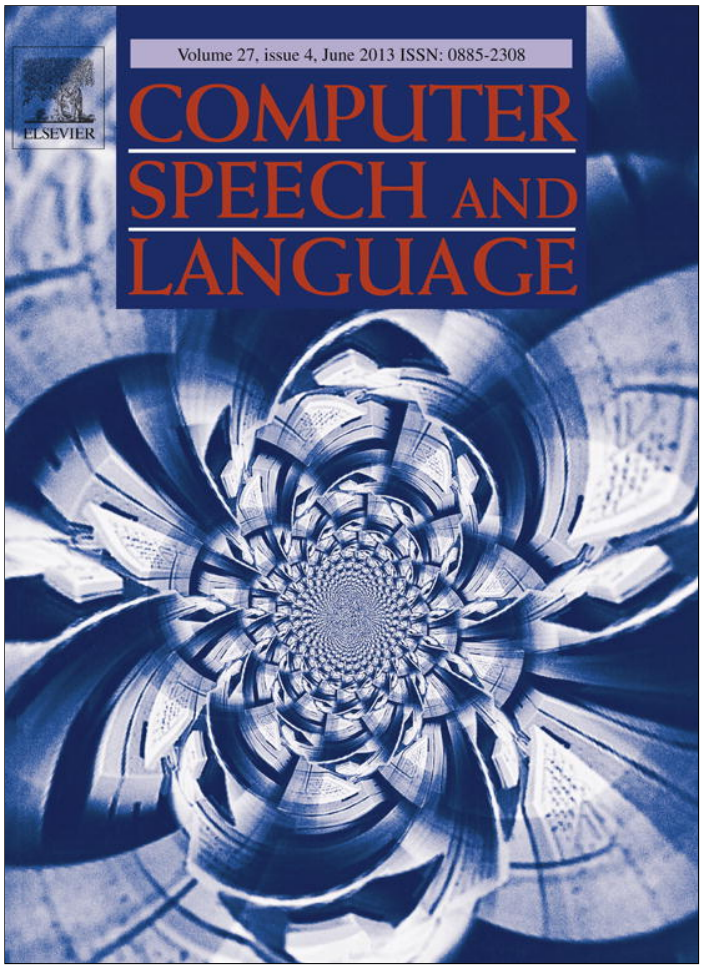
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 989–1010

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Phrase-level speech simulation with an airway modulation model of speech production☆

Brad H. Story*

Speech Acoustics Laboratory, Dept. of Speech, Language, and Hearing Sciences, University of Arizona, 1131 E. 2nd St., P.O. Box 210071, Tucson, AZ 85721, United States

Received 28 June 2012; received in revised form 12 October 2012; accepted 15 October 2012
Available online 26 October 2012

Abstract

Artificial talkers and speech synthesis systems have long been used as a means of understanding both speech production and speech perception. The development of an airway modulation model is described that simulates the time-varying changes of the glottis and vocal tract, as well as acoustic wave propagation, during speech production. The result is a type of artificial talker that can be used to study various aspects of how sound is generated by humans and how that sound is perceived by a listener. The primary components of the model are introduced and simulation of words and phrases are demonstrated.
© 2012 Elsevier Ltd. All rights reserved.

Keywords: Vocal tract; Vocal folds; Modulation; Speech simulation; Speech synthesis

1. Introduction

Speech is produced by transforming the motion of anatomical structures into an acoustic wave embedded with the distinctive characteristics of speech. This transformation can be conceived as a modulation of the human airway system on multiple time scales. For example, the rapid vibration of the vocal folds modulates the airspace between them (i.e., the glottis) on the order of 100–400 cycles per second to generate a train of flow pulses that excites the acoustic resonances of the trachea, vocal tract, and nasal passages. Simultaneous, but much slower movements of the tongue, jaw, lips, velum, and larynx can be executed to modulate the shape of the pharyngeal and oral cavities, coupling to the nasal system, and space between the vocal folds by adduction and abduction maneuvers. These relatively slow modulations shift the acoustic resonances up or down in frequency and valve the flow of air through the system, thus altering the characteristics of the radiated acoustic wave over time, and providing the stimulus from which listeners can extract phonetic information.

The view that human speech is produced by a modulation system was expressed by Dudley (1940) in an article called "The carrier nature of speech." In it he referred to the relatively high-frequency excitation provided by phonation or noise generation as "carrier waves" that are modulated by slowly-varying, and otherwise inaudible, movements of the vocal tract called "message waves." He based this view on experience in developing both the VOCODER (Dudley,

☆ This paper has been recommended for acceptance by Bernd Moebius.

* Tel.: +1 520 626 9528; fax: +1 520 621 9901.
E-mail address: bstory@email.arizona.edu

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.10.005

Author's personal copy

1939) and the human-operated VODER (Dudley et al., 1939), and, in the conclusion, made a curious point that a wide variety of carrier signals – even nonhuman sounds such as instrumental music – could be modulated by the "message waves" and still produce intelligible "speech." This points to the importance of understanding articulatory movement in terms of how it modulates the shape of pharyngeal and oral airspaces over time which, in turn, modulates the acoustic characteristics of the speech signal. Traunmüller (1994) also proposed a modulation theory in which speech signals are considered to be the result of articulatory gestures, common across speakers, that modulate a "carrier" signal unique to the speaker. In this theory, however, the carrier signal is not simply the excitation signal, but includes any aspects of the system that are phonetically neutral and descriptive of the "personal quality" of the speaker. This suggests that embedded within the carrier would be contributions of the biological structure of the vocal tract as well as any idiosyncratic vocal tract shaping patterns, all of which would be modulated during speech production by linguistically meaningful gestures.

Studying speech as a modulation system can be aided by models that allow for enough control of relevant parameters to generate speech or speech-like sounds. Within such models, the shape of the trachea, vocal tract, and nasal passages is usually represented as a tubular system, quantified by a set of *area functions* (cf., Fant, 1960; Baer et al., 1991; Story et al., 1996). This permits computing the acoustic wave propagation through the system with one-dimensional methods in the time domain (cf., Kelly and Lochbaum, 1962; Maeda, 1982; Strube, 1982; Liljencrants, 1985; Smith, 1992) or frequency domain (Sondhi and Schroeter, 1987). Typically for speech, only the vocal tract portion, along with the nasal coupling region, is considered to be time-varying. Thus, the challenge for developing a model that can "speak" is to define a set of parameters that allow efficient, time-dependent control of the shape of the vocal tract area function and coupling to the nasal system.

An articulatory synthesizer is perhaps the most intuitively-appealing approach to controlling the vocal tract because the model parameters consist of positions and movements of the tongue, jaw, lips, velum, etc. These are often represented in the two-dimensional midsagittal plane (cf., Lindblom and Sundberg, 1971; Mermelstein, 1973; Coker, 1976; Maeda, 1990; Scully, 1990) or as more complex three-dimensional models of articulatory structures (Dang and Honda, 2004; Birkholz et al., 2006, 2007), where, in either case, articulatory motion can be simulated by specifying the temporal variation of the model parameters. At any given instant of time, however, the articulatory configuration must be converted to an area function by empirically-based rules in order to calculate acoustic wave propagation, and ultimately produce an acoustic speech signal suitable for analysis or listening (e.g., Rubin et al., 1981; Birkholz et al., 2010; Bauer et al., 2010).

Other approaches consist of parameterizing the area function directly, rather than attending specifically to the anatomical structures. These are particularly useful when precise control of the vocal tract shape is desired. Early examples of this approach are the three-parameter models of Fant (1960) and Stevens and House (1955) in which the area function was described by a parabola controlled by a primary constriction location, cross-sectional area at that location, and a ratio of lip opening length to its area. These models were later modified to include various enhancements (Atal et al., 1978; Lin, 1990; Fant, 1992, 2001). Another type of area function model was proposed by Mrayati et al. (1988). The model parameters were not directly related to articulation but rather to portions of the area function determined to be acoustically-sensitive to changes in cross-sectional area.

Using any of these types of models to produce connected, coarticulated speech requires that the parameters allow for blending of the vowel and consonant contributions to the vocal tract shape. Ohman (1966, 1967) suggested that a consonant gesture (localized constriction) is superimposed on an underlying vowel substrate, rather than considering consonant and vowel to be separate, linearly sequenced gestures. Based on this view, Båvegård (1995), Fant and Båvegård (1997) detailed an area function model in which the vowel contribution was represented by a three-parameter model (Fant, 1992), as mentioned previously, and a consonant constriction function that could be superimposed on the vowel configuration to alter the shape of the area function at a particular location. Ohman's notion of vowel and consonant overlap has also influenced theoretical views of speech motor control. Gracco (1992), for instance, suggested that the vocal tract be considered the smallest unit of functional behavior for speech production, and that the movements of the vocal tract could be classified into "shaping" and "valving" actions. Relatively slow changes of overall vocal tract geometry coincide with the *shaping* category and would generally be associated with vowel production, whereas *valving* actions would impose and release localized constrictions, primarily for consonants.

Story (2005a) introduced an area function model conceptually similar to that of Båvegård (1995), Fant and Båvegård (1997). That is, the model operates under the assumption that consonantal, or more accurately, obstruent-like constrictions can be superimposed on an underlying vowel-like area function to momentarily produce an occlusion or partial

Author's personal copy

occlusion at some location in the vocal tract. It is different, however, in that the vowel substrate is generated by superimposing two shaping patterns, called *modes* (Story and Titze, 1998), on an otherwise neutral vocal tract configuration. Thus, at any instant in time, the shape of the area function and the subsequent acoustic output include contributions from multiple layers of modulation: (1) idiosyncratic characteristics of the neutral vocal tract, (2) the overall shaping influence of the modes, and (3) possible valving (constriction) functions that force some part of area function to become zero or nearly so. The framework for the model is supported by work based on analysis of data from both MRI and X-ray microbeam articulography (Westbury, 1994). Story (2005b, 2007) has shown that the two shaping modes seem to be fairly similar across talkers for vowel production, whereas, the mean or neutral vocal tract shape is generally unique to a talker. In addition, Story (2009b) has shown that a time-varying vocal tract shape representative of a VCV utterance can be separated into a vowel-like substrate and contributions from the obstruent-like constriction.

The purpose of this article is to describe an airway modulation model of speech production based on the parametric system reported in Story (2005a), and to demonstrate that it can be used to simulate word-level and phrase-level utterances. The components of the model consist of kinematic representations of the medial surfaces of the vocal folds (Titze, 1984, 2006) and the shape of vocal tract as an area function (Story, 2005a), as well as static representations of the nasal passages and sinuses, and trachea. The assembly of these components into a system that can generate simulated speech has been used recently to investigate the acoustics and perception of vowels, consonants, and various voice qualities (Bunton and Story, 2009, 2010, 2011; Story and Bunton, 2010; Samlan and Story, 2011). These studies, however, have focused on simple utterances such as isolated vowels and VCVs which are fairly straightforward to simulate with systematic variation of model parameters. The aim in this article is to present a pilot study of several cases in which the model parameters are varied in more complex fashion to simulate a small collection of words and phrases. In the first part, the background and main components of the model are briefly described. The use of the model to generate words and phrases will be demonstrated in the second part.

2. Airway modulation model

The airway modulation model is constructed such that a baseline configuration of the laryngeal, vocal tract, and nasal systems would, without any imposed control, produce a neutral, monotone vowel sound. The model parameters can then be activated such that they alter the baseline configuration in some desired manner such that the acoustic properties of the generated signal is changed. The parameters of the model are controlled by a set of hierarchical tiers as shown in Fig. 1. The leftmost column indicates quantities that define the basic physical structure of the system, the second column contains all the time-varying control parameters, the third column shows the covert intermediate quantities, and the rightmost column indicates the output quantities needed for sound production. Thus, the input parameters in the control tiers impose time-dependent deformations on the shape of the glottis and vocal tract to produce speech. Each control tier will be described in the following sections.

