A model of speech production based on the acoustic relativity of the vocal tract

Brad H. Story[a] and Kate Bunton
Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721, USA

(Received 20 May 2019; revised 10 September 2019; accepted 12 September 2019; published online 17 October 2019)

A model is described in which the effects of articulatory movements to produce speech are generated by specifying relative acoustic events along a time axis. These events consist of directional changes of the vocal tract resonance frequencies that, when associated with a temporal event function, are transformed via acoustic sensitivity functions, into time-varying modulations of the vocal tract shape. Because the time course of the events may be considerably overlapped in time, coarticulatory effects are automatically generated. Production of sentence-level speech with the model is demonstrated with audio samples and vocal tract animations. © *2019 Acoustical Society of America*.
https://doi.org/10.1121/1.5127756

[DDO] Pages: 2522–2528

I. INTRODUCTION

Speech production is often viewed as a process of planning and executing articulatory movements that generate an acoustic signal comprised of a temporally ordered stream of phonetic segments. Movement of the articulators is coordinated, or coarticulated, so that multiple segments overlap in time, thus facilitating rapid and efficient transmission of a message (cf. Kent and Minifie, 1977). Models of speech production are typically designed to emulate this process where the movements of the tongue, jaw, lips, velum, and larynx, or some lower dimensional representation of articulation, are orchestrated to collectively form the time-varying shape of the vocal tract, and transform the voice source into speech (cf., Mermelstein, 1973; Coker, 1976; Rubin *et al.*, 1981; Maeda, 1990; Browman and Goldstein, 1992; Story, 2005, 2009, 2013; Toutios *et al.*, 2011).

In contrast, Story and Bunton (2017) proposed a method, in part inspired by the distinctive region model of Mrayati *et al.* (1988), in which an utterance is planned by specifying directional changes of the resonance frequencies relative to those of the underlying vocal tract configuration. When associated with a temporal "event" function, the specified resonance deflections are transformed, via calculations of acoustic sensitivity functions, into a time-varying modulation of the vocal tract. An advantage of this approach is that an explicit specification of vocal tract characteristics such as constriction location is not required. Rather, the model itself finds a time-dependent vocal tract deformation pattern, containing constrictions and synergistic expansions, that results in the specified acoustic goal.

A limitation of the study reported by Story and Bunton (2017) was that the resonance deflection approach was applied only to stop consonants, whereas the modulation of the vocal tract to produce the underlying vowel substrate was independently generated by a kinematic model (Story, 2005). The aim of the present study was to demonstrate that the resonance deflection modeling approach can be used to generate sentence-level speech where all consonants and vowels are specified as a temporal sequence of *relative* acoustic events, partially overlapped in time, and then transformed to time-varying vocal tract modulations that automatically contain the effects of coarticulation. The scope of the study is limited to description of the vocal tract model and demonstration of sentence-level synthesis. Comparison of the model output to articulatory data and formal perceptual evaluation of the synthesis will be the focus of future research.

II. VOCAL TRACT MODEL CONTROLLED BY RELATIVE ACOUSTIC EVENTS

The structure of the model used in this study was essentially the same as described in Story and Bunton (2017). A time-varying vocal tract area function is generated as the product of a neutral configuration $\Omega(i)$ and a deformation function $D(i,n)$,

$$A(i,n) = \Omega(i)D(i,n), \quad i = [1, N_x], \quad n = [1, N_d], \qquad (1)$$

where, at any given time instant n, $A(i,n)$ consists of $N_x = 44$ contiguous sections or tubelets, each with a length of $L(i) = 0.396\,825$ cm. Although it is not suggested that this level of accuracy is required for the section length, the number is dictated by the wave propagation algorithm used in this study to synthesize speech (Liljencrants, 1985; Story, 1995) such that $L(i)$ is equal to the speed of sound ($c = 35\,000$ cm/s) divided by two times the sampling frequency ($F_s = 44\,100$ Hz). The actual distance from the glottis corresponding to the ith section is then $x(i) = \sum_{z=1}^{i} L(z)$, and results in an overall tract length of 17.46 cm (this length is simply an example used for this study; it could be set to any value appropriate for a human vocal tract by using a different number of tubelet sections or alternate sampling frequency). The time dimension is represented by n, and the total duration of a given utterance is N_d samples.

