A parametric model of the vocal tract area function for vowel and consonant simulation[a)]

Brad H. Story[b)]
Speech Acoustics Laboratory, Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

(Received 1 May 2004; revised 20 January 2005; accepted 21 January 2005)

A model of the vocal-tract area function is described that consists of four tiers. The first tier is a vowel substrate defined by a system of spatial eigenmodes and a neutral area function determined from MRI-based vocal-tract data. The input parameters to the first tier are coefficient values that, when multiplied by the appropriate eigenmode and added to the neutral area function, construct a desired vowel. The second tier consists of a consonant shaping function defined along the length of the vocal tract that can be used to modify the vowel substrate such that a constriction is formed. Input parameters consist of the location, area, and range of the constriction. Location and area roughly correspond to the standard phonetic specifications of place and degree of constriction, whereas the range defines the amount of vocal-tract length over which the constriction will influence the tract shape. The third tier allows length modifications for articulatory maneuvers such as lip rounding/spreading and larynx lowering/raising. Finally, the fourth tier provides control of the level of acoustic coupling of the vocal tract to the nasal tract. All parameters can be specified either as static or time varying, which allows for multiple levels of coarticulation or coproduction. [DOI: 10.1121/1.1869752]

PACS numbers: 43.70.−h, 43.70.Bk, 43.71.Es [AL] Pages: 3231–3254

I. INTRODUCTION

During speech production, coordinated movements of the tongue, jaw, lips, and to some degree the larynx, continuously alter the shape of the vocal tract (i.e., pharynx and oral cavity). Movement of the soft palate varies the acoustic coupling of the vocal tract to the nasal passages, and also may slightly change the shape of the upper pharynx. Integrated actions of individual articulators facilitate the creation of time-varying acoustic resonances that transform the sound generated by vocal-fold vibration or turbulence, into the stream of vowels and consonants that comprises speech. Specifically, it is the articulators' collective effect on the variation in cross-sectional area along the length of the vocal tract (i.e., the *area function*) and coupling to the nasal tract, as well as other possible sidebranch cavities, that is most closely related to the pattern of acoustic characteristics expressed in the speech waveform.

Hence, a simplified view of speech production may consist of a tubular system whose cross-sectional area variation, as a function of time, emulates that of a real vocal tract. This view forms the basis of a certain class of speech production models that operate on a parametric representation of the vocal-tract area function, and allow for calculation of corresponding acoustic characteristics. Area function models contrast with ''articulatory'' models in which the positions of individual articulators or some form of vocal-tract shaping components are represented in the midsagittal plane (e.g., Lindblom and Sundberg, 1971; Coker, 1976; Mermelstein, 1973; Maeda, 1990; Dang and Honda, 2004). These are intuitively appealing because of the physiological correlation between model parameters and human articulatory structures, and their ability to replicate observed articulatory movement. As a result, articulatory-type models are well suited for investigating and establishing speech motor control strategies. The relation of the articulatory parameters to the acoustic characteristics is, however, typically mediated by an empirically based conversion of midsagittal cross dimensions to the area function. Hence, control of the detailed vocal-tract shape at the level of cross-sectional area is less direct than with an area function model. A possible exception is a recently developed model that utilizes midsagittally based control parameters, but avoids the cross-dimension transformation to area by generating vocal-tract shapes based on three-dimensional data obtained from MRI (Badin *et al.*, 1998; Badin *et al.*, 2002).

Though admittedly an abstraction of articulatory reality, the area function is the representation that does provide the most direct theoretical connection between vocal-tract shape and resulting acoustic characteristics. A parametric model of the area function is useful for situations in which precise control of the detailed structure of the vocal-tract shape is desired. For example, such a model may have applications in studying source–tract interactions (Ishizaka and Flanagan, 1972; Titze and Story, 1997), investigating relations between vocal-tract structure and acoustic characteristics that are relevant to phonetic categories (e.g., Fant, 1960; Stevens, 1989) and voice quality (Story, Titze, and Hoffman, 2001; Story and Titze, 2002; Story, 2004), and understanding tract length scaling effects (e.g., Nordström, 1977; Goldstein, 1980; Fitch and Giedd, 1999). In addition, an area function model can be an essential component in synthesizing high-quality speech

[a)]A preliminary version of this paper was presented at the 146th Meeting of the Acoustical Society of America.
[b)]Electronic mail: bstory@u.arizona.edu

for presentation to listeners in perceptual tests when vocal-tract variables are the quantities to be manipulated rather than acoustic characteristics. There are also possible technological applications of synthesis based on area functions (Shadle and Damper, 2001; Sondhi, 2002).

The most straightforward form of an area function model (but perhaps the most inefficient) consists of a direct specification of the cross-sectional areas extending from the glottis to the lips. The parameters in this case are simply the areas themselves. Variation over time requires interpolation from one complete area function (representing one phonetic element) to another. In this approach, area data obtained from imaging studies (e.g., Fant, 1960; Narayanan, Alwan, and Haker, 1995; Story, Titze, and Hoffman, 1996; Baer *et al.*, 1991) can be used directly, but the ability to create realistic time-varying vocal-tract shapes (i.e., area functions that did not exist in the original data set) is limited.

Specification of the area function with a small set of physiologically relevant parameters forms the basis for more parsimonious models. Examples are the well-known ''three-parameter'' models (Fant, 1960; Stevens and House, 1955), where the constriction location X_c (distance from glottis or lips to the constriction) and area A_c are specified along with a ratio of the length of the lip opening to its area (l/A). The areas corresponding to the tongue section are determined by a continuous mathematical function (e.g., parabola) constrained by the three parameters. To be more flexible in the variety of shapes that can be generated, these models have been modified in various ways. Atal *et al.* (1978) extended the number of parameters to five, whereas Lin (1990) incorporated separate continuous functions for the back and front cavities.

Another type of area function model was proposed by Mrayati, Carré, and Guérin (1988), where the parameters were derived purely from acoustic considerations. The vocal tract was divided into separate (distinct) regions, each of which has a sensitivity to formant frequency change that is predictably related to an increase or decrease in cross-sectional area of a particular region. To control the first three formant frequencies, the cross-sectional area of eight regions of unequal length must be specified as parameters. This model is perhaps less interpretable than the previous ones in terms of articulation, but is interesting in the sense that sufficient control parameters could be derived in the absence of articulatory knowledge.

An eventual goal of developing a parametric area function model is to accurately reproduce connected speech. That is, speech created by a vocal tract whose shape alternates between those of vowels and consonants or from one vowel to another. Whereas the models discussed previously are most relevant for vowel articulations (consonant characteristics are not specifically parametrized), it is conceivable that they could be modified or extended to create consonant-like vocal-tract shapes (Lin, 1990), perhaps by allowing the minimum area to approach or become zero. Simulation of connected speech would then be carried out by interpolating a sequence of parameter values over the time course of an utterance. A linear sequencing of vowel and consonant events, however, is limited in its representation of coarticulation in natural speech. For instance, a spectrographic study of vowel–consonant–vowel (VCV) syllables led Öhman (1966) to suggest that a consonant gesture (constriction) is superimposed on an underlying vowel substrate. He concluded that ''A VCV utterance of the kind studied here can, accordingly, not be regarded as a linear sequence of three successive gestures.'' The implication is that speech proceeds as a series of independently controlled vowel-to-vowel transitions, interrupted by superposition of consonant perturbations (Fujimura, 1992). Öhman (1967) subsequently proposed a model that allowed for interpolation of the midsagittal cross distance (width) of one vowel shape to another, over the time course of a syllable. Simultaneously, a consonant constriction function was activated to a degree that also varied over the same time course as the vowel component. At each successive point in time, the consonantal function was superimposed on the modeled vowel substrate to produce a composite tract shape. This view contrasts somewhat with that of Kozhevnikov and Chistovich (1965), who suggested that the consonant–vowel syllable (or C_n V, where *n* denotes multiple consonants) is the primary domain over which coarticulation occurs. In other words, the vocal-tract shape for a consonant or consonant cluster is significantly influenced by the articulatory characteristics of the following vowel, but less so due to the preceding vowel. In either case (and based on much research of coarticulatory processes), it is apparent that the vocal-tract shape at any point in time will be affected by the articulatory demands of adjacent vowels and consonants. Hence, an area function model must be capable of representing and combining the influences of consecutive articulatory events.

Öhman's (1966, 1967) paradigm has influenced various control strategies for articulatory and area function types of models. As discussed by Mattingly (1974), Nakata and Mitsuoka (1965), and Ichikawa and Nakata (1968) implemented the idea of superimposing a consonant on a vowel–vowel transition in a rule-based speech synthesizer. Similarly, Båvegård (1995) and Carré and Chennoukh (1995) have both reported vocal-tract area function models where consonant constrictions are superimposed on an interpolation of a vowel-to-vowel transition. In addition, Browman and Goldstein's (1990) development of ''articulatory phonology'' seems also to be motivated, at least in part, by Öhman's work. In their view, speech is produced by a series of overlapping gestures created by activation of ''tract variables'' such as constriction location and degree of the tongue body and tip.

The purpose of this paper is to describe a kinematic model of the vocal-tract area function that is loosely based on Öhman's concept of a vowel substrate and superposition of a consonantal perturbation. The structure of the model is defined by four perturbation ''tiers,''[1] (see Fig. 1) that together generate a composite time-varying area function. The starting point is a ''neutral'' area function, defined as a vocal-tract shape that produces nearly equally spaced formant frequencies. In tier I, deformation patterns extending from glottis to lips perturb the neutral area function into a specific vowel-like shape, thus forming the vowel substrate. A superposition function is generated in tier II that alters the shape

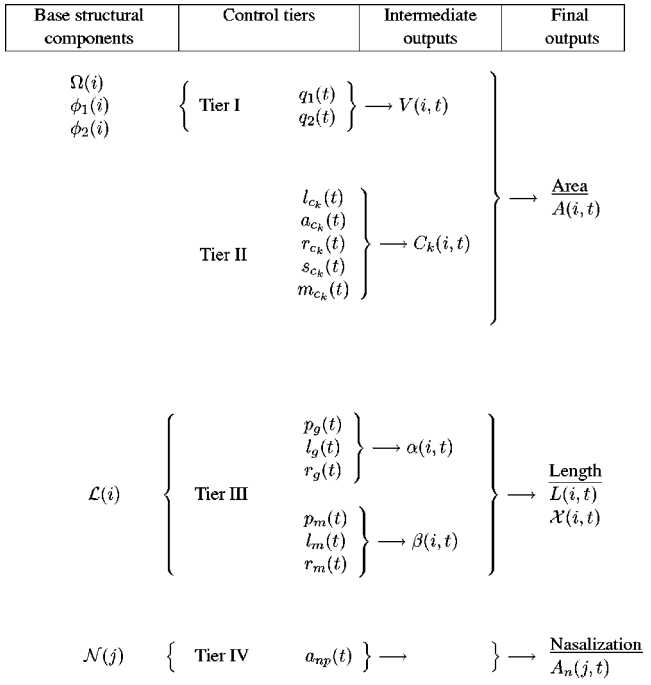

FIG. 1. Diagram of the four-tier area function model. Tier I produces a vowel substrate and tier II generates a superposition function for a consonant. Vocal-tract length changes are generated by tier III, and nasal coupling in tier IV. The ''base structural components'' are dependent only a spatial dimension, whereas the ''final outputs'' are dependent on both space and time.

of the vowel area function in specific regions along the vocal-tract length in order to produce consonantal constrictions. Perturbations of the vocal-tract length can be imposed by tier III, whereas a fourth tier incorporates control of the coupling of the vocal tract to the nasal passages. The parameters within each tier can be time varying; hence, the area function at any instant of time is represented as the combination of the vowel substrate, a superimposed consonantal element, possible lengthening or shortening of various portions of the vocal-tract length, and nasalization. The model is also intended to be flexible enough for easy interchange of components that are characteristic of different speakers. That is, the structure of the underlying vocal tract can be specified independently of the model parameters.

The model is presented here to establish a framework for (1) future studies of the relation between vocal-tract shape and acoustics for connected speech; (2) generating stimuli for perceptual experiments based on manipulation of area function parameters; and (3) eventually producing sentence-level synthetic speech. The specific aim of this paper is limited to a description of the parameters within each of the four tiers and their functional relation to the underlying model. Demonstrations of time-varying area functions and their corresponding acoustic characteristics are also included to verify the concept.

II. AREA FUNCTION MODEL

A schematic representation of the four-tier model is shown in Fig. 1 and descriptions of the components and parameters are given in the Nomenclature. In the first column of the figure are structural components of the vocal tract used to build the foundation of the model. In tiers I and III, these components depend only on the distance from the glottis, as represented by the index *i*, and are modified (by substitution) only if a different speaker's vocal-tract characteristics are desired. The index variable i extends from 1 to N_{vt}, where the area function is assumed to contain N_{vt} cross-sectional areas, concatenated as ''tubelets'' and ordered consecutively from glottis to lips. Similarly, a length function will contain N_{vt} sections representing the length of each tubelet in the area function. Other components of the model contributing to the area or length functions must also contain this same number of sections. Throughout this paper, area functions and associated components contain $N_{vt}=44$ sections.[2] The morphological representation of the nasal tract operates on a different index system *j*, in which the cross-sectional areas are ordered from the point of vocal-tract coupling to the nares.

The control parameters for each tier, shown in the second column of Fig. 1, are used to transform the structural elements (in column 1) into a vocal tract whose shape can be varied over time. Tiers I and II generate time-dependent *area* perturbations in the form of the vowel substrate $V(i,t)$ and consonantal superposition functions $C_k(i,t)$. Together, they produce the composite area function

$$A(i,t)=V(i,t)\prod_{k=1}^{N_c} C_k(i,t) \quad i=[1,N_{vt}], \tag{1}$$

where N_c is the number of consonantal functions. For many cases, only one consonantal function is needed to impose the appropriate constriction. Multiple functions are necessary in cases where simultaneous constrictions may occur. For example, during the production of a consonant cluster such as [sp], there would be a period of time where both the tongue tip and lips are involved in the creation of two separate constrictions. As will be shown in a later section, all C_k's have exactly the same mathematical form, but the control parameters allow for specification of different characteristics of the constriction.

The third tier facilitates *length* perturbations at the glottal and lip ends of the vocal tract. The output is the time-varying composite length function $L(i,t)$, and contains N_{vt} elements representing the length of each tubelet in the area function at a specific instant of time. A cumulative length function $\mathcal{X}$ representing the actual distance from the glottis can be derived from L as

$$\mathcal{X}(i,t)=\sum_{z=1}^{i} L(z,t) \quad i=[1,N_{vt}]. \tag{2}$$

Nasalization is controlled by the fourth tier. At this point, the only parameter is the time-dependent area of the nasal port. It is assigned to a separate tier (rather than embedding it in tiers I or II) to allow acoustic coupling to the nasal tract for either nasal consonant production or nasalization of vowels. Other parameters may be included in the future that more adequately account for the shape of the velopharynx, location of the coupling port, or other changes that may occur during speech production. Additional ''side-branches,'' such as the piriform sinuses and sublingual cavities, also contribute to the overall acoustic character of

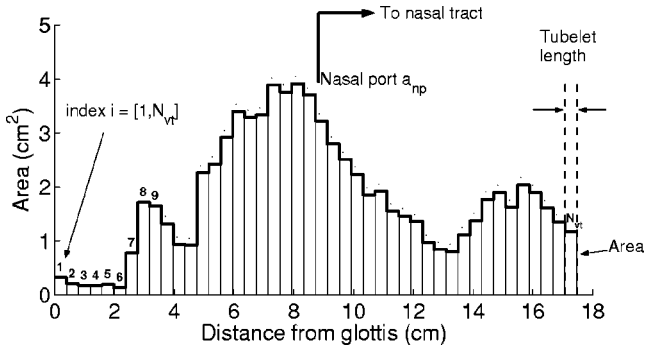

FIG. 2. Example of an area function. It is shown here as a succession of tubelets, denoted by the index *i*, extending from just above the glottis to the lips. Coupling to the nasal passages is indicated by the area a_{np}.

speech. They are not, however, currently included in this model.

An example area function is shown in Fig. 2. It is plotted in stair-step fashion to demonstrate the concatenation of tubelets along the vocal-tract length. Each tubelet has a cross-sectional area as shown by its vertical extent on the *y* axis, and a length as indicated for the *N*th tubelet on the right side of the graph. Note that the index *i* corresponds to the tubelet number; for brevity, these numbers are shown only above the first nine sections. The *x* axis, however, is shown as "distance from the glottis" in units of centimeters, which results from using Eq. (2) to generate the cumulative length function [i.e., $A(i)$ has been plotted against $\mathcal{X}(i)$]. True distance units can be assigned to the *x* axis of the area function and associated perturbation functions, but many of the figures in subsequent sections will simply use *i* as the *x* axis. The nasal coupling location is indicated to be at approximately section 22, or 8.7 cm from the glottis.