2.1. Kinematic model of the medial surfaces of the vocal folds

Modulation of the glottal airspace is accomplished with a kinematic representation of the vibrating portion of the vocal fold medial surfaces in which time-varying surface displacements are superimposed onto a postural configuration (Titze, 1984, 2006). As can be seen in Fig. 2, the prephonatory posture is defined by superior (ξ_{02}) and inferior (ξ_{01}) values of separation at the vocal processes and by a bulging parameter (ξ_b) that provides curvature to the medial surface. The vocal fold length (antero-posterior dimension of the surface along midline) and thickness (inferior-superior extent of the surface) can be specified to be characteristic of a particular talker.

The time-varying vibratory displacement is based on a summation of translational and rotational modes in the vertical dimension and a ribbon mode in the antero-posterior dimension. The amplitude of vibration and mucosal wave velocity are governed by rules as described in Titze (2006). Any of the vibratory, aerodynamic, or structural parameters (e.g., fundamental freq. ($F0$), bronchial pressure P_b, vocal process separation ξ_{02}, etc.) can also be made to be time-varying as indicated in the "Tier 0" portion of Fig. 1. For example, the degree of separation of the vocal processes (ξ_{02}) could be increased to abduct the vocal folds for production of a voiceless consonant and then be reduced again for a voiced production. Thus, airway modulations produced at the level of the vocal folds operate on two time scales, one that is representative of their vibrational frequency (approximately 100–400 Hz) and another for the much slower adductory and abductory movements of the medial surfaces that take place during the unvoiced parts of speech. The

Author's personal copy

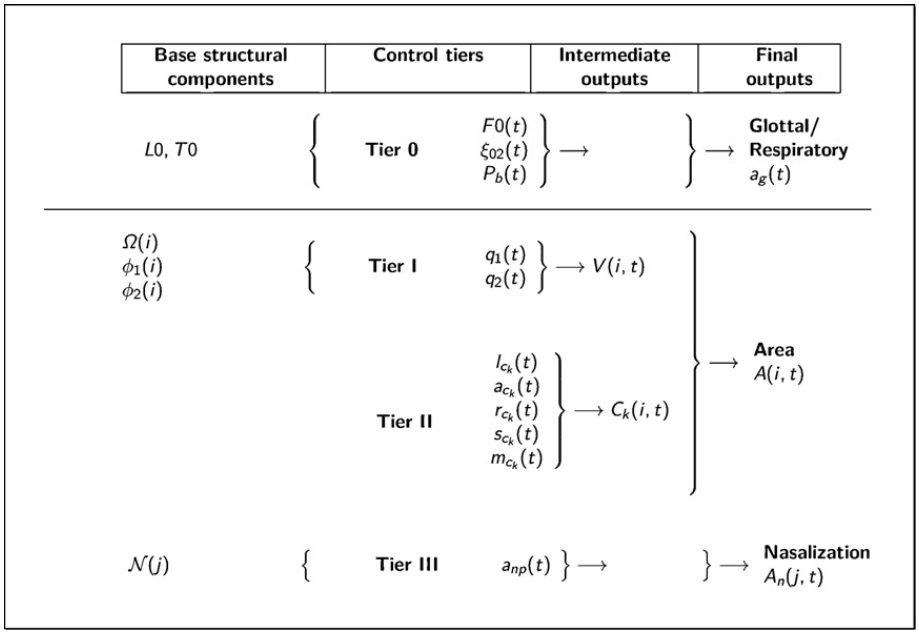


Fig. 1. Diagram of the four-tier model. Tier 0 controls the kinematic vocal fold model; the $L0$ and $T0$ are the initial length and thickness of the vocal folds, respectively. Tier I produces a vowel substrate and Tier II generates a superposition function for a consonant constriction. Time-dependent nasal coupling can be specified in Tier III. The base structural components are dependent only on a spatial dimension (i.e., i and j are indexes corresponding to spatial location), whereas the final outputs are dependent on both space and time.

output of the kinematic vocal fold model is the glottal area $a_g(t)$, calculated as the time-varying sum of the minimum area from each of the vertical channels of the medial surface. To produce sound, the glottal area is aerodynamically and acoustically coupled to the trachea and vocal tract (Liljencrants, 1985; Story, 1995; Titze, 2002). This produces a glottal flow signal which is the primary sound source for vowels.

2.2. Kinematic model of the vocal tract area function

The vocal tract component of the model was described in Story (2005a) and operates according to the view that vowel-like sounds and vowel-to-vowel transitions are produced by modulating a phonetically-neutral vocal tract configuration.

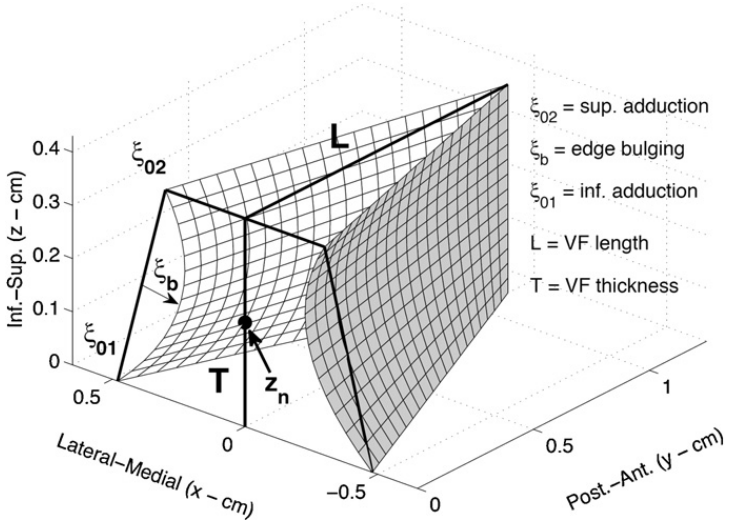


Fig. 2. Kinematic model of the medial surfaces of the vocal folds. The Post-Ant. dimension is the vocal fold length, the Inf.-Sup. dimension is the vocal fold thickness, and the Lateral-Medial dimension is vocal fold displacement. ξ_{02} and ξ_{01} are the prephonatory adductory settings of the upper and lower portions, respectively, of the posterior portion of the vocal folds. ξ_b is the surface bulging, and z_n is a nodal point around which the rotational mode of vibration pivots.

Author's personal copy

In turn, production of obstruent-like sounds, which are the basis of many consonants, results from another level of modulation that imposes severe constrictions on the underlying vowel or evolving vowel transition. This means that the length and other idiosyncratic features of the neutral vocal tract shape set the acoustic background on which vowel transitions are carried, while the vowel transitions provide the acoustic background on which obstruent-like modulations are imposed. Thus, the model allows for independent, parallel control of *shaping actions* to modify the vocal tract for vowel production, and *valving actions* that locally constrict the tract shape to produce obstruent consonants embedded in a surrounding vowel context. Secondary features of articulation such as labialization, palatalization, velarization, etc. can also potentially be produced within the various levels of the model (e.g., Story and Titze, 2002).

The length and shape of the vocal tract are represented in the model by an area function. That is, the cross-sectional area variation along the long axis of the vocal tract is specified at discrete increments of length for a given instant of time. Vowel-like sounds are represented in Tier I (see Fig. 1) as modulations of an underlying neutral tract shape (i.e., formants are nearly equally spaced) and obstruents are imposed in Tier II as modulations of the underlying vowel substrate. Time-dependent nasal coupling is controlled by Tier III. An additional tier was included in Story (2005a) that allows for dynamic length change operations at both the glottal and lip ends. Although, length changes are ultimately necessary to realistically model the vocal tract shape variation in human speech, this tier was not used for the present study because of the added complexity required to change tract length dynamically in the wave-propagation algorithm (Laakso et al., 1996; Mathur et al., 2006) used to generate the speech samples (see Section 2.3). This is a current limitation of the acoustic portion of the system, not of the parametric model of the area function.

In the first tier, vowel-to-vowel transitions denoted as $V(x, t)$ can be produced by perturbing a neutral vocal tract configuration with two shaping patterns called modes such that,

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

where x is the distance from the glottis and $\Omega(x)$, $\phi_1(x)$, and $\phi_2(x)$ are the mean vocal tract diameter function and modes, respectively, as defined in Story (2005a). The time-dependence is produced by the mode scaling coefficients $q_1(t)$ and $q_2(t)$, and the squaring operation and scaling factor of $\pi/4$ convert the diameters to areas. The mean diameter function and modes used for the present study were based specifically on those reported in Story (2009b), however, similar data from other speakers could also be substituted (cf., Story, 2005b). Additionally, the $\Omega(x)$ function can be modified in various ways to alter the overall voice quality (Story et al., 2001; Story and Titze, 2002). In any case, the coefficients [q_1, q_2] have been shown to have an essentially one-to-one relation with the first two formant frequencies [F1, F2]. As shown in Fig. 3a, the white dots indicate [q_1, q_2] pairs that, when used with Eq. (1), would generate area functions roughly representative of the vowels [i, ɑ, u]; the coefficient pair located at the origin produces the mean area function which is essentially a neutral vowel. The resulting [F1, F2] pairs corresponding to these same vowels are similarly shown as white dots in Fig. 3b. The dark grid in part (a) of the figure represents the "available" coefficient space for generating area functions, and the grid in part (b) is the acoustic space that corresponds to the [F1, F2] values of each point in the coefficient grid.[1] This mapping can also be used in the reverse direction to transform formant frequencies measured from acoustic recordings into coefficients suitable for generating time-varying area functions. The red trajectories in both parts of the figure correspond to the temporal characteristics of the coefficients and corresponding formants for the word "Ohio," and will be used in a later section.

Obstruent-like shapes are generated in Tier II with a scaling function $C(x)$ that extends along the length of the vocal tract. The value of $C(x)$ is equal to 1.0 everywhere except in the region of the desired constriction location l_c. The shaping of the constriction around l_c is determined by a Gaussian function that includes control parameters for the extent of the constriction along the tract length called r_c, and a skewing factor s_c that dictates the asymmetry of the constriction. In the velar region, r_c would typically be set to values larger than for more anterior locations to more accurately represent the extent of a constriction produced by the tongue body. When any vowel-like area function is multiplied by $C(x)$, the region in the vicinity of l_c will be reduced in area to the value specified by a_c, thus superimposing the constriction. When $a_c = 0$, the constriction function will generate a complete occlusion of the vocal tract, but it may be set to nonzero values for production of sounds with partial occlusions such as fricatives, liquids, and glides. The constriction function can be made time dependent with a temporal activation parameter called the magnitude $m_c(t)$ which serves to gradually

[1] The formant values were derived from calculations of the frequency response of each area function based on a transmission line technique (Sondhi and Schroeter, 1987; Story et al., 2000.)

