Speech Communication 87 (2017) 1–17

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

An acoustically-driven vocal tract model for stop consonant production

Brad H. Story*, Kate Bunton

Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences, University of Arizona, P.O. Box 210071, Tucson, AZ 85721, United States

ARTICLE INFO

Article history:
Received 29 July 2016
Revised 8 November 2016
Accepted 8 December 2016
Available online 9 December 2016

Keywords:
Vocal tract
Speech modeling
Area function
Formant
Resonance
Speech synthesis

ABSTRACT

The purpose of this study was to further develop a multi-tier model of the vocal tract area function in which the modulations of shape to produce speech are generated by the product of a vowel substrate and a consonant superposition function. The new approach consists of specifying input parameters for a target consonant as a set of directional changes in the resonance frequencies of the vowel substrate. Using calculations of acoustic sensitivity functions, these "resonance deflection patterns" are transformed into time-varying deformations of the vocal tract shape without any direct specification of location or extent of the consonant constriction along the vocal tract. The configuration of the constrictions and expansions that are generated by this process were shown to be physiologically-realistic and produce speech sounds that are easily identifiable as the target consonants. This model is a useful enhancement for area function-based synthesis and can serve as a tool for understanding how the vocal tract is shaped by a talker during speech production.

1. Introduction

The human vocal tract can be modeled as a tubular conduit of varying cross-sectional area extending from glottis to lips. During speech production the shape of the conduit is precisely modulated over time, shifting the acoustic resonances of the system upward and downward in frequency to form a pattern of phonetic information. Although a tubular representation may appear to be a simple structure, understanding how the acoustic resonances are controlled through modulation of the vocal tract shape is a complex problem. Even a slight constriction or expansion along the tract length may affect multiple resonances, and may be enacted by a talker to facilitate a change in the frequency of some resonances while inhibiting change in others. It is, of course, additionally complicated by the fact that the modulation of the vocal tract shape by a talker is produced by the movement of the speech articulators (i.e., tongue, jaw, lips, larynx, and velum), and so any pattern of deformation imposed on the tubular conduit model must be physiologically attainable.

A tubular representation of the vocal tract can be expressed mathematically as an *area function* where cross-sectional area is specified as a function of the distance from the glottis (or lips, depending on the desired coordinate system). Various models have been proposed in which the area function is controlled by a small set of physiologically-relevant parameters whose values may be static, say for a sustained vowel configuration, or interpolated over some temporal duration to generate a time-dependent vocal tract shape such as a vowel-to-vowel (VV) transition. For example, the "three parameter" models of Stevens and House (1955) and Fant (1960) operated by specifying the location and cross-sectional area of a primary constriction, along with the ratio of lip opening length divided by its area. Based on either of their respective mathematical formulations, a given set of the three parameter values would determine the shape of the area function. Although capable of generating reasonable configurations for vowels or time-dependent VVs, such models did not include the capability necessary for representing the coarticulatory effects of a continuous stream of vowels and consonants.

A different type of vocal tract model was largely motivated by Öhman's (1963, 1966) spectrographic analyses of vowel-consonant-vowel (VCV) utterances that suggested a VCV should not be regarded as a linear sequence of successive, independent vocal tract gestures, but rather consists of a constrictive consonant gesture superimposed on an underlying vowel substrate (i.e, vowel-vowel transition). This view of speech production requires that a vocal tract model accommodates at least two channels of parametric instructions operating in parallel (Mattingly, 1981). The first channel would represent time-dependent vocal tract configurations for vowels, whereas the second channel may specify the spatial and temporal characteristics of a consonant superposition function re-

* Corresponding author.
E-mail address: bstory@email.arizona.edu (B.H. Story).

http://dx.doi.org/10.1016/j.specom.2016.12.001
0167-6393/

sulting in a sequence of vocal tract shapes containing the combined, and hence coarticulatory effects of both channels. Öhman (1963) alluded to such a model with regard to tubular representations of the vocal tract, and later refined the idea to allow for interpolation of the midsagittal cross-distance (width) of one vowel shape to another, while a consonant constriction simultaneously was activated over the same time course (Öhman, 1967). This work also influenced Nakata and Mitsuoka (1965) who proposed an area function model that incorporated separate vowel and consonant channels. With the intent to serve as a component in a text-to-speech synthesizer, theirs was a fairly elaborate system that included vowel and consonant target shapes combined with transition and coarticulation functions. Similarly, vocal tract area function models where consonant constrictions are superimposed on an interpolation of a vowel-to-vowel transition have been described, for example, by Båvegård (1995); Fant and Båvegård (1997), and Carré and Chennoukh (1995).

In the same vein, Story (2005a) proposed a model in which multiple tiers of airway modulation operate in parallel to produce a composite time-dependent configuration. In the lowest tier, transitions from one vowel to another can be generated by modulating a phonetically-neutral vocal tract configuration with shaping patterns that affect cross-sectional areas along the entire tract length. Mathematically, this vowel substrate can be expressed as a time-varying area function,

$$V(x,t) = \frac{\pi}{4}[\Omega(x) + q_1(t)\phi_1(x) + q_2(t)\phi_2(x)]^2 \quad (1)$$

where $\Omega(x)$, $\phi_1(x)$, and $\phi_2(x)$ are the mean (neutral) vocal tract diameter function and shaping patterns (referred to as "modes"), respectively, and x is the distance from the glottis. Time-dependence is produced by the coefficients $q_1(t)$ and $q_2(t)$, and the squaring operation and scaling factor of $\pi/4$ convert diameters to areas. Acoustically, the neutral shape supports a set of resonances whose frequencies are widely spaced (similar to a uniform tube), but also unique to a particular talker; thus, modulations of the neutral shape perturb the resonances upward or downward in frequency to generate formants appropriate for the desired vowels. An example $V(x, t)$ is shown in Fig. 1a, where the area function configuration is maintained as neutral (i.e., $q_1 = q_2 = 0$) for the initial 200 ms, and then transitions to an [ɑ]-like shape. The first three resonance frequencies calculated across the duration of $V(x, t)$ are shown in Fig. 1b and deviate as expected from those produced by the neutral configuration (dashed lines) during the period from 200 to 500 ms.

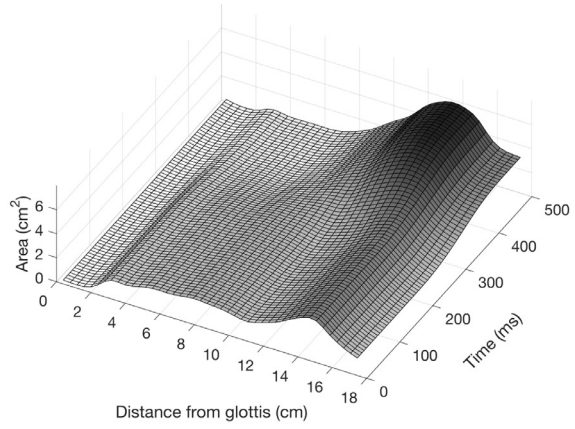
Obstruent-like sounds are produced by a second tier of modulation that imposes localized, and typically severe constrictions on the underlying vowel substrate by specifying constriction parameters such as location, cross-sectional area, and shape. An example consonant superposition function $C(x, t)$ with a constriction located 14.7 cm from the glottis is shown in Fig. 1c; it maintains a value of 1.0 everywhere except in the spatio-temporal region of the constriction. The product $V(x, t)C(x, t)$, carried out along the length of the vocal tract and at each time instant, is plotted in Fig. 1d, and the associated resonance frequencies are shown in Fig. 1e. The superposition of the constriction has the acoustic effect of redirecting the resonance frequencies established by the lower (vowel) tier, at least for a brief period during which the consonant tier is active. Thus, over the time course of an utterance, the modulations imposed by the two tiers generate a sequence of vocal tract area functions whose time-varying resonances effectively encode the phonetic influence of both vowels and consonants, but preserve a quality that is unique to a specific talker.

This model was later used to relate constriction location to perception of voiced stop consonants (Story and Bunton, 2010). Several series of simulated V_1CV_2 utterances, produced by time-varying area functions similar to the one shown in Fig. 1, were presented to listeners in which the constriction location for the consonant was incrementally shifted from the lip termination back to the velar region. It was found that identification of stops coincided with the directions in which the first three resonance frequencies of the vocal tract were deflected away from those resonance frequencies that would exist for the vowel substrate alone. For example, constrictions located anywhere within a 1.5–2.0 cm long section of the area function in the alveolar region (e.g., Fig. 1) produced deflection patterns at the onset of the consonant such that, relative to the underlying vowel, the first resonance frequency, f_{R1}[1], decreased, while the second and third resonances, f_{R2} and f_{R3}, both increased in frequency, regardless of the vowel configuration. This pattern robustly predicted that listeners would identify the consonant as the alveolar /d/, even though the absolute direction of change of the resonance frequencies (i.e., observable formant transitions) varied depending on the vowel context. Identification shifted from one consonant to another at constriction locations where the deflection pattern changed; for instance, when the constriction location was moved into the velar region the deflection pattern for the consonant onset shifted such that f_{R1} decreased, f_{R2} increased, and f_{R3} decreased, and this pattern was predominately associated with identification of. /g/.

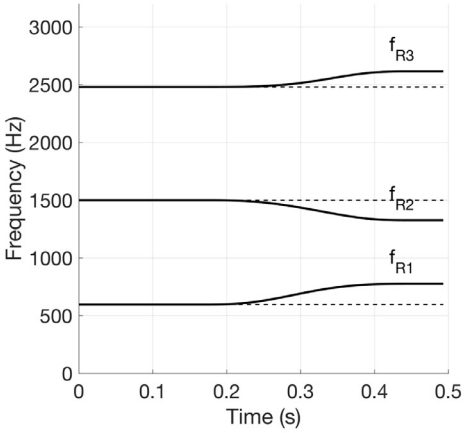
An inherent limitation with this model, however, as well as with the other models discussed previously, is the difficulty in accurately specifying the shape of consonant superposition functions. Perhaps the same might be said for the vowel tier; however, statistical analyses of area function sets measured for vowels have at least provided a reasonable means of estimating realistic vocal tract configurations for vowel-vowel transitions (Story and Titze, 1998; Story, 2005b; Mokhtari et al., 2007). The consonant function is more elusive to define because it doesn't exist as an independently measurable entity, rather it is a modifier of the vowel substrate. In the Story (2005a) model, the consonant function is specified by four parameters related to the constriction: location, cross-sectional area, range, and skew. The latter two parameters dictate the overall shape of the constriction, where "range" sets the amount of vocal tract length that is influenced by the constriction, and the "skewing" value allows the shaping to be asymmetric about constriction location if desired. Considering that bilabial and alveolar consonants are produced with the lips and tongue tip, respectively, the range setting for constrictions located in these regions would need to be relatively small, whereas a velar consonant presumably requires a larger range to account for the involvement of the tongue body. Story and Bunton (2010) did vary the range and skew settings relative to the constriction location to produce the simulations that were presented to listeners. This was fairly successful with regard to the experiment reported, but considering the less than 100% within-category identification of consonants produced in the velar regions, the quality of some of the samples could clearly be improved. Manual trial and error modification of the parameters can enhance the quality of consonants as shown for simulations of words and phrases (Story, 2013), as well as facilitate understanding how the components of a model relate to the output, but in general it is a laborious and inefficient process.

Certainly part of the limitation can be attributed to an insufficient knowledge of articulation in a wide range of phonetic contexts, and how that articulation shapes the vocal tract. Articulatory data and new insightful analyses of such data may be useful in this regard, but perhaps the problem is also rooted in an assumption inherent to any type of vocal tract modeling, whether based

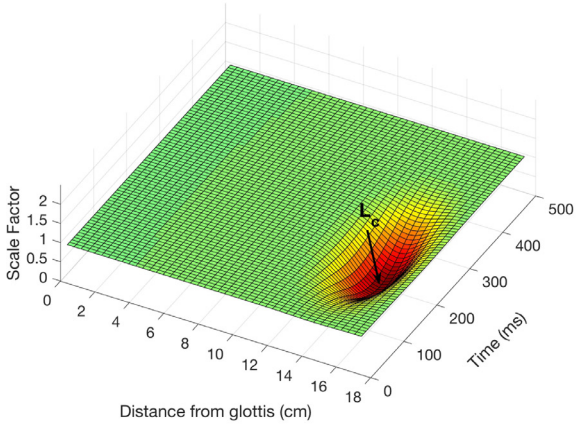
[1] According to the conventions recently proposed by Titze et al. (2015), the vocal tract resonance frequencies determined from a direct calculation of the frequency response are denoted as f_{Rn}, whereas *formant* frequencies measured from the acoustic signal by processing algorithms are denoted as *Fn*.



