TubeTalker: An airway modulation model of human sound production

Brad H. Story
Dept. of Speech, Language, and Hearing
University of Arizona
Tucson, AZ
bstory@email.arizona.edu

Abstract

Artificial talkers and speech synthesis systems have long been used as a means of understanding both speech production and speech perception. This article begins with an overview and history of artificial speech systems which includes mechanical models, electronic devices, and computation-based simulations. The development of an airway modulation model is then 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 two phrases are demonstrated.

Keywords: vocal tract, vocal folds, modulation, speech synthesis.

1. Introduction

Humans have a system of airways that, for purposes of speech production, serves to transform articulatory movements into an acoustic wave that carries the distinctive characteristics of speech. This transformative process is accomplished by modulating the airway system on multiple time scales (cf. Fig. 1). 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. When coupled with respiratory pressure the modulated glottis generates a train of flow pulses that, along with possible turbulence, will excite the acoustic resonances of the trachea, vocal tract, and nasal passages resulting in an acoustic wave radiated at the lips. Simultaneous, but much slower articulatory movements can be executed to modulate the shape of the pharyngeal and oral cavities, the coupling to the nasal system, the space between the vocal folds by abduction, or the magnitude of respiratory pressure. These slow modulations will 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.

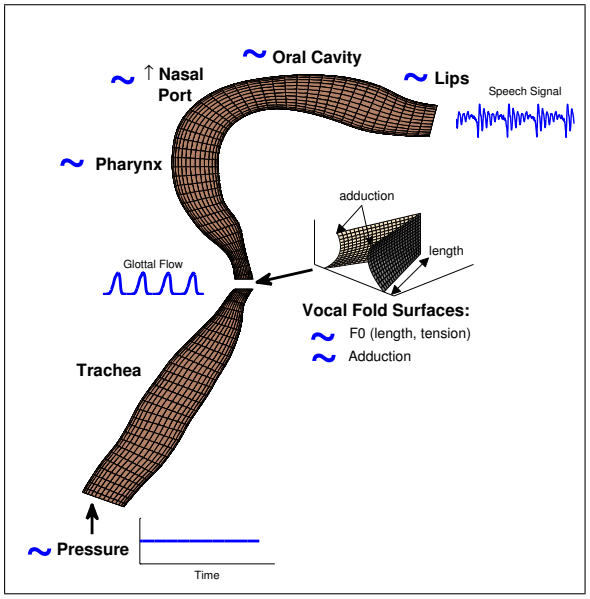

Figure 1. Airway modulation model of the speech production system. The glottal flow signal is the result of high frequency modulation of the glottis (e.g., 100-400 Hz) whereas the components marked with "∼" can be modulated at low frequencies (e.g. 0-15 Hz).

The purpose of this article is to demonstrate how a spoken phrase can be simulated with a model of speech production based on airway modulation. The input parameters impose time-dependent deformations on the airway shape which structure the acoustic signal into speech. Although this work is new in the sense of it being both theoretically and data-based, as well as implemented on a modern computer system, the interest in creating an artificial talker is centuries old. The aims of the article are to 1) provide a brief history of two artificial talking devices that were used in performance situations, 2) describe the background and main components of the airway modulation model, and 3) demonstrate how the model can be used to produce short phrases.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
p3s 2011, March 14-15, 2011, Vancouver, BC, CA.
Copyright remains with the author(s).

2. Abbreviated History of Talking Machines

Development of artificial talkers has long been of interest for purposes of science, technology, art, entertainment, and deception. The first true speaking machines are often attributed to Kratzenstein and von Kempelen[1, 2], both of whom built mechanical devices in which a sound source excited variously shaped resonant cavities resulting in speech-like sounds. Both also exhibited their inventions in the late 1700's; Kratzenstein's device, which could produce five vowels, won a prize in 1779 offered by the Imperial Academy of Sciences at St. Petersburgh. [1] In the early 1800's Charles Wheatstone [4] and Robert Willis [5] both attempted to recreate and improve upon von Kempelen's speaking machine; in each case the primary interest was in acquiring a scientific understanding speech production.

Perhaps no early inventor of artificial talkers exemplified the integration of engineering *and* theatrical presentation better than Joseph Faber. Although little is known about Faber's personal life, according to Lindsay [6], he was born in Germany and first became an astronomer. His interests, however, eventually diverted into anatomy and mechanics. He supposedly began work on his speaking machine in the 1820's, basing his design largely on von Kempelen's published record of his device. A bellows supplied pressure to a reed which, in turn, supplied the acoustic excitation to a set of resonant chambers that could be coupled or decoupled with sliding plates. The components of the machine were controlled by levers connected to a keyboard and foot pedal. Thus, speech could be produced by hand and foot gestures of a trained operator (who apparently was Faber himself). After several years of demonstrating to European audiences that his invention (called the "Amazing Talking Machine") could speak and sing with clarity, Faber traveled to New York City in 1844. The exhibitions there were met with mixed reviews. A newspaper correspondent was impressed enough to write that the only problem with the device was that it had "a strong German accent" [6], although interest from the general public was reportedly low.

