so long as $\mathbf{d}'$ is sufficiently close to $\mathbf{d}$. Thus the following approximation can be written:

$$A_n(\mathbf{d}') \approx \ell_{n,\mathbf{d}}(\mathbf{d}') \quad (3)$$

for $\mathbf{d}'$ sufficiently close to $\mathbf{d}$. $\ell_{n,\mathbf{d}}(\mathbf{d}')$ can be written,

$$\ell_{n,\mathbf{d}}(\mathbf{d}') = \mathbf{d}' \cdot \beta_n(\mathbf{d}) + \gamma_n(\mathbf{d}), \quad (4)$$

where $\beta_n(\mathbf{d})$ is a vector of slopes and $\gamma_n(\mathbf{d})$ is the intercept.

Let $S_n = \{(\mathbf{d}^k, A_n^k); k = 1 \dots M\}$ be the set of data points pertaining to axial location z_n . The slopes and intercepts of the local linear function in Eq. (4) are estimated by weighted multi-linear regression using data points $(\mathbf{d}^k, A_n^k)$ near $\mathbf{d}$, which is not necessarily a data point. The weight function is a decreasing function of the distance from $\mathbf{d}$, specifically the bi-cubic function used by [Cleveland \(1979\)](#). This function has a parameter called the regression bandwidth, b_R , that expresses the neighborhood size as a proportion of the data from 0 for no data to 1 for the whole data set S_n . Let q be the closest integer to $b_R M$, and let $r(\mathbf{d}, b_R)$ denote the distance from $\mathbf{d}$ to the q th closest data point. (Distance refers to Euclidean distance in the $\mathbf{d}$ coordinate space, which is denoted with double vertical bars, $\|\cdot\|$.) The weight assigned to data point with independent variable value $\mathbf{d}^k$ for the locally weighted regression at $\mathbf{d}$ is

$$w(\mathbf{d}^k | \mathbf{d}, b_R) = W\left(\frac{\|\mathbf{d} - \mathbf{d}^k\|}{r(\mathbf{d}, b_R)}\right), \quad (5)$$

where $W(u) = (1 - u^3)^3$ for $0 \leq u \leq 1$ and $W(u) = 0$ otherwise.

Weighted least squares estimation must be performed for every value of $\mathbf{d}$ at which we want to know A_n . For computationally efficient loess, local linear regression is

pre-computed at each of the data points $(\mathbf{d}^k, A_n^k)$, with resulting slopes and intercepts $\bar{\beta}_n(\mathbf{d}^k)$ and $\bar{\gamma}_n(\mathbf{d}^k)$. Neighborhoods and weighting functions of the same form as those used for regression are used to interpolate the pre-computed slope and intercept values, $\bar{\beta}_n$ and $\bar{\gamma}_n$. These neighborhoods are parameterized by an interpolation bandwidth b_I . The final regression slopes and intercepts associated with a vector of cross-distances $\mathbf{d}$ are written,

$$\begin{aligned}\beta_n(\mathbf{d}) &= \frac{\sum_{k=1}^M w(\mathbf{d}^k|\mathbf{d}, b_I) \bar{\beta}_n(\mathbf{d}^k)}{\sum_{k=1}^M w(\mathbf{d}^k|\mathbf{d}, b_I)}, \\ \gamma_n(\mathbf{d}) &= \frac{\sum_{k=1}^M w(\mathbf{d}^k|\mathbf{d}, b_I) \bar{\gamma}_n(\mathbf{d}^k)}{\sum_{k=1}^M w(\mathbf{d}^k|\mathbf{d}, b_I)}.\end{aligned}\quad (6)$$

Because the data set here is so small, the constraint $b_I = b_R$ is employed. The joint bandwidth, called the loess bandwidth, is determined by optimization in a split-sample validation procedure. For each of the speakers there are two tokens for each vowel, and the first set of vowel tokens is used to train the loess model several times over a range of loess bandwidths. The second set of vowel tokens is used to evaluate each of the trained models by comparing the observed cross-sectional areas with those predicted by the models as a function of loess bandwidth. The model with the best test results is selected.

Prior to applying loess, the dimensionality of the cross-distance function is reduced using principal components analysis (PCA) over the 22 vowel productions for each subject. Then, for each vocal tract axis position, a loess regression is computed using the lower-order principal components as the independent variables, and the cross-sectional area at that

position as the dependent variable. Note that the cross-sectional area at a vocal tract location is not presumed to depend only on the midsagittal cross-distance at that location, but on the cross-distances throughout the vocal tract. This is an important difference with the traditional α - β mapping. The fact that the independent variables are cross-distances throughout the vocal tract means that phoneme identity information used by [Ericsson \(2005\)](#) or in the height-sensitive α - β model described in Sec. II C is included implicitly.