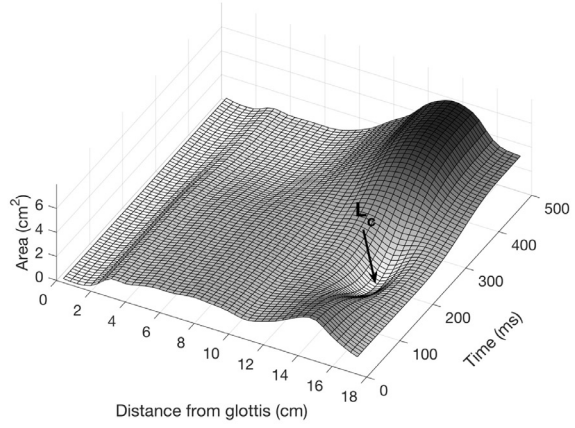
(a) $V(x,t)$



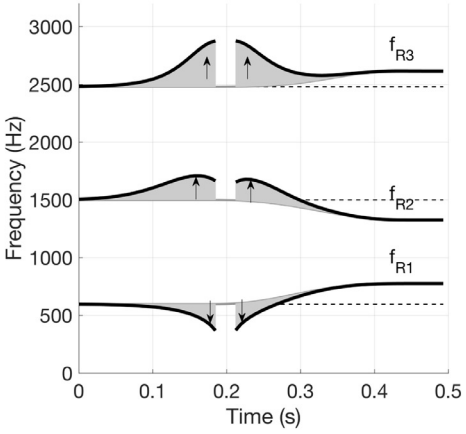
(b) Resonance frequencies of $V(x,t)$



(c) $C(x,t)$



(d) $V(x,t)C(x,t)$



(e) Resonance frequencies of $V(x,t)C(x,t)$

Fig. 1. Demonstration of the multi-tier model of the vocal tract area function and associated resonance frequencies. The 3D surfaces in the left column show how a consonant superposition function can be combined with a vowel-to-vowel (VV) transition to generate a time-varying area function representative of a VCV. The upper and lower panels in the right column are time-varying resonance frequencies calculated for the VV and VCV, respectively. (a) vowel substrate area function, (b) calculated resonance frequencies of the vowel substrate (solid), along with those of the neutral configuration (dashed) for reference, (c) consonant superposition function with constriction location $L_c = 14.7$ cm from the glottis, (d) composite time-varying area function when $C(x, t)$ is superimposed on $V(x, t)$, and (e) resonances frequencies of $V(x, t)C(x, t)$ (solid black), along with those of the vowel substrate (solid gray), and the neutral configuration (dashed).

on the articulators themselves or a derived quantity such as an area function. That is, most approaches to vocal tract modeling emphasize understanding how to generate articulatory movements or patterns of movements that produce a reasonable approximation of the acoustic characteristics observed in natural speech. This process is reminiscent of the first part of Stetson's well-known statement that that "Speech is rather a set of movements made audible... " (Stetson, 1951, p. 203). Benefit may also be derived, however, from considering the second part of Stetson's statement as an alternative view that speech may rather be "... a set of sounds produced by movement." This latter view points toward development of a model in which the vocal tract shape is controlled by acoustic parameters that encode the phonetic information to be transmitted, rather than by direct specification of spatial or geometrical parameters; i.e., the vocal tract shape is molded by the acoustic necessities of achieving a phonetic target.

To some degree the "distinctive region model" proposed by Mrayati et al. (1988) achieved this goal. They showed that if the vocal tract was represented as a tube of uniform cross section along its length, it could be divided into a series of eight regions that, when independently constricted or expanded, would predictably shift the first three resonance frequencies into a distinct pattern. For example, if the phonetic goal was to shift all three resonances downward in frequency, say to produce a bilabial consonant, the model would predict that the region nearest the lip termination should be constricted; other regions may be simultaneously constricted or expanded to enhance achievement of the phonetic goal. Although this model was criticized in its original form for being anthropomorphically weak, as well as for minimizing the limitations of the acoustic calculations used to derive the distinctive regions (Boë and Perrier, 1990), it did demonstrate that specification of targeted changes in resonance frequencies could be directly mapped to specific changes in configuration of a vocal tract area function. Both Carré (2004) and Story (2006) later developed iterative techniques that utilized acoustic sensitivity functions directly as vocal tract deformation patterns rather than as a means of dividing the tract length into distinctive regions. Carré's work focused on how perturbations of a uniform tubular conduit seem to give rise to the vocal tract shapes observed in speech articulation, whereas the goal in Story (2006) was in tuning measured vocal tract shapes such that their calculated resonance frequencies better matched the formant frequencies obtained from analysis of recorded speech. Adachi et al. (2007) and Kreuzer and Kasess (2015) formulated a similar iterative approaches but added the effects of vocal tract length perturbation and nasal branching, respectively, to the calculation of acoustic sensitivity functions.

The purpose of the present study was to utilize the link between acoustic sensitivity functions and vocal tract shape to develop a consonant superposition component of an area function model in which the shape is determined by specifying a resonance deflection pattern rather than geometrical parameters such as location and degree of constriction. That is, the direction and magnitude of change of f_{R1}, f_{R2}, and f_{R3} become the input parameters to the model. The geometrical properties of the consonant superposition function are then derived directly from the sensitivities of the resonances of a given vocal tract shape to perturbations of cross-sectional area along its extent from glottis to lips. When coupled with a time-dependent activation, the consonant superposition function gradually deforms the underlying vowel area function into a shape that shifts the first three resonances in the specified directions to produce the consonant onset and then gradually removes the deformation to release the consonant. The specific aims of the article are to: 1) describe the new components of the model and 2) demonstrate production of stop consonants embedded in four vowel-to-vowel contexts.

2. Method

The structure of the model used in this study was the same as described in Story (2005a); Story and Bunton (2010), and also in the Introduction, where a time-varying vocal tract area function is the product of vowel and consonant components such that,

$$A(x,t) = V(x,t)C(x,t). \tag{2}$$

Representing the distance from glottis to lips, the x dimension is discretized such that an area function at any given time instant consists of $N_x = 44$ contiguous "tubelet" sections each with a length of $L(i) = 0.396825$ cm, where i is the section number[2]. The value of x corresponding to the i^{th} section is then,

$$x(i) = \sum_{z=1}^{i} L(z) \tag{3}$$

and results, in this case, in an overall tract length of 17.46 cm. The time dimension t is also discretized such that at any time sample n within a duration of N_t samples, the composite area function can be written in matrix form as,

$$A(i,n) = V(i,n)C(i,n) \quad i = [1, N_x], \quad n = [1, N_t] \tag{4}$$

where the vowel substrate $V(i, n)$ is defined to be the same as Eq. (1), but in spatially and temporally discretized form. The sample interval[3] used for the area function model throughout this study was $T = 6.866$ ms.

The consonant function $C(i, n)$ developed in Story (2005a), and demonstrated in Fig. 1, consisted of geometrical and temporal parameters built into a Gaussian function that formed the shape of the constriction. In the new approach, the geometrical configuration of $C(i, n)$ is derived directly from the acoustic properties of $V(i, n)$ at each time sample. Thus, because $C(i, n)$ is essentially a function of the vowel substrate, it will be assumed in subsequent sections that a suitable $V(i, n)$, such as shown in Fig. 1a, has been generated with Eq. (1). Specific vowel substrates will be introduced in later sections to demonstrate the method.

2.1. Calculation of acoustic sensitivity functions

Calculation of the acoustic sensitivity of a specific vocal tract configuration is necessary for deriving the new form of $C(i, n)$. Considering an area vector for a vowel-like configuration at a single time instant, n, to be $V(i)$, the sensitivity of a particular resonance frequency to a change in vocal tract cross-sectional area can be defined as the difference between the kinetic energy (KE) and potential energy (PE) within each i^{th} section, divided by the total energy (TE) in the system (Fant and Pauli, 1975). A sensitivity function can be written as,

$$S_j(i) = \frac{KE_j(i) - PE_j(i)}{TE_j} \quad j = 1, 2, 3 \ldots \quad \text{and} \quad i = [1, N_x] \tag{5}$$

where j is the resonance number, and

$$TE_j = \sum_{i=1}^{N_x} [KE_j(i) + PE_j(i)]. \tag{6}$$

[2] The section lengths, $L(i)$, can be allowed to vary like the cross-sectional areas (Story, 2005a), but in this study they were maintained at a constant value of $L(i) = 0.396825$ cm. This particular length is derived from the algorithm used to calculate wave propagation in the vocal tract.

[3] $T = 6.866$ ms is the sampling interval of articulatory data in the University of Wisconsin-Madison X-ray microbeam database (XRMB). Because previous work by the authors has combined vocal tract modeling with XRMB data, the 6.866 ms interval has been adopted for efficiency.

The kinetic and potential energies for each resonance frequency are based on the pressure $P_j(i)$ and volume velocity $U_j(i)$ computed for each section of an area vector and are calculated as,

$$KE_j(i) = \frac{1}{2}\frac{\rho L(i)}{V(i)}|U_j(i)|^2 \tag{7}$$

and

$$PE_j(i) = \frac{1}{2}\frac{V(i)L(i)}{\rho c^2}|P_j(i)|^2 \tag{8}$$

where again $L(i)$ is the length of each tubelet section, ρ is the density of air, and c is the speed of sound.

Calculations of pressures, flows, and frequency response functions were accomplished with a transmission-line type model of the vocal tract (e.g., Sondhi and Schroeter, 1987; Story et al., 2000) that included energy losses due to yielding walls, viscosity, heat conduction, and acoustic radiation at the lips (see Appendix for details). As an example, the neutral area function obtained with $q_1 = q_2 = 0$ in Eq. (1) is shown in Fig. 2a, along with the calculated frequency response. It can be noted that the first three resonances are located at 600, 1504, and 2492 Hz, close to those of a uniform tube of similar overall length. The acoustic sensitivity functions calculated with Eqs. (5)–(8) for this vocal tract configuration are plotted in Fig. 2b. Each line indicates the relative sensitivity of the first, second, and third resonance frequencies (f_{R1}, f_{R2}, f_{R3}) to a small perturbation of the area function, $\Delta V(i)$. Mathematically, this is written as,

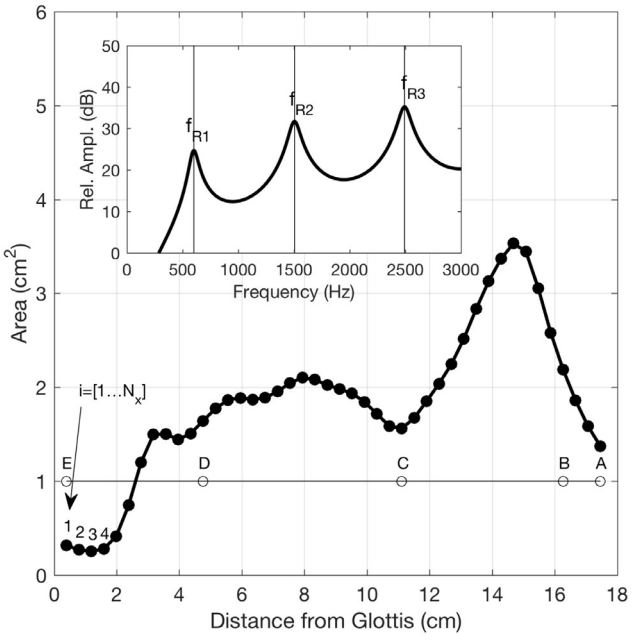
$$\frac{\Delta f_{Rj}}{f_{Rj}} = \sum_{i=1}^{N_x} S_j(i)\frac{\Delta V(i)}{V(i)} \tag{9}$$

where j is again the resonance number. This equation dictates that a positive or upward shift in the resonance frequency will occur when a positive change in area, $\Delta V(i) > 0$, is imposed at values of i where $S_j(i) > 0$, or when a negative change in area, $\Delta V(i) < 0$, is imposed where $S_j(i) < 0$; the opposite shift in resonance frequency occurs if the polarities of $\Delta V(i)$ and $S_j(i)$ oppose each other.