The shape of the deformation function $D(i,n)$ in Eq. (1) is controlled by three parameters representing the polarity and

[a]Electronic mail: bstory@email.arizona.edu

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. When written in a vertical orientation, they form a resonance deflection pattern (RDP) that coincides with the spatial arrangement of formants as observed in a spectrogram. For example, in the equation below,

$$\begin{bmatrix}\delta_3\\ \delta_2\\ \delta_1\end{bmatrix} = \overset{\sim\text{bilabial}}{\begin{bmatrix}-1\\ -1\\ -1\end{bmatrix}} \text{ or } = \overset{\sim\text{alveolar}}{\begin{bmatrix}1\\ 1\\ -1\end{bmatrix}} \text{ or } = \overset{\sim\text{velar}}{\begin{bmatrix}-1\\ 1\\ -1\end{bmatrix}}, \tag{2}$$

the first RDP vector would indicate a downward deflection of all three resonances, typical of a bilabial consonant, whereas the other two RDPs are representative of alveolar and velar consonants, respectively (cf. Story and Bunton, 2017).

Each RDP must be associated with an *event* function $E(n)$ that dictates the time course of the resulting vocal tract modulation required to actually produce the acoustic/phonetic event as a speech signal. This is a smoothly varying curve whose amplitude is constrained to be between 0 and 1. The event functions used for this study were based on a Gaussian pulse shape,

$$E(n) = e^{-\ln(16)\left((n-N_p)/N_w\right)^2}, \quad n = [1, N_d], \tag{3}$$

where n is the current time sample, N_p is the time sample at which a peak amplitude of 1.0 is achieved, and N_w is the width of the Gaussian at half maximum (i.e., at an amplitude of 0.5). The total duration of the event is N_d time samples, the same as in Eq. (1). The sampling frequency of the vocal tract modulation was set to $f_{svt} = 146$ Hz, which is the same as the x-ray microbeam database (Westbury, 1994), and will facilitate efficient comparison of the model output to articulatory data in future studies. Thus, the actual time values represented by the parameters in Eq. (3) are $t_p = N_p/f_{svt}$, $t_w = N_w/f_{svt}$ and $t_d = N_d/f_{svt}$.

III. TRANSFORMATION OF RESONANCE DEFLECTION PATTERNS INTO VOCAL TRACT MODULATION

Processing steps for transforming the RDPs (i.e., δ_j) associated with an event function $E(n)$ into a time-dependent vocal tract deformation function $D(i, n)$ will be described in Secs. III A–III D below, first for the case of a *single* specified acoustic event, and then for multiple events, as are required to produce sentence-level speech.

A. Sensitivity function calculation

The first step is to calculate the frequency response of the neutral area function $\Omega(i)$ [see Eq. (1)], and from it determine the resonance frequencies f_{R1}, f_{R2}, and f_{R3}. The specific $\Omega(i)$ used for this study is shown in Fig. 1(a) and is based on the adult male model described in Story (2005) and Story *et al.* (2018). The piriform sinuses are represented as a single side branch and are coupled to the main vocal tract at 2.4 cm from the glottis (Story, 1995; Dang and Honda, 1997). The frequency response is shown in the upper inset plot where the first three resonances are located at 596, 1401, and 2331 Hz. The sensitivity of each resonance frequency, f_{Rn}, to a change in vocal tract cross-sectional area is the difference of kinetic energy (K_e) and potential energy (P_e) within each ith section, divided by the total energy in the system (e.g., Fant and Pauli, 1975). A sensitivity function can be written as

$$S_j(i) = \frac{K_{e_j}(i) - P_{e_j}(i)}{\sum_{i=1}^{N_x}\left[K_{e_j}(i) + P_{e_j}(i)\right]}, \quad j = 1,2,3;\ i = [1, N_x], \tag{4}$$

where j is the resonance number. The kinetic and potential energies, K_e and P_e, 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. These quantities, along with the frequency response function [Fig. 1(a)], were calculated with a transmission-line type model of the vocal tract (Sondhi and Schroeter, 1987; Story and Bunton, 2017) that included energy losses due to yielding walls, viscosity, heat conduction, and acoustic radiation at the lips. The sensitivity functions calculated for $\Omega(i)$ are shown in the upper panel of Fig. 1(b) where the solid, dotted, and dashed lines indicate the sensitivity of the first, second, and third resonance frequencies (f_{R1}, f_{R2}, f_{R3}), respectively, to a small perturbation of the area function, $\Delta\Omega(i)$. This relation can be written as

$$\frac{\Delta f_{Rj}}{f_{Rj}} = \sum_{i=1}^{N_x} S_j(i)\frac{\Delta\Omega(i)}{\Omega(i)}, \tag{5}$$