It is noted that 14 time-varying control parameters are specified in this model (see column 2 in Fig. 1 and the Nomenclature). Relative to some existing area function models, this is a relatively large number of parameters for which to specify accurate time variations. As will be shown in subsequent sections, however, the parameters do support a precise description of the area function and allow a wide range of flexibility for specifying how the tract shape changes over time. The model has also been designed so that tiers II, III, and IV can be effectively removed, if desired, by setting the parameters to constant values. To model, for example, constant-length, non-nasalized, vowel–vowel transitions, $m_{c_k}(t)$, $L_m(t)$, $L_g(t)$, and $a_{np}(t)$ could be set to zero and all of the other parameters in their respective tiers would become irrelevant, essentially reducing the model to the two parameters in tier I. Similarly, any of tiers II, III, or IV could be utilized independently of the others by providing appropriate parameter values. Eventually some parameters may be found to covary and would not necessarily require a separate specified time variation. For instance, the constriction range $r_{c_j}(t)$ and skewing quotient $s_{c_j}(t)$ are likely to be related to the constriction location $l_{c_j}(t)$; hence, three parameters could perhaps be collapsed into one.

A. Tier I: Vowel substrate

The first tier is based on previous work where a principal components analysis was used to decompose a speaker-specific collection of vowel area functions into a neutral tract shape and a set of basis functions, referred to as *modes* (Story and Titze, 1998). The modes perturb the neutral tract shape according to the following equation:

$$V(i,t)=\frac{\pi}{4}[\Omega(i)+q_1(t)\phi_1(i)+q_2(t)\phi_2(i)]^2$$

$$i=[1,N_{vt}], \tag{3}$$

where the sum of the terms in brackets represents a set of diameters extending from the glottis to the lips. The squaring operation and scaling factor of $\pi/4$ converts the diameters to areas. $\Omega(i)$ is referred to as a neutral diameter function[3] and $\phi_1(i)$ and $\phi_2(i)$ are the modes. The time-dependent parameters $q_1(t)$ and $q_2(t)$ are coefficient values that, when multiplied by the corresponding mode and added to the neutral diameter function as in Eq. (3), construct a desired vowel. The modes have been shown to capture aspects of vowel articulation that allow the model to produce vocal-tract shapes whose acoustic characteristics span a typical $F1-F2$ vowel space (Story and Titze, 1998). Note that when $q_1=q_2=0$, the area function specified as $(\pi/4)\Omega^2(x)$ is expected to produce nearly equally spaced formant frequencies, hence the name "neutral."

This form of the vowel substrate was developed with the assumption that $\Omega(i)$, $\phi_1(i)$, and $\phi_2(i)$ could be derived from an adequate inventory of *any* speaker's vocal-tract area functions. Thus, different speakers' vocal tracts could be modeled by simply interchanging these components. Preliminary data supportive of this assumption were presented in Story (2002), but future analyses of additional MRI-based area function data will need to be performed for verification of the concept, and to provide vowel substrate components for other speakers. The remainder of this paper will utilize the $\Omega(i)$, $\phi_1(i)$, and $\phi_2(i)$ based on MRI-obtained area functions for a single male speaker (Story *et al.*, 1996). They are given in numerical form in Appendix A.

B. Tier II: Consonant perturbation function

The purpose of the second tier is to generate perturbation functions $C_k(i,t)$ that, when multiplied element-by-element with $V(i,t)$, superimpose consonant constrictions on the vowel substrate. The parameters in this tier are the location, area, range, and skewness of the constriction. Location and area roughly correspond to the standard phonetic specifications of place and degree of constriction, whereas the range defines the amount of vocal-tract length over which the constriction will influence the tract shape. The skewness parameter allows for constriction asymmetry along the tract length dimension. An additional parameter is the constriction "magnitude," which is the means by which the constriction is activated or deactivated. Whereas multiple consonant superposition functions $C_k(i,t)$ can be generated [see Fig. 1 and Eq. (1)], they are mathematically identical and their

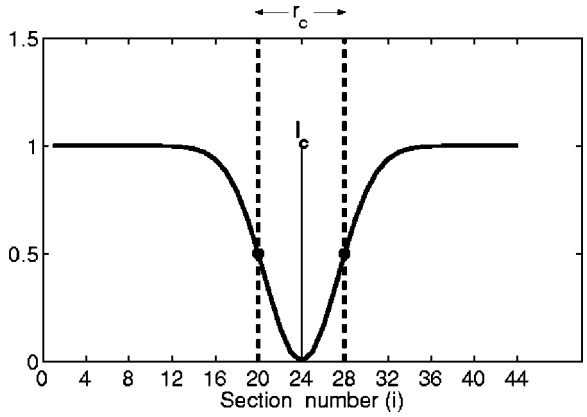

FIG. 3. Example consonantal superposition function $C(i)$ corresponding to Eq. (4). In this case $l_c=24$ and $r_c=8$.

implementation requires only separate specification of the parameters. Hence, only a single constriction will be addressed in the following formulation.

The perturbation has been implemented mathematically with a Gaussian function of the basic form

$$C(i)=1-e^{-\ln(16)[(i-l_c)/r_c]^2}, \tag{4}$$

where l_c is the constriction location. The parameter r_c is the range, and is defined to be the distance between points along the vocal-tract length where the consonant function $C(i)$ is equal to 0.5. This is assured by use of the constant ln (16). In this particular formulation, the parameters must be specified in terms of the index i; however, i could be substituted with $\mathcal{X}(i)$ [Eq. (2)] so they could be specified in actual units of distance. The function will have a value of zero at the point where $i=l_c$, and will asymptotically approach 1.0 on either side of this point. An example is shown in Fig. 3 for the case of $l_c=24$ and $r_c=8$. $C(i)$ is equal to 1 for $i=[1,12]$, after which it decreases continuously and becomes zero at $i=24$ (i.e., $i=l_c$). At locations $i>24$, $C(i)$ gradually increases back to a value of 1. Note that $C(i)=0.5$ at both $i=20$ and $i=28$, due to the range setting of $r_c=8$.

The Gaussian formulation is straightforward to implement and control because it asymptotically returns to the desired value of 1.0 away from the constriction location. Other functions, such as a cosine, can also be used to create the constriction. These, however, require a piecewise concatenation of linear segments with the cosine to complete the function along the entire length of the tract. Also, care must be taken to ensure that a cosine function behaves properly when the constriction location is near the glottal or lip ends. As partial verification, it will be shown in a later section that a Gaussian-based function, superimposed with the vowel substrate, can reasonably approximate consonant area functions obtained directly from imaging experiments.

To accommodate additional parameters for controlling the shape and timing of the constriction, Eq. (4) can be modified to take the form

$$C(i,t)=\begin{cases}1-m_c(t)d_c(t)e^{-\ln(16)([i-l_c(t)]/[r_{cb}(t)])^2} & \text{for } i<l_c(t)\\ 1-m_c(t)d_c(t)e^{-\ln(16)([i-l_c(t)]/[r_{cf}(t)])^2} & \text{for } i>l_c(t).\end{cases} \tag{5}$$

In this equation, $d_c(t)$ is considered to be the "degree" of the constriction, and is determined by the ratio of the desired cross-sectional area $a_c(t)$ at the point of maximal constriction, to the area of the vowel substrate at the location $l_c(t)$ at some specific instant of time. It is calculated by

$$d_c(t)=1-\frac{a_c(t)}{V(l_c(t),t)}. \tag{6}$$

When $a_c(t)$ is equal to zero, $d_c(t)$ will be 1, as was the case implicitly in Eq. (4). But, $a_c(t)$ can also be assigned a value greater than zero to allow for a constriction that does not occlude the vocal tract, as would be necessary for production of fricative and affricate consonants.[4] The parameters $r_{c_b}(t)$ and $r_{c_f}(t)$ in Eq. (5) are determined from the previously defined range $r_c(t)$, *and* a skewing quotient $s_c(t)$

$$r_{c_f}(t)=\frac{s_c(t)r_c(t)}{1+s_c(t)}, \tag{7}$$

$$r_{c_b}(t)=\frac{r_c(t)}{1+s_c(t)}. \tag{8}$$

When $s_c(t)=1$, the total range is distributed equally upstream and downstream of the constriction location, creating a symmetric superposition function. A skewing quotient that is less than or greater than 1 will distribute the specified constriction range asymmetrically around $l_c(t)$, which may be needed to adequately represent some consonant shapes. Shown in Fig. 4 are two examples of $C(i)$ that were generated with different skewing quotients. In both cases, the constriction location is $l_c=24$ and the range was set to $r_c=8$. The first case [Fig. 4(a)] is for a skewing quotient of $s_c=0.3$, where a larger portion of the range is distributed to the downstream side (toward the lip end) of the constriction location. In the second case [Fig. 4(b)], $s_c=3$, and the distribution of the range is reversed; a larger portion of the range is to the upstream side of the constriction location.

The parameter $m_c(t)$ in Eq. (5) is the "magnitude" of the consonant and serves primarily as a timing function to activate and deactivate the consonantal perturbation. In a sense it can be considered a switch, albeit continuous, that allows the constriction to be formed, more or less, depending on its value at a specific point in time. If $m_c(t)=0$, the consonant perturbation is effectively removed because $C(i,t)$ will have a value of 1 over the entire length of the vocal tract, regardless of the other parameter values. In contrast, when $m_c(t)=1$ the cross-sectional area of the constriction specified by $a_c(t)$ is fully realized in the area function. To simulate connected speech at the syllable or word level, $m_c(t)$ would need to continuously vary between zero and 1 to impose and remove consonants at the appropriate instants of time. If $m_c(t)$ is constrained to a maximum value of 1, however, the constriction area will only be realized over

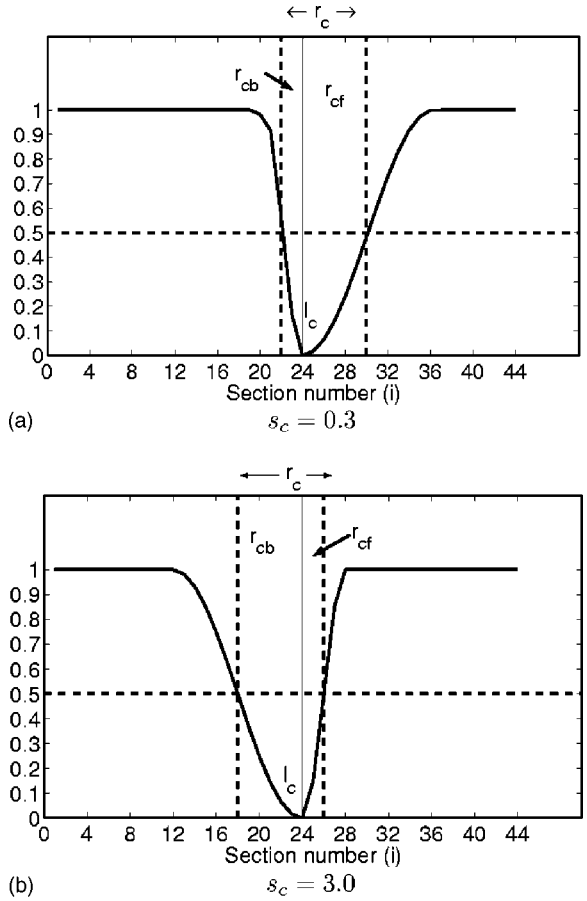

FIG. 4. Demonstration of consonantal superposition functions with asymmetries as specified by the skewing quotient s_c. Both functions were generated with Eq. (5), where $l_c=24$, $a_c=0$, $r_c=8$, and $m_c=1$. (a) Superposition function for $s_c=0.3$. (b) Superposition function for $s_c=3.0$.

a single section along the tract length (i.e., the area of tubelet that is closest to l_c will be zero). For many constriction articulations, an occlusion created by the tongue and lips may consume a larger portion of the tract length than a single tubelet section. Thus, $m_c(t)$ is allowed to exceed 1.0 to force the cross-sectional areas of several consecutive tubelets to be zero, if necessary. With the condition

$$C(i,t)=\max[C(i,t),0], \tag{9}$$

the constriction may be ''spread'' over a greater portion of the vocal-tract length.

An example combination of a single constriction perturbation and vowel substrate is shown in Fig. 5 for a static case (not time dependent). The vowel has been set to the neutral shape $V(i)=(\pi/4)\Omega^2(i)$, and the consonant parameters are, $l_c=24$, $a_c=0$ cm^2, $r_c=8$, $s_c=0.5$, and $m_c=1.1$, where the location and range are specified in terms of the index *i*. The figure contains three plots: the vowel is at the top, the consonant perturbation is in the middle, and the composite area function $A(i)=V(i)C(i)$ is at the bottom. It is observed that $A(i)$ retains the shape of the vowel, except in those sections where the consonant function is less than 1. In these sections the characteristics of both the vowel and the consonant perturbation are expressed in the final output. Note that setting $m_c=1.1$ causes the area to be zero over approximately four sections, effectively spreading the constriction.

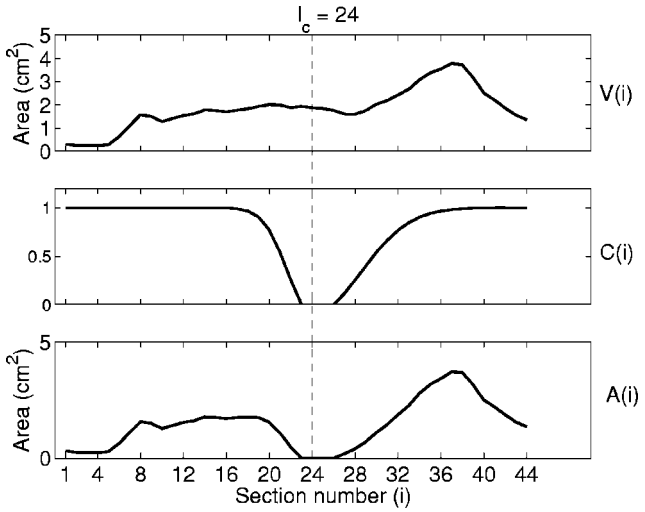

FIG. 5. Combination of the output from tier I and tier II for the case when $q_1=q_2=0$, $l_c=24$, $a_c=0$, $r_c=8$, $s_c=0.5$, and $m_c=1.1$. The top panel shows $V(i)$, the middle panel is $C(i)$, and the bottom panel shows the element-by-element product $A(i)=V(i)C(i)$.

C. Tier III. Length perturbation function

Length modifications are generated in tier III with two superposition functions, similar to those for the consonant constrictions. They are superimposed on a length vector and are designed to either increase or decrease the length of specified portions of the vocal tract.

A nominal or base length vector (length of each tubelet in the area function) consists of N_{vt} *equal* elements

$$\mathcal{L}(i)=\Delta \quad \text{for} \quad i=[1,N_{vt}], \tag{10}$$

where $\Delta=$the tubelet length.[5]

The first function, $\alpha(i,t)$, is intended to produce a length modification near the glottal end of the vocal tract, roughly corresponding to a lowering or raising of the larynx. The function is written as

$$\alpha(i,t)=1+\frac{p_g(t)e^{-K([i-l_g(t)/[2r_g(t)])^2}}{\Delta\Sigma_{i=1}^{N_{vt}}e^{-K([i-l_g(t)]/[2r_g(t)])^2}} \quad i=[1,N_{vt}], \tag{11}$$

where $p_g(t)$ is the amount of larynx lowering ($p_g<0$) or raising ($p_g>0$). The denominator in the second term of the equation is a scaling factor that allows $p_g(t)$ to be specified in actual units of distance (e.g., centimeters). The parameter $l_g(t)$ is the location within the length vector where the length change is centered and maximal. It *must* be specified in terms of the index *i*, much like the constriction location in Eq. (5). The parameter $r_g(t)$ is the number of elements within the length vector over which the length change is distributed. The constant K is set to a value of $2\ln(10\,000)$ to ensure that the length change affects only the number of elements specified by $r_g(t)$. A length perturbation function near the lip end of the vocal tract is needed to represent retraction or protrusion of the lips. Mathematically, this is performed with a function identical to that at the glottal end, except the parameter subscripts are changed. Thus

$$\beta(i,t)=1+\frac{p_m(t)e^{-K([i-l_m(t)]/[2r_m(t)])^2}}{\Delta\Sigma_{i=1}^{N_{vt}}e^{-K([i-l_m(t)]/[2r_m(t)])^2}} \quad i=[1,N_{vt}], \tag{12}$$

where $p_m(t)$ is the amount of lip retraction ($p_m<0$) or protrusion ($p_m>0$) specified in units of distance, $l_m(t)$ is the location where the length change is centered, $r_m(t)$ is the extent over which the length change is distributed, and $K=2\ln(10\,000)$. Typically, the settings for l_g and l_m are 1 and N_{vt}, respectively, so that the maximal length change occurs at the extreme ends of the vocal tract. Equations (11) and (12) are general enough, however, that l_g and l_m can be set to any location along the vocal tract. The perturbed length function is calculated as the product

$$L(i,t)=\mathcal{L}(i)\alpha(i,t)\beta(i,t), \tag{13}$$

resulting in a new length vector with N_{vt} elements, representing modified tubelet lengths.