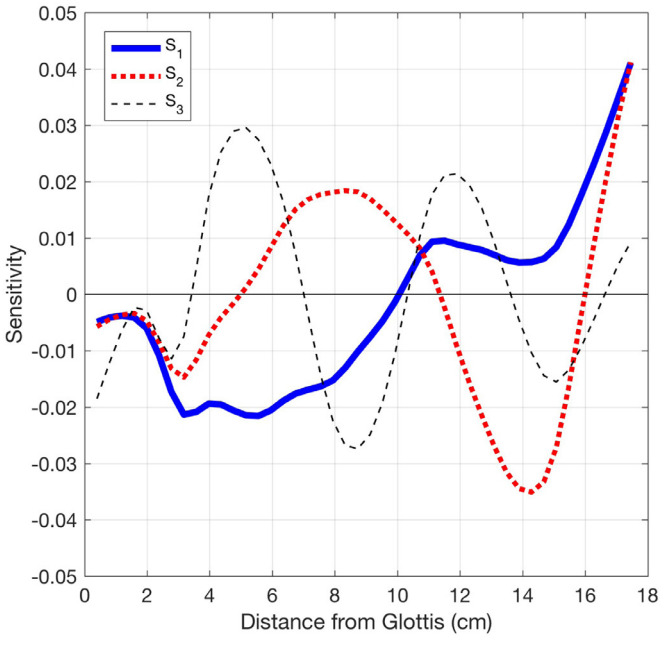
It can be noted that an alternative to the energy-based approach is to calculate the sensitivity functions with a Jacobian formulation (Kreuzer and Kasess, 2015). The advantage is that both the frequency and bandwidth of each vocal tract resonance enter into the calculation which may be particularly useful for enhancing modification of tract configurations that include branching into the nasal passages or subglottal airways. For purposes of the present study, which includes only vowels and stop consonants, the energy-based approach was deemed sufficient, but the Jacobian-based alternative could be implemented in future iterations of the model, especially for vocal tract configurations with greater complexity.

Based on S_1 in Fig. 2b, and using Eq. (9) as a guide, it is observed that f_{R1} would be increased by expanding the area along the vocal tract length from a location 10 cm from the glottis to the lip termination. The first resonance could also be increased in frequency by constricting the region from the glottis to the l0 cm location. Lowering f_{R1} would require the opposite changes in area within the same regions. Based on S_2, an increase in f_{R2} could be produced by expanding the regions between 5–11 cm and from 16 cm to the lip termination, as well as constricting the portions that extend from 0 to 5 cm and 11–16 cm; lowering f_{R2} would require the opposite changes in area. Changes in f_{R3} could be similarly carried out by modifying cross-sectional areas in the positively and negatively valued regions specified by S_3. Sensitivity functions corresponding to higher frequency resonances can also be calculated, but were not used in this study.

Although sensitivity functions can guide manual adjustments of an area function that shift resonance frequencies in specified directions (cf., Mrayati et al., 1988; Story et al., 2001), a more direct



(a)



(b)

Fig. 2. Demonstration of acoustic sensitivity functions for a vocal tract configuration. (a) Area function for a neutral vocal tract shape is shown with a solid line and dots; each dot represents the i^{th} cross-sectional area. For reference, the open circles along the line at 1 cm^2 indicate anatomical landmarks based on the original MR images from which this area function was derived; A=lips, B=incisors, C=junction of hard and soft palate, D=superior aspect of the epiglottis, E = just superior to the glottis. In the inset plot is the frequency response of the area function whose peaks indicate the resonance frequencies needed for calculating the sensitivity functions. (b) Sensitivity functions calculated with Eqs. (5)–(8) that correspond to the first three resonances of the area function in (a).

method was proposed by Story (2006) in which linear combinations of the sensitivity functions were used to incrementally and iteratively deform an area function until the differences between the calculated resonance frequencies and a set of target frequencies were minimized. The latter study suggests that the shape of the sensitivity functions themselves may be utilized to define the overall deformation pattern imposed on the area function rather than simply using their polarities to *suggest* where expansions and constrictions should be manually imposed.

2.2. Acoustically-based consonant superposition function

Sensitivity functions calculated for the vowel substrate $V(i, n)$ at every time sample can be additively combined and normalized to produce a consonant function $C(i, n)$ that, when superimposed on the vowel substrate to modify its shape (i.e., Eq. (4)), shifts the resonance frequencies upward or downward to achieve a phonetic goal. The shape of $C(i, n)$ is controlled by three parameters representing the polarity and normalized magnitude of the resonance deflections required to generate a specific phonetic target. These control parameters are denoted δ_1, δ_2, and δ_3, and can be assigned any value between -1 and 1. Together they form a resonance deflection pattern that, when written in a vertical orientation ordered from bottom to top, coincides with the spatial arrangement of formants as observed in spectrogram. For example,

$$\begin{bmatrix}\delta_3\\ \delta_2\\ \delta_1\end{bmatrix} = \begin{bmatrix}-1\\ -1\\ -1\end{bmatrix} \tag{10}$$

would indicate a targeted downward deflection of each of the three resonances, and thus formants, typical of a bilabial consonant.

The time-dependence of a given resonance deflection pattern across the duration of an utterance is dictated by an associated *event* function $E(n)$. This should be a smoothly varying curve whose amplitude is constrained to be between the values of 0 and 1. The event function used for this study is constructed piecewise, and expressed in discrete-time form as,

$$E(n) = \begin{cases} 0 & n = [1, N_{01}] \\ \frac{1}{2}\left(1 - \cos\left(\frac{\pi(n-N_{01})}{N_{pk}-N_{01}}\right)\right) & n = [N_{01}, N_{pk}] \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(n-N_{pk})}{N_{02}-N_{pk}}\right)\right) & n = [N_{pk}, N_{02}] \\ 0 & n = [N_{02}, N_t] \end{cases} \tag{11}$$

The parameters in this equation represent specific points in time across an utterance where N_{01} is the sample at which the upward trajectory toward the peak begins, N_{pk} is the point at which the peak amplitude of 1.0 is achieved, and at N_{02} the downward trajectory away from the peak becomes zero. The value of N_t is the total number of time samples and represents the duration of the utterance. As an example, the event function shown in Fig. 3 is based on the parameter values $[N_{01}, N_{pk}, N_{02}, N_t] = [5, 30, 55, 73]$ and results in a duration of 500 ms (based on the sampling interval $T = 6.866$ ms). The particular shape of this curve is roughly patterned after the consonant magnitude functions derived from analysis of VCV articulatory data in Story (2009, p. 833).

The input parameters δ_j $(j = 1, 2, 3)$, the event function $E(n)$, and vowel substrate $V(i, n)$ now provide the information required to generate an acoustically-controlled $C(i, n)$ function. The first step, at any time sample n, is to calculate the frequency response of the corresponding vowel area vector $V(i)$, and from it determine the resonance frequencies f_{R1}, f_{R2}, and f_{R3} (e.g., Fig 2a). Using these values, along with $V(i)$, the sensitivity functions $S_1(i)$, $S_2(i)$, and $S_3(i)$ are then calculated with Eqs. (5)–(8).

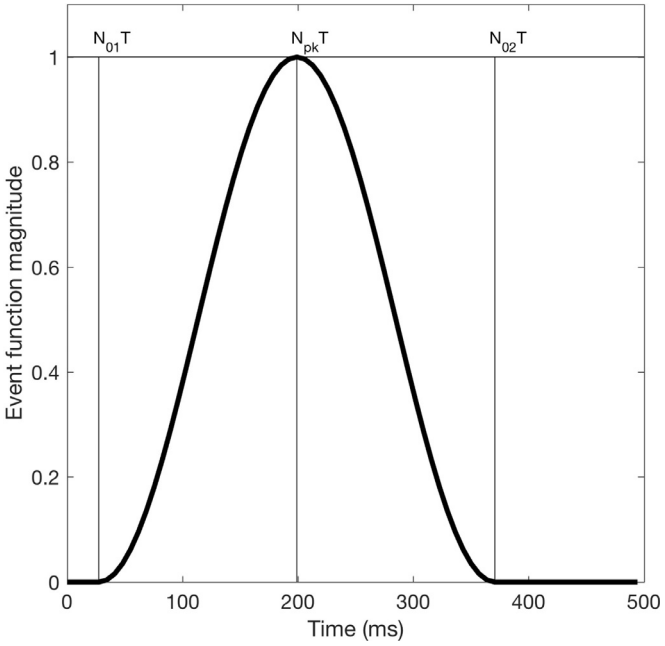


Fig. 3. Event function produced by Eq. (11) when parameters are set to $[N_{01}, N_{pk}, N_{02}, N_t] = [5, 30, 55, 73]$ given in time samples. The horizontal axis of the plot is shown in milliseconds; thus, the parameters are shown multiplied by the sampling interval $T = 6.866$ ms.

The next step is to form a linear combination of the sensitivity functions where the coefficient weights are the input parameters δ_j,

$$y_0(i) = \delta_1 S_1(i) + \delta_2 S_2(i) + \delta_3 S_3(i). \tag{12}$$

The δ_js determine the relative contribution of each sensitivity function to the overall shape of $y_0(i)$. For example, the pattern given in Eq. (10) would dictate that all three sensitivity functions are inverted in polarity and added together with equal magnitudes in order to strongly shift each resonance downward. If, however, the phonetic target called for a strong downward shift of f_{R1}, no change in f_{R2}, and a relatively weak upward shift of f_{R3}, the pattern would be specified as,

$$\begin{bmatrix}\delta_3\\ \delta_2\\ \delta_1\end{bmatrix} = \begin{bmatrix}0.5\\ 0\\ -1\end{bmatrix} \tag{13}$$

thus preventing the S_2 sensitivity function from affecting the shape of y_0.

The epilaryngeal region of the vocal tract located just downstream of the glottis typically does not serve as an articulator for phonetic purposes. To assure that this region is not affected by the consonant function, a constraint $g(i)$ can be imposed that suppresses the effect of $y_0(i)$ near the glottis,

$$g(i) = \begin{cases} 0.0 & \text{for } i = [1, i_{iepi} - 1] \\ 1 - e^{-\ln(16)\left(\frac{i-i_{epi}}{r_{epi}}\right)^2} & \text{for } i = [i_{epi}, N_x] \end{cases} \tag{14}$$

so that,

$$y(i) = g(i)y_0(i) \quad i = [1 \ldots N_x]. \tag{15}$$

The value of i_{epi} represents the number of sections (starting just above the glottis) of the area function attributed to the epilarynx, and r_{epi} is a parameter that adjusts the range of the vocal tract length affected by the exponential function in the second part of the equation. For this study, $i_{epi} = 5$ and $r_{epi} = 6$, but these values can be varied to produce a more or less extensive constraint if

Table 1
Resonance deflection patterns and μ parameters hypothesized to produce bilabial, alveolar, and velar stops. The table is arranged such that each column contains one of the patterns; the bottom to top order of the individual deflections δ_j is intended to mimic the graphical arrangement of formants as observed in spectrogram.

	{b}	-{d}	{g}
δ_3	−1	1	−1
δ_2	−1	1	1
δ_1	−1	−1	−1
μ	1	1	1

desired. Note that a similar constraint was implemented by Carré (2004).

The consonant deformation $C(i, n)$ at each time sample n can now be formed by normalizing $y(i)$ relative to its minimum value, and multiplying by the value of $E(n)$. It is written as,

$$C(i,n) = \frac{-\mu E(n) y(i)}{\min\limits_{i \in [1, N_x]} y(i)} \tag{16}$$

where the minus sign is needed to negate the effect of the denominator always being less than zero. The scaling factor μ is an additional parameter that can be used to either slightly attenuate or amplify the effect of the event function if desired. Whenever the product of μ and $E(n)$ is equal to 1.0, a zero value will be produced in $C(i, n)$ at the location found for the minimum value in $y(i)$, thus generating a complete occlusion of the vocal tract. The μ parameter essentially provides control of the *degree* to which a full occlusion is produced within the vocal tract; if $\mu < 1$, the constriction will only partially occlude the tract, whereas when $\mu > 1$ the extent of the occlusion along the length axis will increase. It can also be noted that because of the normalization process in Eq. (16), the limitation that $\delta_j \in [-1, 1]$ in Eq. (12) is actually unnecessary, but is a convenient constraint for making the specification of deflection patterns uniform (e.g., with normalization, a pattern $\delta_j = [-1, 0.5, 1]$, is identical to $\delta_j = [-2, 1, 2]$). Finally, the composite time-varying area function $A(i, n)$ can be generated from the product of $V(i, n)$ and $C(i, n)$ as indicated by Eq. (4).

To demonstrate the construction of VCV utterances with this model, four different vowel substrates were modified with resonance deflection patterns hypothesized to produce bilabial, alveolar, and velar stops. The first $V(i, n)$ was simply a constant neutral vocal tract configuration held for 500 ms (i.e., $q_1(t) = q_2(t) = 0$ in Eq. (1)). The other three were exactly those presented in Story and Bunton (2010) in which the vocal tract was configured in a neutral shape for the initial 200 ms, then transitioned to the vowel, [i], [ɑ], or [u] in the next 200 ms, and held constant in that configuration for the final 100 ms. The example discussed in the Introduction (Fig. 1a) was, in fact, the neutral-to-[ɑ] vowel substrate from Story and Bunton (2010).

The resonance deflection patterns prescribed for the three stops are shown in Table 1, in the same bottom to top order as introduced in Eq. (10). The parameter μ is shown in the bottom row, and was set to a value of 1.0 for each of the three patterns. This assured that a full occlusion will be achieved, but at only a single location (i.e., the constriction will not spread along the vocal tract length as would result if $\mu > 1$). The IPA symbols embedded within the unconventional curly brackets are used here to differentiate vocal tract area functions and calculated resonance frequencies produced by a model, from actual prescribed phonetic targets or transcriptions of real or synthetic talkers. This notation will be used throughout the remainder of the article.