Author's personal copy

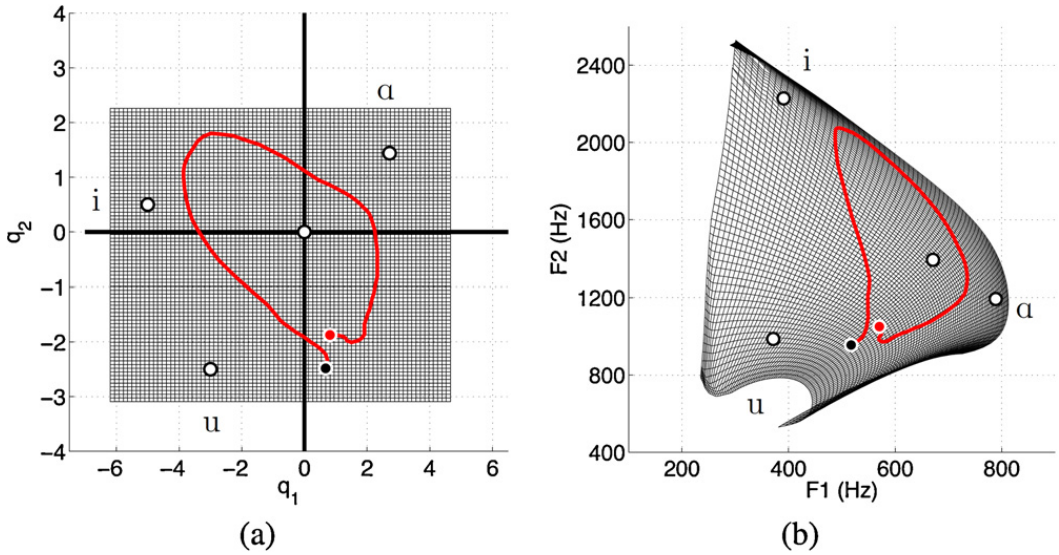


Fig. 3. Mapping of mode coefficients $[q_1, q_2]$ (see Eq. (1)) to formant frequencies [F1, F2]. (a) Coefficient space where the grid indicates the range of q_1 (x-axis) and q_2 (y-axis); the white dots are coefficient pairs that would produce area functions representative of [i, ɑ, u] and the neutral vowel (at the origin); the red trajectory indicates the variation in coefficient values to generate a time-varying area function for the word "Ohio." (b) Vowel space plot where the grid, white dots, and red trajectory represents the [F1, F2] values corresponding to the respective coefficient pairs in part (a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

impose a constriction on the area function and then remove it. When $m_c(t) = 0$, the constriction function has no effect, but when equal to 1.0 the constriction is fully realized within the area function. Multiple constriction functions $C_k(x, t)$ can be specified to allow for either simultaneous or temporally-shifted obstruent-like modulations of the vocal tract shape. Additionally, the l_c, a_c, r_c, and s_c parameters can vary in time if needed to execute a particular type of shape change.

A composite area function $A(x, t)$ is generated by the vocal tract model as the product of each of the N_{vt} elements along the x-dimension of $V(x, t)$ and $C_k(x, t)$ such that at any given time instant,

$$A(x,t) = V(x,t)\prod_{k=1}^{N_c} C_k(x,t) \quad x = \Delta[1, N_{vt}] \tag{2}$$

where N_c is the number of constriction functions. The present study is based on an adult male vocal tract (Story, 2009a) such that all area functions consist of $N_{vt} = 44$ contiguous "tubelet" sections as defined in Story (2005a), each with a length of $\Delta = 0.396825$ cm. This results in a total tract length of 17.46 cm. This particular value of Δ is derived from the characteristics of the wave propagation algorithm and the 44.1 kHz sampling frequency at which it operates. Other sampling frequencies could be chosen that effectively alter the length of the vocal tract.

An example area function is shown in Fig. 4. The glottis is located at the zero point along the x-axis. The tracheal area function, shown in black and based on data from Story (1995), extends from the glottis toward the bronchi in the negative x-direction. The 44-section vocal tract extends toward the lips in the positive x-direction, where the black line indicates the neutral tract shape derived from the mean diameter function as $(\pi/4)\Omega^2(x)$ (i.e., $[q_1, q_2] = [0, 0]$ in Eq. (1)). The blue line is a perturbation of the neutral shape by Eq. (1) when q_1 and q_2 are set to values that produce an [ɑ]-like area function, and the red line demonstrates a composite tract shape with an occlusion in the oral cavity at a location roughly representative of an alveolar stop.

The nasal coupling location is indicated by the upward pointing arrow located at about 8.8 cm from the glottis. If the nasal coupling area is nonzero, the nasal passages and sinuses become part of the overall model; their configuration $\mathcal{N}(j)$ (see Fig. 1) is based on data reported in Story (1995) and Pruthi et al. (2007). In the real human system, nasal coupling results from lowering the soft palate. This has the secondary effect of slightly modifying the shape of the vocal tract in the vicinity of the soft palate. To account for this change, the nasopharyngeal portion of $\mathcal{N}(j)$ (for $j = [1, 6]$) and corresponding region of the vocal tract area function are adjusted based on the "distributed coupling method" explained in Pruthi et al. (2007).

Author's personal copy

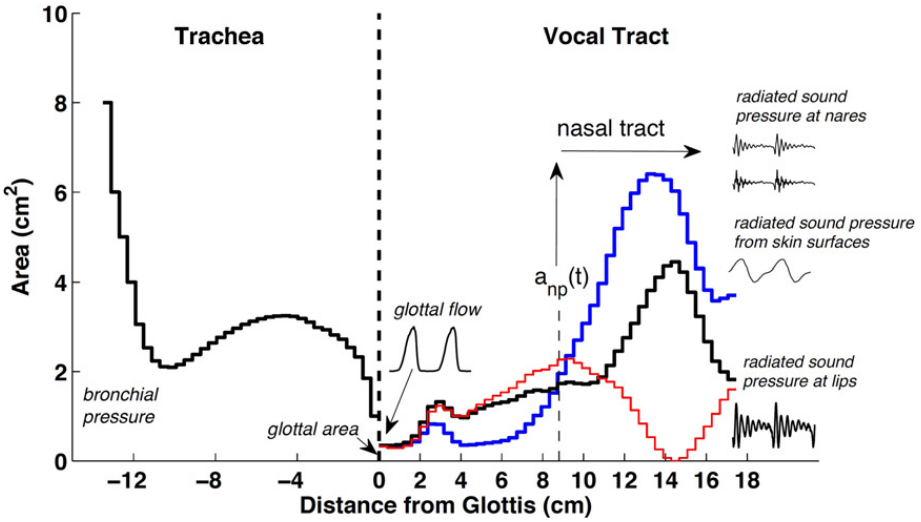


Fig. 4. Area function portion of the model. The glottis is located at 0 cm, the trachea extends from the glottis in the negative direction and the vocal tract extends in the positive direction. The black line in the positive direction indicates the area function for a neutral vowel $(\pi/4)\Omega(x)^2$, the blue line is a perturbation of the neutral shape based on Eq. (1), and the red line demonstrates an area function with an occlusion located 14.5 cm from the glottis. The nasal coupling location is indicated by the dashed line and upward pointing arrow located at about 8.8 cm from the glottis. The stair-step nature of each area function indicates the concatenated tubelet structure of the model. The waveforms shown are samples of glottal flow and radiated sound pressures at the lips, nares, and skin surfaces. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3. *Wave propagation and glottal flow*

To generate a speech signal, the glottal area $a_g(t)$ produced by the vocal fold component of the model is aerodynamically and acoustically coupled to the trachea, vocal tract, and nasal tract. The coupling is realized with a wave-reflection algorithm (Liljencrants, 1985; Story, 1995; Titze, 2002) that computes the acoustic wave propagation in the airways. Interaction of the glottal area with the time-varying acoustic pressures present just inferior and superior to the glottis generates the glottal flow signal. The wave propagation algorithm also includes energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips, nares, and skin surfaces (specifically as described in Story (1995)). A sample glottal flow signal is shown in Fig. 4 near the junction of the trachea and lower vocal tract; the radiated speech signal is shown near the lips and is analogous to a microphone signal recorded from a real talker. If the nasal coupling $a_{np}(t)$ is nonzero, there will be sound radiated at the nares as indicated in the upper right portion of Fig. 4. The sound radiated from the skin surfaces is typically of low amplitude when the vocal tract is unoccluded (e.g., vowels) or if the vocal folds are not vibrating, but becomes the dominant part of the speech signal during voiced productions with a fully or partially occluded vocal tract (e.g., voiced stops). The composite speech signal is the sum of the four signals shown at the right hand side of Fig. 4.

Additionally, aspiration noise produced by glottal turbulence is approximated by adding a noise component to the glottal flow when the Reynolds number exceeds a threshold value. This approach has been often used in speech production modeling for both aspiration and fricative type sounds (e.g., Fant, 1960; Flanagan and Cherry, 1969), although the physical realities of jet structure, vortex shedding, production of dipole and quadrupole noise sources, and potential multiple locations of such sources are not represented. The specific formulation of a noise source added to the glottal flow used in this study is based on Titze (2006). At every time instant the Reynolds number is calculated as

$$Re = \frac{u_g \rho}{L \mu} \tag{3}$$

where u_g is the instantaneous glottal flow, L is the length of the glottis, ρ is the air density, and μ is the air viscosity (Titze, 2006). The noise component of the flow is then generated in the form proposed by Fant (1960) such that

$$U_{nois} = \begin{cases} N_f(Re^2 - Re_c^2)(4 \times 10^{-6}) & \text{for } Re > Re_c \\ 0 & \text{for } Re \leq Re_c \end{cases} \tag{4}$$

Author's personal copy

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 300 and 3000 Hz (2nd order Butterworth), Re is the calculated Reynolds number, and Re_c is a threshold value below which no noise is allowed to be generated. Fant (1960) suggested that Re_c should be on the order of 1800 or less for fricative sounds. Based on spectral analysis of simulated vowels and preliminary listening experiments, a value of Re_c=1200 was chosen, along with the scaling factor of 4×10^{-6}, for all subsequent simulations in this study. The result is a noise source whose amplitude is modulated by the magnitude of instantaneous air flow through the glottis, whether during periods of vocal fold vibration or non-vibratory abductory maneuvers. A similar noise source is used in the vocal tract during production of sounds with severe constrictions such as stops, affricates, and fricatives. If the Reynolds number at any point along the vocal tract length exceeds the threshold given in Eq. (4), a noise source is switched on at a location immediately downstream of that point.

2.4. Model implementation

The complete model is currently implemented with a combination of code written in C and Matlab (2011). The vocal fold motion, flow calculations, and acoustic wave propagation are written in C and compiled as a Matlab mex file. Additional Matlab code is used to generate the time-varying area function and time-dependency of all parameters shown in Fig. 1. In its current form, the model runs with a compute-time to real-time ratio of about 2:1 on an iMac with 3.4 GHz processor, but it would be possible to optimize the code for near real-time operation if desired.

Because the airway modulation model is largely represented by modifying the shape of a tubular structure, it has been given the colloquial name of "TubeTalker" in some earlier publications where it was used to generate simple utterances. In the remainder of this article, *TubeTalker* will be used synonymously with *airway modulation model*.

3. Simulation of word-level and phrase-level speech

In this section, two words and two phrases simulated with the TubeTalker system are demonstrated graphically and with multimedia content. The words are *"Ohio"* and *"Abracadabra"*, and the phrases are *"He had a rabbit"* and *"The brown cow."* These were chosen to show how the various components of the TubeTalker model can be used to produce a range of speech sounds in a connected speech context. The word *Ohio* provides a case where the change in vocal tract shape is based only on vowels, and the consonant [h] requires only a glottal change (i.e., abduction). In contrast, *Abracadabra* contains two clusters ([bɹ]), albeit the same one twice, an unvoiced velar stop, and a voiced alveolar stop. The two phrases increase the level of complexity by requiring multiple abductory maneuvers at the glottis in *He had a rabbit*, as well as production of a nasal consonant directly preceding an unvoiced velar stop in *The brown cow.*

To simplify the presentation, the same rising and falling fundamental frequency (F0) contour was used for all four cases, as shown in Fig. 5a. The F0 range is representative of a typical adult male talker. In addition, the bronchial pressure was kept constant throughout each utterance at a value of 7840 dyn/cm^2 (equivalent to 8 cmH_2O), and the separation of the vocal processes was maintained at $\xi_{02} = 0.1$ cm for all voiced portions. The neutral vocal tract shape is shown in Fig. 5b in a "pseudo-midsagittal" configuration generated by plotting equivalent diameters of the neutral area function (red curve in Fig. 4) along a curved vocal tract profile. This view is more anatomically-intuitive than the area function representation, and will be used to demonstrate changes of vocal tract shape for the subsequent cases. This particular shape was the baseline vocal tract configuration on which all other vocal tract shape changes were superimposed. As indicated by the frequency response curve in the inset plot, it has resonances corresponding to F1, F2, and F3 located at 680, 1405, and 2442 Hz, respectively. A simulation based on the neutral tract shape and the F0 contour can be heard by playing Audio Clip 0.