As a demonstration, length changes of $p_g=-1$ cm at the glottal end, and $p_m=+2.0$ cm at the lip end, were generated with the length perturbation functions. The locations of maximal length change were $l_g=1$ and $l_m=44$, while both r_g and r_m were set equal to 8. The product of $\alpha(i,t)\beta(i,t)$ alone is shown in Fig. 6(a), where it is less than 1 at the glottal end for the decrease in length, equal to 1 from $i=9$ to $i=35$ for no length change, and greater than 1 near the lips to increase the length. The composite length function $L(i,t)$ is plotted in Fig. 6(b). It has an identical shape to that in Fig. 6(a), but the amplitude has been scaled by the nominal length vector. Thus, the plot shows the tubelet length for every element of the length function. The effect of the modified length vector on an area function is shown in Fig. 6(c). It is plotted in stair-step form as a function of distance measured from the *middle* of vocal-tract length, so that the increase or decrease in tubelet length can be easily observed. The lip protrusion can be seen at the right side of the plot, where the length of the lip end of the vocal tract has been increased by 2 cm. The contributions to this overall change come from the gradual length increases of tubelets 36 to 44, where the maximum change is at tubelet 44. The length change at the glottal end can be seen at the left side of Fig. 6(c), where the lengths of tubelets 1 to 8 have been shortened to create the 1 cm reduction in length.

D. Tier IV: Nasalization

As shown previously in Fig. 1, the cross-sectional area of the nasal port is the sole parameter in tier IV. In this simple implementation, it is assumed that the area function of the nasal tract is essentially static (unchanging) during speech production except for the nasal port area $a_{np}(t)$. Other than the first section, the basic cross-sectional area morphology of the nasal tract contained in base function $\mathcal{N}(j)$ (e.g., Dang and Honda, 1994; Story, 1995) will essentially pass unchanged through tier IV to the final output as $A_n(j,t)$. The first section, $A_n(1,t)$, is set equal to the coupling area $a_{np}(t)$ and will be zero, except during production of nasal consonants and nasalized vowels. Although beyond the scope of the present study, cross-sectional area changes in the velopharynx and main vocal tract when the nasal port is open could be more accurately represented by including additional parameters (e.g., Maeda, 1982). Volumetric imaging studies of nasal consonants and nasalized vowels would be an ideal method for providing data to establish the appropriate parametric representation.

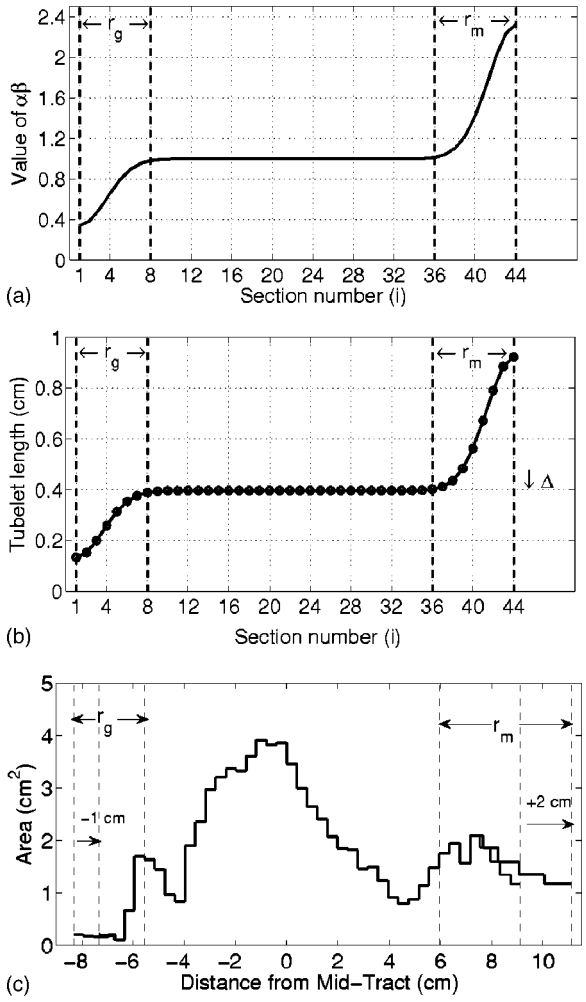

FIG. 6. Example of vocal-tract length change produced by tier III, based on Eqs. (11), (12), and (13). Parameter p_g, representing length change at the glottal end, was -1 cm, and p_m, representing length change at the lip end, was $+2$ cm. The range at both ends (r_g and r_m) was set to be 8 sections. (a) Product of $\alpha(i,t)\beta(i,t)$; (b) length function $L(i,t)$; (c) effect of modified length vector on an area function. The x axis is shown as distance from the center of the vocal-tract length so that length changes at both ends of the vocal tract can be easily observed.

III. STATIC CONSONANTS

A. Stops, nasals, and fricatives

Consonant area functions measured with MRI were reported by Story *et al.* (1996) for the same speaker on which the ''modes'' in Appendix A are based. In addition, four fricative area functions were collected at the same time, but have not been previously published. They are given in numerical form in Appendix B. All were ''static'' consonant shapes because the image acquisition methods required the speaker to maintain a particular vocal-tract configuration for approximately 10 seconds, and repeat it numerous times. Image sets for consonants with an occlusion of the vocal tract were necessarily acquired in their voiceless form, but it is

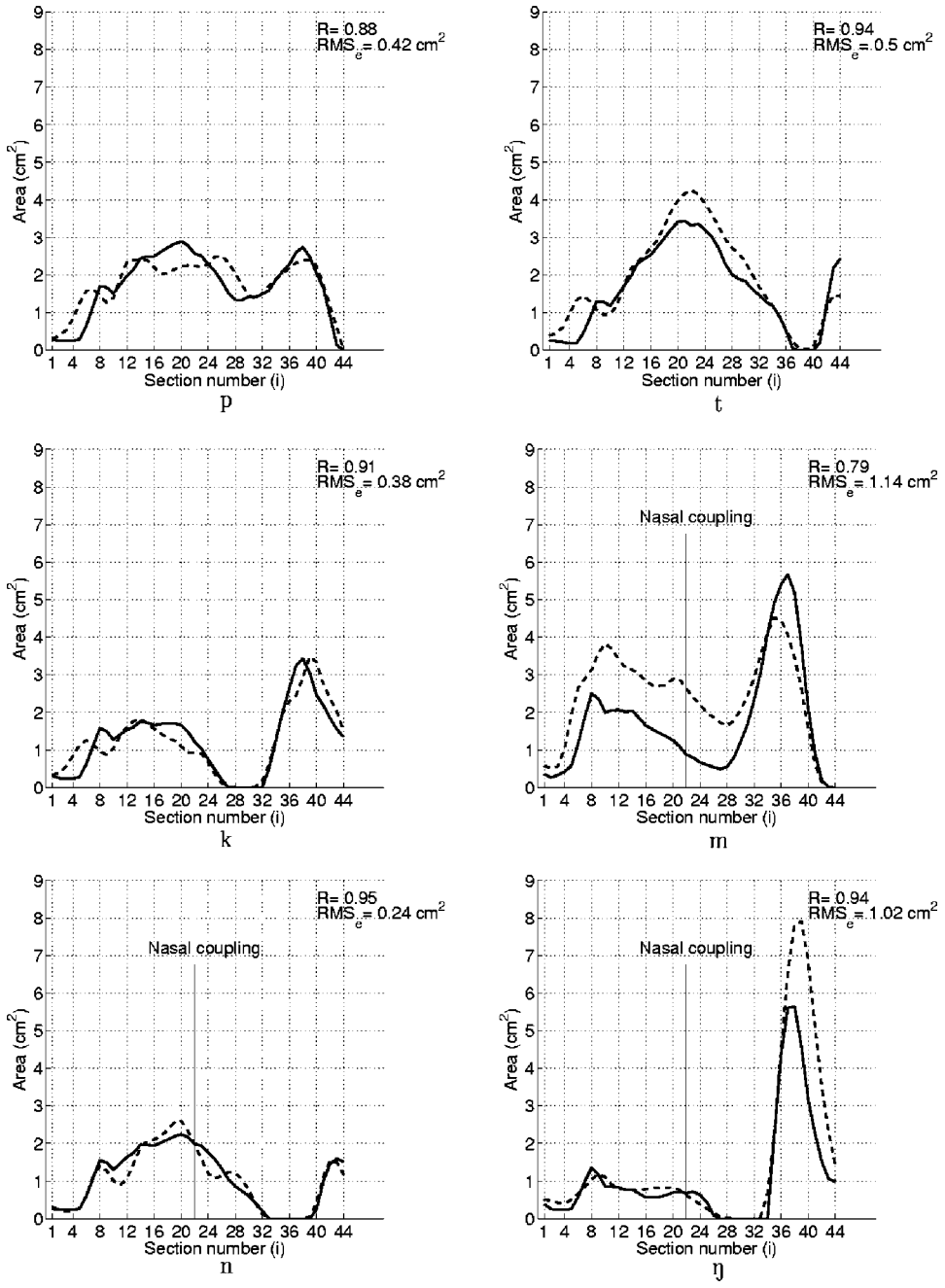

FIG. 7. Comparison of measured area functions (dashed lines) for six consonants with those generated by the area function model (solid lines). In each plot, a correlation coefficient and the rms error is shown at the upper right-hand corner, providing an indication of the fit between measured and modeled area functions. The model parameters for each consonant are given in Table I. For the three nasal consonants, the coupling point between the vocal tract and nasal passages is indicated with a vertical line. Each measured area function shown in this figure has been smoothed prior to fitting the model parameters.

assumed that these tract shapes could be used to guide the synthesis of either a voiced or voiceless consonant. In an attempt to create a neutral vowel context, the speaker was also asked to produce each consonant as if it were preceded and followed by a schwa [ə]. Under these conditions, the resulting area functions do not provide information about coarticulation, but they do specify the constriction location and spatial variations for a variety of consonants. Throughout this section, the measured consonant area functions are used to test the ability of the model described in Sec. II to generate area functions suitable for consonant production. Because the area functions are static, the time dependence of the parameters will be eliminated for the following explanation.

Model parameters for each consonant were first adjusted with an optimization algorithm until a ''best fit'' was determined by minimizing the squared difference between the area functions generated by the model and obtained from measurement. Additional manual tuning was needed to ensure that the portion of the area function dominated by the constriction was fit closely. Comparisons of the original measured and modeled area functions are shown in Figs. 7 and 8. From observation, the fit of the model to the measured consonants appears to be reasonably good, at least in the region of the constrictions. The gross shape in the portions of the area function away from the constrictions also appears well represented in most cases, although there are some large local deviations. For [m] and [f], the model captures the appropriate shape variation along the tract length, but the actual areas were considerably different.

To provide an assessment of the similarity of the area function shape produced by the model parameters relative to

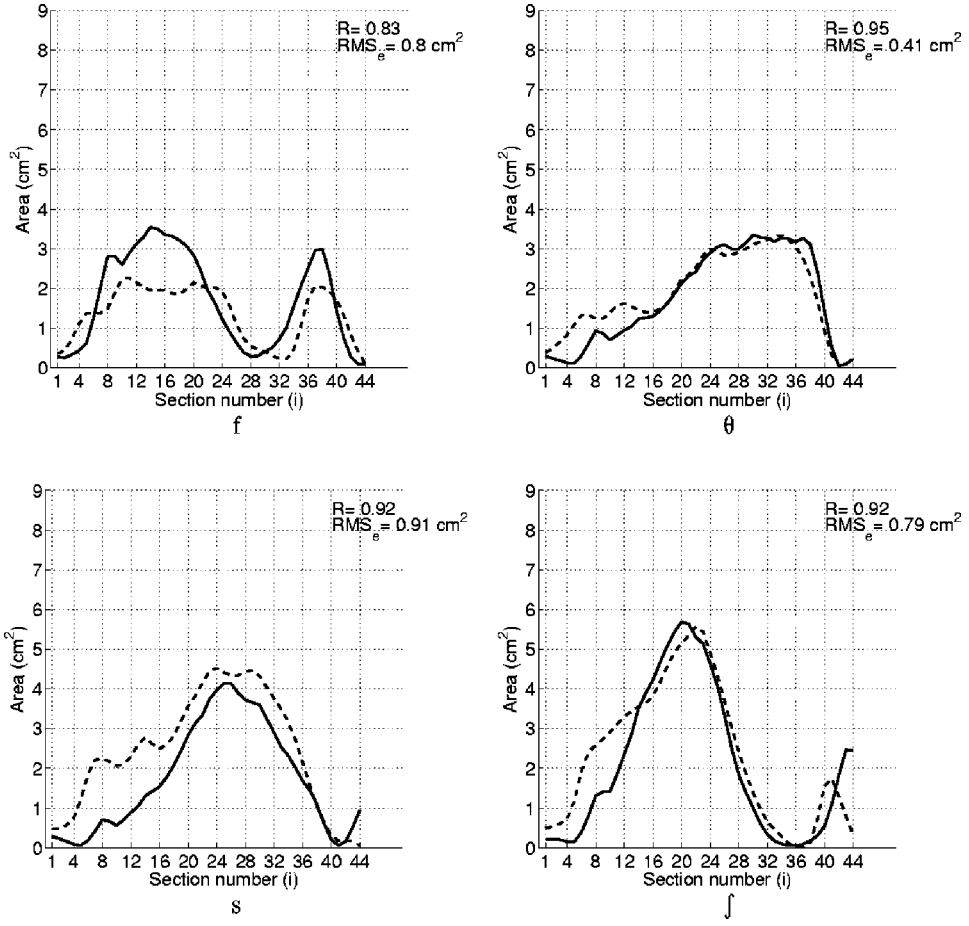

FIG. 8. Comparison of measured area functions (dashed lines) for four fricative consonants with those generated by the area function model (solid lines). Again, in each plot, a correlation coefficient and the rms error is shown at the upper right-hand corner, providing an indication of the fit between measured and modeled area functions. The model parameters for these consonants are also given in Table I and each measured area function shown in this figure has been smoothed prior to fitting the model parameters.

the original measurements, a correlation coefficient (R) was calculated for each pair of measured and modeled area functions. The calculation was performed by dividing the covariance of a given pair of area functions by the product of their standard deviations (e.g., Taylor, 1982). As an indication of the absolute differences in cross-sectional area between each measured and modeled pair of area functions, an rms error was also calculated. The correlation coefficients and rms values are shown at the upper-right side of each plot in Figs. 7 and 8. For seven of the ten consonants $R>0.9$, indicating that the overall shapes of the model-generated area functions were well correlated with the measured versions. The other three consonants generated with the model are somewhat less correlated with $R=0.88$, 0.79, 0.83, for [p], [m], and [f], respectively. The maximum rms error was 1.14 cm^2 for [m], and the minimum was 0.24 cm^2 for [k]. The rms error was generally lowest for the consonants with high correlation coefficients. The exception was [ŋ], for which $R=0.95$ and the rms error was 1.02 cm^2.

Parameter values for each consonant resulting from the optimization process are given in Table I. These data are arranged in three groups consisting of stops, nasals, and fricatives. Within each group they are ordered in terms of their location within the vocal tract. The consonants with a complete occlusion of the vocal tract ([p,t,k,m,n,ŋ]) all required the magnitude setting m_c to be greater than 1. This is to account for the extended region within these area functions where the constriction area a_c is zero. In addition to the

TABLE I. Model parameter values for consonantal area functions.