2.3. Consonant identification test

Although the focus of this study was not specifically on perception, a limited experiment was conducted to determine if simulated audio samples based on the constructed VCVs would by identified by listeners according to the hypothesized deflection patterns. Simulated acoustic waveforms were generated with the TubeTalker system (Story, 2013) in which a kinematic model of the vocal fold surfaces provided the voice source, and a digital waveguide algorithm was used to compute the propagation of acoustic waves in the vocal tract. The time-varying area functions constructed for each combination of resonance deflection pattern and vowel substrate were incorporated into TubeTalker to generate a VCV which was stored as an audio file. The fundamental frequency of the voice began at 95 Hz, increased over the course of 200 ms to a peak of 130 Hz, and then decreased gradually over the final 300 ms down to a frequency of 77 Hz. These samples were simulated as voiced consonants, and with no release burst present. The latter condition was imposed to maintain similarity to the Story and Bunton (2010) samples, and also to assure that the identification was based primarily on the temporal characteristics of the formants.

There were 12 files total (3 deflection patterns × 4 vowel substrates) that were presented to listeners via the Alvin interface (Hillenbrand and Gayvert, 2005). Each listener was seated in a sound booth and samples were played over a loudspeaker (Yamaha MSP3) set at a comfortable listening level. After hearing each sample, a listener used a computer mouse to choose, "b", "d", "g" or "ambiguous" from buttons displayed on the computer screen. The samples were played in random order, and each was repeated four times. Five listeners were recruited to participate in the experiment, thus a total 240 responses were collected (12 samples × 4 repetitions × 5 listeners).

All twelve simulated VCVs are included with this article as a set of audio files in ".wav" format.

3. Results

3.1. Neutral vowel substrate

Consonant superposition functions, composite area functions, and calculated resonance frequencies resulting from the three deflection patterns in Table 1, are shown in Fig. 4 for case where the vowel substrate was held constant as a neutral vowel. The figure is arranged such that the plots in each of the three columns correspond to a specific deflection pattern.

In the first column are the results of the {b} deflection pattern in which the three resonances were prescribed to decrease in frequency during the time course of the event function, relative to the resonances of the underlying neutral vowel configuration. The superposition function $C(i, n)$ (Fig. 4a) is plotted as a three-dimensional surface where the x and y axes represent distance from the glottis and duration of the utterance. The z-axis is the scaling factor; when it is less than one and approaching zero a red valley is produced in the surface indicating a constriction. Expansions are generated when the scaling factor exceeds one, and these are shown as blue prominences. In this case, the primary constriction is located at the last section of the area function (i.e., section 44, $L_c = 17.5$ cm), as would be expected for a bilabial consonant. The time course of the constriction is emphasized with the thick black curve. A simultaneous expansion can be observed at location $L_e = 14.7$ cm, just 2.8 cm posterior to the lips, and is emphasized with the thick white line. This is followed posteriorly by additional "ripples" that impose fairly minor constrictive and expansive effects. When $C(i, n)$ is imposed on the underlying vowel substrate the result is the composite time-varying area function

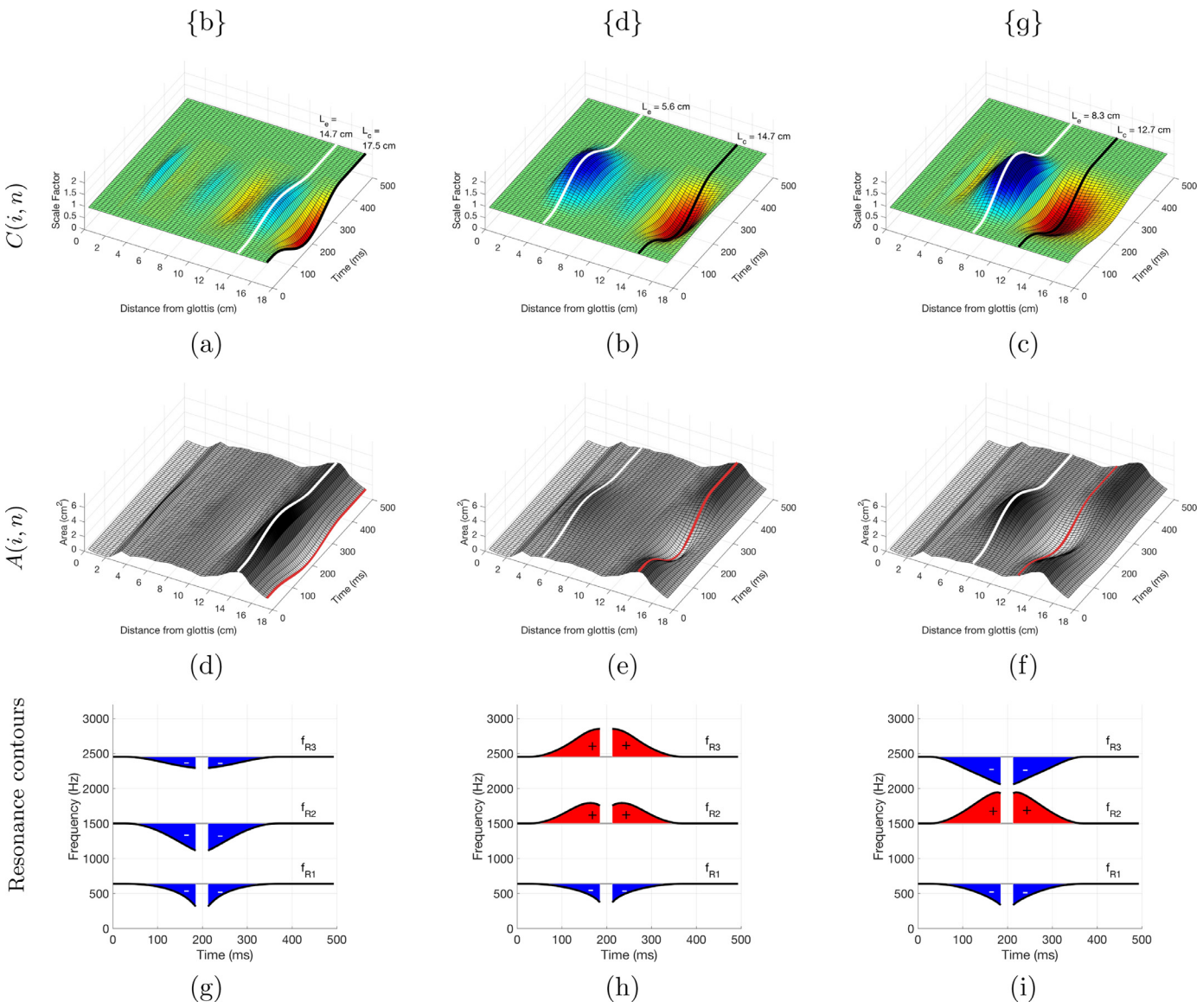


Fig. 4. Consonant superposition functions, composite area functions, and calculated resonance frequencies resulting from the three deflection patterns in Table 1. Each column of plots corresponds to one of the three patterns for stop consonants. The top row contains the derived $C(i, n)$ functions shown as 3D surfaces, and the middle row shows the composite area functions $A(i, n)$. In the bottom row are the calculated resonance frequencies for each case, where the blue and red shaded regions indicate the downward or upward deflection of the resonance frequencies, respectively, due to the consonant deformation of the underlying vowel substrate.

shown in Fig. 4d. The time course of the vocal tract shape at the lips, traced out by the red line, shows a gradually decreasing area that becomes an occlusion at 200 ms and then gradually returns to the lip area of the neutral vowel; the simultaneous expansion shown with the white line follows the same time course except with an increase in area prior to 200 ms followed by a decrease. The three time-varying resonance frequencies calculated for $A(i, n)$ are shown in Fig. 4g. They are plotted relative to the underlying resonances of the vowel substrate; the gap centered at 200 ms is the period of time in which the vocal tract is essentially occluded, and is based on the time samples where the event function E has a value greater than 0.99. The shaded regions mark the influence of the consonant deformation; in this plot the shaded regions are all blue to indicate that the deflection of the three resonances is downward in frequency.

The second deflection pattern in Table 1 produced the consonant superposition function plotted in Fig. 4b. In this case, the primary constriction was located at $L_c = 14.7$ cm from the glottis, as marked by the red valley and the black line. This point is 2.8 cm posterior to the lips, roughly in the alveolar region of the vocal tract according to the anatomical landmarks given previously in Fig. 2. Although there is a slight expansion just posterior to the constriction at L_c, the largest expansion occurred at a location $L_e = 5.6$ cm, shown as the blue region, which would be just superior to the epiglottis. The superposition of this surface on the neutral vowel configuration is shown in Fig. 4e and associated resonance frequencies are plotted in Fig. 4h. The shaded regions again indicate the effect of the consonant deformation on the vocal tract resonances relative to the vowel substrate; the blue regions correspond to a downward deflection whereas the red regions indicate an upward shift of a resonance frequency. As prescribed by the {d} pattern in Table 1, the first resonance frequency did indeed decrease, and the frequency of both the second and third resonances increased.

Fig. 4c shows the consonant function generated by the {g} pattern in Table 1. The location of the constriction is now at $L_c = 12.7$

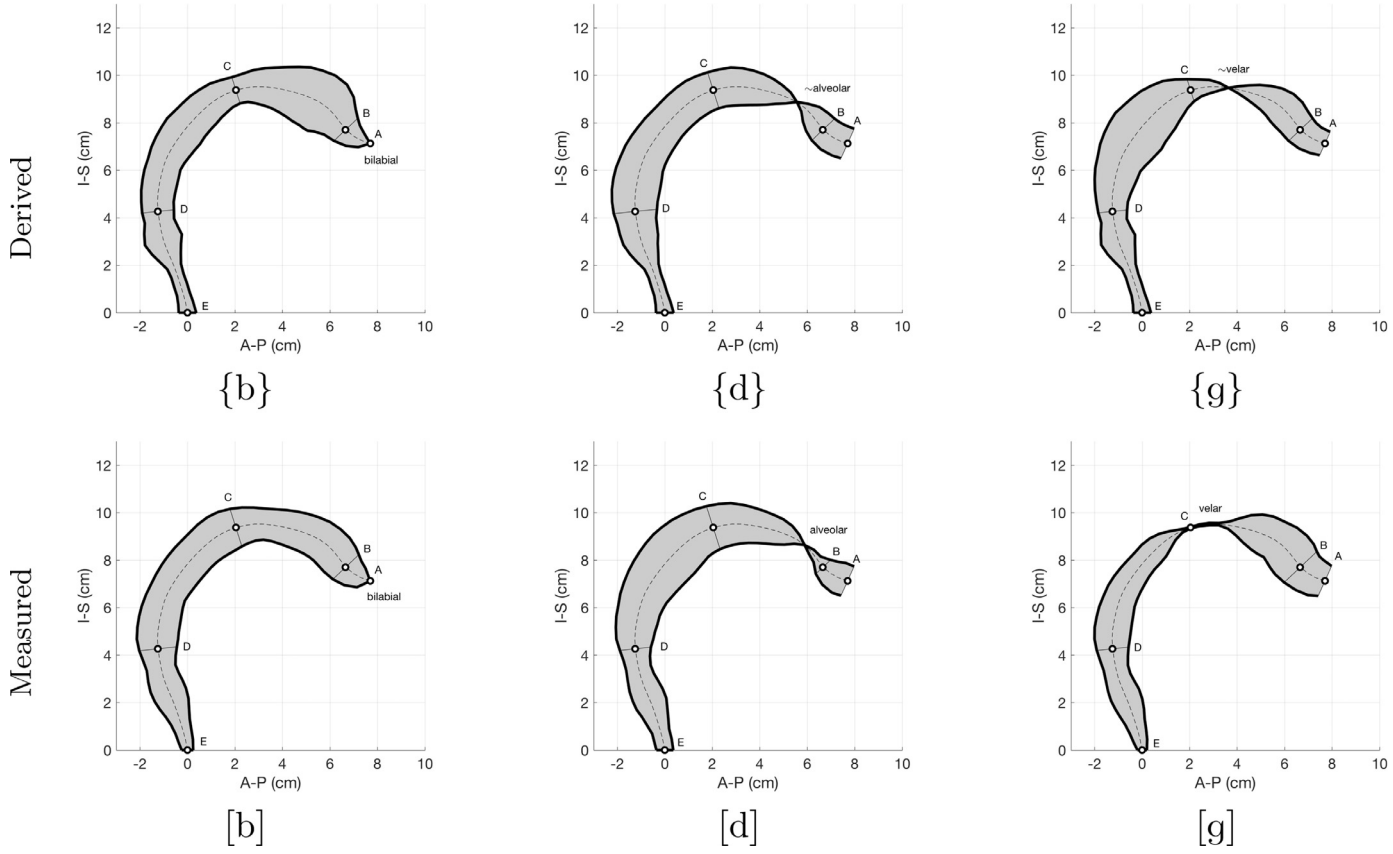


Fig. 5. Pseudo-midsagittal plots of vocal tract configurations for three stop consonants. The plots in the top row are based on the area functions generated at $t = 200$ ms for each case shown previously in Fig. 4. The bottom row contains similar plots based on measured area functions reported in Story (1995, 2005a). For reference, the open circles along the centerline in each plot indicate anatomical landmarks based on the original MR images from which these area functions in the bottom row were measured; A = lips, B=incisors, C=junction of hard and soft palate, D=superior aspect of the epiglottis, E = just superior to the glottis.

cm, exactly 2 cm back from the constriction location derived for the {d} pattern. An expansion at $L_e = 8.3$ cm follows immediately posterior to the constriction. In contrast to either of the previous constrictions, the extent, or "range," along the vocal tract axis is larger for this deflection pattern. The resulting time-varying area function in Fig. 4f shows the effect of the simultaneously imposed constriction and expansion. These two regions appear to work synergistically to produce the pattern of resonance frequencies plotted in Fig. 4i. In this case, the first and third resonances shaded in blue are shifted downward in frequency, and the second resonance shaded in red is shifted upward, as prescribed by the {g} pattern.