Temporal variation of the model parameters for each word or phrase was based on audio recordings of the same talker from whom the vocal tract shape was derived (Story, 2009a). Formant frequencies were extracted from the vowel portions of each recording using an LPC tracking algorithm programmed in Matlab (2011). Any breaks in the formant tracks due to the presence of consonants were replaced with an interpolation from the vowel preceding the break to the vowel following the break. The formant tracks for F1 and F2 were then converted to mode coefficients via the mapping

Author's personal copy

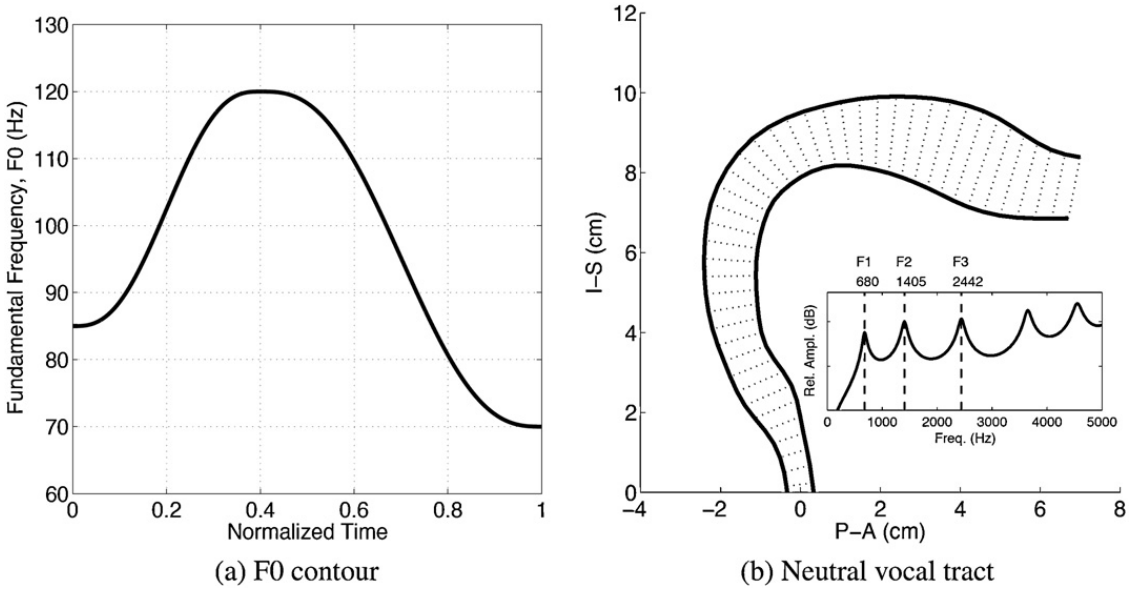


Fig. 5. Baseline settings for all simulations. (a) F0 contour. (b) Neutral vocal tract configuration on which all shape perturbations were imposed. The inset plot shows the calculated frequency response this shape, where the first three formant frequencies are indicated by the dashed lines.

shown previously in Fig. 3 to provide the Tier I parameters for vowel production. The timing of the glottal, consonant, and nasal port parameters were determined largely by inspection of the audio record and subsequent trial and error.[2]

3.1. Word 1: "Ohio"

The word *Ohio* was chosen as a first demonstration because it requires a time-varying specification of only the glottal parameter $\xi_{02}(t)$ (vocal process separation) and the vowel parameters $[q_1(t), q_2(t)]$. As can be seen in the top panel of Fig. 6, the value of $\xi_{02}(t)$ was set to be constant at 0.1 cm throughout the time course of the word except for a period from 0.15 to 0.32 s where it was increased to 0.3 cm to produce the aspiration and reduced vocal fold vibration for the [h] sound. The mode coefficient trajectory from Fig. 3a is shown as two time-dependent functions in the second panel of Fig. 6; there is little change in the mode coefficients from 0 to 0.2 s while the initial [o] and [h] onset are produced, but then change rapidly for the diphthong [ɑɪ] and then final vowel [o].

Shown in the next three panels of Fig. 6 are selected waveforms produced by the simulation. They are glottal flow $u_g(t)$, intraoral pressure $P_{oral}(t)$, and radiated pressure P_{out} (i.e., the speech signal). The glottal flow waveform is a train flow pulses whose periods vary according to the F0 contour discussed previously. In addition the DC (non-oscillatory) component increases during the period of time that the distance between the vocal processes $\xi_{02}(t)$ increases due to the abductory maneuver for the [h] and generates aspiration noise. This has the effect of decreasing the amplitude of the radiated pressure. The intraoral pressure P_{oral} signal is determined as the low-pass filtered total pressure at a point near the velar region in the vocal tract area function, and is nearly zero during the entire utterance because there are no periods of vocal tract occlusion. In the bottom panel is the wide-band spectrogram of the speech signal. It clearly shows devoiced and turbulent nature of the [h] sound, and also the rapid change in the F2 formant frequency as the [h] is released into the following diphthong [ɑɪ].

The temporal change in the vocal tract shape during production of *Ohio* is shown in Fig. 7. The red lines represent the configuration at the beginning of the utterance whereas the blue lines represent the utterance end; the dotted black lines indicate the variation in shape that occurs during the utterance. In this case, the initial and final vocal tract configurations were nearly the same, thus the red and blue lines fall almost on top of each other. The inset plot shows the variation of first three formant frequencies calculated directly based on the modulation of the neutral vocal tract shape; they follow essentially same path as the formant bands in the spectrogram (Fig. 6 bottom panel) except that the

[2] Story (2009b) describes a method for extracting model parameters for VCVs from articulatory fleshpoint data, but is not yet developed to the point of providing adequate information at the word or phrase level.

Author's personal copy

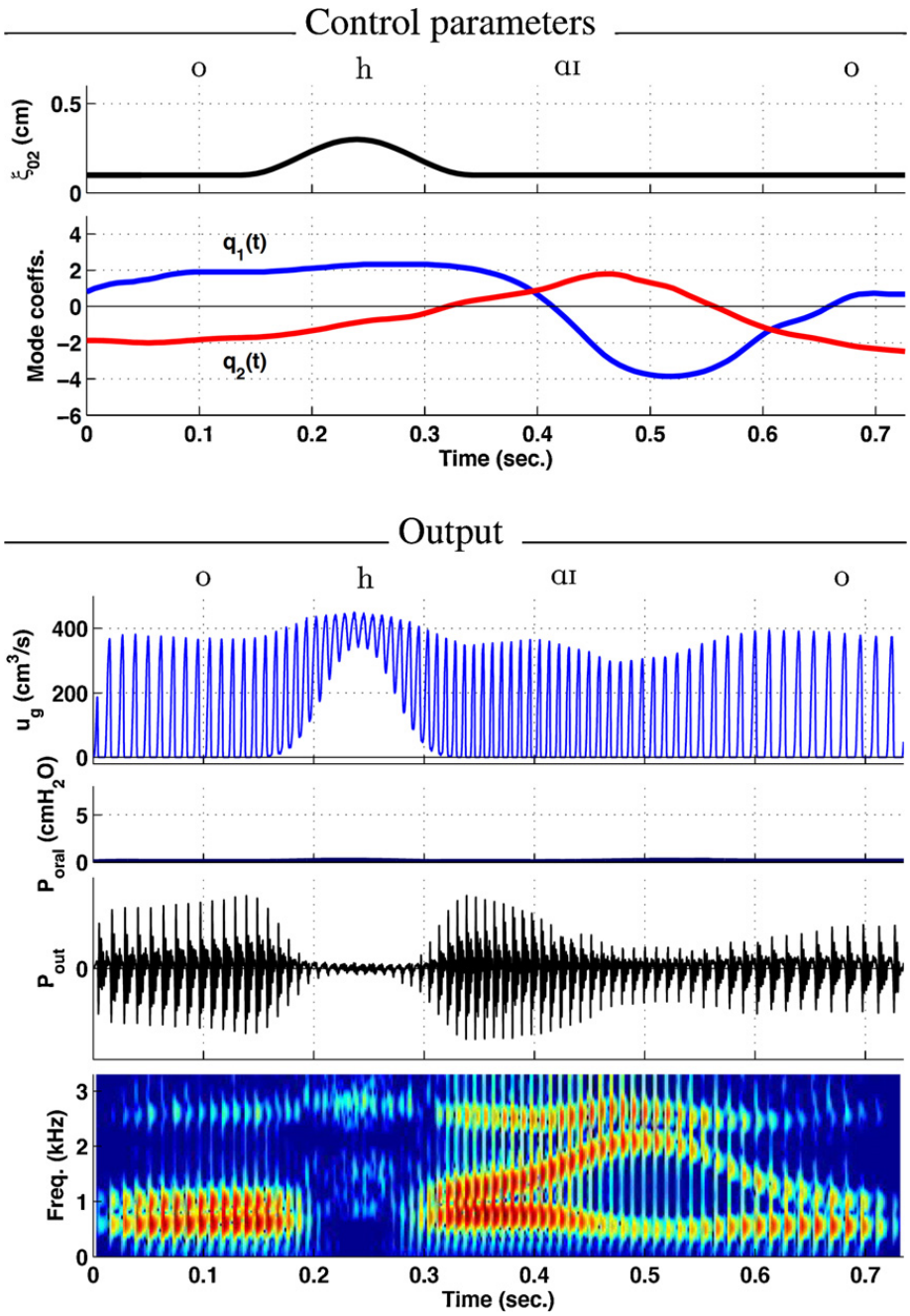


Fig. 6. Simulation of the word *Ohio*. Time-dependent control parameters are shown in the upper two panels; for this word, only ξ_{02} and the mode coefficients $[q_1, q_2]$ were varied. Glottal flow u_g, intraoral pressure P_{oral}, total radiated pressure P_{out}, and the wide-band spectrogram of P_{out} are shown in the lower four panels.

break in the formants due to the unvoiced [h] is not present. The speech signal can be heard by playing Audio Clip 1. An animation of the vocal tract producing *Ohio* is available for viewing in Video Clip 1; it shows a slow motion (without audio) view of the vocal tract movement along with a running wide-band spectrogram.

3.2. Word 2: "Abracadabra"

The second simulation was the word *Abracadabra*. The control parameters shown in Fig. 8 indicate a more complex specification of temporal variation than the previous case. At the glottal level, the ξ_{02} value is maintained at 0.1 cm throughout the utterance except for a period from 0.35 to 0.42 s where it is briefly increased to 0.5 cm for the voiceless [k]. The mode coefficients in the second panel dictate a modulation of the vocal tract shape that produces the continuous

Author's personal copy

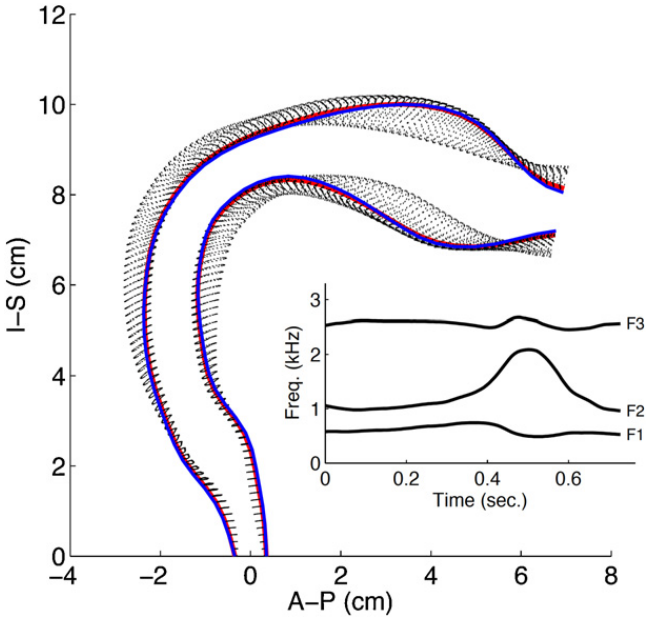


Fig. 7. Vocal tract modulation for *Ohio*. The red lines represent the configuration at the beginning of the utterance, the blue lines represent the utterance end, and the dotted black lines indicate the variation in shape that occurs during the utterance. The inset plot shows the variation of first three formant frequencies calculated directly based on the changing vocal tract shape. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

vowel sequence [æ→ə→ɪ→æ→ə] while the third panel shows the magnitude functions m_{c_k} that are responsible for imposing and removing constrictions at specific locations along the vocal tract length.