Faber's talking machine did, however, attract the attention of Princeton scientist Joseph Henry [2] who requested a private demonstration. Afterward Henry wrote to a colleague that "The plan of the machine is the same as that of the human organs of speech, the several parts being worked by strings and levers instead of tendons and muscles." He also noted that "The German [i.e., Faber] was studying the lesson for the [requested] exhibition, ... he speaks but little English and [we were] ... obliged to make him repeat the sentences several times before they were properly articulated. With a little practice the figure [i.e., the talking machine] really pronounced the words better than the operator[;] its organs were under more readily control than his own" [7]. Henry also suggested that the keystrokes used to operate Faber's talking machine could perhaps be transmitted over electrical wires and used to activate a similar device at a remote location (reconfigured with electro-mechanical actuators). Although this was essentially an idea for a kind of telephone, Henry did not return to an interest in such matters for another 30 years. In the meantime Faber took his talking machine back to Europe as an exhibition with P.T. Barnum's enterprise and had moderate success as an entertainer [6].

Although the comments about Faber's talking machine having a German accent were intended to highlight an objectionable aspect of the device, they provide some insight into how the operator's patterns of speech production were imposed on the speech output. That is, the hand and foot gestures used for operating the machine could be not generated independently of the operator's native speaking abilities; in some sense, listeners still heard Joseph Faber speaking, just not by sound production with the typical human speech articulators. There was nothing inherently "German" about the talking machine; had a native speaker of a different language learned to operate the device the output would have taken on the characteristics of that person.

Nearly 100 years later, Homer Dudley[8] introduced an *electronic* artificial talker called the "VODER." [3] The VODER was essentially a human-operated version of the vocoder [9] that was, in many ways, similar to Faber's earlier mechanical creation. The VODER included a foot pedal and a keyboard with 14 keys and wrist bar. The operator controlled the excitation source with the wrist, the fundamental frequency with the foot pedal, and the amplitude envelopes of 10 fixed bandwidth filters with the finger keys; remaining keys were used for consonant production and silent periods[8]. The VODER was demonstrated, with some fanfare, to the general public at the 1939 World's Fair in New York and at the Golden Gate Exposition in San Francisco the same year. It is noteworthy that twenty-four people were trained to operate the VODER at these exhibitions. Their formal training lasted around *six months* but in general about one year was required to develop the ability to produce intelligible speech with the VODER; in fact, Dudley wrote[8] "... the first half [of the year of training was] spent in acquiring the ability to form any and all sounds, the second half being devoted to improving naturalness and intelligibility." Once learned, however, the ability to "speak" with the VODER was apparently retained for years afterward, even without

[1] It is possible, however, that these devices were preceded by a speaking machine designed by a French academician named Charles Sorel [3]. In the third volume of the 1667 edition of his textbook *La Science Universelle*, Sorel describes an intricate speaking machine consisting of pipes and a keyboard used to control their actions. He also wrote extensively about the "conjunction" of speech sounds as they were produced by the machine, which was apparently a precursor to the modern notion of coarticulation [3].

[2] For whom the unit of electrical inductance is named.

[3] An acronym comprised of the capitalized letters in "Voice Operation DEmonstratoR."

continued practice. Some twenty years after the World's Fair exhibitions, one of the original trained operators was invited back to Bell Labs for an "encore performance" of sorts (the occasion was Homer Dudley's retirement) with a restored version of the VODER. It was reported that "She sat down and gave a virtuoso performance..." [10].

The examples of both Faber's Talking Machine and Dudley's VODER, suggest that the operator of the device learns and internalizes a set of rules for generating speech with a new sound producing system. Although the ability of a human operator to acquire such rules is highly desirable for performance-driven artificial speech (or song), much research has been to devoted to explication of those rules in an attempt to develop text-to-speech synthesis systems, as well as provide general knowledge about the speech planning and production process. A theoretical point of view concerning how such rules might be defined is to consider the speech signal as the result of multiple layers of modulation imposed on the underlying sound production device, whether it be of a mechanical, electrical, or biological nature. Indeed, based in part on the experience with the VODER, Dudley[12] expressed this view in an article called "The carrier nature of speech" by referring 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 noted that this view applied across human speech, the VODER, and the vocoder system (and could be extended to apply to Faber's talking machine as well.).