The first two eigenvectors of the PCA of the resampled cross-distance functions, PC1 and PC2, are found to account for more than 90% of the variance in the cross-distance measurements, with a root-mean-square (rms) error of 0.15 cm for the female speaker and 0.19 cm for the male speaker. The third and higher principal components make negligible contributions to the variance in the cross-distance measurements and are therefore excluded. Figure 6 shows a plot of the mean cross-distance function over the 22 vowel tokens together with the first two principal component eigenvectors scaled by their corresponding principal values.

The rms errors on the validation set for the female speaker in the split-sample validation procedure for obtaining loess bandwidth are plotted in Fig. 7. Figure 7 shows that at the low end of the range of bandwidths tested, below about 0.4, the loess models derived from the training set fit the validation training set poorly. This is interpreted as over-fitting idiosyncratic token variation in the training set. As the loess bandwidth increases, the rms errors decrease, with the rms error of the model fit to the validation set nearly plateauing at larger values. However, as the loess bandwidth increases, the set of tokens included in each neighborhood increases, loess regression becomes more similar to a global linear regression, and physically impossible negative predicted area values begin to occur. Therefore, an optimum loess bandwidth is found by finding the largest value that yielded no negative

FIG. 6. The mean cross-distances and the first two principal components, PC1 and PC2, resulting from a principal components analysis of resampled cross-distance functions for the (a) female and (b) male in the [Ericsson \(2005\)](#) data. The principal components have been scaled according to their corresponding singular values.

FIG. 7. An example of rms error as a function of loess bandwidth and image plane. This is for the female in the [Ericsson \(2005\)](#) data.

areas over the observed vocal tract cross-distance values. For the female speaker, the optimum value of the loess bandwidth is 0.66; for the male, 0.43.

The final loess regression parameters for each speaker are estimated over the speaker's entire dataset (all 22 vowel tokens) using the loess bandwidth determined by the constrained split-sample validation procedure. While different loess models are constructed for each image plane, the same loess bandwidth, b_R , is the same for all of them.

In some image planes, the results of loess regression are not appreciably different from that which would result from global regression. On the other hand, some planes show substantial differences between the loess and global regression. Figure 8 shows the results of loess regression at image plane 5, in the mid-pharyngeal region, for both speakers. The results in this region show substantial non-linearity in the map from the cross-distance principal components to area. It can be seen that the predicted areas are greatest toward the left side of the PC1 axis. However, for both speakers, the loess regression surface is somewhat cupped near the center of the PC1–PC2 plane. Furthermore, it is evident that the region to the right along the PC1 axis is less sloped than to the left; the trend of increasing area with decreasing PC1 is not uniform, and capturing this non-linearity results, for instance, in a better fit for the tokens of /u/ at large negative values of PC1.

E. Comparing the reanalyses of the Ericsson data

The reanalyses of the Ericsson data are compared in terms of rms error in each of the image planes in Fig. 9. For both the male and the female and in each image plane the smallest rms error is attained with either loess or the height-sensitive α – β model. Overall, the height-sensitive α – β model performs better on the female subject than did loess, and the reverse is true for the male subject. The linear α – β model has rms errors greater than or equal to the non-linear α – β model. There is a tendency for these two models to produce larger rms error toward the front of the vocal tract than in the rear of the vocal tract.

III. DISCUSSION

Results of the linear and non-linear α – β models illustrated in Fig. 4 can be compared with previous studies. The

(a)

(b)

FIG. 8. Examples of the loess regression surface at image plane 5 (upper-pharynx) for the (a) male and the (b) female from the [Ericsson \(2005\)](#) data.

results of [Baer et al. \(1991, Figs. 20 and 23\)](#) exhibit substantial differences between individual subjects in the magnitudes and trends of α and β . Our data also exhibit substantial differences in the subjects' α values. The [Baer et al. \(1991\)](#) β values are near or greater than two, except in the lower pharynx. In this study the power β rarely attained a value greater than two. The discrepancy might be explained by a trading relation between α and β in either study or the language difference (English versus Swedish), as well as individual variation. Further, there was no attempt to keep the image planes at right angles to the vocal tract centerline in the mouth in [Baer et al. \(1991\)](#), while there was some compensation for differences in the axial orientation between the male and female subject in [Ericsson \(2005\)](#).