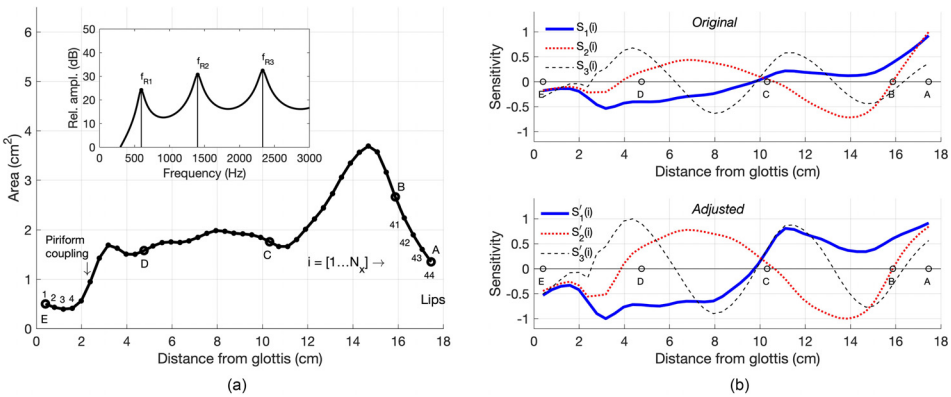

FIG. 1. (Color online) Area function and sensitivity functions for the neutral vocal tract $\Omega(i)$. The points marked with A–E are the approximate locations of the lips (A), incisors (B), hard-palate/soft-palate junction (C), superior aspect of the epiglottis (D), and glottis (E), respectively. (a) Area function for $\Omega(i)$ is shown with a solid line and dots; each dot represents the ith cross-sectional area; inset plot shows the frequency response where peaks are the vocal tract resonances. (b) Sensitivity functions calculated with Eqs. (4) and (5) that correspond to the first three resonances of the area function.

where j is again the resonance number. Equation (5) dictates that an upward shift in the resonance frequency will occur when a positive change in area, $\Delta\Omega(i) > 0$, is imposed at values of i where $S_j(i) > 0$, or when a negative change in area, $\Delta\Omega(i) < 0$, is imposed where $S_j(i) < 0$; the opposite shift in resonance frequency occurs if the polarities of $\Delta\Omega(i)$ and $S_j(i)$ oppose each other.

B. Adjustments to the sensitivity functions

To avoid disproportionate influence of any particular region of the vocal tract on the deformation function, the second step is an adjustment that balances the magnitude of each sensitivity function $S_j(i)$ from the glottis to the lips. The adjustment is carried out by first storing the polarities of each ith section of the jth sensitivity function in a vector such that $Q(i) = 1$ for $S_j(i) \geq 0$ and $Q(i) = -1$ for $S_j(i) < 0$, where $i = [1, N_x]$. Next, $|S_j(i)|$ is low-pass filtered (second order Butterworth) with a normalized cutoff frequency of 0.1, assigned to a vector $R(i)$, and then used to generate the trend function $R_o(i) = R(i) + \max[|S_j(i)| - R(i)]$. An intermediate adjusted sensitivity function is determined by removing the trend such that $R_a(i) = (|S_j(i)|/R_o(i))Q(i)$, where the multiplication by $Q(i)$ restores the polarity of each section to be the same as the original $S_j(i)$. The final adjusted and normalized sensitivity function is $Z_j(i) = R_a(i)/\max(|R_a(i)|)$. Note that the $Q(i)$, $R(i)$, $R_o(i)$, and $R_a(i)$ are not assigned j indices because they are all temporary vectors used only during the adjustment process of each jth sensitivity function.