The time-dependent shape of each $m_{c_k}(t)$ was generated with a minimum jerk interpolation (cf. Story, 2005a) for which the time points and amplitudes of the onsets, peaks, and offsets are specified. The function labeled m_{c_1} in the figure occurs first in time and generates a constriction at the lips for the initial [b]. The second function m_{c_2} begins just slightly later in time to impose the appropriate constriction for [ɹ]. The large amount of overlap of these two functions is necessary to generate the two consonants as a cluster. These same two functions are repeated, but with slightly longer durations, near the end of the word to generate the second [bɹ] cluster (hence, they are not separately labeled in the figure). The m_{c_3} and m_{c_4} functions impose constrictions in the tract for [k] and [d], respectively.

The locations l_c, cross-sectional areas a_c, and extents r_c of each constriction associated with the magnitude functions are shown in the upper part of Table 1. Both the locations and extents are given in centimeters; for l_c this denotes distance from the glottis, and r_c values denote a section of the vocal tract around the specified location that is affected by the imposed constriction.[3] For example, the location associated with m_{c_1} is $l_c = 17.5$ cm meaning that the constriction is imposed at the lips, produces a cross-sectional area of $a_c = 0.0\,\text{cm}^2$, and affects 2.8 cm of the vocal tract length around l_c. For the [ɹ] generated by m_{c_2}, two constriction locations are specified, one in the oral cavity at 13.4 cm from the glottis and another in the pharyngeal portion at 4.7 cm; also, because the vocal tract does not become fully occluded during an [ɹ], the cross-sectional areas associated with m_{c_2} are nonzero.

Simulated waveforms and the spectrogram of the speech signal are shown in the lower part of Fig. 8. The glottal flow is periodic throughout the word except during the [k] where there is both abduction of the vocal fold processes and occlusion of the vocal tract. In addition, the amplitude of the glottal flow is reduced when constrictions are imposed on the tract shape, and coincides with the increases in intraoral pressure that can be seen just below the glottal flow. The speech signal P_{out} includes pressure radiated at both the lips and skin surfaces, where the latter can be observed as the low amplitude periodicity during the [b] and [d] consonants. The wide-band spectrogram in the bottom panel

[3] If the baseline vocal tract configuration were modified such that the overall tract length were longer or shorter, both l_c and r_c could be specified as percentage of the total tract length rather than absolute length values as used here.

Author's personal copy

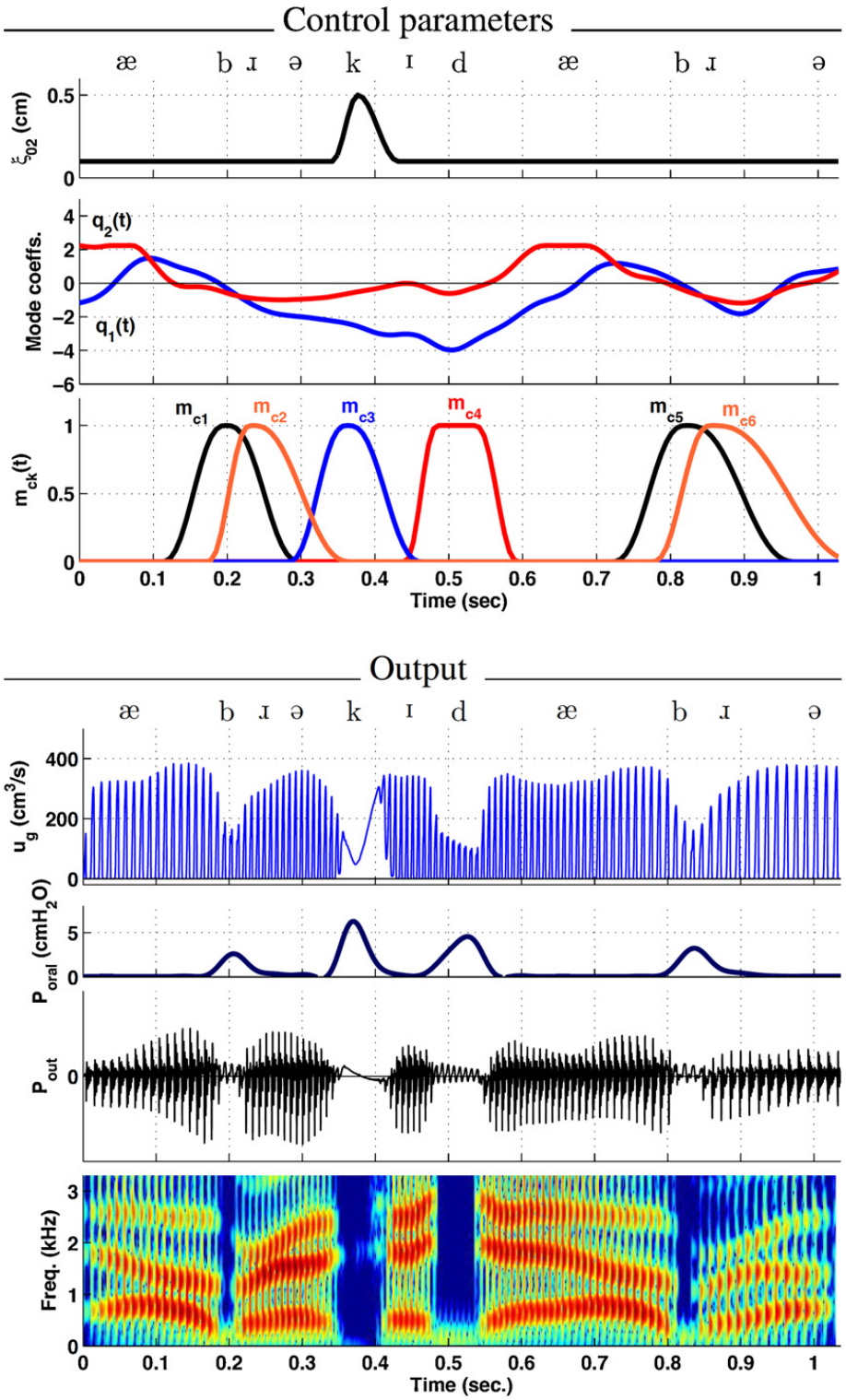


Fig. 8. Simulation of the word *Abracadabra*. Time-dependent control parameters are shown in the upper three panels: vocal process separation ξ_{02}, mode coefficients $[q_1, q_2]$, and consonant magnitude functions m_{c_k}. Additional parameters associated with the m_{c_k} functions are provided in Table 1. Glottal flow u_g, intraoral pressure P_{oral}, total radiated pressure P_{out}, and the wide-band spectrogram of P_{out} are shown in the lower four panels.

Author's personal copy

Table 1
Control parameter values for simulating utterances. Each is a time-dependent consonant magnitude function that imposes a constriction based on the parameter values shown; l_c is the constriction location as distance from the glottis (see Fig. 4); a_c is the cross-sectional area at l_c when the constriction is fully expressed in the vocal tract shape; r_c is the extent of the vocal tract length around l_c that is affected by the constriction. Two values separated by a comma indicates that two locations within the vocal tract are simultaneously affected by the magnitude function, whereas if connected by an arrow the first parameter value is interpolated to the second over the time course of the simulation.

	m_{c_1}	m_{c_2}	m_{c_3}	m_{c_4}	m_{c_5}
Abracadabra					
l_c (cm)	17.5	[13.4, 4.7]	12.6	15.1	
a_c (cm^2)	0.0	[0.40, 0.20]	0.0	0.0	
r_c (cm)	2.8	[1.8, 1.8]	4.7	3.2	
He had a rabbit					
l_c (cm)	[15.1→12.8]	17.5	4.7	15.1	
a_c (cm^2)	[0.0→0.5]	[0.5→0.0]	0.4	0.0	
r_c (cm)	[1.9→3.2]	1.6	1.9	1.9	
The brown cow					
l_c (cm)	15.6	17.5	[13.1, 4.7]	14.7	12.8
a_c (cm^2)	0.02	0.0	[0.1, 0.4]	0.0	0.0
r_c (cm)	1.6	1.9	[2.5, 1.6]	1.9	7

indicates the time variation of the first three formants and their clear interruption during time periods where the vocal tract is occluded, or nearly so.

The temporal change in the vocal tract shape during production of *Abracadabra* is shown in Fig. 9 along with the formant tracks in the inset plot. The phonetic symbols mark the approximate locations in space and time where the associated constrictions occur. The formant tracks illustrate the acoustic manifestation of Tiers I and II of the model (see Fig. 1). In Tier I, the mode coefficients produce continuously varying shaping actions that modulate the

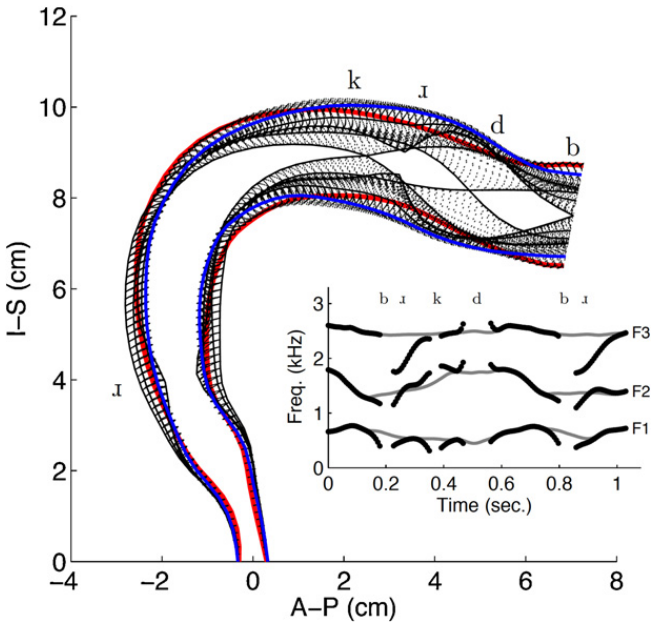


Fig. 9. Vocal tract modulation for *Abracadabra*. The red lines represent the configuration at the beginning of the utterance, the blue lines represent the utterance end, and the dotted black lines indicate the variation in shape that occurs during the utterance. The solid black lines denote tract shapes at points in time where an imposed constriction is fully expressed. The phonetic symbols are shown at the approximate locations where the associated constrictions occur. The first three formant frequencies are shown in the inset plot as a series of dark points and the gray lines are time-variations of the three formants that would occur in the absence of any imposed constrictions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Author's personal copy

configuration neutral vocal tract shape, and consequently generate the smoothly varying formant frequencies shown as gray lines in the background. The magnitude functions in Tier II rise quickly to impose a constriction and rapidly decrease to remove it. The acoustic effect, shown as a series of black dots in figure, is to deflect the formants away from the underlying vowel transition, in directions determined by the constriction location, for short periods of time and then allow them to return.

The speech signal resulting from the simulation of *Abracadabra* can be heard by playing Audio Clip 2a. The second clip (Audio Clip 2b) is the same simulation but with the magnitude functions and changes in ξ_{02} removed; it is essentially like listening to speech produced with the formants shown as gray lines in Fig. 9. A vocal tract animation is available for viewing in Video Clip 2. It again shows a slow motion (without audio) view of the vocal tract movement along with a running wide-band spectrogram.