Consonant	q_1	q_2	l_c (i)	a_c (cm^2)	r_c (i)	s_c	m_c	a_{np} (cm^2)
p	−1.5	0.5	44	0.0	4	1	1.1	0
t	−1.5	2.0	39	0.0	6	1.1	1.3	0
k	0.0	0.0	30	0.0	10	1.5	1.2	0
m	0.0	−3.0	44	0.0	4	1	1.1	1.04
n	−0.5	0.4	38	0.0	11	3	1.3	1.09
ŋ	3.0	−1.0	31	0.0	8	1.5	2.0	1.26
f	−3.0	−2.0	43	0.1	4	1.0	1.0	0
θ	1.0	2.0	42	0.05	9	0.3	1.0	0
s	0.5	3.5	41	0.05	9	0.6	1.0	0
ʃ	−4.0	3.5	36	0.05	11	0.4	1.0	0

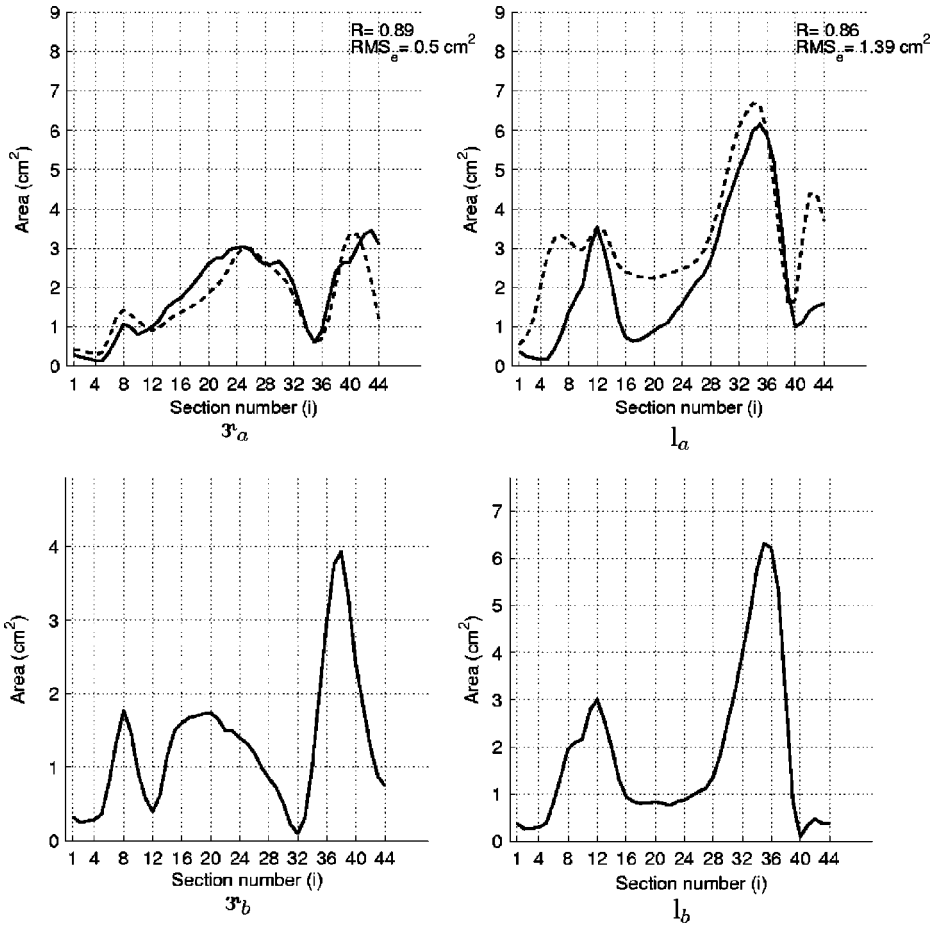

FIG. 9. Area functions for [ɝ] and [l]. Shown in the upper two plots are comparisons of measured area functions (from Story *et al.*, 1996) (dashed lines) to those generated with the area function model (solid lines). In the bottom two plots are model-based area functions that generate acoustic characteristics in line with reported values.

constriction, the nasals also required the nasal coupling a_{np} to be greater than zero. However, a_{np} was not included in the optimization process and the values for it shown in Table I were taken directly from Story *et al.* (1996). For the four fricatives, m_c was equal to 1, and the constriction area was set slightly greater than zero to allow for a narrow orifice connecting the back and front cavities. In general, the range of the constrictions r_c increased as the location was moved from the lips toward the glottis. The exception is in the nasal series where the range of 11 sections for [n] exceeds the 8 sections for [ŋ]. Skewing quotients (s_c) were determined to be 1.0 for the consonants with constriction locations at or near the lips (i.e., [p,m,f]). For the stops and nasals with constriction locations farther back in the vocal tract ([t,k,n,ŋ]), the skewing quotient needed to be greater than 1. In the fricative series the reverse occurred; s_c was less than 1 for all locations posterior to the lips.

If the measured area functions could have been determined from image sets collected during the actual production of a VCV instead of a static posture, q_1 and q_2 should, presumably, have been equal to zero since the speaker was asked to produce each consonant as if it were preceded and followed by a schwa [ə] [roughly equivalent to $(\pi/4)\Omega^2(x)$]. As shown in Table I, however, q_1 and q_2 were set to values different than zero for most of the consonants; only [k] had $q_1=q_2=0$. It must concluded that either the speaker did not accurately produce the consonant configuration within an [əCə] utterance, or these are the shape changes that need to be imposed on an [ə] to accommodate a particular constriction. Nonetheless, the important result is that the consonant perturbation model can generate area functions that are reasonably well matched to those measured from volumetric imaging.

TABLE II. Model parameter values for [ɝ] and [l] area functions. The "a" versions represent the best fit to the measured area functions, where the "b" parameters produce area functions with formant frequencies closer to reported values.

Consonant	q_1	q_2	$[l_{c1},l_{c2}]$ (i)	$[a_{c1},a_{c2}]$ (cm^2)	$[r_{c1},r_{c2}]$ (i)	$[s_{c1},s_{c2}]$	$[m_{c1},m_{c2}]$
ɝ$_a$	0.0	2.0	[12, 35]	[1.0, 0.6]	[4, 4]	[1, 1]	[1, 1]
l$_a$	3.0	0.0	[12, 40]	[3.5, 1.0]	[4, 4]	[1, 1]	[1, 1]
ɝ$_b$	0.0	−1.0	[12, 32]	[0.4, 0.1]	[4, 6]	[1, 1]	[1, 1]
l$_b$	2.0	−2.0	[12, 40]	[3.0, 0.1]	[4, 4]	[1, 1]	[1, 1]

B. Liquids

In addition to the ten consonants discussed previously, Story *et al.* (1996) also reported area functions for static productions of [ɝ] and [l] (for [l], the cross-sectional areas of the two lateral pathways were summed and incorporated into the area function). These were similarly fit with the area function model, but each required two consonant superposition functions to produce an adequate representation of the original. The two upper plots of Fig. 9 show comparisons of the original measured and modeled area functions; the corresponding parameter values are given in the upper part of Table II. For both [ɝ] and [l], the first consonantal function was centered at element 12 and set to areas of 1.0 and 3.5 cm^2, respectively. (The sizes of these cross-sectional areas are large enough that they could be considered consonantal ''settings'' rather than constrictions.) The second constriction for the [ɝ] was centered at element 35 and set to an area of 0.6 cm^2; for [l] a similar constriction was imposed at element 40 with area equal to 1 cm^2. The combination of the two constrictions and the settings of the two vowel substrate parameters given in Table II generates area functions for both consonants that are reasonably close to the originals, as determined both by visual comparison and calculated correlation coefficients of 0.89 for [ɝ] and 0.86 for [l]. The calculated rms error values are also similar to those determined for the previous consonants.

As noted in Story *et al.* (1996), the measured area functions for [ɝ] and [l] produced formant frequency patterns that were not closely representative of those determined from recorded speech. This was primarily due to constriction cross-sectional areas that were too large. Two additional area functions were generated with the area function model that better represent the appropriate acoustic characteristics for these consonants. The parameters are given in the lower part of Table II and the area functions are shown in the bottom two plots of Fig. 9. Specifically, the shift of the second constriction location for [ɝ] to element 32, the change in cross-sectional areas of a_{c1} and a_{c2}, and overall reshaping of the area function with different vowel substrate parameters lowers the third formant ($F3$) from 2.3 kHz for the original to approximately 1.75 kHz, which is more in line with reported values (Peterson and Barney, 1952; Lee, Potamianos, and Narayanan, 1999; Espy-Wilson, 1992). The primary modification for [l] was a decrease in the second constriction area as well as a change in the vowel substrate parameters. These changes combine to increase the third formant frequency from 2.7 kHz for the original area function to about 3 kHz, also similar to reported values (Espy-Wilson, 1992).

It is noted that the area function model ignores the presence and possible effects of lateral pathways for the [l] and sublingual cavities for both [l] and [ɝ] (Espy-Wilson, 1992; Espy-Wilson *et al.*, 2000; Alwan, Narayanan, and Haker, 1997; Narayanan, Alwan, and Haker, 1997). For more accurate representations, the model may eventually need to be augmented with parameters and associated structural components that more closely replicate these differences. In addition, the attempt in this section has been to use the measured area functions as a starting point for the many possible variants of [l] and [ɹ] (Alwan and Haker, 1997; Narayanan and Haker, 1997). This would require a variety of possible settings of the constriction parameters.

The measured area functions used in this section represent only limited instances of consonant production, but they do provide reasonable test cases for assessing the capability of the area function model. The tests demonstrate that the model does have the flexibility to generate realistic area functions within the consonant superposition paradigm and provides parameter values that can be used as a starting point for simulation of consonants.

IV. TIME-VARYING AREA FUNCTIONS

Throughout the description of the area function model in Sec. II, the parameters were shown as time-dependent variables. In this section, a series of time-varying area functions was generated that simulates possible vowel–vowel (VV) and vowel–consonant-vowel (VCV) utterances. For each case, the duration of the utterance was 0.5 s and vocal-tract area variations were accomplished by allowing three parameters to vary with time: two mode coefficients, q_1 and q_2, that create the vowel substrate, and the consonantal magnitude m_c. Other parameters such as the area, location, and

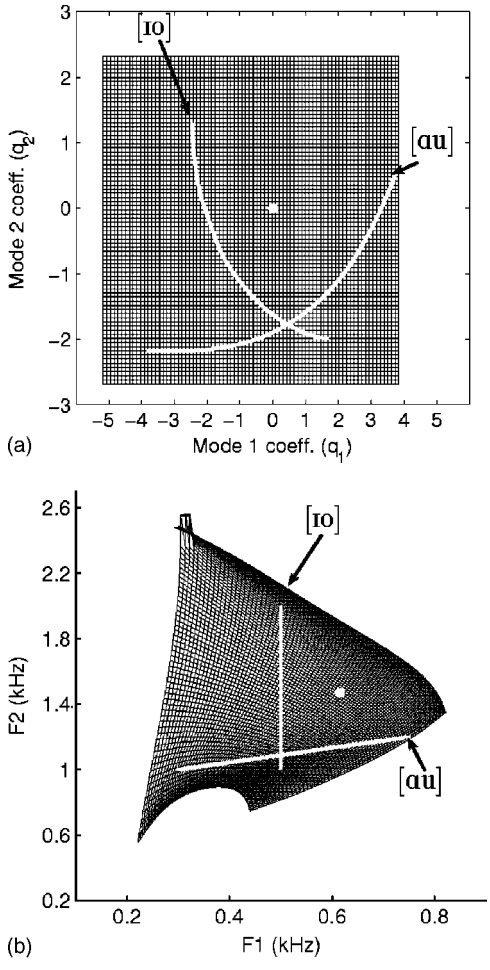

FIG. 10. Mapping between mode coefficients (q_1 and q_2) in (a) and formant frequencies ($F1$ and $F2$) in (b). The curved white lines in the upper plot are the coefficient variations that would produce the corresponding linear formant trajectories for [ɪo] and [au] in the lower plot. The white dot located at $q_1=q_2=0$ in the coefficient plot corresponds to the white dot in the $F1-F2$ plot.

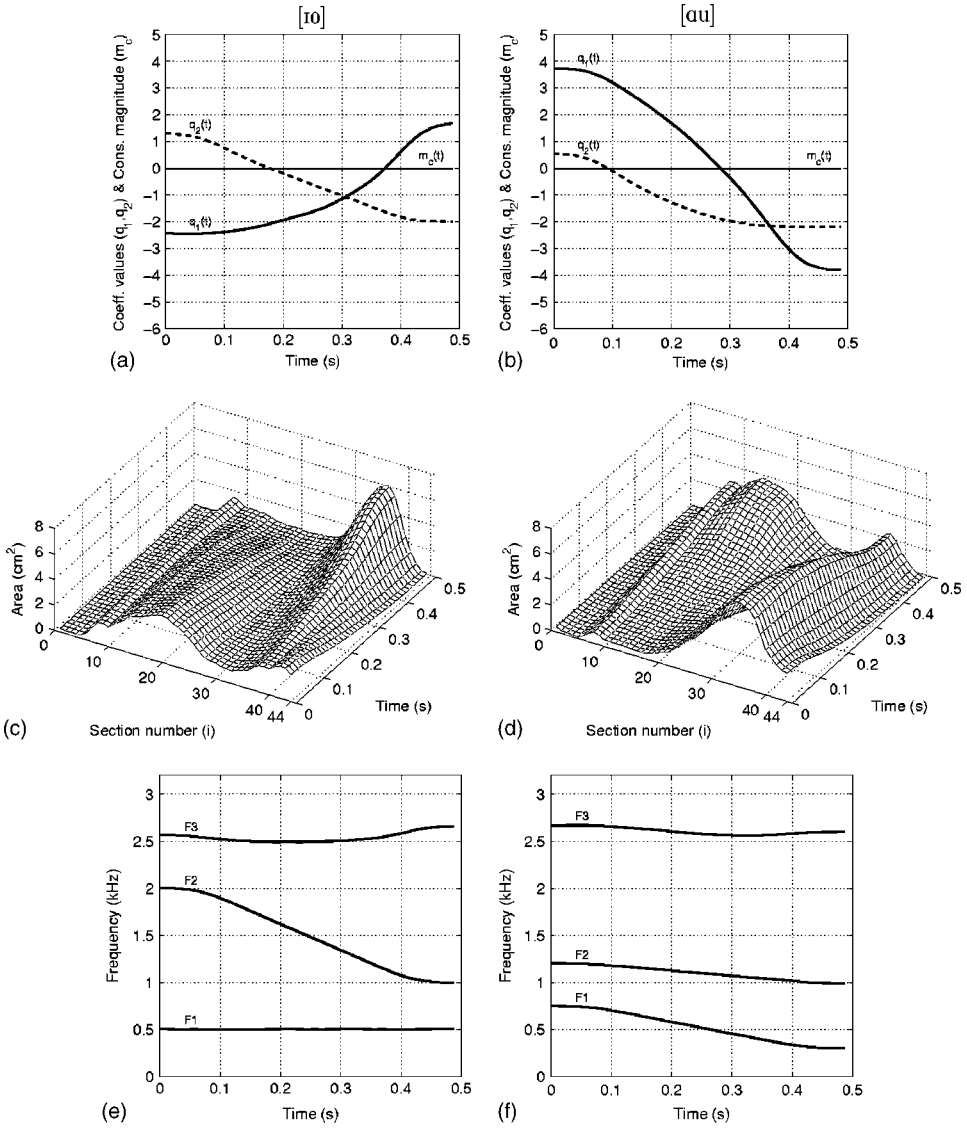

FIG. 11. Area function simulations for two VV transitions. Each *column* of plots shows, in descending order, the time variation of mode coefficients $q_1(t)$ and $q_2(t)$ and consonant magnitude $m_c(t)$ [(a) and (b)], time-varying area functions [(c) and (d)], and time-varying formant frequencies [(e) and (f)]. In the left column the transition is approximately [ɪo], and on the right, [ɑu].

range of the constriction were changed for each case but were held constant over the duration of the simulated utterance. Additional cases are presented that utilize more parameters to simulate a two-consonant cluster (VCCV), vocal-tract length change during a VV transition, and a VCV with a nasal consonant.

A coefficient-to-formant mapping (Story and Titze, 1998, 2002) was used to determine the time variations of the tier I parameters, $q_1(t)$ and $q_2(t)$, that would approximate the VV transitions [ɪo] and [ɑu]. The mapping is shown in Fig. 10. In the upper panel [Fig. 10(a)] is a set of 6400 pairs of q_1 and q_2 coefficients, bounded by the maximum and minimum values given in Table IV. The point where q_1 and q_2 are both equal to zero is indicated with the white dot. The two curves are coefficient trajectories that, when sampled with an appropriate time step ($\Delta T = 0.0125$ for the present examples), can produce time-varying area functions for [ɪo] or [ɑu], respectively. The mesh in Fig. 10(b) is comprised of first and second formant frequencies corresponding to the area functions produced with the coefficient pairs in the upper panel mesh. The straight lines in this figure are $F1-F2$ formant trajectories that correspond to the q_1-q_2 coefficient curves for [ɪo] and [ɑu] in Fig. 10(a). The $F1-F2$ trajectories were deliberately chosen to be linear VV transitions; however, any $F1-F2$ trajectory measured from natural, recorded speech could be mapped to corresponding q_1-q_2 coefficient curves, as long as the formant trajectory remains with the boundaries of the $F1-F2$ space (black mesh) (Story and Titze, 2002). The coefficients $q_1(t)$ and $q_2(t)$ for each vowel transition are shown as functions of time in Figs. 11(a) and (b).