C. Calculation of the deformation function

A linear combination of the three sensitivity functions from the previous step can now be formed as

$$y(i) = \delta_1 Z_1(i) + \delta_2 Z_2(i) + \delta_3 Z_3(i), \tag{6}$$

where the coefficient weights are the δ_j components of the specified RDP vector, and determine the relative contribution of each sensitivity function to the overall shape of $y(i)$. The deformation function, $D(i, n)$, at each time sample n can now be formed by normalizing $y(i)$ relative to its minimum value, and multiplying by $E(n)$,

$$D(i,n) = \frac{-\mu E(n) y(i)}{\min\limits_{i \in [1,N_x]} y(i)}, \tag{7}$$

where the minus sign is needed to negate the effect of the denominator always being less than zero. The μ parameter controls the *degree* to which the deformation constricts the vocal tract; if $\mu < 1$, constrictions will only partially occlude the tract as is characteristic of vowels, liquids, glides, and fricatives; when $\mu = 1$ a complete closure will be formed at the location of the minimum value in $y(i)$; and if $\mu > 1$ the extent of the complete closure will spread along the length axis of the vocal tract. The final operation is to use Eq. (1) to generate the composite time-varying area function $A(i,n)$ from the product of $\Omega(i,n)$ and $D(i,n)$.

D. Sequencing multiple acoustic events

Word and sentence-level speech can be generated by sequencing multiple acoustic events along a time axis. Because there may be considerable temporal overlap of the event functions, some additional considerations are needed to generate a deformation function. At every time sample n, the steps described previously in Secs. III A–III C are executed in a loop where each iteration k attends to one event function. The order of execution is carried out in ascending order of the values of μ, and the output of each iteration replaces the original $\Omega(i)$ vector. That is, when multiple event functions are specified, the RDP associated with the smallest μ is used to generate the initial deformation at a given time sample producing a "temporary" $A_k(i,n)$ which is then fed back through each of the three steps in Secs. III A–III C, where the next iteration will attend to the RDP with next smallest μ value.

IV. SENTENCE-LEVEL SPEECH PRODUCTION

In this section, use of the vocal tract model to produce sentence-level speech is demonstrated by generating synthetic versions of "*a dog ate a bug*" and "*a frog ate a fly*." The first sentence contains only vowels and stop consonants, whereas the second includes the added complexity of fricative-liquid clusters.

A. Sentence 1

The RDPs, μ values, and event functions for "*a dog ate a bug*" are shown positioned sequentially along a timeline in the upper panel of Fig. 2(a). Distributed along the top of the plot are the ARPAbet phonetic symbols (Shoup, 1980; Klatt, 1987) associated with each acoustic event (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). An exception is the unstressed neutral vowel {ax} which is expressed in the output signal simply by the absence of any other specified event. An {ax} occurs at the beginning of the sentence, and again at about 1.05 s. The peak of the first event, corresponding to {d}, occurs at $t_p = 0.14$ s, and has a half-width $t_w = 0.1$ s [see Eq. (3)]. The RDP associated with the first event specifies a downward deflection of the first vocal tract resonance and upward deflections of the second and third resonances, respectively. The value of μ is 1.1 which assures that the vocal tract will be fully occluded when the event function reaches its peak, and the occlusion will spread spatially along the vocal tract axis because $\mu > 1$.

The second event in the sequence, whose peak is located at 0.34 s, is intended to be the vowel {ao} ("aw") and has an RDP that directs the first resonance upward in frequency and the second resonance downward. The deflection of the third resonance is left unspecified as indicated by $\delta_3 = 0$; this does not mean that δ_3 must always be zero for this vowel, but was deemed sufficient for this particular case. With its location in time and a width of $t_w = 0.21$ s, the {ao} event function generates considerable temporal overlap with the previous {d} event, as well as with the subsequent {g} event, as can be seen by the darker shading in the figure. During these

intervals of overlap, the multiple-event sequencing process described in Sec. III is used to determine the vocal tract area configuration at each point in time.

The next seven events are specified in similar fashion where the μ values for the stop consonants are 1.0 or greater and the vowels are less than 1.0. It can be noted that the event function for the {g} includes a period of time, denoted as t_h, where the peak value is held constant at 1.0 in order to sustain the occlusion; this is not a necessary condition to produce {g} but was useful in the timing of the events for this sentence. The extensive temporal overlap of the {eh} and {ih} vowels (peaks located 0.79 and 0.86 s, respectively) produces the diphthong in the word "ate," which is, in turn, heavily overlapped with the event function for the {t}. The final three events generate "bug," again with extensive overlap in time. The timing parameters for all specified events are given in Table I.