3.3. Phrase 1: "He had a rabbit"

The control parameters and output quantities for the phrase *He had a rabbit* are shown in Fig. 10. This simulation is more complex than the previous cases because there are multiple abductory/adductory maneuvers required at the glottis, and heavily overlapped vocal tract constrictions. At the beginning of the utterance ξ_{02} is constant at a value 0.4 cm in order to generate the initial [h], while the vocal tract is configured in an [i]-like shape in preparation for the following vowel. This produces a high steady glottal flow with turbulence and results in a low-amplitude noise signal radiated at the lips, as can be seen in the P_{out} signal. It is noted that there is some increase in intraoral pressure during this [h] because of the constricted oral cavity already in place for the following [i] vowel. The spectrogram during this period of time shows a band of noise at about 2.7–3.0 kHz, followed by the appearance of three formant bands arranged in a pattern indicative of an [i].

At 0.3 s, ξ_{02} increases slightly to 0.2 cm to produce the second [h] and then returns to 0.1 cm for the next voiced portion extending to 0.8 s. During this span of voicing, multiple constrictive actions are generated in rapid succession by magnitude functions m_{c_1}, m_{c_2}, and m_{c_3}. The first function, m_{c_1}, is used to generate a brief constriction for what is effectively an alveolar tap at 0.48 s. The amplitude of the function then drops slightly but then returns to a value of one by 0.55 s to produce a constriction for the [ɹ]. This is realized in the model by shifting the constriction location l_c, and modifying the cross-sectional area a_c and the extent r_c, over the time course of m_{c_1} as indicated in the middle part of Table 1. The second and third magnitude functions (m_{c_2} and m_{c_3}) create additional constrictions in the pharynx and at the lips, respectively, for production of the [ɹ]. The overall effect of these constrictions on the output waveforms is to modulate the amplitudes of the glottal flow and radiated pressure signal. It can be seen in the spectrogram that, from about 0.35 to 0.8 s, the formant bands are fairly evenly spaced for the [æ], briefly lose some amplitude during the tap, and then F3 is shifted downward during [ɹ] before rising again for the second [æ].

There is an additional nonzero portion of m_{c_2} extending from 0.75 to 1.1 s that executes a bilabial constriction for the [b]. During this time, a slight abduction is imposed, even though it is a "voiced" consonant, but was needed in order to match the devoicing observed in the audio recording. It could likely be removed without any loss of intelligibility, however. The final magnitude function m_{c_4} generates a constriction for the [t] and coincides with an abduction of the vocal processes reflected as an increase of ξ_{02} from 0.1 cm to 0.4 cm. Both constrictions produce a full occlusion of the vocal tract and consequently allow the intraoral pressure to build up prior to their release. The glottal flow in this time period is reduced in amplitude when the constrictions are present and then increases rapidly following the release of the [t] constriction. The effects are displayed in the spectrogram as very low frequency energy between 0.82 and 0.9 s, distinct formant frequencies from 0.9 to 0.97 s, silence during 0.97–1.3 s, and finally a strong band of noise centered around 2.9 kHz.

The modulation of the vocal tract prescribed by the control parameters for *He had a rabbit* is shown in Fig. 11. The first three formant frequencies are shown in the inset plot in the same manner as in Fig. 9 where the gray lines are formants that would exist without the constrictions imposed. The constrictions deflect the formants away from the vowel sequence for short periods of time and then return to match the gray lines during periods of vowel production. The speech signal for *He had a rabbit* can be heard by playing Audio Clip 3a, whereas Audio Clip 3b is the same simulation without consonants and glottal changes. A vocal tract animation along with a running wide-band spectrogram is available for viewing in Video Clip 3.

Author's personal copy

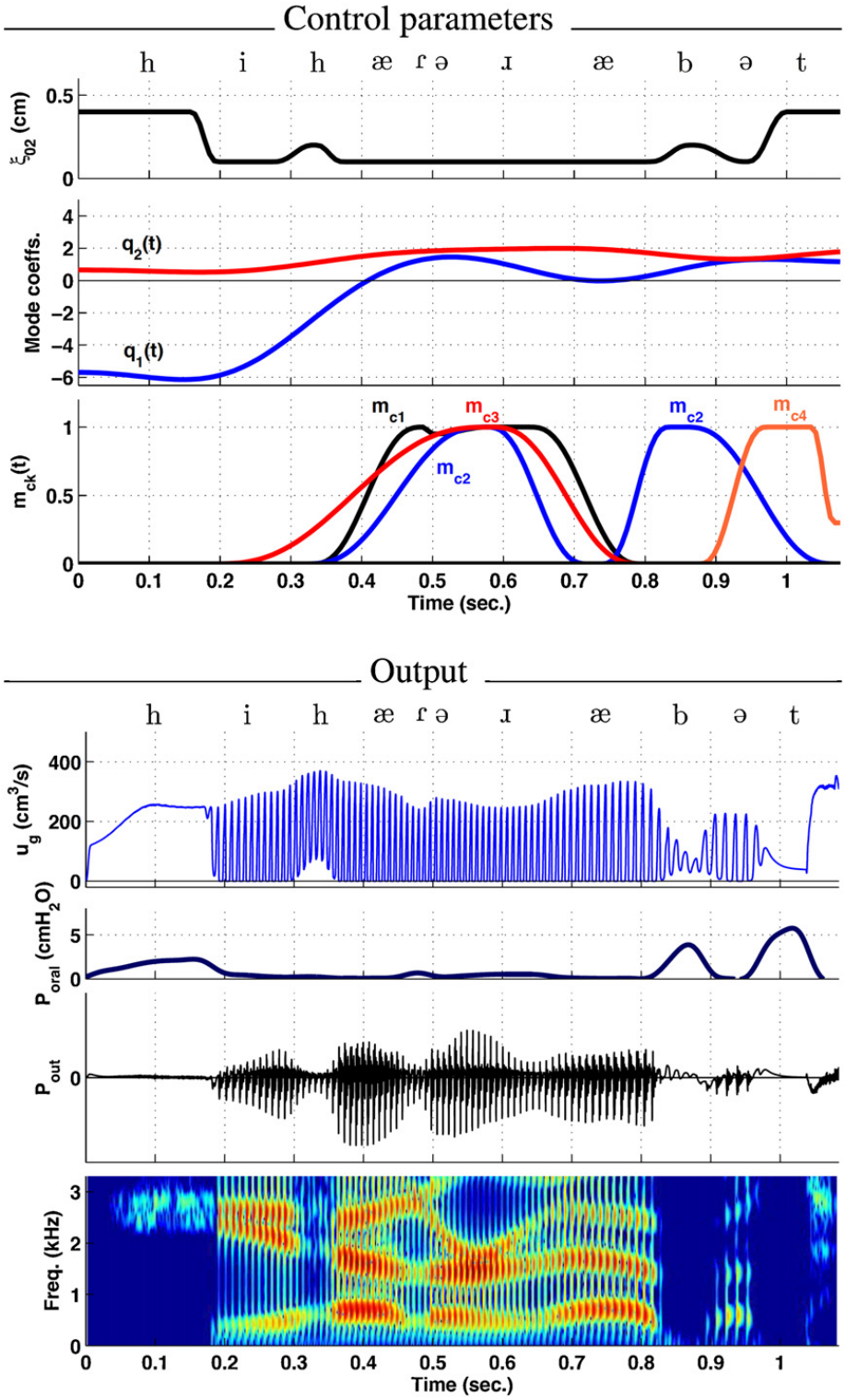


Fig. 10. Simulation of the phrase *He had a rabbit*. Parameters and waveforms are displayed in the same order as in Fig. 8. Additional parameters associated with the m_{c_k} functions are provided in Table 1.

3.4. Phrase 2: 'The brown cow"

The second phrase simulated was *The brown cow.* The control parameters, output waveforms, and spectrogram are presented in Fig. 12. In comparison to the previous simulations, the novel of part of this phrase is the succession of the

Author's personal copy

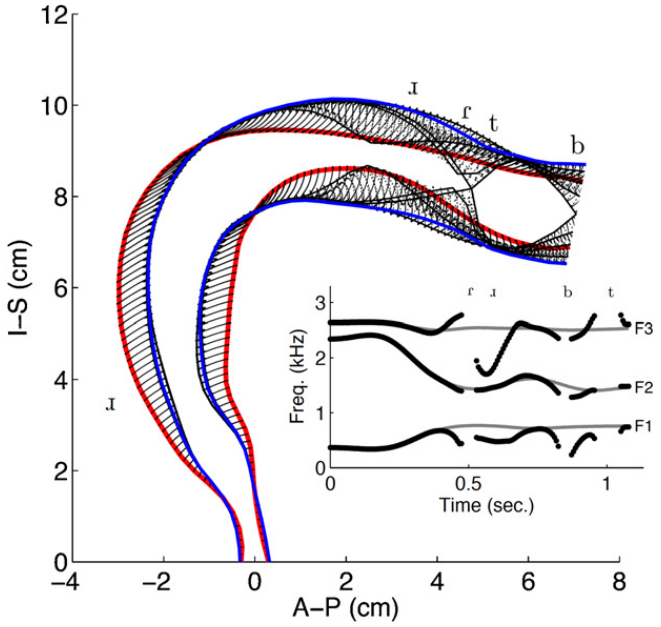


Fig. 11. Vocal tract modulation and formant frequencies for *He had a rabbit* shown in the same format as in Fig. 9.

nasal consonant [n] immediately followed by the velar stop [k]. This means that in addition to the adductory, vowel, and constriction parameters, the temporal variation of the nasal port area must also be specified.

The phrase begins with ξ_{02} set to 0.1 cm to allow for vocal fold vibration, and a constriction is imposed in the vocal tract by function m_{c_1} just posterior to the lips for production of the voiced fricative [ð]. The constriction location, cross-sectional area, and extent are given in lower part of Table 1; it is noted that the cross-sectional area of the constriction is nonzero ($a_c = 0.02$ cm^2) to allow for frication turbulence to be generated. From the waveform plots, the constriction can be seen to reduce the amplitude of the glottal, increase the intraoral pressure, and produce low-amplitude radiated pressure. As this function is released, m_{c_2} is already increasing in amplitude to impose a constriction at the lips for the [b]. There is a short period of time between m_{c_1} and m_{c_2} during which the vocal tract is unoccluded. It is labeled phonetically in the figure as a neutral vowel, but can be seen in the spectrogram as formants transitioning from one consonant to another.

The m_{c_2} and m_{c_3} functions are largely overlapped in order to create the [bɹ] cluster. That is, as the constriction is produced at the lips for the [b], constrictions for the following [ɹ] are already moving into place so that the [b] releases into the [ɹ] sound. This can be observed in the spectrogram from 0.1 to 0.3 s where the low amplitude portion is followed by the appearance of three formants, F3 being at a low frequency for the [ɹ] and then rising into the following vowel.

Function m_{c_4} imposes a constriction with zero cross-sectional area at a location roughly corresponding to the alveolar ridge (14.7 cm from the glottis in Table 1). Simultaneously, the nasal port area is increased from zero to 0.2 cm^2, allowing air to flow and sound to propagate through the nasal tract, resulting in production of the consonant [n]. The glottal flow waveform indicates a slight reduction in amplitude during the nasal consonant. The amplitude of the radiated pressure is also low and, for this period of time, is based on the signals exiting the nares (and skin). The spectrogram shows that the formants become somewhat ambiguous from about 0.4 to 0.46 s due to the additional poles and zeros imposed by the coupling of the nasal tract.