The subsequent sections of this article are intended to describe speech production in the framework of an *airway modulation model*. It will be shown that the "carrier nature of speech" perspective can be applied not just in terms of the excitation source and vocal tract movements, but at multiple levels of the speech production system. The primary purpose of this model is to facilitate understanding the acoustic characteristics of time-dependent changes in airway shape and ultimately how those characteristics are related to the perception of speech. The control parameters of the model, however, do lend themselves to real-time control, hence the model could potentially be configured as a performance-driven device.

3. Airway Modulation Model

The airway modulation model, also referred to as "Tube-Talker," is a combination of two primary components: 1) a kinematic representation of the medial surfaces of the vocal folds, and 2) a kinematic area function representation of the trachea, nasal tract, and vocal tract airways. A multi-tier technique has been developed for controlling the model parameters in both components as is shown in Fig. 2. 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

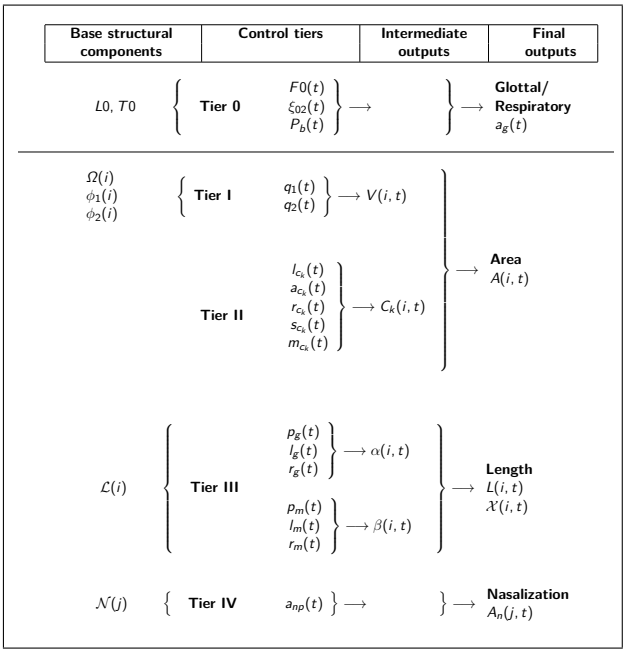

Figure 2. Diagram of the five-tier model. Tier 0 controls the kinematic vocal folds model; the $L0$ and $T0$ are the initial length and thickness of the vocal folds, respectively. Tier I produces a vowel substrate and Tier II generates a superposition function for a consonant constriction. Vocal-tract length changes are generated by Tier III, and nasal coupling in Tier IV. The base structural components are dependent only a spatial dimension, whereas the final outputs are dependent on both space and time.

the covert intermediate quantities, and the rightmost column indicates the output quantities needed for sound production. Each tier will be described in the following sections.

3.1. 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 [13, 14] . As can be seen in Fig. 3, 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 [14]. Any of the vibratory, aerodynamic, or structural parameters (e.g.,

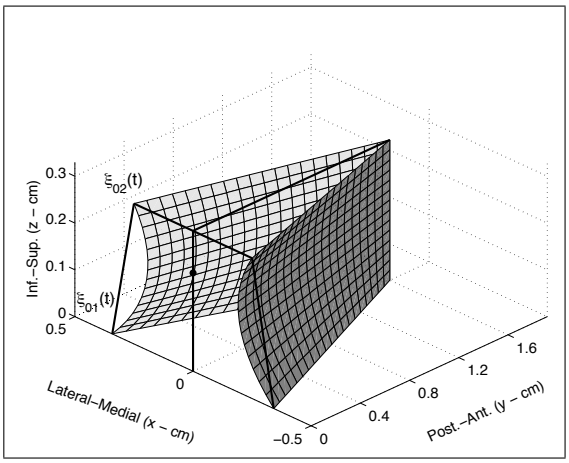

Figure 3. Kinematic model of the medial surfaces of the vocal folds. The Post-Ant. dimension is the vocal fold length, the Inf.-Sup. dimension is the vocal fold thickness, and the Lateral-Medial dimension is vocal fold displacement. ξ_{02} and ξ_{01} are the prephonatory adductory settings of the upper and lower portions, respectively, of the posterior portion of the vocal folds.

fundamental freq. ($F0$), bronchial pressure P_b, vocal process separation, etc.) can also be made to be time-varying as indicated in the "Tier 0" portion of Fig. 2. 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 output of the kinematic source model is the glottal area function $a_g(t)$, calculated as the time-varying sum of the minimum area from each of the vertical channels of the medial surface. The glottal area is aerodynamically and acoustically coupled to a wave-reflection model of the trachea, vocal tract, and nasal tract [16, 17, 15]. The resulting glottal flow is determined by the interaction of the glottal area with the time-varying acoustic pressures present just inferior and superior to the glottis. A noise component is added to the glottal flow signal if the calculated Reynolds number within the glottis exceeds a threshold value (> 1200).