The time variation for the consonantal parameters was generated by a fifth-order polynomial function that produced a "minimum jerk" movement (Hogan, 1982). Other functions such as cosine, damped second-order system, or minimum energy (e.g., Nelson, 1983) could also be used. A minimum jerk transition from one position to another can be specified mathematically as

$$u(t) = u_o + (u_f - u_o)\left(10\left(\frac{t}{T}\right)^3 - 15\left(\frac{t}{T}\right)^4 + 6\left(\frac{t}{T}\right)^5\right) \quad \text{for } 0 \leq t \leq T, \tag{14}$$

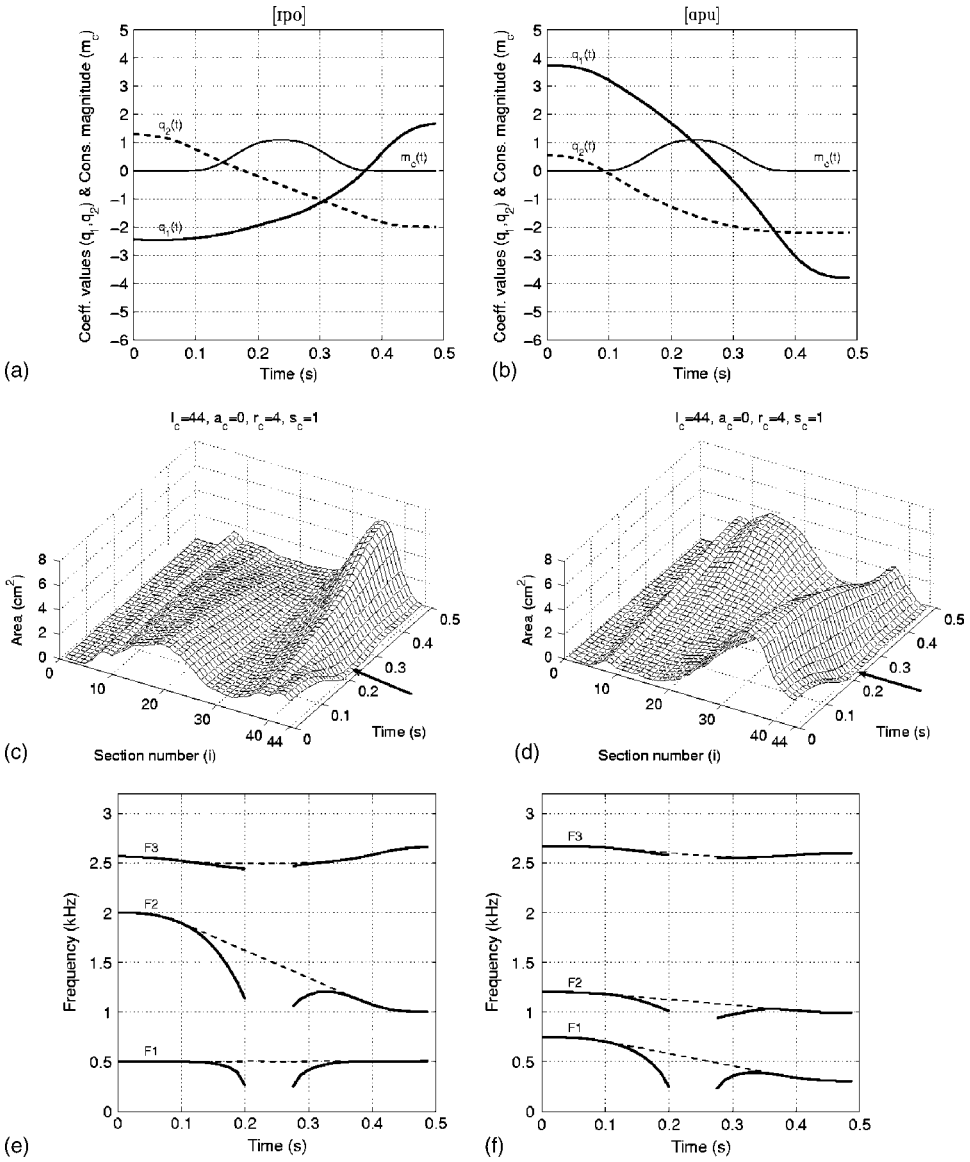

FIG. 12. Area function simulation for two VCVs. In the left column the VCV is approximately [ɪpo], and on the right, [ɑpu]. Shown in descending order are the time variations of mode coefficients $q_1(t)$ and $q_2(t)$ and consonant magnitude $m_c(t)$ [(a) and (b)]. Next are the time-varying area functions [(c) and (d)], where the constant consonantal parameters are shown at the top of the plot, and the point of maximum constriction is indicated with the arrow. Finally, the time-varying formant frequencies are shown at the bottom of the figure [(e) and (f)].

where u_o and u_f are the initial and final positions, respectively, and T is the duration of the movement. Thus, any of the time-varying parameters of the area function model could replace the general variable u, and Eq. (14) would determine its time course of change from one specified value to another. This method of specifying the time variation of parameters is perhaps overly simplistic, but at this point serves the purpose of demonstrating some of the capabilities of the model.

A. VV simulations

The first case is shown in Fig. 11(a), where $q_1(t)$ and $q_2(t)$ initially specify an approximation of the vowel [ɪ] and change over time to values representative of the vowel [o]. Also shown is $m_c(t)$, which is zero across the entire utterance. This means that the consonant tier (tier II) is effectively shut off, and the result is a VV transition. The variation of the area function over time is presented as a three-dimensional plot in Fig. 11(c), where the transition from [ɪ] to [o] can be observed in terms of 40 successive area functions, spaced 0.0125 s apart. For each of the area functions within the [ɪo] transition, a frequency response function was calculated (Sondhi and Schroeter, 1987), and from it the first three formant frequencies were determined with a peak-picking algorithm. Figure 11(e) shows the variation of $F1$, $F2$, and $F3$ over the time course of the vowel transition. The spacing between the first two formants is initially large for the [ɪ]. $F2$ then decreases by about 1 kHz in its transition to [o], while $F1$ remains constant at 0.5 kHz, and $F3$ changes only slightly.

The second column of Fig. 11 presents an analogous case, but with the mode coefficients set to approximate the transition from [ɑ] to [u] [Fig. 11(b)] and $m_c(t)$ is set to zero over the entire utterance. Figures 11(d) and (f) show the time variation of the area function and formant frequencies, respectively. Again, 40 successive area functions are plotted, and the formant frequencies were determined from calculation of the frequency response of each area function. Note that the choice of a linear $F1-F2$ trajectory [see Fig. 10(b)] causes the cross-sectional area at and near the lips to be reduced prior to the formation of the midtract constriction for the [u].

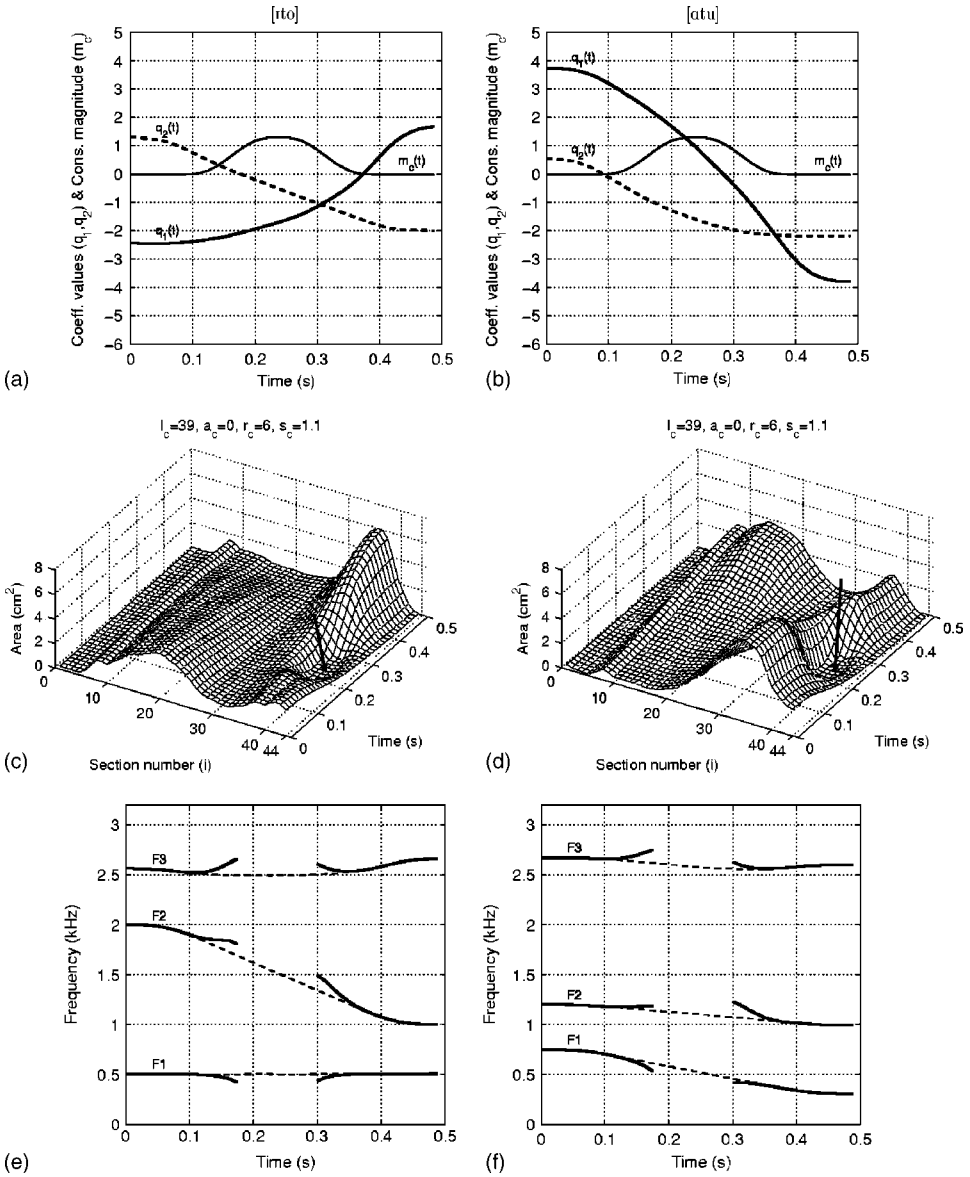

FIG. 13. Area function simulation for two VCVs. In the left column the VCV is approximately [ɪto], and on the right, [ɑtu]. The ordering of the individual plots is identical to those in Fig. 12.

B. VCV simulations

Presented in the next six figures are cases in which the same two vowel substrates shown in Fig. 11 for [ɪo] and [ɑu] were used, but the consonant tier (tier II) was also activated by a nonzero time variation of $m_c(t)$ to produce a VCV. Parameters in tier II were set to approximate the consonants [p,t,k,θ,ɹ] and [l].

Shown in Fig. 12 are two cases where a consonant constriction was imposed at the lips (section 44) with a cross-sectional area of zero, and a range of 4 sections. The time course of $q_1(t)$, $q_2(t)$, and $m_c(t)$ for each case is plotted in Figs. 12(a) and (b). The variations of the mode coefficients are the same as the previous case, but $m_c(t)$ now rises from zero to 1.1 to activate the constriction, and then decreases back to zero to release it. Whereas the choice to have $m_c(t)$ reach its peak 0.25 s into the utterance was arbitrary for these demonstrations, displacing the peak of the consonantal time variation to an earlier or later time point would likely create significant, and potentially interesting, changes in the formant frequency characteristics. In addition, if the goal were to match the formants to a specific production (recording) of the utterance, $m_c(t)$ would likely need to follow some other time course as well.

The parameters displayed at the top of the area function figures [Figs. 12(c) and (d)] were kept constant, and, with the exception of q_1 and q_2, are the same values as those shown in Table I for [p]. The resulting area functions are shown in Figs. 12(c) and (d) and are essentially the same as those in Fig. 11, except in the region near the lip end, where the cross-sectional area is decreased to zero over a time period from approximately 0.2 to 0.27 s. Note that $m_c(t)$ is greater than zero from 0.1 to 0.4 s, but the vocal tract is occluded for only the period of time where $m_c(t)$ is greater than or equal to 1.

The effect of the constriction on the formant frequencies can be seen in Figs. 12(e) and (f). The dashed lines indicate formant variations for the VV transitions of the previous case, whereas the solid lines show the time-varying formant frequencies with the presently imposed constriction. The break in the time course of the formants occurs during the

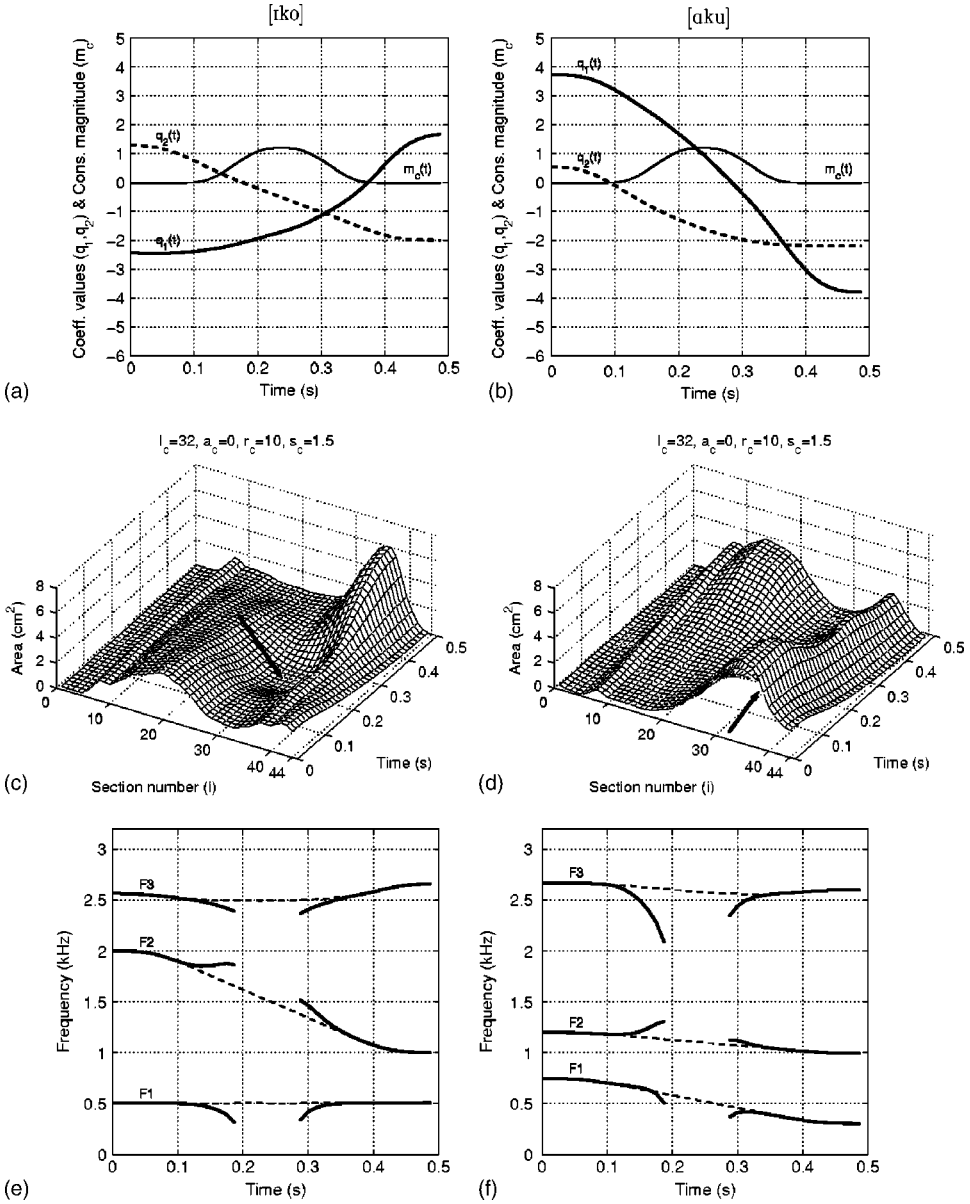

FIG. 14. Area function simulation for two VCVs. In the left column the VCV is approximately [ɪko], and on the right, [ɑku]. The ordering of the individual plots is identical to those in Fig. 12.

period of time in which the vocal tract was fully occluded by the constriction. In both cases, the constriction perturbs all three formant frequencies downward just prior to the occlusion, and upward after its release. Formant characteristics like these for a bilabial consonant are well known (Stevens, 1998) and the figures themselves are reminiscent of those reported by Öhman (1966) for similar initial and final vowels.

Time-varying parameters for the next two cases [Figs. 13(a) and (b)], appear similar to those in the previous figure. The exception is that $m_c(t)$ rises to maximum value of 1.3. More significant, however, are the changes imposed on the constant parameters, where the constriction location is set to section 39, the range is set to 6 sections, and the skewing quotient has been increased to 1.1. These settings are roughly representative of an alveolar stop consonant (see Table I). The constriction can be seen in both time-varying area functions [Figs. 13(c) and (d)] by following element 39 along the time axis, where the occlusion begins at about 0.18 s and is released at 0.3 s. For the case in the left column (≈[ɪto]), the constriction causes $F1$ to decrease by a small amount prior to the vocal-tract occlusion and then it rises following the release of the consonant. During the same time period $F2$ and $F3$ both rise prior to the occlusion and then fall after it is released. The case in the right column (≈[ɑtu]) generates the same directions of formant frequency change, even though the underlying vowel transition is different.