Collectively, the relative acoustic events specified for the sentence generate the time-varying vocal tract area function $A(i, n)$ shown in Fig. 2(b). The lips are labeled as point A, the glottis as point E, and the white lines labeled B, C, and D indicate the approximate anatomic landmarks of the incisors, hard-palate/soft-palate junction, and superior aspect of the epiglottis, respectively. At every point in time, the shape of the area function is influenced by multiple events, and thus represents the coarticulation of the phonetic segments. The complete occlusions indicated by the arrows in the figure are located at points along the vocal tract length that are fairly typical for the bilabial, alveolar, and velar stop consonants they are intended to produce, even though their specification was based entirely on relative deflections of the vocal tract resonances.

Using an algorithm to calculate wave propagation in the vocal tract coupled with a kinematic model of vocal fold vibration (cf. Story, 2013), the $A(i,n)$ in Fig. 2(b) produced the speech signal plotted in the middle panel of Fig. 2(a). The input parameters of the vocal fold model were set to generate a rising and falling fundamental frequency contour, and an abductory maneuver to assure that the {t} in "ate" was unvoiced in the output signal. Aspiration noise produced by glottal turbulence was emulated by adding a noise component to the glottal flow when the Reynolds number within the glottis exceeded a threshold value (Story, 2013). The noise component of the flow was generated in the form proposed by Fant (1960) such that

$$U_{nois} = \begin{cases} N_f(Re^2 - Re_c^2)(1 \times 10^{-6}) & \text{for } Re > Re_c \\ 0 & \text{for } Re \le Re_c \end{cases} \quad (8)$$

where N_f is a broadband noise signal (random noise generated with values ranging in amplitude from -0.5 to 0.5) that has been band-pass filtered between 500–2500 Hz (second order Butterworth), Re is the calculated Reynolds number, and $Re_c = 1200$ is the threshold value below which no noise is allowed to be generated. A similar noise source is used in the