3.2. Kinematic model of the vocal tract area function

The vocal tract component of the model was described in Story[18] and is based on a perspective that vowels and vowel-to-vowel transitions are produced by modulating a phonetically-neutral vocal tract configuration. In turn, production of 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 consonant modulations are imposed. Thus, the neutral tract shape serves as a carrier for vowel modulations which, in turn, serve as a carrier for consonant constriction modulations.

The length and shape of the vocal tract is 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. Based on the multi-tier approach previously developed for controlling the shape of the vocal tract[18], vowels are represented in Tier I (see Fig. 2) as modulations of an underlying neutral (in an acoustic sense) tract shape and consonants are imposed in Tier II as modulations of the underlying vowel substrate. Tier III allows for dynamic length change operations at both the glottal and lip ends, and nasal coupling is controlled by Tier IV.

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

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

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 [18]. The time-dependence is produced by the mode scaling coefficients $q_1(t)$ and $q_2(t)$. The squaring operation and scaling factor of $\pi/4$ converts the diameters to areas.

Production of consonants 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 constriction extent along the tract length, and skewness. The extent r_c is defined as the distance between the half maximum points of $C(x)$ and a skewing factor s_c dictates the degree of asymmetry of the constriction. When any vowel-like area function is multiplied by $C(x)$, the region in the vicinity of l_c will be reduced in area, thus superimposing the constriction. For velar consonants r_c would typically be set to larger values to more accurately represent the extent of a constriction produced by the tongue body. The constriction function can be made time dependent with a temporal activation parameter called the consonant magnitude $m_c(t)$, such that it will impose the constriction to produce the specified cross-sectional area $a_c(t)$ at a specific time and location in the vocal tract.

A composite area function $A(x,t)$ is generated by the vocal tract model as the product of each element along the x-dimension of $V(x,t)$ and $C(x,t)$ such that at any given

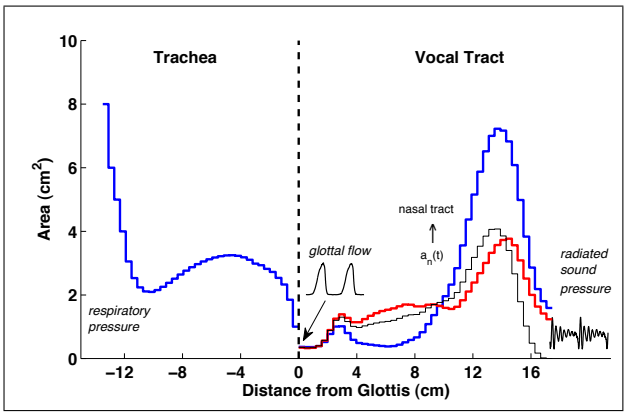

Figure 4. Demonstration of the area function model. The glottis is located at 0 cm, the trachea extends from the glottis in the negative direction and the vocal tract extends in the positive direction. The red line indicates the neutral area function $\Omega(x)$, the blue line is a vowel modulation of the neutral shape based on Eqn. (1), and the black line demonstrates an area function with an occlusion located at the lips. The nasal coupling location is indicated by the upward pointing arrow located at about 8.8 cm from the glottis.

time sample t_n,

$$A(x_i, t_n) = \prod_{i=1}^{N_x} V(x_i, t_n) C(x_i, t_n) \qquad (2)$$

where N_x is the number of cross-sectional areas representing the complete area function. All area functions used in the present study consisted of $N_x = 44$ contiguous "tubelet" sections as defined in [18]. 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 extends from the glottis toward the bronchi in the negative x-direction, and the 44-section vocal tract extends toward the lips in the positive x-direction. The red line indicates the neutral area function $\Omega(x)$, the blue line is a vowel modulation of the neutral shape based on Eqn. (1), and the black line demonstrates an area function with an occlusion located at the lips. The nasal coupling location is indicated by the upward pointing arrow located at about 8.8 cm from the glottis.

The TubeTalker model can potentially simulate any type of speech utterance if the appropriate information concerning the structure and timing of the input parameters is known. In the next section, simulation of two phrases is demonstrated based on extracting such information from articulatory data.