The [n] is followed immediately by the unvoiced velar stop [k] imposed by m_{c_5}, and the location, cross-sectional area, and extent are given in Table 1. This requires that the nasal port area return to zero quickly for proper build-up of intraoral pressure and to avoid an unwanted nasal emission. In addition, because the velar stop is unvoiced, the vocal processes must be abducted to halt vocal fold vibration and to allow the subglottal airspace to be coupled to the intraoral space. Thus, ξ_{02} undergoes a rapid increase from 0.1 cm to 0.7 cm to simulate the abductory maneuver. There is a slight noise burst at the release of m_{c_5} which is most easily seen in the spectrogram at about 0.58 s. This is followed by the final diphthong [ɑʊ].

Author's personal copy

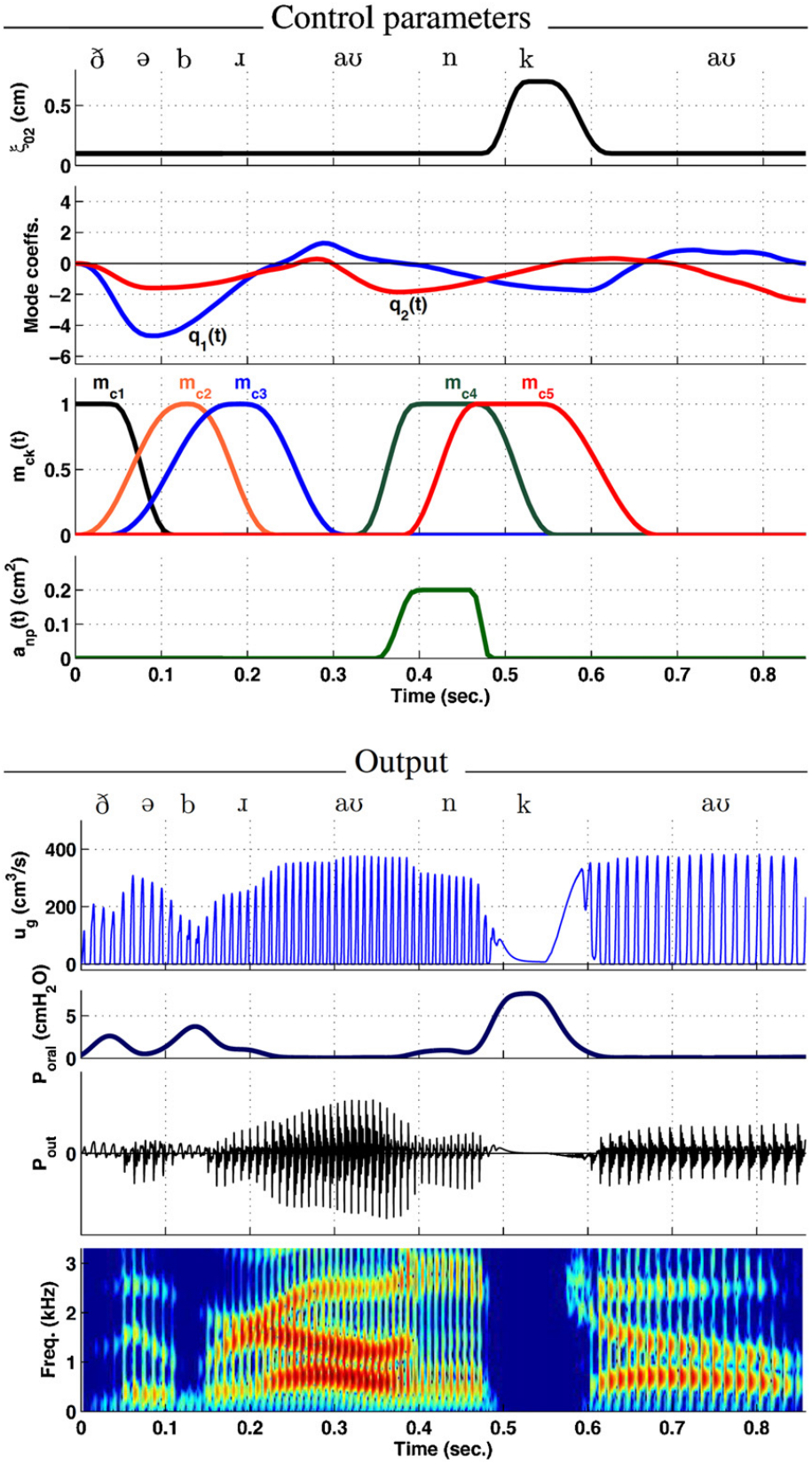


Fig. 12. Simulation of the phrase *The brown cow*. Time-dependent control parameters are shown in the upper four panels as in previous figures but with the addition of nasal port area a_{np}. Parameters associated with the m_{c_k} functions are provided in Table 1. Glottal flow u_g, intraoral pressure P_{oral}, total radiated pressure P_{out}, and the wide-band spectrogram of P_{out} are shown in the lower four panels.

Author's personal copy

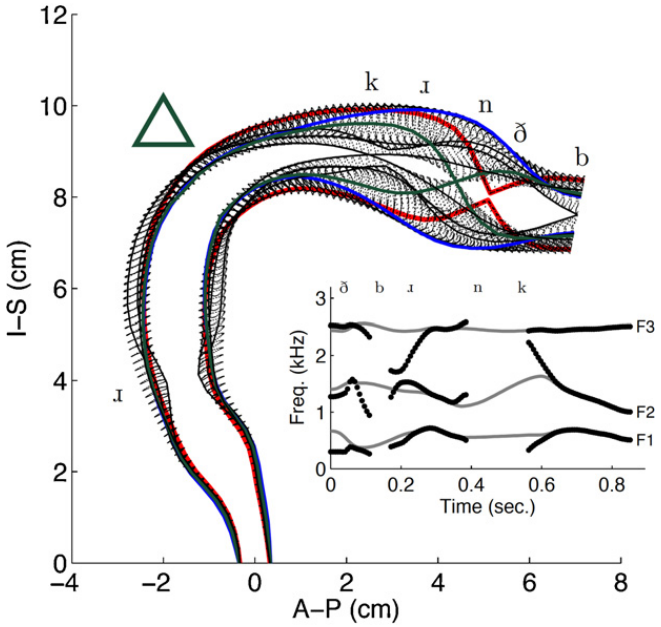


Fig. 13. Vocal tract modulation for *The brown cow* shown in the same format as in Figs. 9 and 11.

Shown in Fig. 13 is the vocal tract movement imposed by the control parameters for *The brown cow.* This phrase required the vocal tract to be shaped for five *different* consonants, and utilized the entire extent of the oral cavity for imposing constrictions as can be seen by the placement of the phonetic symbols along the vocal tract profile. The green triangle in the upper left part of the figure denotes that the nasal port was opened during the time that the tract was configured as indicated by the solid green line; i.e., during production of the [n]. The first three formants are again plotted in the inset but the effect of the nasal coupling between 0.4 and 0.5 s is not shown. The simulation of *The brown cow* can be heard by playing Audio Clip 4a, and Audio Clip 4b is the vowels-only version. Animation of the vocal tract movements can be viewed in Video Clip 4 along with the associated wide-band spectrogram.

3.5. *Intelligibility and naturalness*

The primary focus of the present study was to describe the model and demonstrate that it can simulate speech suitable for future studies of speech acoustics and speech perception. It is of interest to know, however, to what degree listeners can recognize the simulated words and phrases, and if they think they sound like a human talker. Toward this goal, a rudimentary test of intelligibility and naturalness of the two words and two phrases was conducted. Graduate students were asked to participate as listeners in a classroom environment. They were told that they would hear four audio samples produced by a speech synthesizer and their task was to write the word or phrase. In addition they were asked to rate the naturalness of each sample on a scale where 1 was “machine-like” and 10 was “human-like.” All samples were presented at a comfortable listening level via a single monitor loudspeaker (Yamaha, MSP3) placed near the front and center of the classroom. Prior to presentation of the first speech sample, the vowel sound produced by the neutral vocal tract (Audio Clip 0) was played for the listeners to familiarize them with the voice quality. The words and phrases were then played once each with a pause between to allow the listeners to complete the recognition and rating tasks. There were a total of 39 listeners, all of whom reported that American English was their primary spoken language. Intelligibilty was scored as the number of words correct for each sample.

The percent correct scores for *Ohio, Abracadabra, He had a rabbit* and *The brown cow* were 100, 97, 85, and 97 percent, respectively. The phrase *He had a rabbit* had the lowest intelligibility score, but this was primarily because several listeners transcribed it as “Peter the rabbit,” the well known character in a series of children’s stories; this is not surprising considering that the phrase was played to the listeners without any context.

The naturalness ratings spanned the entire scale from 1 to 10 for each of the four samples, however the median values leaned toward the “machine-like” end of the scale with a rating of 3 for *Abracadabra* and a rating of 4 for the other three samples. Discussions with the listeners following data collection did not reveal any particular aspect of the

Author's personal copy

simulations that were regarded as "unnatural," rather there was simply a sense that the sounds they were hearing were not fully human. It is possible that informing the listeners a priori that the samples were *synthesized* speech biased them toward rating closer to the machine-like end of the scale. Highly controlled tests are needed in the future to more thoroughly determine both naturalness and intelligibility of the speech produced by the model.

3.6. Modifications to the phrases

Once the temporal variation of the control parameters for a given word or phrase has been established, various aspects of the model can be altered. For example, the underlying structure of the vocal tract could be modified with respect to its length and distribution of cross-sectional areas; the timing of the constrictions could be stretched or compressed; the vocal folds could be set to produce a breathy or pressed quality, or contain a tremor; the nasal port could be kept open to simulate hypernasal speech. The ability to make such modifications is useful for conducting simulation and perceptual studies related to speech disorders or speech qualities. As a demonstration, the phrases presented in the previous two sections have been modified in several ways. These are described below and can be heard by playing the associated audio files.

1 The neutral vocal tract shape, specified as $\Omega(x)$ in the model, was modified to produce two different sound qualities for the phrase *"He had a rabbit."* In the first case, the pharyngeal portion of $\Omega(x)$ was slightly constricted, and the oral cavity was slightly expanded. The second case was the opposite – the pharyngeal portion was slightly expanded and the oral cavity was slightly constricted. In addition, to create the impression of a large talker in both cases, the vocal tract length was increased to 19.8 cm and the mean fundamental frequency was lowered by 15 percent. The phrase was simulated with exactly the same modulations of the vocal tract and voice source as discussed previously, except they were superimposed on the two new neutral vocal tract shapes. The first modification to "He had a rabbit" can be heard in Audio Clip 5a, and the second in Audio Clip 5b.
2 The timing of all control parameters was altered such that the first half of each phrase was increased in duration by 25 percent and the latter half decreased by 25 percent. The total duration of each phrase is the same as the original. The modified versions of "He had a rabbit" and "The brown cow" can be heard in Audio Clip 6a and Audio Clip 6b, respectively.
3 The timing of all control parameters was altered such that the first half of each phrase was decreased in duration by 25 percent and the latter half increased by 25 percent. The total duration of each phrase is again the same as the original. The modified versions of the two phrases can be heard in Audio Clip 7a and Audio Clip 7b, respectively.
4 The baseline separation of the vocal processes was increased from $\xi_{02} = 0.1$ cm to $\xi_{02} = 0.15$ cm. This change has the effect of allowing a greater non-oscillatory component of the glottal flow during voicing, and results in increased glottal turbulence. The perceptual effect is a breathier voice quality. The modified versions of the two phrases can be heard in Audio Clip 8a and Audio Clip 8b, respectively.
5 The nasal coupling area was maintained at a minimum value of 0.2 cm^2 throughout the duration of each phrase. The effect is to nasalize all portions of the phrases resulting in a hypernasal quality. The modified versions of the two phrases can be heard in Audio Clip 9a and Audio Clip 9b, respectively.
6 The entry area to the vocal tract was increased to effectively widen the epilaryngeal tube. This modification alters the voice quality in two ways – the first three formants are shifted slightly downward in frequency and the glottal flow waveform is altered. The perceptual effect is a darker voice quality. The modified versions of the two phrases can be heard in Audio Clip 10a and Audio Clip 10b, respectively.