To get a sense of how the stop consonant configurations relate to anatomical landmarks, the area functions that occurred at 200 ms (the time instant at which the vocal tract was fully occluded) in each of the $A(i, n)$ plots in Fig. 4 have been replotted as "pseudo-midsagittal" profiles in the top row of Fig. 5. These are generated by first converting each cross-sectional area to an equivalent diameter. They are then plotted perpendicular to a vocal tract centerline function that was measured from MRI-based data (Story et al., 1996). In each plot, the centerline is dashed, and the open circles denote the same anatomical landmarks that were indicated in Fig. 2; that is, A is located at the lips, B roughly indicates the location of the incisors, C is the junction of the hard and soft palate, D is the superior aspect of the epiglottis, and point E is just superior to the glottis. For purposes of comparison, similar plots are shown in the second row based on area functions *measured* for the same stop consonants using MRI (Story, 1995; 2005a). The talker who produced these vocal tract configurations is the same person from whom the neutral vowel shape (Fig. 2a) was derived. It should be noted, however, that because of the nature of collecting volumetric MR image data, these area functions represent a configuration held constant by the talker while anticipating release into a neutral vowel; they were not snapshots within a dynamic event. Nonetheless, they do give an indication of the location and extent of the constriction along the tract length that can be compared to the derived shapes.

The {b} constriction is clearly aligned with the lips as expected for a bilabial stop. The overall shape is slightly less uniform than that of the measured configuration below it, but is quite similar. The constriction for the {d} is positioned about 1.5 cm posterior to the incisor landmark, which is just slightly farther away than the location in the measured shape. Away from the constriction, the derived and measured configurations are again similar. The location of the {g} constriction is 1.6 cm anterior to the junction of the hard and soft palate, whereas in the measured case the constriction begins at the hard/soft palate junction and spreads forward by about 1.6 cm. Thus, the anterior edge of the measured constriction is at essentially the same location as that in the derived tract shape. The extent of the occlusion along the tract length, however, is much shorter in the derived shape because using a value of $\mu = 1.0$ assures a vocal tract closure at a single section (i.e., Eqn. 16). For comparison, three additional derived vocal tract shapes for {g} are shown in Fig. 6 where μ was set to 1.05, 1.1, and 1.2, respectively. In these cases, the length of the occluded portion becomes 0.85 cm, 1.3 cm, and 2.0 cm for the three different values of μ, and are more in line with the measured shape.

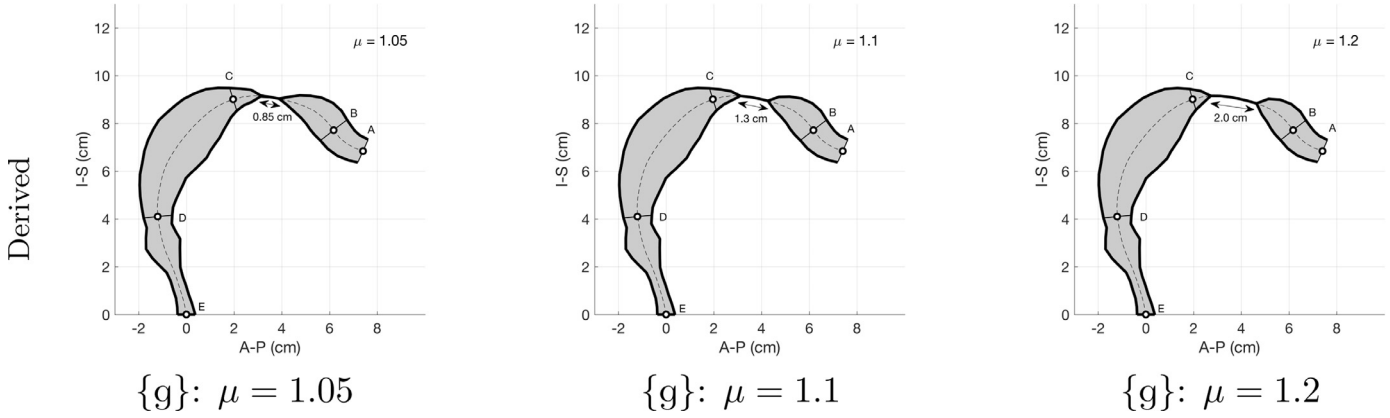


Fig. 6. Pseudo-midsagittal plots of three additional vocal tract configurations for the derived {ɡ} condition (see Fig. 2c) where the value of μ was set to be 1.05, 1.1, and 1.2. The open circles along the centerline in each plot indicate the same anatomical landmarks as in Fig. 2.

3.2. Vowel substrates: {əi}, {əɑ}, and {əu}

The results of applying the three resonance deflection targets to the three time-varying vowel substrates from Story and Bunton (2010), are shown in Figs. 7, 8, 9. Similar to Fig. 4, consonant superposition functions, composite time-varying area functions, and resonance plots are included for each case. In this section, however, each figure demonstrates the effect of *one* deflection pattern on *each* of the three vowel substrates. This arrangement allows for comparison of similarities and differences of consonant deformations derived for the same targeted deflection pattern imposed on different vowel contexts.

As can be seen in Fig. 7, the {b} pattern as a target generates a $C(i, n)$ function for each vowel substrate that imposes a primary constriction at the lips, just as it did for the neutral substrate in Fig. 4. In addition, there are subtle expansions and constrictions extending from the lips back into the vocal tract in each case that are, by visual inspection, nearly indistinguishable from one another. The composite area functions shown in the middle row each produce a set of time-varying resonance frequencies (bottom row) that are identical during the onset portion of the utterance (i.e., 0–200 ms) and then follow a unique path in the latter portion determined by the final vowel. In all three cases, however, the resonance frequencies are deflected downward relative to the vowel-only resonances during the entirety of the consonantal duration. This occurs regardless of the direction of the vowel-only resonances; that is, the shaded portions for all cases are blue even though the absolute direction of change in the vowel-only resonances may be upward or downward in frequency.

The plots associated with the {d} target pattern (Fig. 8) show that the $C(i, n)$ functions are likewise quite similar to that derived for neutral substrate. In fact, the primary constrictions and expansions for all three vowel substrates are located at $L_c = 14.7$ cm and $L_e = 5.6$ cm, respectively, identical to the neutral substrate conditions. There are some subtle differences that are visible across the three $C(i, n)$ surfaces, mostly at the lips and between L_c and L_e. The resonance frequencies determined from the composite area functions are again identical in the onset portion, but, of course, with f_{R1} deflected downward and f_{R2} and f_{R3} deflected upward as prescribed by the {d} pattern. In the offset portion (i.e., 200–500 ms), the resonances maintain the same deflection pattern relative to the underlying changes in vowel substrate, regardless of their direction.

For the {ɡ} pattern (Fig. 9), the primary constriction is identically located at $L_c = 12.7$ cm across the three vowel substrates, and the primary expansion was at $L_e = 8.3$ cm, both of which are also the same as was derived for the neutral case. Because it occurs in the upper portion of the pharynx, the effect of the primary expansion on the composite area function is perhaps more visibly prominent than for the previous two deflection patterns. This can be seen clearly in the {əgɑ} case where the upper pharynx expands synergistically with the primary constriction. The plots of the resonance frequencies again show that the deflection directions prescribed by the {ɡ} pattern are maintained regardless of vowel substrate.

The identical locations of the constrictions and expansions for each deflection pattern, regardless of vowel substrate, were due to the constant neutral configuration present for the initial 200 ms of each of the three vowel-vowel transitions, combined with the timing of the event function (Fig. 3) which achieved its peak value at 200 ms. Thus, the shaping of the vocal tract up to the point in time where the occlusion occurred was identical for all of the cases, whereas the shaping after the release of the occlusion was somewhat influenced by the specific vowel context. The subtle differences in the $C(i, n)$ functions generated for the three deflection patterns in Figs. 7, 8, 9 can be more easily observed by computing a difference function,

$$D^p(i,n) = C^p_{vv}(i,n) - C^p_0(i,n) \qquad p = [\{b\}, \{d\}, \{g\}] \tag{17}$$

where $C^p_0(i,n)$ is the consonant superposition function derived for the neutral vowel substrate for each deflection pattern p (i.e., Fig. 4a–c), and $C^p_{vv}(i,n)$ is the consonant function for same deflection pattern imposed on one of the other three vowel substrates. The difference functions for all nine cases are shown in Fig. 10 as 3D surfaces. Portions in green showed no differences, whereas the blue and yellow regions indicate where in the surface a given $C^p_{vv}(i,n)$ was either larger or smaller, respectively, than the corresponding $C^p_0(i,n)$. Because each vowel substrate begins with the neutral vowel configuration, all difference functions are zero-valued (i.e., green) from 0 to 200 ms. For a period of time after the 200 ms point, there are clear differences indicated by the alternating blue and yellow pattern. The {b} target pattern produced the smallest differences, followed in order by {d} and {ɡ}. The latter two patterns presumably produce greater differences across the vowel substrates because they require larger deformations in the underlying area function to realize the consonant.

3.3. Consonant identification

The results of the consonant identification experiment are summarized in the confusion matrix of Table 2. The samples were al-