4. Simulation of phrase-level speech

Demonstrated in this section are the steps required to simulate a phrase with the TubeTalker model. The information needed in order to define the input parameters are the time-varying coefficients, $q_1(t)$ and $q_2(t)$, that define the vowel-to-vowel variation in Eqn. 1, the locations and extent of the

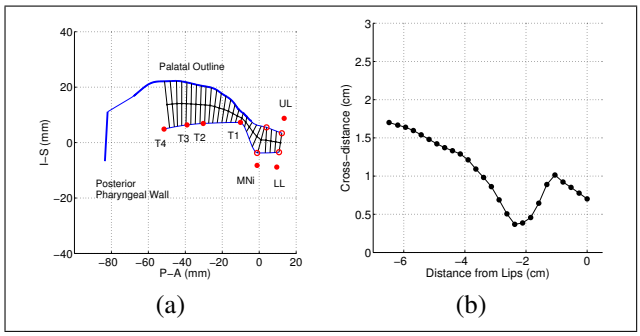

Figure 5. (a) Midsagittal representation of the ð in *"The black cat"* based on articulatory data collected with the XRMB system. (b) Cross-distance function of the same time shown in (a).

primary consonant constrictions, and the temporal patterns of those same constrictions. At this point, such information is extracted from articulatory data collected with the University of Wisconsin-Madison's X-ray microbeam system(XRMB) [19].

The first phrase simulated for this article was *"The black cat"* (broadly transcribed as /ðəblækːæt/); this was chosen in part because it contains a variety of consonant types including a voiced fricative, bilabial, alveolar, and velar stops (both voiced and unvoiced), and a liquid. The XRMB system consists of tracking gold pellets affixed to the tongue, lips, and mandible during production of speech. Shown in Fig. 5a is the midsagittal configuration of the pellets at the first time frame in the phrase (i.e. during production of /ð/). The solid red points represent the actual pellet positions, whereas the open circles are phantom pellets that are an estimate of air tissue interface in the teeth and lip regions [20]. Most directly related to the TubeTalker model is the shape of the airspace itself. Based on the algorithm reported in [20], the airway shape in the oral cavity can be estimated by fitting a centerline and subsequently measuring the cross-distances at successive locations, anterior to posterior, from the lips toward the velar region. The *cross-distance function* for this time frame is shown in 5b. Performing this same analysis on each successive time frame over the duration of the utterance results in a time-varying cross-distance function, as plotted in Fig. 6, where the temporal and structural changes in oral cavity shape can be observed from left to right. Each of the consonant constrictions in the phrase are indicated by arrows.

The spatial location and temporal variation of both vowel and consonant components were determined based primarily on a technique reported in [21]. This technique first attempts to generate a best fit to the cross-distance in a given time frame, based on a linear combination of the modes and neutral vocal tract shape as presented previously in Eqn. 1, but configured for the XRMB data. When collected over all successive time frames of the phrase, this process produces the time-dependent coefficients, $q_1(t)$ and $q_2(t)$. These are

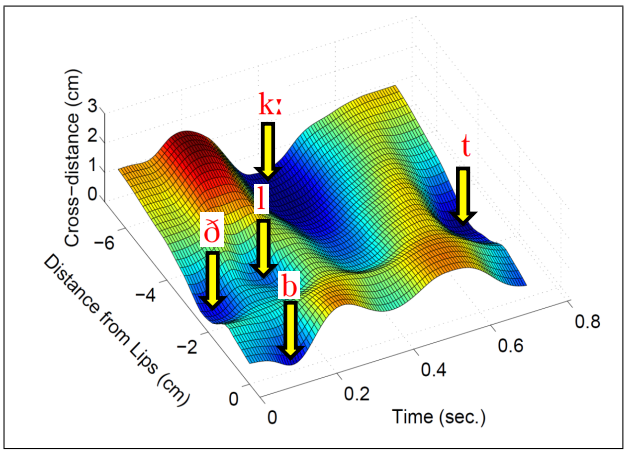

Figure 6. Time-varying cross-distance function of *"The black cat."* Arrows indicate both the spatial and temporal locations of the primary consonant constrictions.

plotted in the upper panel of Fig. 7a and indicate the vowel-to-vowel modulations that need to be imposed on the neutral vocal tract shape. The spatial location of each consonant constriction can be estimated directly from the cross-distance function, as indicated by the arrows. The temporal variation of each constriction (i.e. the onset, sustained portion, and offset), was determined from an error function based on the ratio of the original cross-distance function (Fig. 6) to a vowel-only version reconstructed with the $q_1(t)$ and $q_2(t)$ coefficients (see top panel, Fig. 7a). The timing functions for all the consonant constrictions in the phrase are shown together in second panel of Fig. 7a. The first three consonants are heavily overlapped in time, whereas the /k/ and /t/ are well separated from the other constrictions. The lower two panels indicate the time course of the nasal coupling area and the vocal fold separation distance (i.e. adduction/abduction), both of which have been determined by trial and error. For the *"The black cat,"* the nasal coupling is maintained at zero throughout, but the vocal folds need to move apart during the voiceless portions so that the vibration ceases and pressure builds up prior to the constriction release.