[Sundberg et al. \(1987\)](#) derived a power law for the region along the hard palate for three female and three male subjects. According to their Table IV, α ranges from 1.6 to 2.6 and β from 1.33 to 1.48. These ranges appear to agree well with those shown in Fig. 4 from about 0.6 to about 0.8

FIG. 9. Comparison of various methods used to calculate the mapping from cross-distance to cross-sectional area in the [Ericsson \(2005\)](#) data in terms of rms error in area estimation in each of the image planes, (a) female subject and (b) male subject. The planes are ordered by distance from the glottis.

of normalized vocal tract length. The bi-parabolic model of [Perrier *et al.* \(1992\)](#) provides a fixed β of 1.5. Our results indicate that this is a reasonable model in many places along the vocal tract.

The height-sensitive α - β model results for the α of the male and female shown in Figs. 5(a) and 5(b) compare favorably with Ericsson's results for the female ([Ericsson, 2005](#), Figs. 7-5) for those of her vowel groups that can be associated completely with either low, mid, or high vowels. In particular, there is a maximum for α in the pharynx and α grows substantially in the front of the vocal tract for low vowels for one subject in each study. Also the magnitudes of β that are derived in the height-sensitive analysis agree with those derived by Ericsson, along with the increasing trend in β with axial position [Figs. 5(c) and 5(d)]. Sometimes β possesses a dip in the oro-pharyngeal region ([Ericsson, 2005](#), Figs. 7-5 and 7-6).

It would seem best to employ more sophisticated models of cross-distance to area function mappings than the traditional α - β models. Either, or both, the loess and the height-sensitive α - β model produce lower rms errors than the traditional α - β models, particularly in the mouth region of the vocal tract (Fig. 9). Both the loess and height-sensitive

α - β model use global knowledge of the independent variables, the cross-distances as a function of distance from the glottis, to determine cross-sectional areas. The former uses principal components of cross-distance functions and the latter a vowel feature. The loess and vowel height-sensitive α - β models are hindered here by the size of the data set on which they are based. For instance, there are only six tokens of low vowels for each speaker, and only 22 vowel tokens overall for each speaker. Only half of these are used to validate loess bandwidth, and validation uses rms error over all the image planes simultaneously. The loess analysis shows larger rms error than the linear α - β model in the lower part of the female vocal tract [Fig. 9(a)] because the optimum choice of the female loess bandwidth is based on only 11 tokens in the test set and rms error over all the image planes simultaneously. With more data we could justify choosing different loess bandwidths at each image plane. Similarly, it is probably the lack of data that makes the rms error of the height-sensitive α - β model large in planes 9–11 of the male speaker [Fig. 9(b)].

The fact that the ANCOVA shows vowel height to be a significant factor, and that one, or both, loess and height-sensitive α - β model exhibit smaller rms error for the mouth

region of the vocal tract can be understood with a brief consideration of the anatomy of the tongue in relation to hard structures in the mouth. As the vowel raises, the tongue starts to contact the mesial surface of the teeth and the palate, first with the front of the tongue body and rear of the tongue blade coming in contact with the molars. For the high front vowels, much of the tongue blade is in contact with the lateral portions of the palate and/or the mesial surface of the teeth. For image planes in the mouth region, this will appear as a discontinuity in the cross-sectional area as the tongue moves from low to mid to high as air space between the teeth and inner surface of the cheeks is excluded. Ericsson (2005) makes reference to this phenomenon in chapter five of her dissertation. Thus it is not surprising that there is a significant vowel height feature in mappings from cross-distance to cross-sectional areas that is captured with loess as well.

How should vowel height effects be incorporated into an articulatory synthesizer? The height-sensitive α - β model as it is implemented here is inappropriate for articulatory synthesis, because a categorical feature serves as an independent variable. One way to overcome this is with physical tongue-to-teeth distance measures. This would involve estimating aspects of dental structures, which would best be done by including a scan of the inferior surfaces of the upper teeth in flesh point measurements. With knowledge of the teeth, separate α - β models could be constructed, depending on whether the tongue is below or above the inferior surface of the upper teeth in the cross-sectional plane. In a similar way, Ericsson (2005) in chapter seven proposed a set of midsagittal geometrical features that do not depend on phonological tags at the end of chapter seven of her dissertation. These were qualitative suggestions that captured aspects of cross-sectional tongue shape that were observed to help group vowel tokens for area prediction, such as the concavity of the tongue cross section. While these geometrical features capture more than vowel height, they were not evaluated directly for efficacy in predicting cross-sectional areas.