Two simulations of a VCV with an approximation of a velar stop consonant are shown in Fig. 14. The time-varying parameters are, again, nearly identical to the previous cases, but with a maximum value of $m_c(t)=1.2$. The constriction location is at section 32 with a range of 10 sections, and the skewing quotient is set to 1.5. (Note that in Table I, $l_c=30$ was specified for this consonant; in the two vowel contexts used for these demonstrations it was necessary to set $l_c=32$ to produce formant transitions representative of a velar consonant.) The time-varying area function in Fig. 14(c) indicates the constriction forming along the time axis at section 32; the occlusion is indicated by the arrow. In Fig. 14(d), the constriction is more difficult to see because the oral cavity

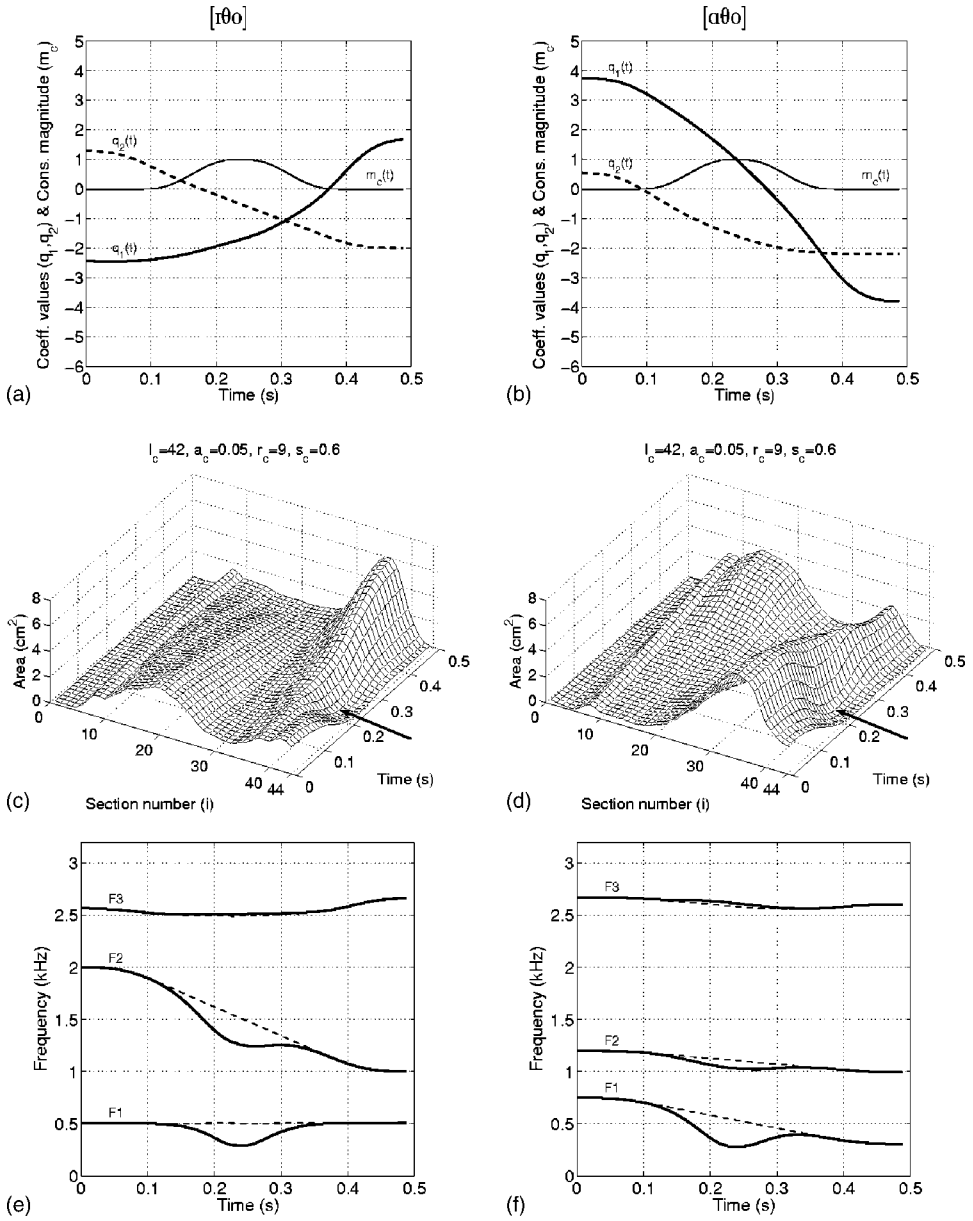

FIG. 15. Area function simulation for two VCVs. In the left column the VCV is approximately [ɪθo], and on the right, [ɑθu]. The ordering of the individual plots is identical to those in Fig. 12.

portion of the vowel area function is expanded in the early stage of the utterance. Following section 32 along the time axis, however, indicates a depression in the area function that occurs at about 0.19 s. Figure 14(e) shows the time-varying characteristics of $F1$, $F2$, and $F3$ for [ɪko], where $F1$ and $F3$ both fall prior to the occlusion and rise after its release. $F2$ exhibits the opposite behavior, rising during the onset of the consonant and falling as it is released. The formant characteristics are similar for [ɑku] [Fig. 14(f)], at least in terms of the direction of change for each formant.

The next two cases contain a constriction area that is nonzero, representative of a fricative consonant. Time-varying parameters are shown in Figs. 15(a) and (b). The consonant magnitude $m_c(t)$ has a maximum value of 1, but, because $a_c = 0.05\ \text{cm}^2$, the minimum area that is achieved during the simulated utterance will be greater than zero. The constriction parameters are taken directly from Table I for the consonant [θ]. The formation of the constriction can be seen along element 42 in the time-varying area function plots [Figs. 15(c) and (d)], and the corresponding formant frequency patterns are displayed in Figs. 15(e) and (f). Because an occlusion does not occur during the time course of the two VCVs, the formant frequencies are continuous. In both cases, $F1$ and $F2$ are perturbed downward in frequency during the presence of the consonant, whereas $F3$ is barely affected. Because the area function is not occluded in this case, the formant frequencies could be calculated over the entire utterance as shown in Figs. 15(e) and (f), but an actual production of [θ] would be unvoiced and a spectrogram would show a discontinuity in the formant frequencies. The present ''simulation'' of the VCV is only of the time-varying area function and the resulting formant frequencies, not speech itself. Hence, the appearance of formants is not affected by the presence or absence of voicing.

Area function simulations that included [ɹ] and [l], are given in Figs. 16 and 17, respectively. The vowel substrate was either [ɪo] or [ɑu], and each consonant require *two* magnitude functions [$m_{c1}(t)$ and $m_{c2}(t)$] that, in this case, both followed the same time course with a maximum value of 1. The constant parameters are those presented in Table II (for

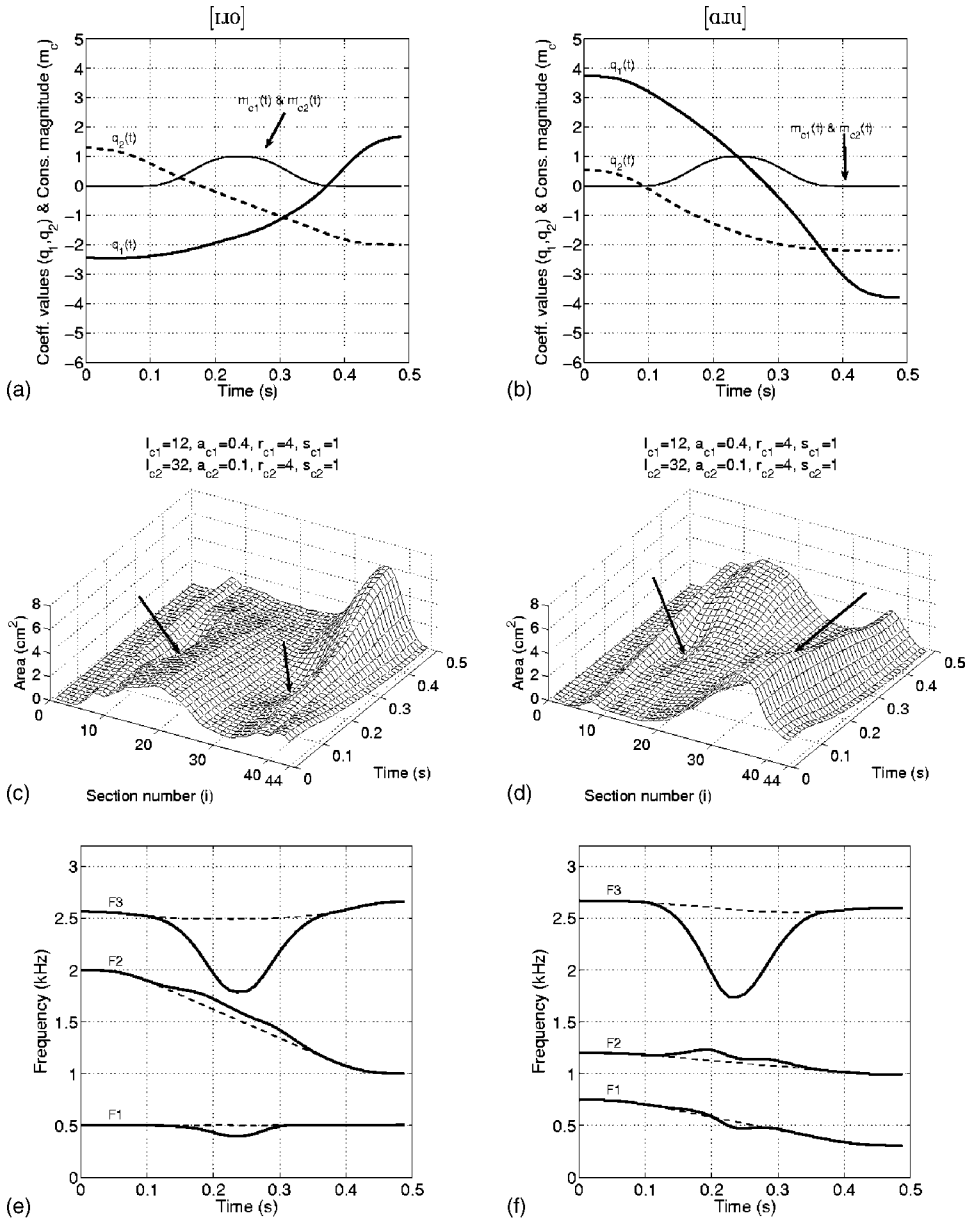

FIG. 16. Area function simulation for two VCVs. In the left column the VCV is approximately [ɪo], and on the right, [ɑɪu]. The ordering of the individual plots is identical to those in Fig. 10. There are, however, two simultaneous constrictions required to produce the [ɹ]. They are each indicated by the arrows and the parameters are specified at the top of the plots in (c) and (d).

the "b" versions) and are shown at the top of the area function plots in each figure. The phonetic symbol [ɹ] is used here because the constriction function creates a time-varying utterance, rather than a sustainable sound such as [ɜˠ].

For the [ɹ], a depression in the time-varying area functions [Figs. 16(c) and (d)] can be seen along the time axis of section 12, most significantly at time 0.25 s. Similarly, at the same point in time, the second constriction can be observed along the time axis of section 32. In either vowel substrate context, the primary effect of the constrictions on the formant frequencies [Figs. 16(e) and 15(f)] is to lower $F3$ by almost 0.8 kHz to bring it momentarily to a value of about 1.77 kHz. In addition, the constrictions perturb $F2$ upward in frequency, whereas $F1$ is perturbed downward.

Area functions for the two simulations with [l] [Figs. 17(c) and (d)] show a slight expansion along the time axis for section 12, opposite of the constrictive effect imposed by the [ɹ] on this same section. At section 40, the second constriction can be seen to take effect during the same time period as the first. In both vowel substrate contexts, the constrictions displace $F3$ upward in frequency, although the pattern of variation over time is different.

C. VCCV simulations

During speech production, two or more consonants may occur consecutively in a cluster without an intervening vowel. For example, in the word /splɪt/, multiple constrictions are rapidly formed and released prior to production of the vowel. Two area function simulations of a two-consonant cluster are presented in Fig. 18. The vowel substrates [Figs. 18(a) and (b)] were again the same [ɪo] and [ɑu] transitions used in all of the previous examples. The two intended consonants superimposed on the vowel substrates were, in sequence, [p] and [ɹ]. Their production requires specification of three constrictions, one for [p] and two for [ɹ], whose parameters will be the same as those in Tables I and II, respectively. The time variation of each constriction magni-

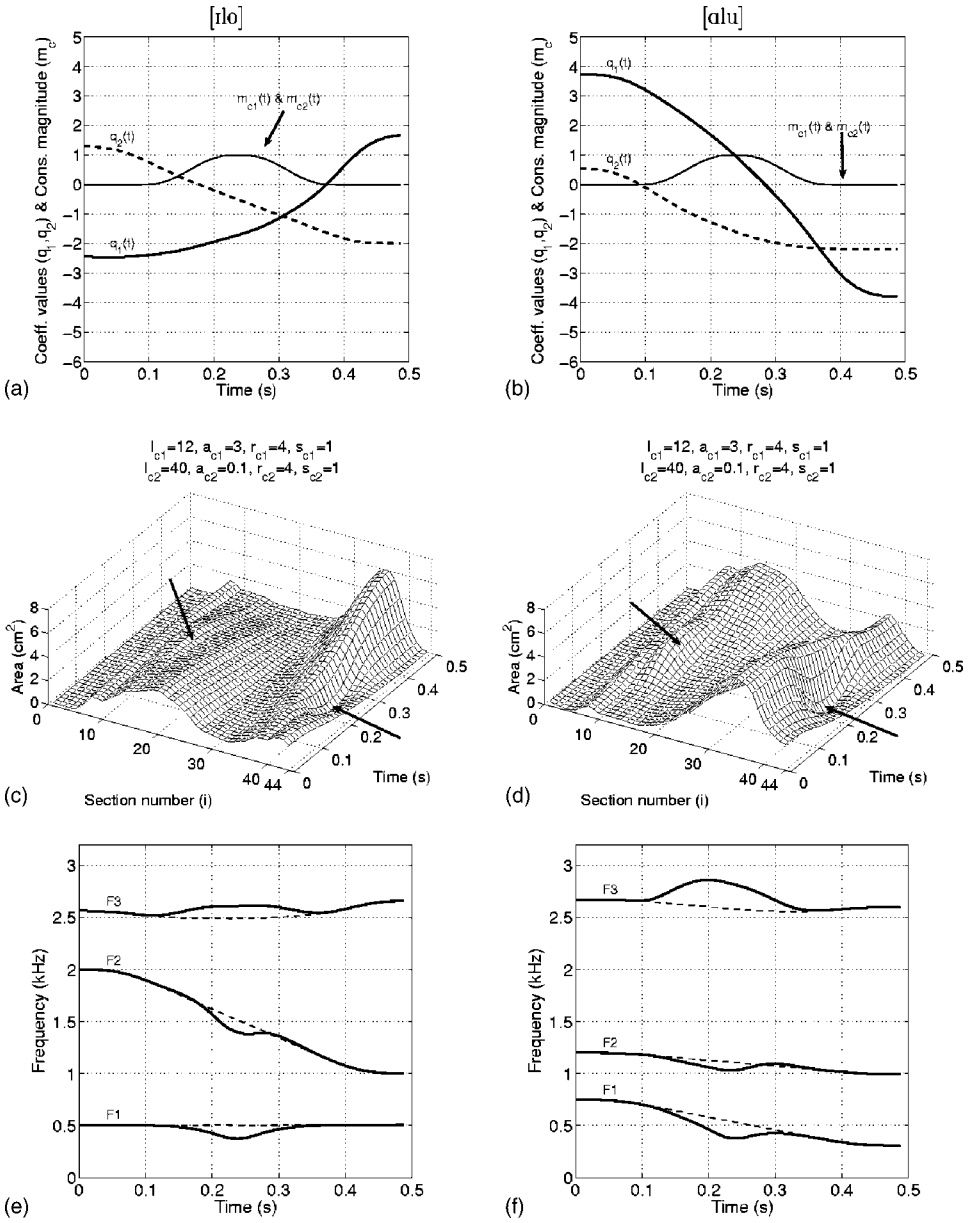

FIG. 17. Area function simulation for two VCV transitions. In the left column the VCV is approximately [ɪlo], and on the right, [ɑlu]. The ordering of the individual plots is identical to those in Fig. 10. Like the previous figure, there are two simultaneous constrictions required to produce the [l]. Again, they are each indicated by the arrows and the parameters are specified at the top of the plots in (c) and (d).

tude is shown in Figs. 18(c) and (d) (these are identical figures but were duplicated to maintain continuity along each column). The magnitude for [p] (solid line) becomes nonzero at 0.08 s, rises to its peak of 1.1 at 0.19 s, and then decreases to zero at 0.33 s. The two constrictions for the [ɹ] follow exactly the same time course as each other, and are identically plotted as the dashed line in the figures. Their magnitudes begin to rise at 0.13 s, which is just slightly delayed relative to the [p] magnitude. The peak occurs at 0.25 s and the constrictions are completely released ($m_{c2}=m_{c3}=0$) at 0.39 s.

The resulting time-varying area functions are shown in Figs. 18(e) and (f), where the three arrows indicate the location and time of each constriction. As dictated by the time course of the constriction magnitudes, there is considerable temporal overlap of the consonants, creating more complex coarticulation than in any of the previous cases. Specifically, the constrictions for the [ɹ] begin to form during production of the [p], and are fully in place just shortly after the occlusion is released.

The formant frequencies for each case [Figs. 18(g) and (h)] are perturbed downward by the [p] constriction, although by different absolute amounts. In the [ɪo] context, $F1$ decreases by 0.25 kHz, while $F2$ drops by 0.8 kHz. The decrease in $F1$ for the [ɑu] context is more than 0.45 kHz, whereas the change in $F2$ is only 0.12 kHz. In both cases, $F3$ drops by about 0.2 kHz. At the point in time when the [p] constriction is released, the vocal tract is nearly configured to produce the [ɹ], and in both cases, the third formants are located at about 1.85 kHz and decrease as the [ɹ] becomes fully expressed in the area function. As the [ɹ] constrictions fade, all of the formants rise to their locations for the final vowel. Most notably is $F3$, which increases in both cases by nearly 0.8 kHz.