Shown in Fig. 8 is a time sequence of vocal tract area functions generated by the vowel and consonantal input parameters (upper two panels of Fig. 7b). The portion of the area functions extending from about 10-17.5 cm corresponds to the cross-distance function of the oral cavity plotted previously in Fig. 6; the constrictions are similarly located in space and time but the magnitudes appear somewhat different because this figure shows cross-sectional *area* rather than cross *distance*. Time-varying resonance frequencies (formants) calculated directly from two versions of the area function sequence are plotted in Fig. 7b. The blue lines indicate the formants that would exist in the absence of consonant constrictions (i.e., the case if all $m_c(t) = 0$) and the red lines are those calculated when the constrictions are

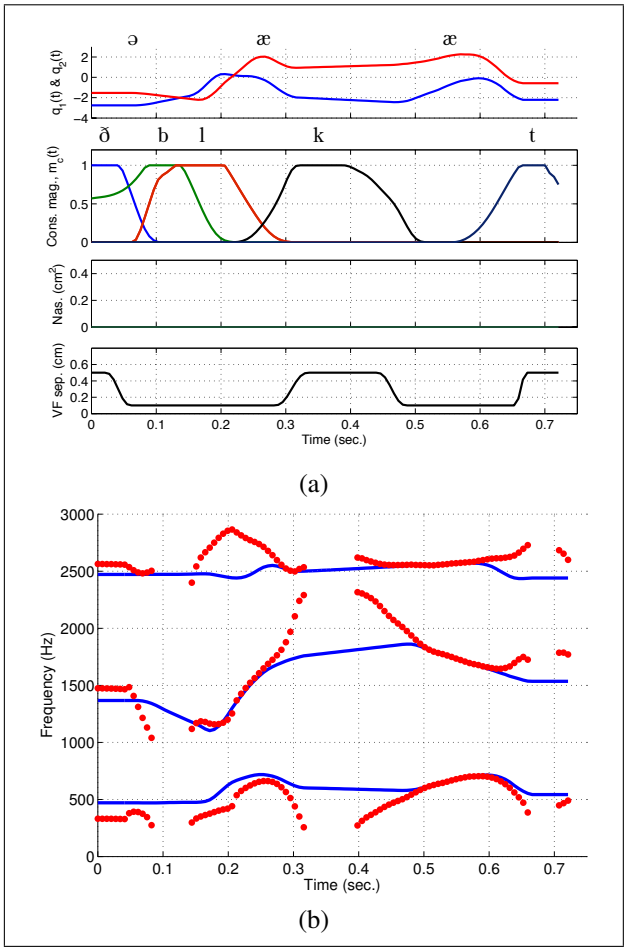

Figure 7. (a) Model input parameters for *"The black cat"*. (b) Calculated formant frequencies for the area function sequence generated with the parameters in (a); the blue lines are the formants in the absence of the consonant constrictions, whereas the red lines are those calculated with the constrictions imposed.

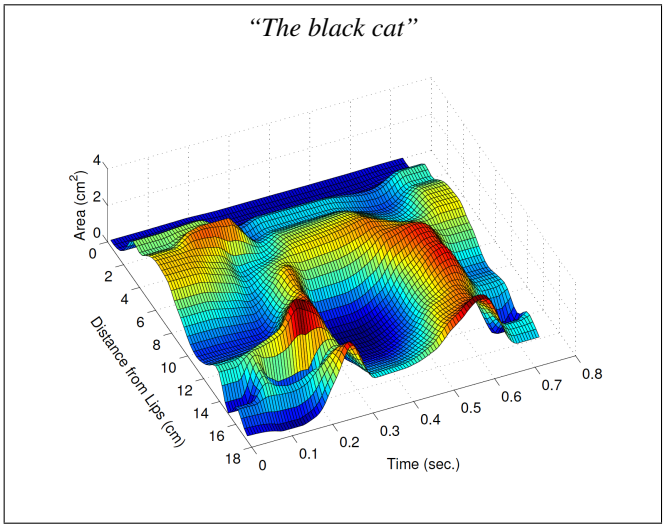

Figure 8. Time-varying sequence of area functions generated by the model parameters in Fig. 7.

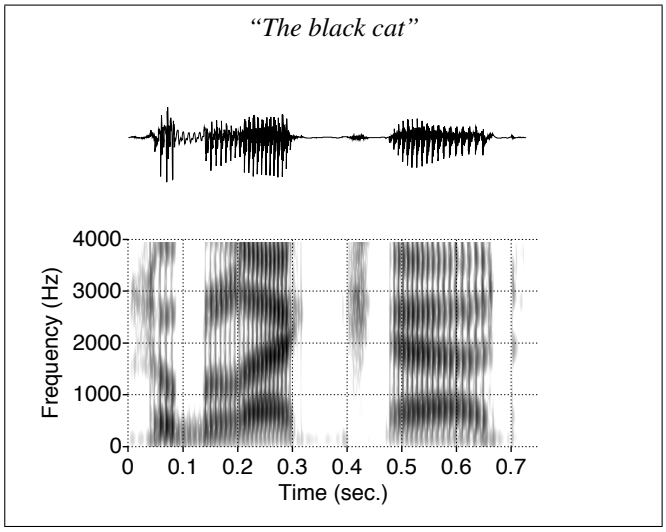

Figure 9. Simulated acoustic waveform and spectrogram for the *"The black cat."*

imposed. The effect of modulating the vowel substrate provided by the Tier I parameters ($q_1(t)$ and $q_2(t)$) with the constrictions generated by Tier II is a "deflection" of the formants away from their vowel-only paths, where the direction is determined by the location of the constriction along the vocal tract length.