Without knowledge of dental structure, a non-parametric model such as loess can be implemented. In fact, such a non-parametric approach may be preferable because of other, unquantified vowel-dependent features of a speaker's vocal tract. An example is provided by the advanced tongue root feature in Akan, where this feature is accompanied by expansion of the pharyngeal width (Tiede, 1996). Using a non-parametric approach would require more vowel tokens to attain robust estimates of the mapping between cross-distance and cross-sectional areas. On the other hand, all the models could be improved with more data.

IV. CONCLUSION

The traditional models for mapping from midsagittal cross-distances to cross-sectional areas that use only local cross-distance information are not optimal models. We have presented two alternative approaches—the height-sensitive α - β model and a non-parametric model called loess, that generally performed better than the traditional linear and

nonlinear α - β models. Information about dentition is necessary to implement the height-sensitive α - β model, if phonemic information cannot be used. If dental information is available, then this model is appropriate for articulatory synthesis, while the loess approach can be implemented without such information. However, implementation of loess in this work suffered because of a paucity of vowel tokens, as did the other models.

Speech data collection with MRI involves much skill on the part of experimentalists, including the measurement of dimensions once the images have been obtained. Measurements of cross-distances and cross-sectional areas are best done in planes that are orthogonal to the vocal tract axis, which changes as articulation changes. At least approximate corrections when the planes are not orthogonal to the axis should be applied, and even curvature of the vocal tract itself taken into account (i.e., Milenkovic *et al.*, 2010). As more data are obtained a much better understanding of the individual variation in the relation between cross-distance and cross-section area of the vocal tract will be possible.

ACKNOWLEDGMENTS

This work was supported by Grant No. NIDCD-001247 to CRESS LLC.

¹These distances should be adjusted by the cosine of the average angle between the gridline and the normals to the inner and outer outline of the midsagittal vocal tract outline.

- Baer, T., Gore, G. C., Gracco, L. C., and Nye, P. W. (1991). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging," *J. Acoust. Soc. Am.* **90**, 799–828.
- Cleveland, W. S. (1979). "Robust locally weighted regression and smoothing 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, computational algorithms," *J. Econometr.* **37**, 87–114.
- Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. (Erlbaum, Mahwah, NJ), pp. 362–363, 507–508.
- Engwall, O., and Badin, P. (1999). "Collecting and analysing two- and three dimensional MRI data for Swedish," *Quarterly Progress and Status Report KTH, TMH-QPSR*, Vol. 40, pp. 11–38.
- Ericsson, C. (2005). "Articulatory-acoustic relationships in Swedish vowel sounds," Ph.D. dissertation, Stockholm University, Sweden, available at <http://www2.ling.su.se/staff/ericsson/thesis/index.html>
- Heinz, J. M., and Stevens, K. N. (1964). "On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech," *Quarterly Progress Report, Research Laboratory of Electronics (Massachusetts Institute of Technology, Cambridge)*, Vol. 74, p. 192.
- Hill, T., and Lewicki, P. (2006). *Statistics: Methods and Applications* (StatSoft, Tulsa, OK), p. 268.
- Maeda, S. (1972). *On the Conversion of Vocal Tract X-Ray Data Into Formant Frequencies* (Bell Laboratories, Murray Hill, NJ).
- Maeda, S. (1982). "A digital simulation of the vocal-tract system," *Speech Commun.* **1**, 199–229.
- Maeda, S. (1990). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shaping using an articulatory model," in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, The Netherlands), pp. 131–149.
- McGowan, R. S., and Berger, M. A. (2009). "Acoustic-articulatory mapping in vowels by locally weighted regression," *J. Acoust. Soc. Am.* **126**, 2011–2032.

- Mermelstein, P. (1973). "An articulatory model for the study of speech production," *J. Acoust. Soc. Am.* **53**, 1070–1082.
- Milenkovic, P. H., Yaddanapudi, S., Vorperian, H., and Kent, R. D. (2010). "Effects of a curved vocal tract with grid-generated tongue profile on low-order formants," *J. Acoust. Soc. Am.* **127**, 1002–1013.
- Perrier, P., Boë, L.-J., and Sock, R. (1992). "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients," *J. Speech Hear. Res.* **35**, 53–87.
- Rubin, P., Baer, T., and Mermelstein, P. (1981). "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Am.* **70**, 321–328.
- Story, B. H., Titze, I. R., and Hoffman, E. A. (1996). "Vocal tract area function from magnetic imaging," *J. Acoust. Soc. Am.* **100**, 537–554.
- Sundberg, J., Johansson, H., Wilbrand, H., and Ytterbergh, C. (1987). "From sagittal distance to area," *Phonetica* **44**, 76–90.
- Tiede, M. K. (1996). "An MRI-based study of pharyngeal volume contrast in Akan and English," *J. Phonetics* **24**, 399–421.