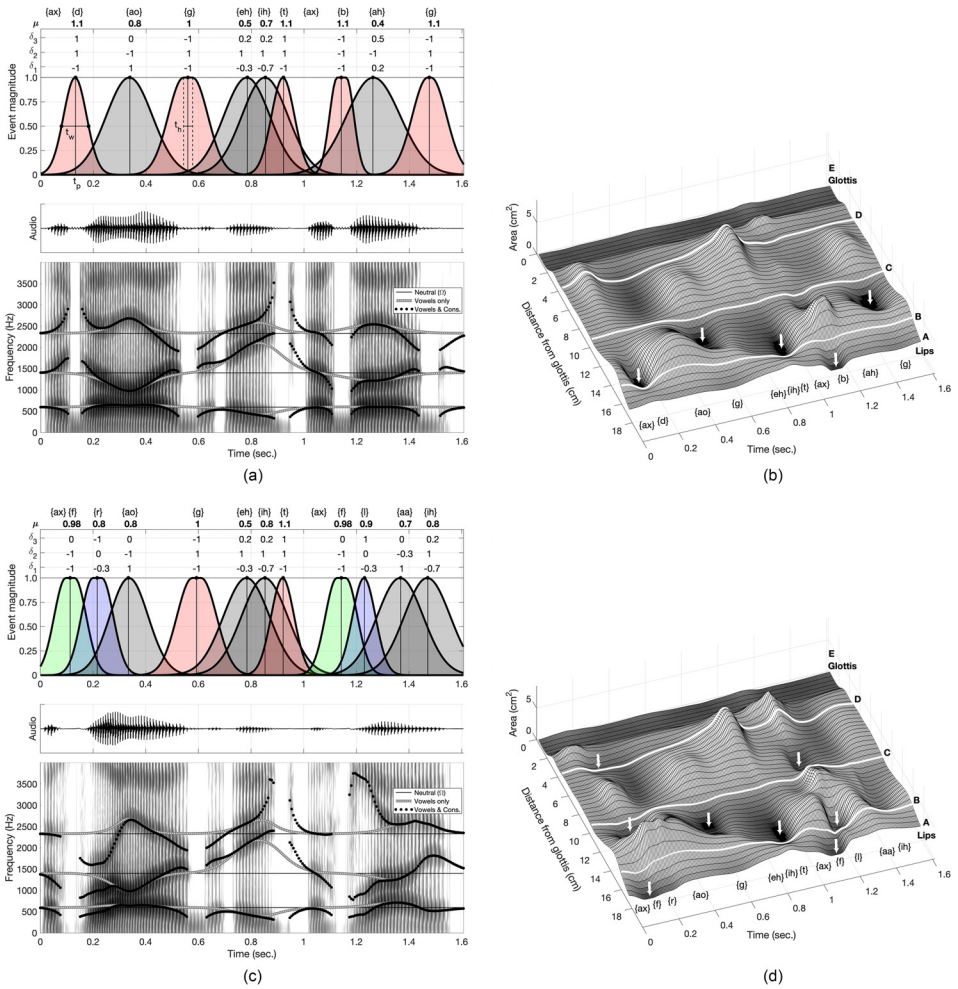

FIG. 2. (Color online) Specification of events to produce two sentences and model output. (a) RDPs, event functions, waveform, spectrogram, and calculated vocal tract resonances for the sentence "*a dog ate a bug*." Overlap of the events is indicated by darkened shading. The t_p, t_w, and t_h are examples of temporal values of peak, width, and hold, respectively. (b) 3 D surface plot of time-varying area function $A(i, n)$ where the arrows mark the stop consonant constrictions and the points marked with (A)–(E) are the anatomic landmarks described in the text and caption for Figs. 1(c) and 1(d) are identical to (a) and (b) but for the sentence "*a frog ate a fly*."

TABLE I. Timing parameters t_p, t_w, and t_h for the two sentences. All values are in seconds but can be converted to samples by multiplying by $f_{svt} = 146$.

Sentence 1						*"a dog ate a bug"*					
Event number	1	2	3	4	5	6	7	8	9		
Symbol	d	ao	g	eh	ih	t	b	ah	g		
t_p (s)	0.14	0.34	0.55	0.79	0.86	0.92	1.13	1.27	1.47		
t_w (s)	0.10	0.21	0.14	0.21	0.21	0.10	0.07	0.24	0.14		
t_h (s)	0	0	0.03	0	0	0	0.03	0	0.01		
Sentence 2						*"a frog ate a fly"*					
Event number	1	2	3	4	5	6	7	8	9	10	11
Symbol	f	r	ao	g	eh	ih	t	f	l	aa	ih
t_p (s)	0.10	0.21	0.34	0.58	0.79	0.86	0.92	1.13	1.23	1.37	1.47
t_w (s)	0.10	0.10	0.17	0.14	0.21	0.21	0.10	0.10	0.10	0.21	0.21
t_h (s)	0.03	0.03	0	0.03	0	0	0	0.03	0	0	0

vocal tract where the Reynolds number is calculated in each ith section at every time sample n, and if it exceeds the threshold value Re_c, noise is switched on at a location immediately downstream of that point (cf. Flanagan, 1972, p. 54).

The corresponding wideband spectrogram is shown in the bottom panel of Fig. 2(a), and is overlaid with three sets of calculated resonance frequencies. The first set consists of the resonance frequencies calculated from the neutral area function, $\Omega(i)$, and are shown as the thin, static, horizontal lines extending from the beginning to the end of the sentence. These are the reference values for the deflections imposed by the RDP specifications. A second set, shown as thick gray lines, represents a special case for which only the vowel events in the sentence were allowed to influence the area function (i.e., μ values were set to zero for all consonant events). These show how the resonances are deflected away from the horizontal lines (neutral resonances) according to the specified RDPs. The third set, shown as black dots, tracks the resonances generated from the time-varying area function with all vowel and consonant events included. The breaks indicate time intervals during which the vocal tract was fully occluded or nearly so; these lines also track the formant frequencies in the wideband spectrogram. Viewing the thick gray lines (vowel events only) along with the black dots (all events) shows the relative and coarticulated nature of the overlapped events. For example, between about 0.7–0.9 s both f_{R2} and f_{R3} are sweeping upward in frequency due to the {eh}-{ih} diphthong events, but the RDP for the subsequent {t} also specifies an upward deflection of the same two resonances. The model does indeed assure that both f_{R2} and f_{R3} are deflected above those of the diphthong alone, even though they were already deflected well above the resonances of the neutral vocal tract shape $\Omega(i)$.