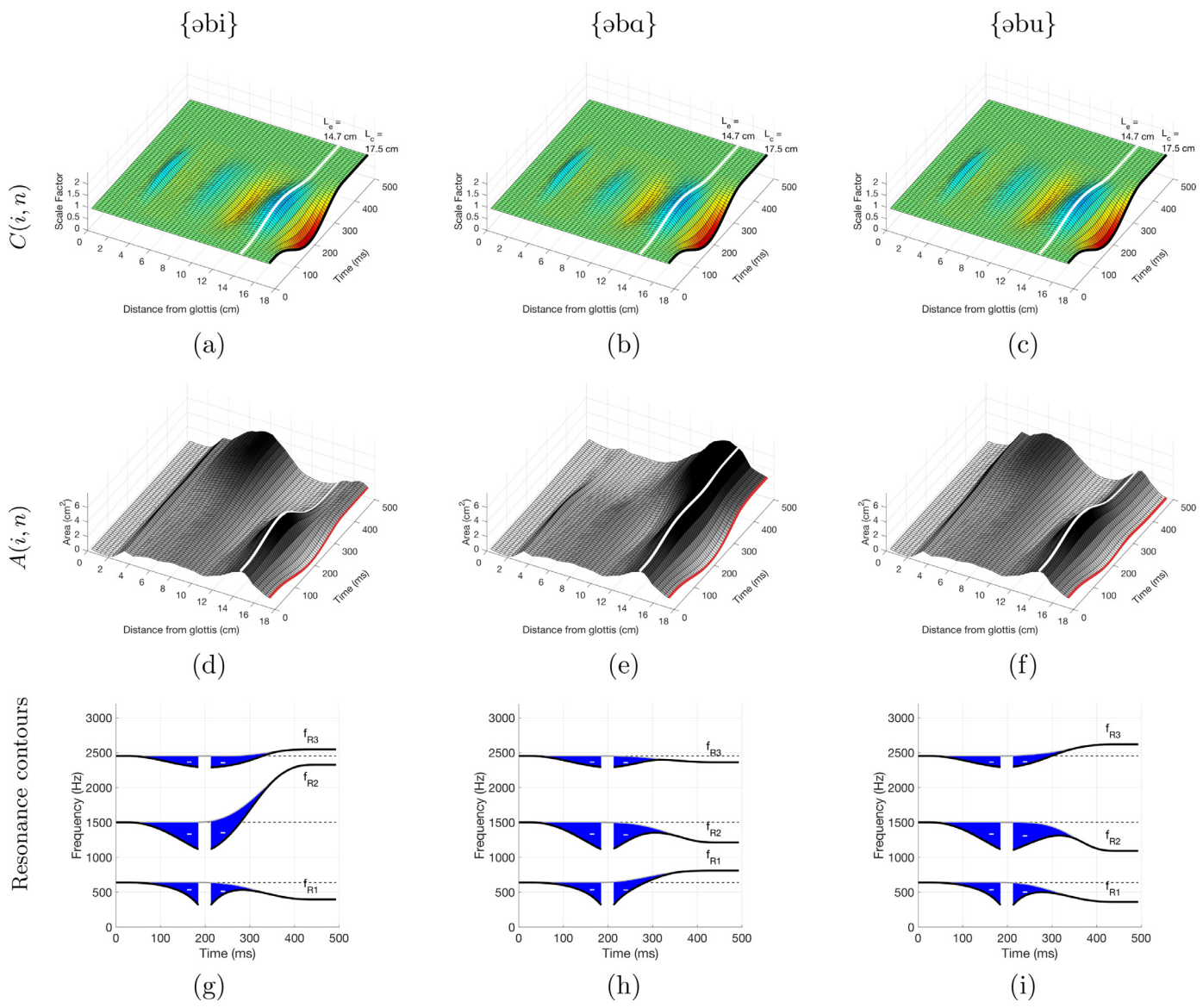


Fig. 7. Consonant superposition functions, composite area functions, and calculated resonance frequencies resulting from the {b} deflection pattern in Table 1 combined with the three vowel substrates {əi}, {əɑ}, and {əu}. Each column of plots corresponds to one of the three patterns for stop consonants. The top row contains the derived $C(i, n)$ functions shown as 3D surfaces, and the middle row shows the composite area functions $A(i, n)$. In the bottom row are the calculated resonance frequencies for each case, where the blue and red shaded regions indicate the downward or upward deflection of the resonance frequencies, respectively, due to the consonant deformation of the underlying vowel substrate.

Table 2
Confusion matrix for identification of simulated consonants by five listeners.

	[b]	[d]	[g]	Ambiguous
{b}	**80**	0	0	0
{d}	0	**80**	0	0
{g}	0	3	**75**	2

most always identified as the intended target based on deflection patterns in Table 1. The exceptions were three intended {g} samples marked as [d], two of which occurred in the {əu} vowel context, and one in the {əɑ} context. Two other intended {g} samples, both in the {əi} context, were marked as "ambiguous." Two confusions of {g} with [d] occurred in VCVs where the final vowel was {u}, whereas the other occurred with {i} as the final vowel. The two stimuli marked as ambiguous both had {ɑ} as the final vowel. Overall, nearly 98 percent of the samples were identified as the intended target.

4. Discussion

The identification test was not an extensive perceptual evaluation, but did demonstrate that listeners recognized the intended consonants in nearly 98% of the samples. Considering that the vocal tract was occluded for only a brief period of time, and there was no mechanism present to generate a release burst in the simulated VCVs, it is somewhat surprising that the identification of the consonants was so robust. Perhaps this is because the new approach to modeling the consonant deformation assures that the resonances follow a path that maintains their target deflection away from the underlying vowel substrate throughout the duration of the event function. In contrast, an approach where geometrical characteristics (e.g., constriction location and extent) are specified may successfully produce the target deflection of the resonances

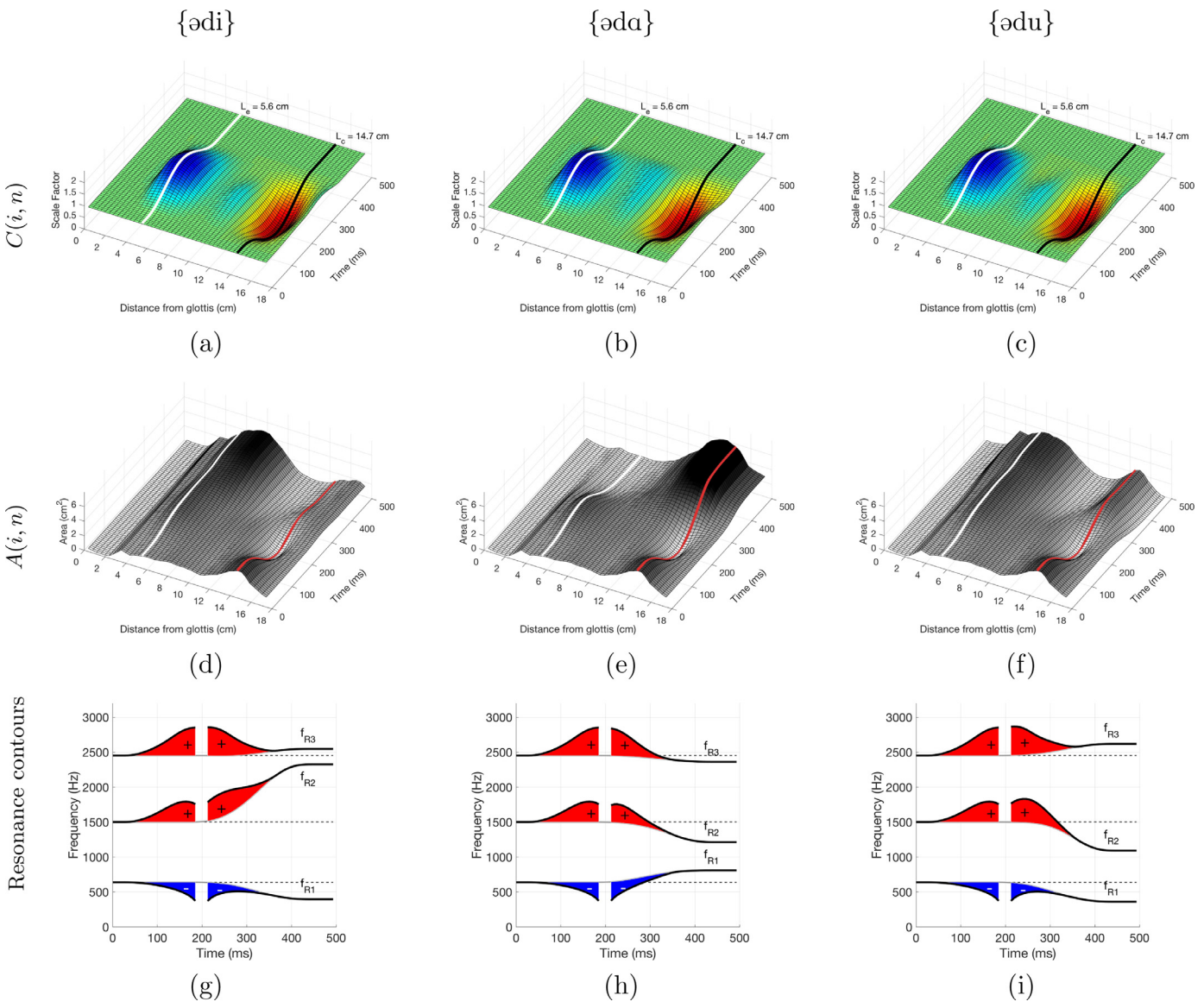


Fig. 8. Consonant superposition functions, composite area functions, and calculated resonance frequencies resulting from the {d} deflection pattern in Table 1 combined with the three vowel substrates {əi}, {əɑ}, and {əu}. Arrangement is identical to Fig. 7.

for part of the consonant duration but not across its entirety. It is noted that the consonant identification functions reported in Story and Bunton (2010), based on a continuum of constriction location, showed that the most robustly identified samples were based on locations aligned almost exactly with the L_c values that were measured from the derived superposition functions (Figs. 4, 7, 8, and 9). It may be no surprise then that the simulated samples here were identified with high accuracy, but it is interesting that the acoustically-based consonant superposition functions automatically set the constriction location at these particular distances from the glottis.

Eliminating direct control of the constriction shape in an area function model, and instead allowing acoustic properties to prescribe the vocal tract deformation may seem like a recipe for producing physiologically unrealistic or unusual configurations. The superposition functions presented in Section 3, however, indicated that the primary constrictions generated for the target consonants, {b}, {d}, and {g} were located, regardless of vowel substrate, at 17.5, 14.7, and 12.7 cm from the glottis, respectively. As compared to measured vocal tract configurations (cf., Fant, 1960; Perkell, 1969; Story, 2005a, 2009), these are typical constriction locations for production of these consonants. Additionally, the range, or extent of the tract length affected by these constrictions is smallest for {b} , larger for {d} , and largest for {g} , also in line with measured vocal tract shapes.

At first glance, it is tempting to reconcile these physiologically-reasonable results with the anatomical location and structure of the tongue and lips. For example, a constriction for a prototypical /g/ consonant produced by the tongue body could be expected to occlude the vocal tract in the velar region and affect a fairly broad extent of tract length due to the tongue body size. Similarly, constrictions for /d/ and /b/ could be expected to occur in the alveolar and lip regions, respectively, and affect lesser extents of the tract length because of the smaller sizes of the tongue tip and lips. The problem with this interpretation is that the consonant superposition functions that produced the constrictions shown in Section 3 are based purely on the acoustic properties of a vocal tract area function (i.e., a tubular conduit), and not on anatomical structures. Thus, the geometrical configuration of each constriction can only be attributed to the properties of the acoustic sensitiv-

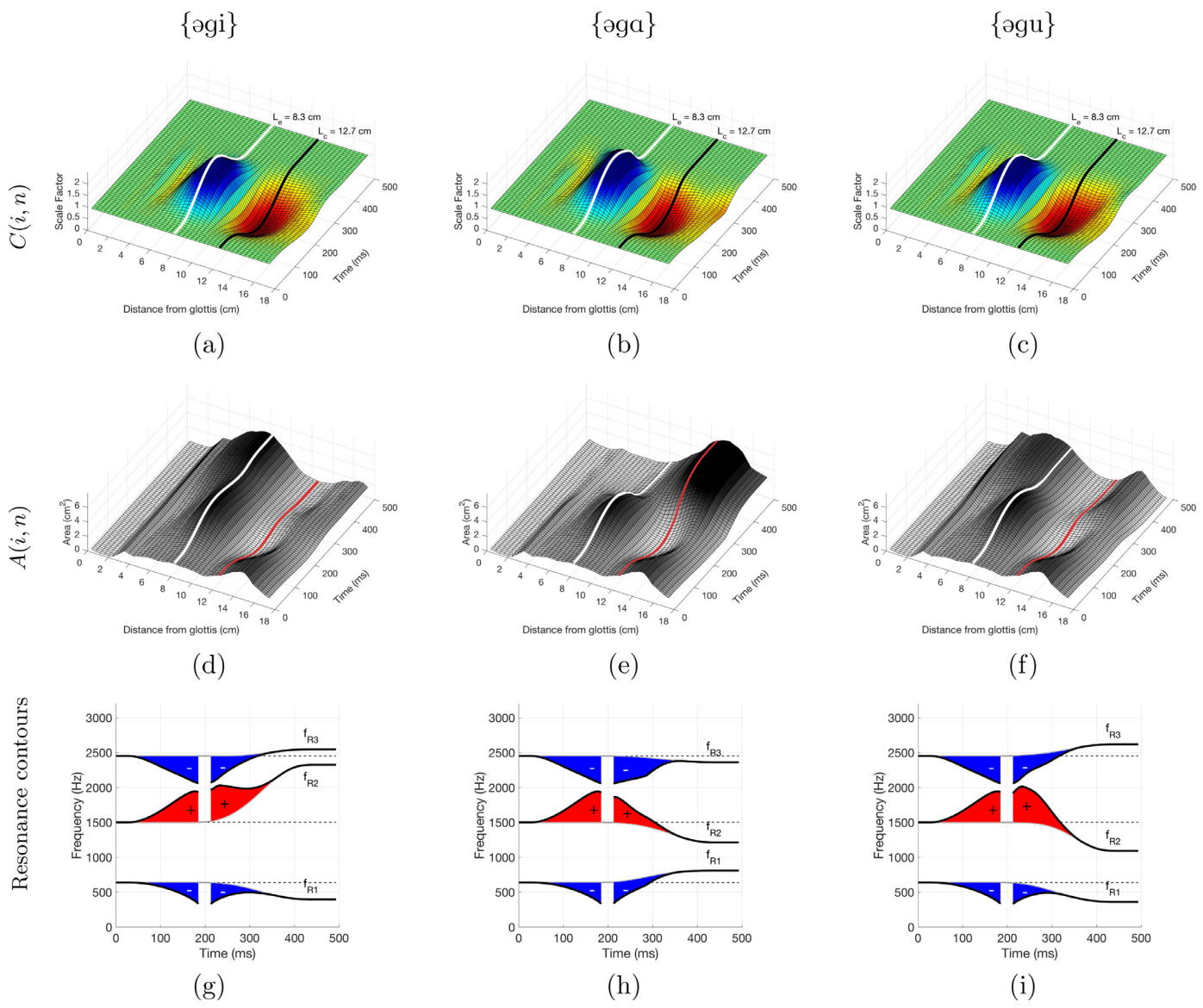


Fig. 9. Consonant superposition functions, composite area functions, and calculated resonance frequencies resulting from the {ɡ} deflection pattern in Table 1 combined with the three vowel substrates {əi}, {əɑ}, and {əu}. Arrangement is identical to Figs. 7 and 8.

ity functions. It is curious, however, that the tongue and lips seem particularly well suited to produce the types of the constrictions generated by linear combinations of the sensitivity functions.