The final step is to simulate the acoustic pressures and volume velocities as the vocal tract is modulated. The simulated acoustic waveform and corresponding wide-band spectrogram are shown in Fig. 9. The periods of silence or unvoiced sound are due to the increase in vocal fold separation that occurs at the beginning, at about 0.3-0.45 seconds, and at the end of the utterance. The periodic, but low-frequency portion that can be seen at about 0.1 seconds is due to acoustic radiation from the skin surfaces during the vocal tract occlusion for /b/.

For comparison, a second phrase was also simulated by the same methods described previously. The XRMB version of *"The brown cow"* (broadly transcribed as /ðəbraʊnkaʊ/) was analyzed and the model input parameters determined. These are shown in Fig. 10a, where again there is temporal overlap of all parameters. In this case, there is also nasal coupling required during the production of the /n/, but must be brought to zero rapidly in order to adequately allow pressure to build up for the release of the velar /k/. The calculated formant frequencies for both the vowel-only and vowel+consonant cases are shown in Fig. 10b as the blue and red lines, respectively.

The simulated acoustic waveform for *"The brown cow"* is shown along with the corresponding wide-band spectrogram in Fig. 11.

5. Conclusion

An airway modulation model called *TubeTalker* was introduced as a system for generating artificial speech. The voice

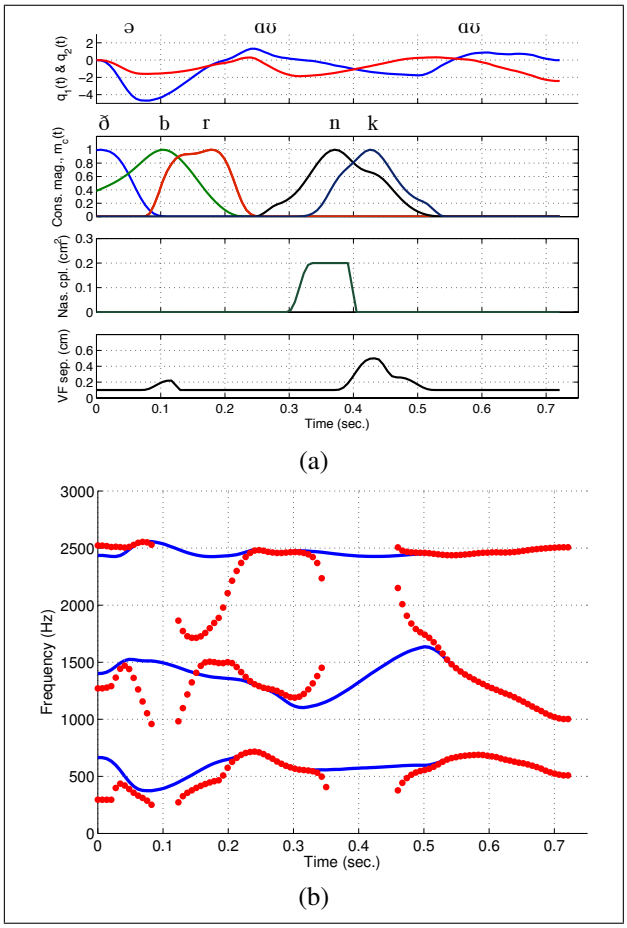

Figure 10. (a) Model input parameters for *"The brown cow"*. (b) Calculated formant frequencies for the area function sequence generated with the parameters in (a); the blue lines are the formants in the absence of the consonant constrictions, whereas the red lines are those calculated with the constrictions imposed.