An audio file of the synthesized sentence and a slow-motion animation of the time-varying vocal tract shape are available as multimedia files Mm. 1 and Mm. 2, respectively. The vocal tract animation is a projection of the equivalent radii of the time-varying area function onto a 2D profile (Story *et al.*, 2018), and the inset plot shows the calculated resonance frequencies.

Mm. 1. Synthesized sentence "*a dog ate a bug*." This is a file of type "wav" (142 Kb).

Mm. 2. Animation of the time-varying vocal tract and resonance frequencies for the sentence "*a dog ate a bug*." This is a file of type "mov" (1 Mb).

B. Sentence 2

Figures 2(c) and 2(d) show event functions, time-varying area function, spectrogram, and calculated resonances for the second synthesized sentence, "*a frog ate a fly*." The total duration is the same as the first sentence, and the temporal characteristics of the events {ao}, {g}, {eh}, {ih}, {t} are either the same or quite similar (some slight adjustments were made to accomodate different consonant events). The two unstressed neutral vowels {ax} are again produced during the absence of any other specified event. What is different from the first sentence is that the first and second events specify a cluster consisting of the fricative {f} and liquid {r}, and the eighth and ninth events specify a similar cluster of {f} and {l} followed by the diphthong {aa}-{ih} (see Table I for timing parameters). The RDP for each {f} deflects the first two resonances downward and, with $\mu = 0.98$, will almost fully occlude the vocal tract, but not quite, as is needed for a fricative consonant. In addition, the event functions for both {f}'s include a 0.03 s hold duration (t_h) at the peak value to generate a fricative sound. The liquids were specified primarily by δ_3, which was set to -1 for {r} and $+1$ for {l}. The other two parameters had the same value for both liquids and were set to $\delta_1 = -0.3$ and $\delta_2 = 0$. The value of μ was set to 0.8 for {r} and 0.9 for {l}, both of which generate a large deflection of the third resonance, but a less severe constriction of the vocal tract than the fricative {f}.

The time-varying area function in Fig. 2(d) shows that the primary constriction generated for both {f}'s is essentially located at the lips, the location expected for a speech sound typically produced by contacting the upper incisors with the lower lip (i.e., "labio-dental"). As the first fricative blends into the {r}, two constrictions appear in the area function, one just anterior of hard-palate/soft palate junction (point C), and the other near the superior aspect of the epiglottis (point D). Similarly, the {l} that is produced around 1.2 s also contains two constrictions, but located just posterior to the incisors (point B) and posterior to hard-palate/soft-palate junction (point C), respectively.

The speech signal was generated in the same manner as the first sentence, but with two additional abductory maneuvers of the vocal folds to assure that the fricatives were unvoiced. The spectrogram [Fig. 2(a)] shows frication noise at about 0.14 and 1.16 s for the two {f}'s, respectively, followed by a lowering of the third resonance frequency for the {r} and raising of the same resonance for the {l}. Synthesis of "a frog ate a fly" and corresponding vocal tract animation are available as multimedia files Mm. 3 and Mm. 4, respectively.

Mm. 3. Synthesized sentence "*a frog ate a fly*." This is a file of type "wav" (142 Kb).

Mm. 4. Animation of the time-varying vocal tract and resonance frequencies for the sentence "*a frog ate a fly*." This is a file of type "mov" (1 Mb).

V. DISCUSSION AND CONCLUSION

The model described here was shown to accept, as input, discrete, relative specifications of acoustic speech events and transform them into modulations of the vocal tract to produce sentence-level speech. Although the two sentences synthesized for this study (included as multimedia files) are likely intelligible to many listeners, there are undoubtedly some segments that may sound unusual relative to natural human speech. This is largely due to the heuristic approach taken with regard to using the model. Other than estimating overall sentence duration, neither of the synthesized sentences were, in any way, based on analysis of audio recordings of human speech production. Rather, the events, as shown in Fig. 2, were laid out along a timeline and manually adjusted until a version of each sentence was deemed reasonable by informal listening. Most difficult, and perhaps noticeable from the audio files, was setting the timing of both the vocal tract and laryngeal events for the fricative-liquid clusters in the second sentence in order to generate a plausible voiceless {f} followed by either the {r} or {l}, which, of course, are both voiced.