In addition to exhibiting a primary constriction, each $C(i, n)$ function generated a simultaneous, upstream expansion whose locations were also consistent across all vowel substrates. Whether or not these expansions are physiologically realistic is difficult to assess from articulatory or vocal tract data because their effect is to increase the cross-sectional area of an already unoccluded part of the vocal tract, as opposed to the obvious point of occlusion produced by a constriction. Comparison with the measured vocal tract shapes shown in Fig. 2 indicates that although differences exist within the unoccluded regions, the derived expansions certainly do not generate unrealistic or absurd shapes. Perkell's (1969) analysis of X-ray movies collected during one talker's production of /həCɛ/ utterances also offers some hint of expansions occurring simultaneously with constrictions. Two measurements of vocal tract cross-distance in the pharynx, one referred to as "lower pharynx width" and the other "upper pharynx width," were made at points similar to the locations of the primary expansions in the {d} and {ɡ} superposition functions, respectively. Comparison of these measurements shows that, just prior to and during an occlusion, the relative increase in lower pharynx width is greater than in the upper pharynx for production of /həɾɛ/, but the opposite for /həkɛ/ (Perkell, 1969, p. 75 & 85). Although not conclusive evidence, these data do provide some support that the expansion portions of the consonant superposition functions are potentially realistic too.

Together the constrictions and expansions operate synergistically to enhance the targeted deflection of resonance frequencies. Although quite similar for each vowel context, the $C(i, n)$ functions were shown to possess subtle differences that were vowel dependent, suggesting a "fine tuning" of the deformation pattern to assure the appropriate direction of resonance frequency shift. Perhaps these effects are akin to the global nature of phonetic gestures proposed many years ago by Mattingly (1990). In an attempt to define the term "gesture" he argued against traditional phonetic descriptors that focused on single articulators because they "*make it easy to forget that some other important things may be happening... that are essential to the performance of the phonetic task.*" In this view, the $C(i, n)$ functions shown in Figs. 4, 7, 8, and 9 could be re-

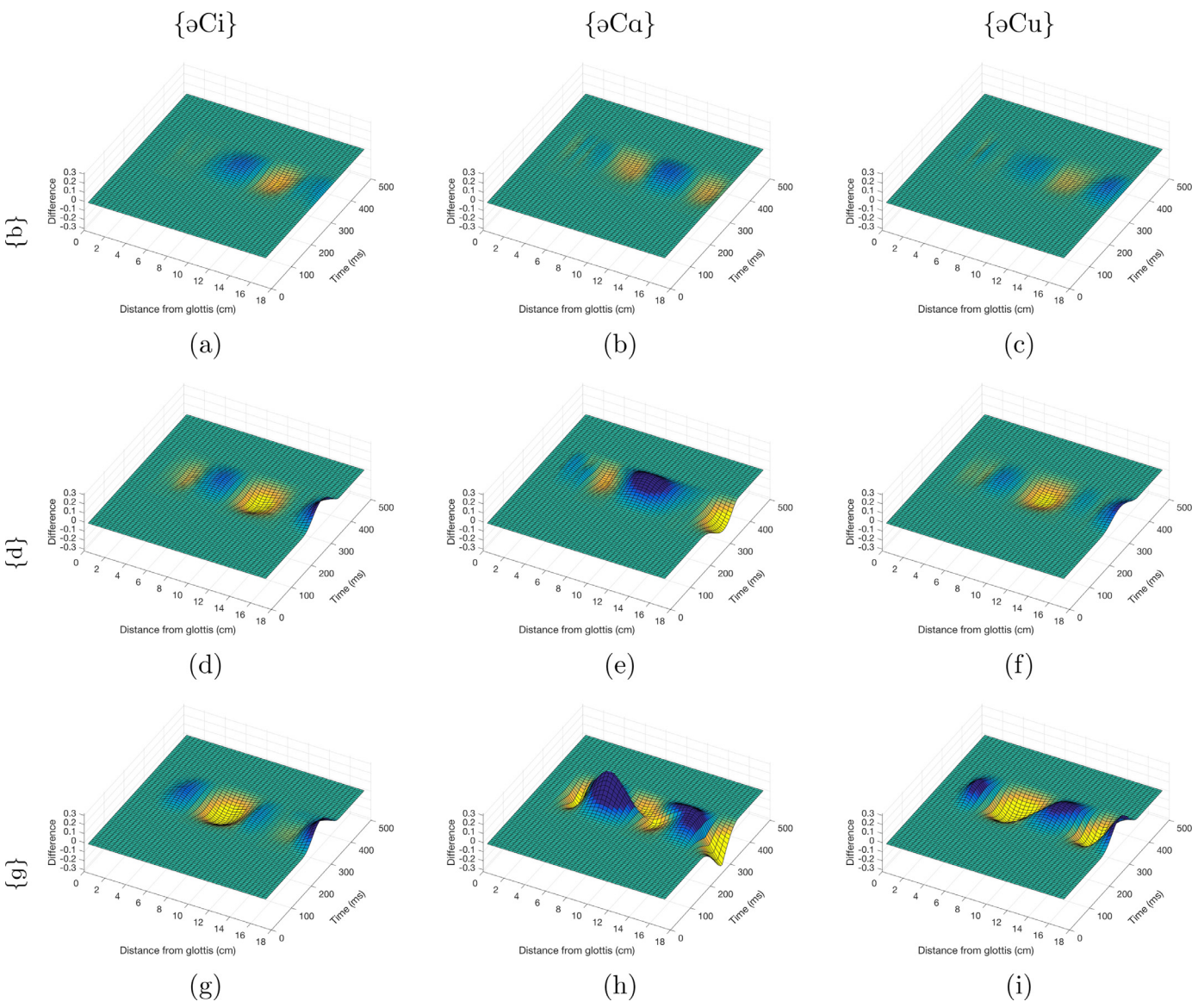


Fig. 10. Difference functions based on Eq. (17), relative to the neutral vowel substrate condition. All plots are equally scaled along the difference dimension.

garded as a phonetic "event" that is global due to its effect on the overall vocal tract shape, in contrast with the localized constriction discussed in the Introduction (i.e., Fig. 1). It must be considered, however, that any given consonant superposition function represents an "acoustically-ideal" deformation pattern that may not be entirely attainable by a talker. Constraints imposed on the system relative to factors such as anatomy, motor control, developmental stage for child talkers, as well as competing external demands (e.g., clenching a pen between the teeth) could limit a talker's ability to fully utilize the phonetic event prescribed by a $C(i, n)$ function. These effects could be incorporated into the model with constraint functions similar to Eq. (14) that would attenuate or even eliminate the upstream expansions.

A next step in developing this model is to investigate the effect of vocal tract length change, both in terms of growth during speech development as well as dynamic changes in length the occur during speech production. With regard to the former, it is of interest to scale the vocal tract length according to anatomical measurements (e.g., Vorperian et al., 2009 and determine if the specification of normalized resonance deflections patterns produces the same effects as in the current study. This may be a useful approach for understanding some aspects of speech acquisition skills. Dynamic length changes that occur during speech are also important to eventually include in the model. For example, a more realistic vowel substrate for {əu} (cf., Figs. 7, 8, 9) would allow the length vector $L(i)$ to vary with time so that the overall tract length increases as the vowel configuration transitions from the neutral to the {u}. This could potentially alter the sensitivity function calculation and may shift the location of the constriction during the time course of the event function; this may be an effect similar to the forward movement of the tongue during velar consonants observed from articulatory data (cf., Kent and Moll, 1972; Moosshammer et al., 1995).

This model is also potentially capable of producing a wide range of vowel and consonant combinations simply by specifying multiple event functions that are associated with resonance deflection patterns. Consider the example in Fig. 11a in which two event functions are shown to occur within a 500 ms duration, one near the beginning and one near the end. Above each event function is a resonance deflection pattern, and includes a value of 1.1 for the scaling factor μ. The first is the intended target for {d} whereas the second pattern is for {g}. If the underlying vowel substrate is set be

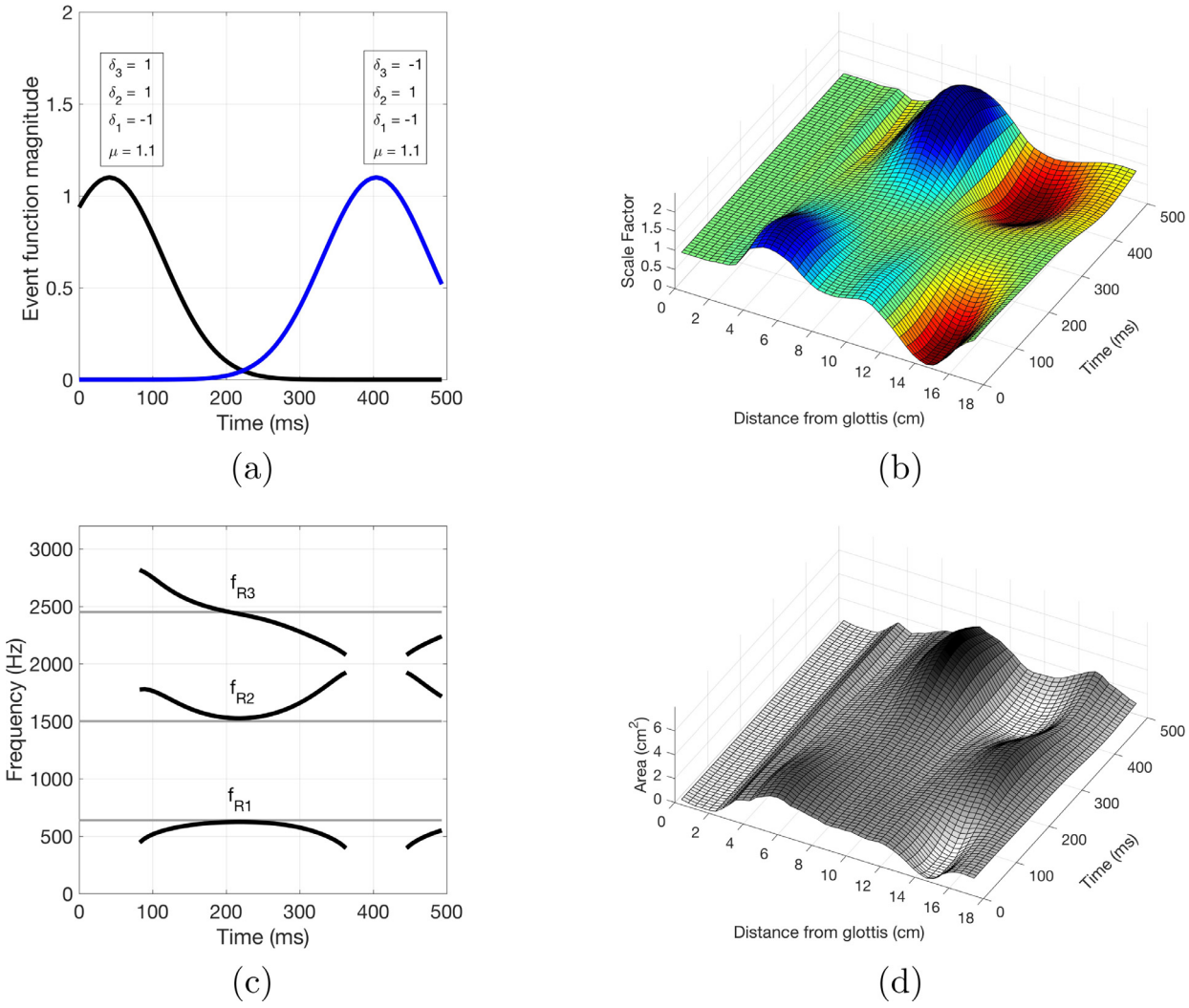


Fig. 11. Simulation of the word "dug." (a) Two event functions and associated resonance deflection patterns representing the target {d} at the beginning, and {g} at the end. (b) Consonant superposition function, (c) resonance frequencies calculated for the composite area function in (d).

a constant neutral vowel across the entire 500 ms duration (as in the examples in Fig. 4), the event functions would be expected to produce the word "dug." The global effect of the constrictions and expansions imposed on the vocal tract by the consonant superposition function in Fig. 11b can be seen in the composite area function of Fig. 11d. The calculated resonance frequencies are plotted in Fig. 11c and clearly indicate that the target deflection patterns have been achieved in the output. A simulation of the word "dug" based on the composite area function in Fig. 11d is included as an audio file in the supplementary material. The same file also contains several other simulations generated with the same event functions but with various combinations of resonance deflection patterns.