D. VV simulations with vocal-tract length change

In this section, a vowel-to-vowel transition is simulated with the area function model while localized changes to the vocal-tract length are simultaneously imposed. A transition

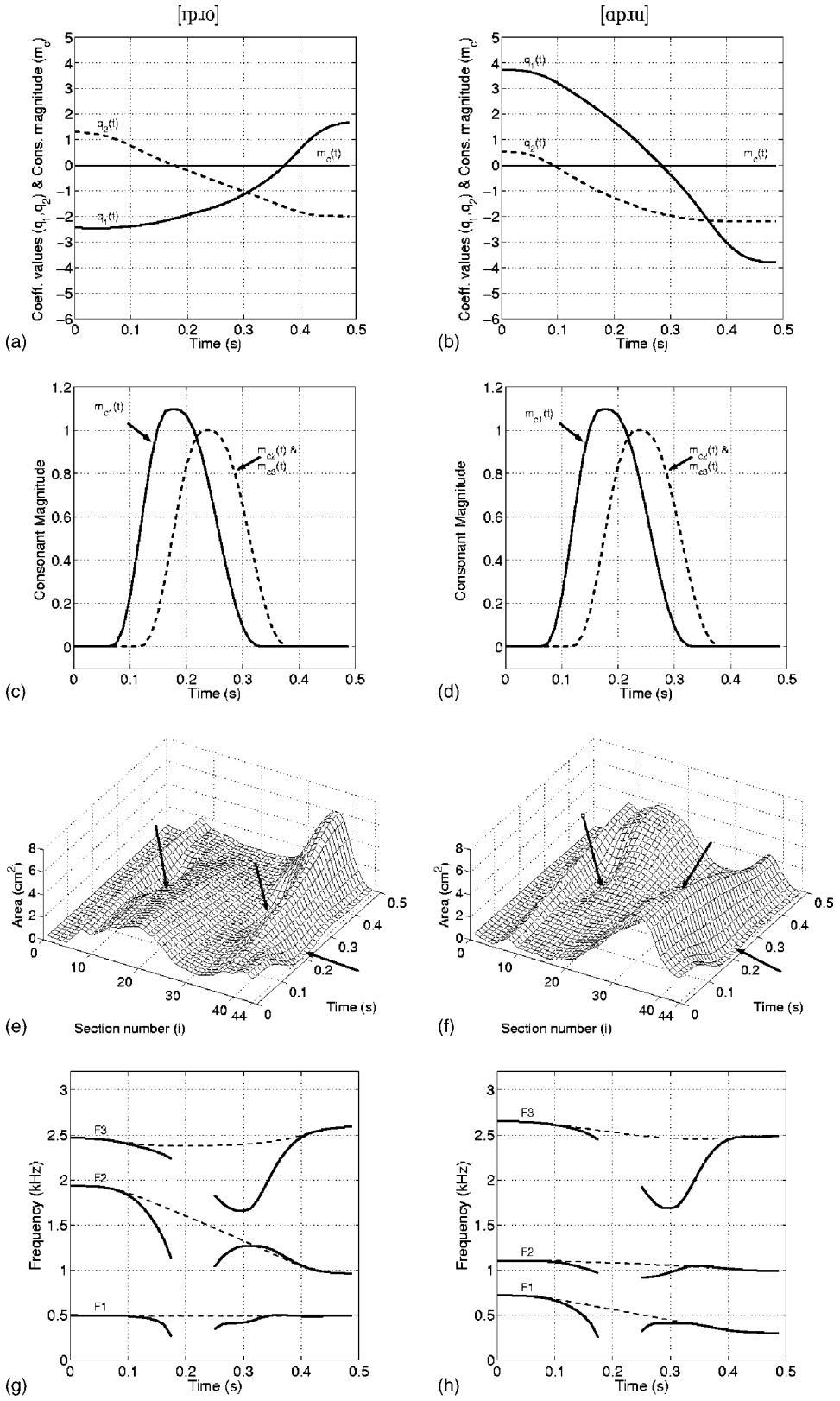

FIG. 18. Area function simulation for two VCCVs. In the left column the VCCV is approximately [ɪpɹo], and on the right, [apɹu]. Shown in descending order are the time variations of mode coefficients $q_1(t)$ and $q_2(t)$ [(a) and (b)], and consonant magnitudes for $m_{c1}(t)$, $m_{c2}(t)$, and $m_{c3}(t)$ [(c) and (d)]. Next are the time-varying area functions [(e) and (f)], where the three constrictions imposed during the time course of this utterance are indicated by arrows. Finally, the time-varying formant frequencies are shown at the bottom of the figure [(g) and (h)].

from [ɪ] to [o] was generated that had a duration of 0.5 s and followed a time course dictated by the same mode coefficients as in the previous cases. Length variations were imposed by having $p_g(t)$ and $p_m(t)$ in Eqs. (11) and (12) follow the time courses shown in Fig. 19(a). Initially, $p_g(t)$ was set to -0.5 cm to simulate a shortening at the glottal end of the vocal tract and then was increased to $+0.5$ cm. Similarly, at the lip end $p_m(t)$ was set to initially generate a 1-cm decrease in length, followed by the same amount of increase to lengthen the tract. Figure 19(b) shows the length of each section (tubelet) over the time course of the utterance. Most of the 44 sections are maintained at a constant length throughout, with the exception of the glottal end (near section 1) and lip end (near section 44), which are first short-

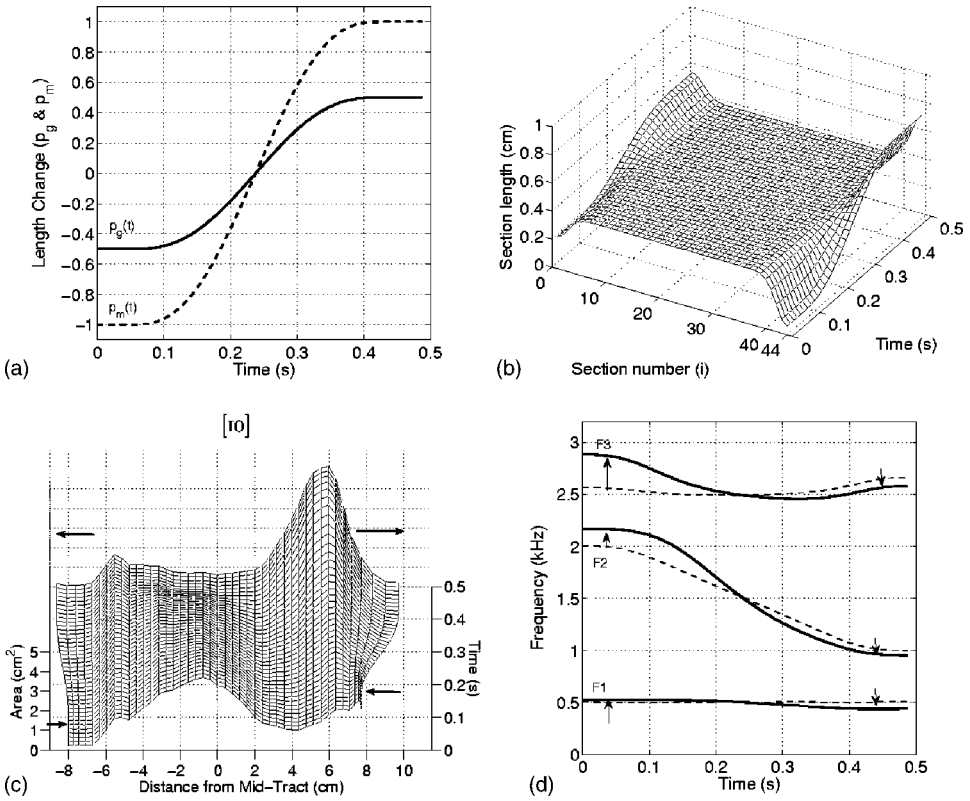

FIG. 19. Area function simulation for a VV transition with simultaneous vocal-tract length change. (a) Time course of the length change specified near the glottis by $p_g(t)$ and near the lips by $p_m(t)$. (b) Time-varying length function [see $L(i,t)$ in Eq. (13)]. (c) Time-varying area function approximating a transition from [ɪ] to [o] with the length changes included. The x axis indicates the distance from the middle of the vocal tract. (d) Time-varying formant frequencies for conditions with the length changes imposed (solid line) and for constant vocal-tract length (dashed line).

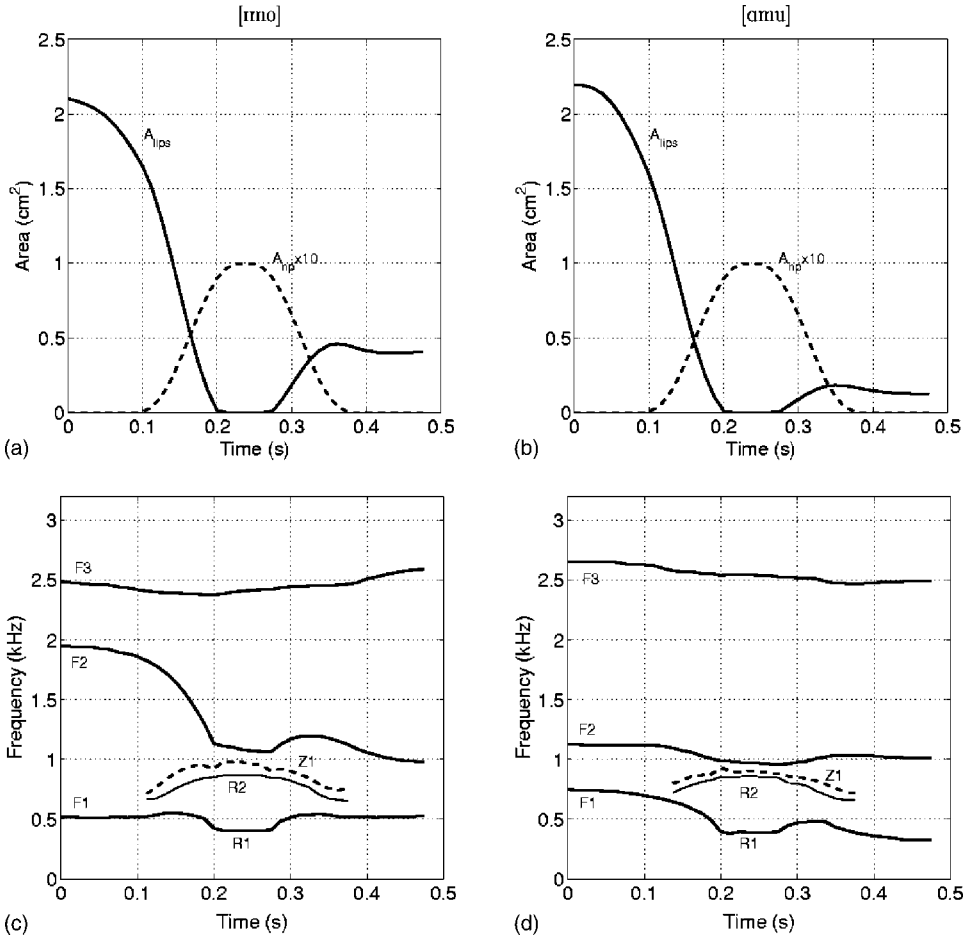

FIG. 20. Simulation of two VCVs with a nasal consonant. The upper plots show the time course of the area change at the lip section (section N_{vt}) (solid) and the nasal coupling area a_{np}, for (a) [ɪmo] and (b) [ɑmu]. The lower plots show the acoustic characteristics for each VCV.

ened and then lengthened over time. The resulting time-varying area function is shown in Fig. 19(c), where the view angle has been set so that the length variations can be observed. The x axis indicates the distance from the middle of the vocal tract, as was used in Fig. 6.

Formant frequencies were determined in the same manner as the previous simulations, except that the length variations were included in the calculations. Resulting time-varying formants are plotted in Fig. 19(d); for comparison, the dashed lines represent formants calculated in the case of constant length. In the initial 0.2 s of the utterance, the shortened tract creates an expected increase in all three formants. At 0.24 s, both $p_g(t)$ and $p_m(t)$ pass through zero, hence, the formants at this point are the same as in the constant length case. As $p_g(t)$ and $p_m(t)$ increase, the three formants are observed to decrease, as expected.

E. VCV simulations with a nasal consonant

The final two simulations were produced with the bilabial nasal consonant [m] imposed on the same [ɪo] and [ɑo] vowel contexts used in the previous examples. The consonant parameters were set to be identical to those for the [ɪpo] and [ɑpu] examples (see Fig. 12), except that the nasal coupling area a_{np} was allowed to be nonzero so that sound could be coupled to the nasal passages. The time variation of a_{np} is shown in Figs. 20(a) and (b), along with the area variation of the lip termination section in the main vocal tract (i.e., area of section N_{vt}). The nasal coupling area begins to increase from 0 cm^2 prior to the occlusion of the vocal tract, and remains nonzero until after the occlusion is released. The maximum area of a_{np} was equal to 0.1 cm^2, but to show the lip area on the same plot, a_{np} was multiplied by 10. The combined resonances of the vocal and nasal tracts, over the time course of the two utterances, are shown in Figs. 20(c) and (d). At the beginning of both [ɪmo] and [ɑmu], only the vocal tract formants ($F1$, $F2$, and $F3$) are present. As soon as a_{np} becomes nonzero, resonances (poles) $R1$ and $R2$ emerge, along with a zero, labeled $Z1$ (notation for the poles and zero during the nasal consonant are the same as in Stevens, 1998). During the time period when the lip section is occluded (from 0.2 to 0.27 s), $R1$ is approximately 0.4 kHz, $R2$ is 0.85 kHz, and $Z1$ is nearly 0.9 kHz. These are roughly in line with expected values for a bilabial nasal consonant (Stevens, 1998).

V. DISCUSSION

The results presented in Secs. III and IV demonstrate that the four-tier parametric model is capable of producing area functions representative of a wide variety of vowels and consonants, in both static and time-varying situations. By itself, the model could be used for studying the relation between the vocal-tract shape and formant frequencies. The separation of the input parameters for the vowel substrate, consonant constriction(s), length variation, and nasalization allows for controlled studies of the effect of each tier alone, or simultaneously. Further, the ''base structural components'' that were specified in Fig. 1 are the only part of the model that is speaker specific. Hence, as data like those given in Appendix A become available for additional speakers, these could be used as the structural components, and the model would represent a different speaker, at least with regard to the underlying physical structure.

As mentioned in the Introduction, the eventual goal of this type of modeling is not only to generate area functions, but to synthesize connected speech. That is, to create an acoustic waveform that would result from wave propagation through the time-varying vocal tract. This requires coupling computational models that simulate both the voice source and propagation of acoustic waves, with the parametric area function model presented here that would specify the vocal-tract shape at any specific point in time. This type of synthesis could provide stimuli to test the effect of various vocal-tract parameters on the perception of speech. For example, the effect of constriction location on the identification of stop consonants could be tested by presenting to listeners, a series of audio samples synthesized with successive changes in l_c (cf. Li and Story, 2003). Through such experiments a connection may be made between vocal-tract shape, acoustics, and perception, rather than just the latter two, which is typical of many speech perception experiments [although this is similar to Stevens' (1989) approach to speech production and perception].

At this point, there are still some significant limitations of the area function model. In particular, specification of the appropriate time course for the parameters is not well established, especially for the parameters in tier II, III, and IV. A next step is to pursue studies that could result in development of mappings between model parameters and acoustic characteristics, similar to that already developed for tier I (see Fig. 10), but would include the parameters for constriction formation, length change, and nasalization. The precise nature of the time variation of the model parameters is also likely to be speaker specific. Hence, the development of such mappings would need to be based on the speech of many speakers.

ACKNOWLEDGMENTS

A preliminary version of this paper was presented at the 146th Meeting of the Acoustical Society of America. This work was supported by NIH R01-DC04789.

NOMENCLATURE Independent variables

N_{vt}	number of x-sect. areas contained in the vocal-tract area function
N_{nt}	number of x-sect. areas contained in the nasal tract area function
N_c	number of consonantal functions
i	index of x-sect areas in an area function of the vocal tract $i=[1,N_{vt}]$
j	index of x-sect areas in an area function of the nasal tract $j=[1,N_{nt}]$
k	index of consonant superposition functions $k=[1,N_c]$
t	time

Base structural components

$\Omega(i)$	mean vocal-tract shape (diameter function)

$\phi_1(i), \phi_2(i)$ area function *modes* (basis functions)
$\mathcal{L}(i)$ nominal (default) length vector (length of each tubelet in an area function)
$\mathcal{N}(j)$ nasal tract area function

Time-varying control parameters

$q_1(t), q_2(t)$ amplitude coefficients for the basis functions
$l_{c_k}(t)$ location (place) of the kth consonant constriction specified as distance from the glottis
$a_{c_k}(t)$ minimum cross-sectional area of the kth consonant constriction
$r_{c_k}(t)$ range of the kth consonant superposition function along the tract length
$s_{c_k}(t)$ skewing quotient of consonant superposition function
$m_{c_k}(t)$ magnitude (activation) of the consonant superposition function
$p_g(t)$ amount of length change at the ''glottal'' end of the vocal tract
$l_g(t)$ location within the length vector $\mathcal{L}(i)$ where the length change p_g is centered
$r_g(t)$ range of the length change p_g
$p_m(t)$ amount of length change at the ''lip'' end of the vocal tract
$l_m(t)$ location within the length vector $\mathcal{L}(i)$ where the length change p_m is centered
$r_m(t)$ range of the length change p_m
$a_{np}(t)$ cross-sectional area of nasal coupling port

Intermediate outputs

$V(i,t)$ time-varying vowel substrate
$C_k(i,t)$ kth time-varying consonantal overlay
$\alpha(i,t)$ vocal tract length modification function for the ''glottal'' end
$\beta(i,t)$ vocal tract length modification function for the ''lip'' end

Final outputs

$A(i,t)$ composite vocal tract area function
$L(i,t)$ composite length (of each tubelet) function
$\mathcal{X}(i,t)$ cumulative length function
$A_n(j,t)$ composite nasal tract area function

APPENDIX A: TIER I STRUCTURAL COMPONENTS

The structural components for the vowel substrate are shown numerically in Table III. The first column is the index i, which denotes successive sections along the length of the vocal tract. Section 1 is located just above the glottis and section 44 at the lips. The other three columns are the neutral diameter function $\Omega(i)$, and the two modes, $\phi_1(i)$ and $\phi_2(i)$, that deform the vocal tract. When multiplied by the corresponding mode and combined with $\Omega(i)$ as specified in Eq. (3), the coefficients in Table IV will produce area functions representative of the indicated vowels.