A next step is to perform perceptual experiments that explore listeners' sensitivity to variations in the RDP values and timing of events. For example, all of the stop consonants in the two synthesized sentences were specified by a set of 1's with either a positive or negative polarity. Perhaps those same consonants could be more naturally generated with magnitudes less than 1.0 depending on the surrounding vowel context. That is, coarticulation may be more naturally produced with flexibility in the magnitudes of the RDPs. Also, considering that each vowel event is a syllable nucleus, it would be of interest to understand how much variability can be imposed on their temporal locations and still retain the same perceptual response. The effects of compressing or expanding the acoustic events in time on the resulting vocal tract modulations could provide additional insights into articulatory variability due to speech rate.

The vocal tract modulations generated by the model produced constrictions and expansions at locations along the vocal tract length axis that are roughly similar to those expected based on general knowledge of articulation, even though an utterance was *planned* entirely by specifying the deflection patterns of the vocal tract resonances. The output of the model, however, both in terms of time-varying area functions and speech waveforms, also needs to be compared to articulatory (articulography, MRI, etc.) and acoustic data collected from human talkers. This will allow for an evaluation of whether the vocal tract modulations are physiologically realistic in a wide variety of phonetic contexts.

Although the model demonstrated here was based on an adult male speech production system, the process of planning an utterance by specifying relative acoustic events along a time line is independent of the talker. This means that the same two sentences generated in this study (or other words, phrases, and sentences) could be produced with qualities of a completely different talker (e.g, sex, age, size, etc.) simply by substituting a different vocal tract and voice source. Of interest would be whether the speech production system of a variety of talkers generates the same or different vocal tract modulations for the same set of acoustic events. The model is also independent of language. The two sentences synthesized by the model were English; however, this is only the case because the acoustic events were arranged according to the phonological rules of English. Sentences in another language could be generated by using a different set of phonological rules.

ACKNOWLEDGMENTS

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

Browman, C. P., and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**, 155–180.
Coker, C. H. (**1976**). "A model of articulatory dynamics and control," Proc. IEEE **64**(4), 452–460.
Dang, J., and Honda, K. (**1997**). "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101**, 456–465.
Fant, G. (**1960**). *The Acoustic Theory of Speech Production* (Mouton, The Hague).
Fant, G., and Pauli, S. (**1975**). "Spatial characteristics of vocal tract resonance modes," in *Proceedings of Speech Communication Seminar 74*, Stockholm, Sweden (August 1–3), pp. 121–132.
Flanagan, J. L. (**1972**). *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, Berlin).
Kent, R., and Minifie, F. D. (**1977**). "Coarticulation in recent speech production models," J. Phon. **5**(1), 15–133.
Klatt, D. H. (**1987**). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am. **82**(3), 737–793.
Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type Line analog," DS dissertation, Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech., Stockholm, Sweden.
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. L. Hardcastle and A. Marcha (Kluwer Academic, Dordrecht), pp. 131–149.
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.
Rubin, P., Baer, T., and Mermelstein, P. (**1981**). "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. **70**, 321–328.

Shoup, J. E. (**1980**). "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, edited by W. A. Lea (Prentice-Hall, Englewood Cliffs, NJ), pp. 125–138.
Sondhi, M., and Schroeter, J. (**1987**). "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust. Speech Sign. Process. **35**(7), 955–967.
Story, B. H. (**1995**). "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, University of Iowa, Iowa City, IA.
Story, B. H. (**2005**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am. **117**(5), 3231–3254.
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," Comput. Speech Lang. **27**(4), 989–1010.
Story, B. H., and Bunton, K. (**2017**). "An acoustically-driven vocal tract model for stop consonant production," Speech Commun. **87**, 1–17.
Story, B. H., Vorperian, H., Bunton, K., and Durtschi, R. (**2018**). "An age-dependent vocal tract model for males and females based on anatomic measurements," J. Acoust. Soc. Am. **143**(5), 3079–3102.
Toutios, A., Ouni, S., and Laprie, Y. (**2011**). "Estimating the control parameters of an articulatory model from electromagnetic articulograph data," J. Acoust. Soc. Am. **129**(5), 3245–3257.
Westbury, J. R. (**1994**). X-ray microbeam speech production database user's handbook (version 1.0) (UW-Madison).