There is much that can be done to further develop and understand how the vocal tract shape can be controlled with this system to produce connected speech. For example, what is the effect of setting one or more of the deflection pattern components to fractional values between −1 and 1? Can a combination of event functions and deflection patterns generate fricatives, liquids, glides, affricates, and consonant clusters? Perhaps a bigger question is whether vowels could similarly be specified by their own set of resonance frequency changes such that an entire utterance (syllable, word, sentence, etc.) might be "coded" as a sequence of resonance deflection patterns activated by event functions occurring in serial temporal order, but potentially with much overlap.

The form of the model developed in this study does admittedly push the traditional theoretical limitations of acoustic sensitivity functions. In a strict sense, they can only be accurately used to impose small variations in cross-sectional area on a vocal tract shape to obtain a change in resonance frequencies predicted by Eq. (9) (Schroeder, 1967; Boë and Perrier, 1990). Further modifications to the shape require that sensitivity functions be calculated anew in order to prescribe the next small variations in cross-sectional area (cf., Story, 2006). In the current model, the sensitivity functions are calculated for a vowel-like vocal tract configuration at an instant of time within the substrate $V(i, n)$ and used to form a superposition function that, in turn, deforms that same vowel-like configuration. At the next time step, this process is repeated by calculating the sensitivity functions for the subsequent vowel-like configuration in $V(i, n)$, not the deformed shape from the previous time step. The reason is that the degree to which the underlying shape is deformed is controlled by the amplitude of the event function $E(n)$ (e.g., Eq. (11) & Fig. 3), and as this value approaches 1.0, the deformation relative to the vowel-like shape can be fairly large. Using the deformed shapes may result in sensitivity function magnitudes (in either negative or positive direction) that are very high in some locations of the vocal tract, typically where the cross-sectional area has become quite small due to the previous deformation. This can have the effect of producing

unrealistic subsequent deformation patterns as well as generating possible numerical instabilities. Thus, the current approach avoids some of these potential problems. It is also the case that the combination of sensitivity functions is used only to provide a deformation pattern that shifts resonance frequencies in specific directions, rather than to a specific set of frequency values. The resulting time-varying area functions and calculated resonance frequencies for the simulated VCVs suggest that the approach can be used with some success.

5. Conclusion

The results of this study demonstrate that the geometrical configuration of the consonant superposition function $C(i, n)$ in a multi-tier area function model, can be derived from the acoustic sensitivity of the vowel substrate $V(i, n)$. The process consists of specifying input parameters for a target consonant as a set of directional changes in the resonance frequencies of $V(i, n)$. These are then transformed into time-varying deformations of the vocal tract shape without any direct specification of location or extent of the consonant constriction along the vocal tract. The configuration of the constrictions and expansions that are generated by this process are physiologically-realistic and produce speech sounds that are easily identifiable as the target consonants. This model is a useful enhancement for area function-based synthesis and can serve as a tool for understanding how the vocal tract is shaped by a talker during speech production.

Acknowledgements

Research supported by NIH R01-DC011275 and NSF BCS-1145011.

Appendix A

The pressure and volume velocity within each section of an area function are required for calculation of the sensitivity functions as prescribed by Eqs. (5)–(9). In this study, these calculations were performed in the frequency domain with a lossy transmission line model (Sondhi and Schroeter, 1987; Story et al., 2000) that relates the input quantities on the glottal side to the output quantities at the lips via a chain matrix,

$$\begin{pmatrix} P_{out} \\ U_{out} \end{pmatrix} = \begin{pmatrix} A(f) & B(f) \\ C(f) & D(f) \end{pmatrix} \begin{pmatrix} P_{in} \\ U_{in} \end{pmatrix}. \tag{A1}$$

The frequency-dependent matrix elements A, B, C, and D collectively represent the wave propagation through all sections of the vocal tract. This matrix is the product of N_x similar matrices representing each individual tublet section, thus the pressure and volume velocity in the first tublet ($i = 1$) downstream of the glottis can be computed by,

$$\begin{pmatrix} P(1) \\ U(1) \end{pmatrix} = \begin{pmatrix} A^1(f) & B^1(f) \\ C^1(f) & D^1(f) \end{pmatrix} \begin{pmatrix} P_{in} \\ U_{in} \end{pmatrix} \tag{A2}$$

and in all ith subsequent tublets,

$$\begin{pmatrix} P(i) \\ U(i) \end{pmatrix} = \begin{pmatrix} A^i(f) & B^i(f) \\ C^i(f) & D^i(f) \end{pmatrix} \begin{pmatrix} P(i-1) \\ U(i-1) \end{pmatrix}. \tag{A3}$$

Using the same notation for area and length as in Eqs. (5)–(9) (i.e., $V(i)$ and $L(i)$), the matrix elements for each tubelet are written,

$$\begin{aligned} A^i(f) &= \cosh\left(\tfrac{\sigma L(i)}{c}\right) & B^i(f) &= -\tfrac{\rho c}{V(i)}\gamma \sinh\left(\tfrac{\sigma L(i)}{c}\right) \\ C^i(f) &= -\tfrac{V(i)}{\rho c}\tfrac{1}{\gamma}\sinh\left(\tfrac{\sigma L(i)}{c}\right) & D^i(f) &= \cosh\left(\tfrac{\sigma L(i)}{c}\right) \end{aligned}. \tag{A4}$$

where c is the speed of sound and ρ is the density of air. The other variables are defined to be

$$\gamma = \sqrt{\frac{r + j\omega}{\beta + j\omega}} \tag{A5}$$

and

$$\sigma = \gamma(\beta + j\omega) \tag{A6}$$

where

$$\beta = \frac{j\omega(2\pi F_T)^2}{(j\omega + r)j\omega + (2\pi F_w)^2} + \alpha \tag{A7}$$

and

$$\alpha = \sqrt{j\omega q}. \tag{A8}$$

The radian frequency ω is equivalent to $2\pi f$ where f is frequency in Hz and j is equal to $\sqrt{-1}$. The yielding properties of the wall are set by ratio of wall resistance to mass $r = 408$ rad/s and the mechanical resonance frequency of the wall $F_w = 15Hz$ (Sondhi and Schroeter, 1987). F_T is the lowest resonance frequency of the vocal tract when both the glottal and lip end are closed; it was set to $F_T = 200$ Hz. The parameter q in Eqn. A8 is a correction for thermal conductivity and viscosity, and was set to $q = 4$ rad/s.

References

Adachi, S., Takemoto, H., Kitamura, T., Mokhtari, P., Honda, K., 2007. Vocal tract length perturbation and its application to male-female vocal tract shape conversion. J. Acoust. Soc Am. 121 (6), 3874–3885.
Båvegård, M., 1995. Introducing a parametric consonantal model to the articulatory speech synthesizer. In: Proceedings Eurospeech 95, Madrid, Spain, pp. 1857–1860.
Boë, L.J., Perrier, P., 1990. Comments on Distinctive regions and modes: A new theory of speech production by M. Mrayati, R. Carre' and B. Guerin. Speech Comm. 9 (3), 217–230. doi:10.1016/0167-6393(90)90058-h.
Carré, R., 2004. From an acoustic tube to speech production. Speech Comm. 42, 227–240. doi:10.1016/j.specom.2003.12.001.
Carré, R., Chennoukh, S., 1995. Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures. J. Phonetics 23, 231–241. doi:10.1016/s0095-4470(95)80045-x.
Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague.
Fant, G., Båvegård, M., 1997. Parametric model of VT area functions: vowels and consonants. TMH-QPSR 38 (1), 1–20.
Fant, G., Pauli, S., 1975. Spatial characteristics of vocal tract resonance modes. In: Proc. Speech Comm. Sem. 74., Stockholm, Sweden, Aug 1–3, pp. 121–132.
Hillenbrand, J.M., Gayvert, R.T., 2005. Open source software for experiment design and control. J. Spch. Lang. Hear. Res. 48 (1), 45–60. doi:10.1044/1092-4388(2005/005).
Kent, R.D., Moll, K.L., 1972. Cinefluorographic analyses of selected lingual consonants. J. Spch. Lang. Hear. Res. 15 (3), 453–473. doi:10.1044/jshr.1503.453.
Kreuzer, W., Kasess, C.H., 2015. Tuning of vocal tract model parameters for nasals using sensitivity functions. J. Acoust. Soc. Am. 137 (2), 1021–1031.
Mokhtari, P., Kitamura, T., Takemoto, H., Honda, K., 2007. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients. J. Phonetics 35 (1), 20–39. doi:10.1016/j.wocn.2006.01.001.
Mooshammer, C., Hoole, P., Kühnert, B., 1995. On loops. J. Phonetics 23 (1), 3–21. doi:10.1016/s0095-4470(95)80029-8.
Mrayati, M., Carré, R., Guérin, B., 1988. Distinctive regions and modes: a new theory of speech production. Speech Comm. 7, 257–286. doi:10.1016/0167-6393(88)90073-8.
Nakata, K., Mitsuoka, T., 1965. Phonemic transformation of control aspects of synthesis of connected speech. J. Radio Res. Lab. 12 (61), 171–186.
Öhman, S.E.G., 1963. Coarticulation of stops with vowels. STL-QPSR 4 (2), 1–8.
Öhman, S.E.G., 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. 39, 151–168. doi:10.1121/1.1909864.
Öhman, S.E.G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Am. 41, 310–320. doi:10.1121/1.1910340.
Perkell, J.S., 1969. Physiology of speech production: Results and implications of a quantitative cineradiographic study, 53. MIT Press.
Sondhi, M.M., Schroeter, J., 1987. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans. ASSP ASSP-35 (7), 955–967. doi:10.1109/tassp.1987.1165240.
Stetson, R.H., 1951. Motor Phonetics: A Study of Speech Movements in Action. North Holland, Amsterdam, p. 203. doi:10.1007/978-94-015-3356-0.

Stevens, K.N., House, A.S., 1955. Development of a quantitative description of vowel articulation. J. Acoust. Soc. Am. 27 (3), 484–493. doi:10.1121/1.1907943.

Story, B.H., 1995. Physiologically-based Speech Simulation Using an Enhanced Wave-Reflection Model of the Vocal Tract. University of Iowa Ph. D. dissertation.

Story, B.H., 2005a. A parametric model of the vocal tract area function for vowel and consonant simulation. J. Acoust. Soc. Am. 117 (5), 3231–3254. doi:10.1121/1.1869752.

Story, B.H., 2005b. Synergistic modes of vocal tract articulation for american english vowels. J. Acoust. Soc. Am. 118 (6), 3834–3859. doi:10.1121/1.2118367.

Story, B.H., 2006. Technique for tuning vocal tract area functions based on acoustic sensitivity functions. J. Acoust. Soc. Am. 119 (2), 715–718. doi:10.1121/1.2151802.

Story, B.H., 2009. Vowel and consonant contributions to vocal tract shape. J. Acoust. Soc. Am. 126, 825–836.

Story, B.H., 2013. Phrase-level speech simulation with an airway modulation model of speech production. Comp. Spch. Lang. 27 (4), 989–1010. doi:10.1016/j.csl.2012.10.005.

Story, B.H., Bunton, K., 2010. Relation of vocal tract shape, formant transitions, and stop consonant identification. J. Spch. Lang. Hear. Res. 53, 1514–1528. doi:10.1044/1092-4388(2010/09-0127).

Story, B.H., Laukkanen, A.-M., Titze, I.R., 2000. Acoustic impedance of an artificially lengthened and constricted vocal tract. J. Voice 14 (4), 455–469. doi:10.1016/s0892-1997(00)80003-x.

Story, B.H., Titze, I.R., 1998. Parameterization of vocal tract area functions by empirical orthogonal modes. J. Phonetics 26 (3), 223–260. doi:10.1006/jpho.1998.0076.

Story, B.H., Titze, I.R., Hoffman, E.A., 1996. Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am. 100 (1), 537–554. doi:10.1121/1.415960.

Story, B.H., Titze, I.R., Hoffman, E.A., 2001. The relationship of vocal tract shape to three voice qualities. J. Acoust. Soc. Am. 109 (4), 1651–1667. doi:10.1121/1.1352085.

Vorperian, H.K., Wang, S., Chung, M.K., Schimek, E.M., Durtschi, R.B., Kent, R.D., Ziegert, A.J., Gentry, L.R., 2009. Anatomic development of the oral and pharyngeal portions of the vocal tract: an imaging study. J. Acoust. Soc. Am. 125 (3), 1666–1678. doi:10.1121/1.3075589.