4. Conclusion

An airway modulation model called TubeTalker was introduced as a system for generating artificial speech. The overall goal in developing the model is to facilitate an understanding of how modulations of the basic structure of the glottis and vocal tract are acoustically *encoded* in the time variation of the speech signal, and perceptually *decoded* by a listener into phonetic elements. The model encodes the speech signal by assuming that: (1) an acoustically neutral state of the vocal tract generates resonance characteristics unique to a speaker or particular style of speaking; (2) the speech signal is produced by a hierarchy of modulations that are imposed on the acoustically-neutral state of the vocal tract and glottis; and (3) relations between the levels in the modulation hierarchy generate specific acoustic characteristics

Author's personal copy

that provide cues to phonetic categories. The result is a speech signal in which the acoustic characteristics of the voice source and neutral vocal tract serve as a carrier for the acoustic perturbations produced by vowel-like and obstruent-like modulations. Perceptual decoding of speech may then be thought to consist of demodulating the acoustic signal to acquire the information embedded by each layer of the hierarchy.

TubeTalker is not an *articulatory* model per se, since it does not allow for direct control of the position and movement of the tongue, velum, lips, and jaw as do other speech production models (e.g., Rubin et al., 1981; Birkholz et al., 2006). The mode-based control tier for vowel production (Tier I in Fig. 1), however, was derived from statistical analyses of vocal tract shapes measured from real talkers (cf. Story and Titze, 1998; Story, 2005a, 2007), and the characteristics of the consonant constrictions were based on measured area functions (Story, 2005a). Thus, the model is related to physiological aspects of human speech production, just not directly via the obvious articulatory structures. An advantage is that it does allow direct access to the shape of the glottal airspace and the vocal tract with a control system constructed to modulate the airspaces on multiple time scales and at multiple levels.

Such a model is useful for investigating questions about how the speech signal is affected by airway modulations of different types and at various locations within the speech production system. For example, if the underlying vocal fold or vocal tract structure is altered but the temporal patterns for vowel production and constriction formation remain the same, does the speech signal remain intelligible, even though it will be acoustically different? Are there phase relations of modulations among the control tiers that are necessary in order to maintain or preserve acoustic characteristics that cue perception? What would be the acoustic and perceptual effect of shifting one or more of the consonant constriction functions slightly forward or backward in time, say, for example, in the *He had a rabbit* phrase? How much could they be shifted without altering the percept? What effect would slight changes in constriction location or extent have on the acoustic characteristics of the speech signal? Future studies using TubeTalker may yield insights into how the structure and systematic movement of the vocal tract and vocal folds supports the emergence of phonetically-relevant acoustic properties, while simultaneously allowing the speech signal to be highly variable.

Acknowledgments

This research was supported by NIH R01 DC04789 and NIH R01 DC011275. A preliminary version of this work was presented at the 2011 International Workshop on Performative Speech and Singing Synthesis in Vancouver, BC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.csl.2012.10.005.

References

Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W., 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. Journal of the Acoustical Society of America 63, 1535–1555.

Baer, T., Gore, J.C., Gracco, L.C., Nye, P.W., 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. Journal of the Acoustical Society of America 90, 799–828.

Bauer, D., Birkholz, P., Kannampuzha, J., Kröger, B.J., 2010. Evaluation of articulatory speech synthesis: a perception study. In: 36th Deutsche Jahrestagung fr Akustik (DAGA 2010), Berlin, Germany, pp. 1003–1004.

Båvegård, M., 1995. Introducing a parametric consonantal model to the articulatory speech synthesizer. In: Proceedings Eurospeech 95, Madrid, Spain, pp. 1857–1860.

Birkholz, P., Jackel, D., Kröger, B.J., 2006. Construction and control of a three-dimensional vocal tract model. In: Proc. Intl. Conf. Acoust., Spch., and Sig. Proc. (ICASSP 2006), Toulouse, France, pp. 873–876.

Birkholz, P., Jackel, D., Kröger, B.J., 2007. Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15 (4), 1218–1226.

Birkholz, P.D., Kröger, B.J., Neuscheafer-Rube, C., 2010. Articulatory synthesis and perception of plosive-vowel syllables with virtual consonant targets. In: Proc. Interspch, Makuhari, Japan, pp. 1017–1020.

Bunton, K., Story, B.H., 2009. Identification of synthetic vowels based on selected vocal tract area functions. Journal of the Acoustical Society of America 125 (1), 19–22.

Bunton, K., Story, B.H., 2010. Identification of synthetic vowels based on a time-varying model of the vocal tract area function. Journal of the Acoustical Society of America 127 (4), EL146–EL152.

Author's personal copy

Bunton, K., Story, B.H., 2011. The relation of nasality and nasalance to nasal port area based on a computational model. The Cleft Palate-Craniofacial Journal, http://dx.doi.org/10.1597/11-131.
Coker, C.H., 1976. A model of articulatory dynamics and control. Proceedings of IEEE 64 (4), 452–460.
Dang, J., Honda, K., 2004. Construction and control of a physiological articulatory model. Journal of the Acoustical Society of America 115, 853–870.
Dudley, H., 1939. Remaking speech. Journal of the Acoustical Society of America 11 (2), 169–177.
Dudley, H., 1940. The carrier nature of speech. Bell System Technical Journal 19 (4), 495–515.
Dudley, H., Riesz, R.R., Watkins, S.S.A., 1939. A synthetic speaker. Journal of the Franklin Institute-Engineering and Applied Mathematics 227 (6), 739–764.
Fant, G., 1960. The Acoustic Theory of Speech Production. Mouton, The Hague.
Fant, G., 1992. Vocal tract area functions of Swedish vowels and a new three-parameter model. In: Proc. ICSLP-92, 1, pp. 807–810.
Fant, G., 2001. Swedish vowels and a new three-parameter model. TMH-QPSR 42 (1), 43–49.
Fant, G., Båvegård, M., 1997. Parametric model of VT area functions: vowels and consonants. TMH-QPSR 38 (1), 1–20.
Flanagan, J.L., Cherry, L., 1969. Excitation of vocal-tract synthesizers. Journal of the Acoustical Society of America 45 (3), 764–769.
Gracco, V.L., 1992. Characteristics of speech as a motor control system. The Haskins Laboratories Status Report on Speech Research SR-109/110, 13–26.
Kelly, J.L., Lochbaum, C.C., 1962. Speech synthesis. In: Proc. Fourth Intl. Cong. Acous., PaperG42, pp. 1–4.
Laakso, T.I., Valimaki, V., Karjalainen, M., Laine, U.K., 1996. Splitting the unit delay. IEEE Transactions on Signal Processing Magazine 13 (1), 30–61.
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.
Lin, Q., 1990. Speech production theory and articulatory speech synthesis. DS Dissertation. Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech. Stockholm, Sweden.
Lindblom, B., Sundberg, J., 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. Journal of the Acoustical Society of America 4 (2), 1166–1179.
Maeda, S., 1982. A digital simulation method of the vocal-tract system. Speech Communication 1, 199–229.
Maeda, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modeling. Kluwer Academic Publishers, Dordrecht, 131–149.
Mathur, S., Story, B.H., Rodriguez, J.J., 2006. Vocal-tract modeling: fractional elongation of segment lengths in a waveguide model with half-sample delays. IEEE Transactions on Audio, Speech, and Language Processing 14 (5), 1754–1762.
The Mathworks, 2011. MATLAB, Version 7.13.0.564 (R2011b).
Mermelstein, P., 1973. Articulatory model for the study of speech production. Journal of the Acoustical Society of America 53 (4), 1070–1082.
Mrayati, M., Carré, R., Guérin, B., 1988. Distinctive regions and modes: a new theory of speech production. Speech Commnication 7, 257–286.
Öhman, S.E.G., 1966. Coarticulation in VCV utterances: spectrographic measurements. Journal of the Acoustical Society of America 39, 151–168.
Öhman, S.E.G., 1967. Numerical model of coarticulation. Journal of the Acoustical Society of America 41, 310–320.
Pruthi, T., Espy-Wilson, C., Story, B.H., 2007. Simulation and analysis of nasalized vowels based on magnetic resonance imaging data. Journal of the Acoustical Society of America 121 (6), 3858–3873.
Rubin, P., Baer, T., Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. Journal of the Acoustical Society of America 70 (2), 321–328.
Samlan, R., Story, B.H., 2011. Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling. Journal of Speech, Language, and Hearing Research 54, 1267–1283.
Smith, J.O., 1992. Physical modeling using digital waveguides. Computer Music Journal 16 (4), 74–91.
Scully, C., 1990. Articulatory synthesis. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modeling. Kluwer Academic Publishers, Dordrecht, 151–186.
Sondhi, M.M., Schroeter, J., 1987. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Transactions on ASSP ASSP-35 (7), 955–967.
Stevens, K.N., House, A.S., 1955. Development of a quantitative description of vowel articulation. Journal of the Acoustical Society of America 27 (3), 484–493.
Story, B.H., 1995. Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract. Ph.D. Dissertation. University of Iowa.
Story, B.H., 2005a. A parametric model of the vocal tract area function for vowel and consonant simulation. Journal of the Acoustical Society of America 117 (5), 3231–3254.
Story, B.H., 2005b. Synergistic modes of vocal tract articulation for American English vowels. Journal of the Acoustical Society of America 118 (6), 3834–3859.
Story, B.H., 2007. Time-dependence of vocal tract modes during production of vowels and vowel sequences. Journal of the Acoustical Society of America 121 (6), 3770–3789.
Story, B.H., 2009a. Vocal tract modes based on multiple area function sets from one speaker. Journal of the Acoustical Society of America 125 (4), EL141–EL147.
Story, B.H., 2009b. Vowel and consonant contributions to vocal tract shape. Journal of the Acoustical Society of America 126, 825–836.
Story, B.H., Bunton, K., 2010. Relation of vocal tract shape, formant transitions, and stop consonant identification. Journal of Speech, Language, and Hearing Research 53, 1514–1528.

Author's personal copy

Story, B.H., Laukkanen, A.-M., Titze, I.R., 2000. Acoustic impedance of an artificially lengthened and constricted vocal tract. Journal of Voice 14 (4), 455–469.

Story, B.H., Titze, I.R., Hoffman, E.A., 1996. Vocal tract area functions from magnetic resonance imaging. Journal of the Acoustical Society of America 100 (1), 537–554.

Story, B.H., Titze, I.R., Hoffman, E.A., 2001. The relationship of vocal tract shape to three voice qualities. Journal of the Acoustical Society of America 109 (4), 1651–1667.

Story, B.H., Titze, I.R., 1998. Parameterization of vocal tract area functions by empirical orthogonal modes. Journal of Phonetics 26 (3), 223–260.

Story, B.H., Titze, I.R., 2002. A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function. Journal of Phonetics 30, 485–509.

Strube, H.W., 1982. Time-varying wave digital filters for modeling analog systems. IEEE Transactions on Acoustics, Audio, Speech, and Language Processing ASSP-30 (6), 864–868.

Titze, I.R., 1984. Parameterization of the glottal area, glottal flow, and vocal fold contact area. Journal of the Acoustical Society of America 75, 570–580.

Titze, I.R., 2002. Regulating glottal airflow in phonation: application of the maximum power transfer theorem to a low dimensional phonation model. Journal of the Acoustical Society of America 111, 367–376.

Titze, I.R., 2006. The myoelastic aerodynamic theory of phonation. National Center for Voice and Speech, 197–214.

Traunmüller, H., 1994. Conventional, biological and environmental factors in speech communication: a modulation theory. Phonetica 51, 170–183.

Westbury, J.R., 1994. X-ray microbeam speech production database user's handbook (version 1.0) (UW-Madison).