These data were derived from the vowel area functions reported in Story, Titze, and Hoffman (1996). Each area function was first normalized to the mean vocal-tract length across the vowels, and then a principal components analysis (PCA) was applied to area function set. The ''modes'' are the

TABLE III. Neutral diameter function and two modes that form the vowel substrate in tier I of the area function model. i is an index extending from glottis to lips.

i	$\omega(i)$	$\phi_1(i)$	$\phi_2(i)$
1	0.636	0.018	−0.013
2	0.561	0.001	−0.007
3	0.561	−0.013	−0.029
4	0.550	−0.025	−0.059
5	0.598	−0.036	−0.088
6	0.895	−0.048	−0.108
7	1.187	−0.062	−0.120
8	1.417	−0.076	−0.123
9	1.380	−0.093	−0.118
10	1.273	−0.111	−0.107
11	1.340	−0.130	−0.092
12	1.399	−0.149	−0.075
13	1.433	−0.167	−0.056
14	1.506	−0.183	−0.035
15	1.493	−0.196	−0.014
16	1.473	−0.204	0.008
17	1.499	−0.207	0.032
18	1.529	−0.203	0.057
19	1.567	−0.193	0.084
20	1.601	−0.175	0.111
21	1.591	−0.151	0.138
22	1.547	−0.119	0.164
23	1.570	−0.082	0.188
24	1.546	−0.041	0.206
25	1.532	0.004	0.218
26	1.496	0.051	0.221
27	1.429	0.097	0.214
28	1.425	0.141	0.195
29	1.496	0.181	0.164
30	1.608	0.214	0.121
31	1.668	0.240	0.070
32	1.757	0.257	0.013
33	1.842	0.264	−0.046
34	1.983	0.260	−0.100
35	2.073	0.246	−0.143
36	2.123	0.224	−0.167
37	2.194	0.194	−0.165
38	2.175	0.159	−0.132
39	2.009	0.122	−0.066
40	1.785	0.087	0.031
41	1.675	0.057	0.148
42	1.539	0.038	0.264
43	1.405	0.034	0.346
44	1.312	0.048	0.338

TABLE IV. Mode coefficients that reconstruct the indicated vowels when used with Eq. (3).

Vowel	q_1	q_2
i	−5.176	−2.556
ɪ	−1.092	0.677
ɛ	2.561	3.849
æ	3.471	1.729
ʌ	0.102	−3.565
ɑ	0.981	1.043
ɔ	0.331	2.315
ʊ	0.128	1.356
o	−0.493	−1.910
u	−2.685	−2.065

TABLE V. Area functions for the fricative consonants [f, θ, s, ʃ]. *i* is an index that numbers cross-sectional areas from glottis to lips. The last line indicates the length of each tubelet within the area function.

i	f	θ	s	ʃ
1	0.37	0.39	0.47	0.49
2	0.32	0.40	0.47	0.49
3	0.65	0.76	0.54	0.63
4	1.22	0.71	0.68	0.64
5	1.46	1.10	0.94	0.83
6	1.43	1.50	2.23	2.26
7	1.34	1.34	2.18	2.51
8	1.32	1.23	2.39	2.53
9	1.84	1.15	2.17	2.77
10	2.55	1.34	1.95	2.85
11	2.21	1.72	2.12	3.11
12	2.16	1.65	2.19	3.42
13	2.09	1.52	2.61	3.31
14	1.82	1.50	3.08	3.69
15	2.01	1.32	2.51	3.54
16	1.95	1.42	2.32	3.74
17	1.91	1.49	2.70	4.18
18	1.77	1.49	2.65	4.56
19	1.88	1.94	3.21	4.98
20	2.40	2.41	3.71	5.23
21	2.04	2.29	3.72	5.18
22	1.87	2.37	4.14	5.80
23	2.13	2.99	4.58	5.67
24	2.05	3.09	4.63	4.82
25	1.65	2.98	4.32	4.39
26	1.05	2.75	4.38	3.82
27	0.64	2.82	4.21	2.89
28	0.52	2.96	4.52	2.48
29	0.45	2.95	4.53	1.81
30	0.45	3.13	4.40	1.38
31	0.29	3.24	3.98	0.98
32	0.24	3.13	3.78	0.57
33	0.16	3.36	3.52	0.45
34	0.21	3.43	3.20	0.34
35	1.04	3.21	2.77	0.03
36	1.88	3.08	2.16	0.00
37	2.27	2.82	1.53	0.00
38	2.00	2.39	1.03	0.00
39	1.93	1.75	0.60	0.34
40	1.89	0.90	0.26	2.13
41	1.31	0.09	0.11	1.96
42	0.83	0.00	0.15	1.24
43	0.33	0.09	0.24	0.76
44	0.11	0.22	0.06	0.33
Δ	0.361	0.356	0.353	0.377

two most significant components (in terms of the explained variance) resulting from the PCA (Story and Titze, 1998).

APPENDIX B: AREA FUNCTIONS FOR FRICATIVE CONSONANTS

Four fricative area functions are provided in Table V. These were derived from image sets obtained with magnetic resonance imaging (MRI). These images were collected at the same time as those reported in Story *et al.* (1996), but were not published. The image analysis and determination of the area function were carried out using exactly the same techniques as the previous study. A problem is that the images were acquired in the axial plane. Hence, the detailed structure of the constricted air channels were not well defined, as they may have been with, say, a coronal imaging plane (e.g., Narayan *et al.*, 1995). Thus, they are not adequate to provide information for aerodynamic modeling of turbulence generation (Shadle, 1991; Sinder, 1999), but they do indicate the types of shape information needed for the present area function model.

[1]The term ''tiers'' is used to denote multiple levels of area function control. This is similar, but not identical, to the use of the term by Browman and Goldstein (1990).

[2]Representation of an area function by 44 elements is not a requirement of the general structure of the model. The number derives from the approximate spatial resolution obtained in MRI-based reconstructions of vocal-tract shape (Story *et al.*, 1996). It is also convenient to use 44 elements for simulating speech with acoustic waveguide models because it allows for a sampling frequency of 44.1 kHz when the tract length is approximately 17.5 cm (typical adult male). The formulation of the model, however, is general enough that area functions with any number of elements can be produced.

[3]The symbol Ω is used as a *mathematical* variable describing the ''neutral'' diameter function. It should not be confused with the rarely used *phonetic* symbol [Ω] for a midback rounded vowel (Pullum and Ladusaw, 1996).

[4]It is acknowledged that the details of fricative sound production will not be adequately modeled by a simple constriction of the type proposed here. Rather, it is assumed that a model of turbulent noise generation (e.g., Narayanan, 1995; Sinder, 1999) could be inserted just downstream of the point of minimum area.

[5]For the set of male vocal-tract data used in this paper, $\Delta = 0.396\,825$ cm. Combined with a 44-section area function, the vocal-tract length will be 17.46 cm. It is not suggested that this level of accuracy is required for the tubelet length. Rather, this number originates from the particular type of wave propagation algorithm used by author to synthesize speech (Liljencrants, 1985; Story, 1995), where Δ = speed of sound(c)/2 × sampling frequency (F_s). Choosing $c = 35\,000$ cm/s and $F_s = 44\,100$ Hz, $\Delta = 0.396\,825$.

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

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

Badin, P., Bailly, G. Raybaudi, M., and Segebarth, C. (**1998**). ''A three-dimensional linear articulatory model based on MRI data,'' in Proceedings of the 5th International Conf. on Spoken Language Proc., edited by R. H. Mannell and J. Robert-Ribes 2, 417–420.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C. (**2002**). ''Three-dimensional linear articulatory modeling of tongue, lips, and face, based on MRI and video images,'' J. Phonetics **30**, 533–553.

Båvegård, M. (**1995**). ''Introducing a parametric consonantal model to the articulatory speech synthesizer,'' in Proceedings Eurospeech 95, Madrid, Spain, 1857–1860.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). ''Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels,'' J. Acoust. Soc. Am. **90**, 799–828.

Browman, C., and Goldstein, L. (**1990**). ''Gestural specification using dynamically defined articulatory structures,'' J. Phonetics **18**, 299–320.

Carré, R., and Chennoukh, S. (**1995**). ''Vowel–consonant–vowel modeling by superposition of consonant closure on vowel-to-vowel gestures,'' J. Phonetics **23**, 231–241.

Coker, C. H. (**1976**). ''A model of articulatory dynamics and control,'' Proc. IEEE **64**(4), 452–460.

Dang, J., and Honda, K. (**1994**). ''Morphological and acoustical analysis of the nasal and the paranasal cavities,'' J. Acoust. Soc. Am. **96**, 2088–2100.

Dang, J., and Honda, K. (**2004**). ''Construction and control of a physiological articulatory model,'' J. Acoust. Soc. Am. **115**, 853–870.

Espy-Wilson, C. Y. (**1992**). ''Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English,'' J. Acoust. Soc. Am. **92**, 736–757.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., and Alwan, A. (**2000**). ''Acoustic modeling of American English /r/,'' J. Acoust. Soc. Am. **108**, 343–356.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).
Fitch, W. T., and Giedd, J. (**1999**). ''Morphology and development of the human vocal tract: A study using magnetic resonance imaging,'' J. Acoust. Soc. Am. **106**, 1511–1522.
Fujimura, O. (**1992**). ''Phonology and phonetics—A syllable based model of articulatory organization,'' J. Acoust. Soc. Jpn. **13**, 39–48.
Goldstein, U. G. (**1980**). ''An articulatory model for the vocal tracts of growing children,'' Doctoral dissertation, Department of Electrical Engineering and Computer Science, MIT.
Hogan, N. (**1982**). ''An organizing principle for a class of voluntary movements,'' J. Neurosci. **4**(11), 2745–2754.
Ichikawa, A., and Nakata, K. (**1968**). ''Speech synthesis by rule,'' Reports of the 6th International Congress on Acoustics, edited by Y. Kohasi, Tokyo, International Council of Scientific Unions, 171–1744.
Ishizaka, K., and Flanagan, J. L. (**1972**). ''Synthesis of voiced sounds from a two-mass model of the vocal cords,'' Bell Syst. Tech. J. **51**, 1233–1268.
Kozhevnikov, V. A., and Chistovich, L. A. (**1965**). *Speech: Articulation and Perception* (trans. US Dept. of Commerce, Clearing House for Federal Scientific and Technical Information), No. 30, 543 (Joint Publications Research Service, Washington, D.C.).
Lee, S., Potamianos, A., and Narayanan, S. (**1999**). ''Acoustics of children's speech: Developmental changes of spectral and temporal parameters,'' J. Acoust. Soc. Am. **105**(3), 1455–1468.
Li, K., and Story, B. H. (**2003**). ''An investigation of perceptual tolerance limits of stop constriction regions along the vocal tract,'' J. Acoust. Soc. Am. **114(4) pt 2**, 2337.
Liljencrants, J. (**1985**). ''Speech synthesis with a reflection-type line analog,'' DS dissertation, Dept. of Speech Comm. and Music Acoust., Royal Inst. of Tech., Stockholm, Sweden.
Lin, Q. (**1990**). ''Speech Production Theory and Articulatory Speech Synthesis,'' Doctoral dissertation, Royal Inst. of Tech. (KTH), Stockholm.
Lindblom, B., and Sundberg, J. (**1971**). ''Acoustical consequences of lip, tongue, jaw, and larynx movement,'' J. Acoust. Soc. Am. **4**(2), 1166–1179.
Maeda, S. (**1982**). ''The role of the sinus cavities in the production of nasal vowels,'' in IEEE Proceedings, 911–914.
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 W. J. Hardcastle and A. Marchal (Kluwer Academic, Dordrecht), pp. 131–149.
Mattingly, I. G. (**1974**). ''Speech synthesis for phonetic and phonological models,'' in *Current Trends in Linguistics*, edited by T. A. Sebeok (Mouton, The Hague), Vol. 12, pp. 2451–2487.
Mermelstein, P. (**1973**). ''Articulatory model for the study of speech production,'' J. Acoust. Soc. Am. **53**(4), 1070–1082.
Mrayati, M., Carré, R., and Guérin, B. (**1988**). ''Distinctive regions and modes: A new theory of speech production,'' Speech Commun. **7**, 257–286.
Nakata, K., and Mitsuoka, T. (**1965**). ''Phonemic transformation and control aspects of synthesis of connected speech,'' J. Radio Res. Labs. **12**, 171–186.
Narayanan, S. S. (**1995**). ''Fricative consonants: An articulatory, acoustic, and systems study,'' Ph.D. thesis, UCLA, Los Angeles, CA.
Narayanan, S., Alwan, A., and Haker, K. (**1997**). ''Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. I. The laterals,'' J. Acoust. Soc. Am. **101**, 1064–1077.
Narayanan, S. S., and Alwan, A. A. (**1995**). ''An articulatory study of fricative consonants using magnetic resonance imaging,'' J. Acoust. Soc. Am. **98**, 1325–1347.
Nelson, W. L. (**1983**). ''Physical principles of economies of skilled movements,'' Biol. Cybern. **46**, 135–147.
Nordström, P-E. (**1977**). ''Female and infant vocal tracts simulated from male area functions,'' J. Phonetics **5**, 81–92.
Öhman, S. E. G. (**1966**). ''Coarticulation in VCV utterances: Spectrographic measurements,'' J. Acoust. Soc. Am. **39**, 151–168.
Öhman, S. E. G. (**1967**). ''Numerical model of coarticulation,'' J. Acoust. Soc. Am. **41**, 310–320.
Peterson, G. E., and Barney, H. L. (**1952**). ''Control methods used in a study of the vowels,'' J. Acoust. Soc. Am. **24**(2), 175–184.
Pullum, G. K., and Ladusaw, W. A. (**1996**). *Phonetic Symbol Guide* (University of Chicago Press, Chicago), p. 148.
Shadle, C. H. (**1991**). ''The effect of geometry on source mechanisms of fricative consonants,'' J. Acoust. Soc. Am. **19**, 409–424.
Shadle, C. H., and Damper, R. I. (**2001**). ''Prospects for Articulatory Synthesis: A Position Paper,'' in Proceedings of 4th ISCA Workshop on Speech Synthesis, August/September 2001, 121–126, Pitlochry, Scotland.
Sinder, D. J. (**1999**). ''Speech synthesis using an aeroacoustic fricative model,'' Doctoral dissertation, Rutgers University, New Brunswick, NJ.
Sondhi, M. M. (**2002**). ''Articulatory modeling: A possible role in concatenative text-to-speech synthesis,'' Proceedings of IEEE Workshop on Speech Synthesis, 73–78.
Sondhi, M. M., and Schroeter, J. (**1987**). ''A hybrid time–frequency domain articulatory speech synthesizer,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**(7), 955–967.
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).
Stevens, K. N., and House, A. S. (**1955**). ''Development of a quantitative description of vowel articulation,'' J. Acoust. Soc. Am. **27**(3), 484–493.
Story, B. H. (**1995**). ''Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract,'' Ph.D. dissertation, University of Iowa.
Story, B. H. (**2002**). ''A parametric area function model of three female vocal tracts based on orthogonal modes,'' Presented at the Acoustical Society Meeting, J. Acoust. Soc. Am. **112(5)**, pt. 2, 2418.
Story, B. H. (**2004**). ''Vowel acoustics for speaking and singing,'' Acust. Acta Acust. **90**(4), 629–640.
Story, B. H., and Titze, I. R. (**1998**). ''Parametrization of vocal tract area functions by empirical orthogonal modes,'' J. Phonetics **26**(3), 223–260.
Story, B. H., and Titze, I. R. (**2002**). ''A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function,'' J. Phonetics **30**, 485–509.
Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). ''Vocal tract area functions from magnetic resonance imaging,'' J. Acoust. Soc. Am. **100**(1), 537–554.
Story, B. H., Titze, I. R., and Hoffman, E. A. (**2001**). ''The relationship of vocal tract shape to three voice qualities,'' J. Acoust. Soc. Am. **109**, 1651–1667.
Taylor, J. R. (**1982**). *An Introduction to Error Analysis* (University Science, Mill Valley, CA).
Titze, I. R., and Story, B. H. (**1997**). ''Acoustic interactions of the voice source with the lower vocal tract,'' J. Acoust. Soc. Am. **101**(4), 2234–2243.