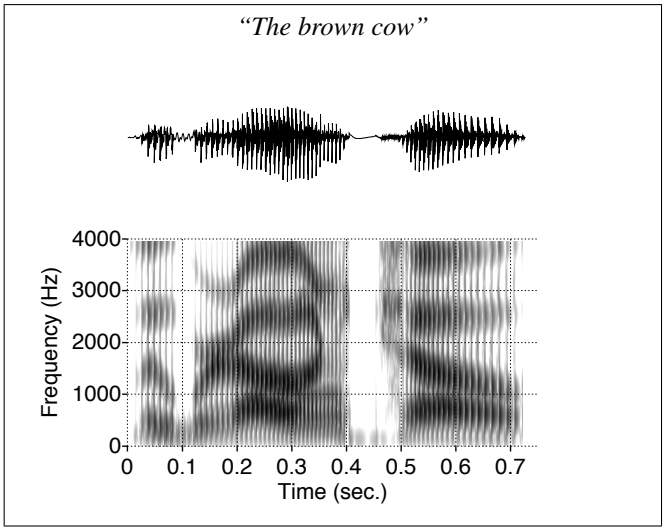

Figure 11. Simulated acoustic waveform and spectrogram for the *"The brown cow."*

source component is based on a kinematic representation of the medial vocal fold surfaces that can be set into vibration for voicing and abducted/adducted for voiceless portions of an utterance. The vocal tract is represented by an area function whose time-dependent shape is generated by vowel and consonant modulations. The result is a speech signal that can be analyzed in the same way as recorded, natural speech, and can be presented to listeners for formal or informal evaluation.

The model is currently implemented with a combination of code written in C and Matlab [22]. 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. 2. In its current form, the model runs with a compute-time to real-time ratio of about 8:1 on a PC laptop or Macbook Air, but it would possible to optimize the code for near real-time operation if desired.

6. Acknowledgments

This research was supported by grant NIH R01 DC04789. The X-ray microbeam data used for this article were collected by Kate Bunton, Carl Johnson, and Rick Konapacki. Thanks also to John Westbury and Gary Weismer for facilitating the data collection.

References

[1] D. Brewster, *Letters on Natural Magic*, Chatto and Windus, Piccadilly, London, pp. 267-271, 1883.

[2] H. Dudley and T. H. Tarnoczy, "The speaking machine of Wolfgang von Kempelen," *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 151-166, 1950.

[3] Z. Fagyal, "Phonetics and speaking machines: On the mechanical simulation of human speech in the 17th century," *Historiographia Linguistica,* vol. 28, no. 3, pp. 289-330, 2001.

[4] C. Wheatstone, "On Speaking Machines," London & Westminster Review 28, 1837, printed in *The Scientific Papers of Sir Charles Wheatstone*, pp. 348-367, 1879.

[5] R. Willis, "On vowel sounds, and on reed-organ pipes," *Trans. Camb. Phil. Soc.* vol. 3, pp. 231-268, 1838.

[6] D. Lindsay, "Talking heads" *Invent. & Tech. Mag.*, vol. 13, no. 1, last viewed at americanheritage.com on 03.03.2011, 1997.

[7] J. Henry , "1846 Letter from Joseph Henry to Henry M. Alexander," in *The Papers of Joseph Henry*, Marc Rothenberg, Ed., vol. 6, pp. 359-364, Smithsonian Inst., 1992.

[8] H. Dudley, R. R. Riesz, and S. S. A. Watkins,"A synthetic speaker," *J. Franklin Inst.* vol. 227, no. 6, pp. 739-764, 1939.

[9] H. Dudley,"Remaking speech," *J. Acoust. Soc. Am.* vol. 11, no. 2, pp. 169-177, 1939.

[10] J. Flanagan,"An interview conducted by Frederik L. Nebeker," *IEEE History Center,* Interview 332, 8 April 1997.

[11] F. Cooper, "An interconversion of audible and visible patterns as a basis for research in the perception of speech" *Proc. Nat. Acad. Sci.* vol. 37, pp. 318-325, 1951.

[12] H. Dudley, "The carrier nature of speech," *Bell Sys. Tech. J.* vol. XIX, no. 4, pp. 495-515, 1940.

[13] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area, *J. Acoust. Soc. Am.*, vol. 75, pp. 570-580, 1984.

[14] I. R. Titze, "*The Myoelastic Aerodynamic Theory of Phonation*, National Center for Voice and Speech, pp. 197-214, 2006.

[15] I. R. Titze, "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," *J. Acoust. Soc. Am.*, vol. 111, pp. 367-376, 2002.

[16] J. Liljencrants, *Speech Synthesis with a Reflection-Type Line Analog*, DS Dissertation, Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech., Stockholm, Sweden, 1985.

[17] B. H. Story, *Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract*, Ph. D. Dissertation, University of Iowa, 1995.

[18] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *J. Acoust. Soc. Am.,* vol. 117, no. 5, pp. 3231-3254, 2005.

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

[20] B. H. Story, "Time-dependence of vocal tract modes during production of vowels and vowel sequences," *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3770-3789, 2007.

[21] B. H. Story, "Vowel and consonant contributions to vocal tract shape," *J. Acoust. Soc. Am.*, vol. 126, pp. 825-836, 2009.

[22] The Mathworks, MATLAB, Version 7.11.0.584 (R2010b).