Published in final edited form as:

Perspect Speech Sci Orolfac Disord. 2014 August ; 24: 7–20. doi:10.1044/ssod24.1.7.

Structure, Movement, Sound, and Perception

Brad H. Story

Speech, Language, and Hearing Sciences, University of Arizona

Abstract

Models that take the form of artificial talkers and speech synthesis systems have long been used as a means of understanding both speech production and speech perception. The article begins with a brief history of two artificial speaking devices that exemplify the representation of speech production as a system of modulations. The development of a recent airway modulation model is then described that simulates the time-varying changes of the vocal tract and acoustic wave propagation. 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.

Introduction

The focus of this article is on the development and use of models to understand the relation of anatomical structure and articulatory movement to the acoustics and perception of speech sounds. Often realized as speech synthesizers or artificial talkers, such models simplify and emulate the human speech production system. The type of simplification presented here is to view speech production as a set of simultaneously imposed modulations of the airway system. Specifically, the vibratory motion of the vocal fold structure modulates the glottal airspace, while slower movements of other structures such as the tongue, jaw, lips, and velum modulate the shape of the pharyngeal and oral cavities, and coupling to the nasal system. The precise timing of these modulations produces an acoustic wave that is the stimulus from which listeners extract phonetic and talker-specific information.

The first aim of the article is to briefly review two historical models of speech production that both exemplify a system in which structure is modulated with movement to produce intelligible speech. The second aim is to describe some theoretical aspects of a computational model that allows for simulation of speech based on precise spatio-temporal modulations of an airway structure. 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.

Modulation Models of Speech Production

The first true speaking machines are often attributed to Kratzenstein and von Kempelen (c.f., Brewster, 1883; Dudley & Tarnoczy, 1950) who, in the late 1700's, both built mechanical devices in which a sound source excited variously shaped resonant cavities resulting in

Disclosure: Brad Story has no financial or nonfinancial relationships related to the content of this article.

speech-like sounds. In the early 1820's, a German inventor named Joseph Faber began work on a speaking machine based largely on the public record of von Kempelen's device (Lindsay, 1997). As a simulation of human speech production, the machine included a bellows that supplied pressure to a reed which, in turn, supplied 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 of the machine 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" (Lindsay, 1997). Faber's talking machine also attracted the attention of Princeton physicist Joseph Henry who requested a private demonstration. Afterward Henry wrote to a colleague "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" (as qtd. in Rothenberg et al. 1992, p. 362). He also noted that Faber's production of English sentences was actually better with the machine than when Faber attempted to speak them himself (Rothenberg et al. 1992).

Although the comments about Faber's talking machine having a German accent when speaking English 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 not be 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 a century later, Homer Dudley and colleagues introduced an electronic artificial talker called the "VODER," an acronym comprised of the capitalized letters in "Voice Operation DEMonstratorR" (Dudley et al., 1939). In many ways, it was similar to Faber's earlier mechanical creation in that an operator controlled the system with the hands and feet. The difference was that the VODER utilized electrical circuits to emulate the acoustic characteristics of speech rather than attempting to mechanically replicate the human airways and sound sources. This device was demonstrated at the 1939 World's Fair in New York and at the Golden Gate Exposition in San Francisco the same year. Twenty four people were trained to operate the VODER for these exhibitions and typically about 1 year was required to develop the ability to produce intelligible speech with it. In fact, Dudley et al. wrote "[...] 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" (Dudley et al., 1939, p. 763). Once learned, the ability to "speak" with the VODER was apparently retained for years afterward, even without continued practice. Some 20 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 device. As recalled by James Flanagan, a Bell Labs engineer and speech scientist, that “She sat down and gave a virtuoso performance on the Voder” (As qtd. in Pieraccini, 2012, p. 55).

The examples of both Faber’s Talking Machine and Dudley’s VODER suggest that the operators of the devices learned and internalized 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 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 (Mattingly, 1974). A theoretical 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. Dudley (1940) expressed this view in an article called “The carrier nature of speech” in which 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.”

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 at multiple levels of the speech production system. The primary purpose of this model is to understand the acoustic characteristics of time-dependent changes in airway shape and ultimately how those characteristics are related to the perception of speech.

Airway Modulation Model

Description

The starting point for development of the model, described in most detail in Story (2005a), is based on a view of speech production where vowels and transitions from one vowel to another are produced by modulation of a “phonetically-neutral” vocal tract configuration. In turn, consonants result from another layer of modulation that imposes severe, localized constrictions on the underlying vowel or evolving vowel transition. Thus, the speech signal carries information concerning each of the successive layers of structure or movement on which it is built. Perception of speech may then be thought to consist of “peeling away” the layers of modulation in their acoustically-coded form to recover the embedded information.

The concept of a neutral shape is a typical starting point for vocal tract modeling and is often represented as a tube with uniform cross-sectional area along the entire length of the vocal tract (Fant, 1960; Schroeder, 1967; Mermelstein, 1967). A uniform tube is assumed to approximate both the articulatory configuration and the equally-spaced formant frequency characteristics of a phonetically-neutral vowel. Tubular shapes representative of other vowels are obtained by perturbing the uniform configuration (e.g., Schroeder, 1967). A more physiologically-realistic view of the neutral vocal tract shape has emerged from analysis of area function inventories collected from numerous talkers. For example, Story and Titze

(1998) showed that the mean area function for a male talker, based on a 10-vowel collection, produced a pattern of acoustic resonances with nearly equally-spaced formant frequencies, similar to those of an equivalent-length uniform tube. Its shape, however, was approximately uniform only within the pharyngeal portion. Other parts of the vocal tract were non-uniformly narrowed or expanded. Similarly, mean area functions determined from vocal tract data collected for six talkers were shown to be of nonuniform shape, even though their formant frequency patterns were approximately “neutral” in the range of the first two formants (Story, 2005b). Differences in the shape of these mean area functions across talkers were reflected acoustically by wide variations in the upper formant frequencies. These studies suggested that a talker’s mean vocal tract shape provides acoustic patterns that are both phonetically neutral and idiosyncratic to the speaker. Other studies have supported the idea that speakers may modify their neutral tract shape to produce different idiosyncratic characteristics while maintaining the phonetically-neutral quality (Story, Titze, and Hoffman, 2001; Story and Titze, 2002). A similar view was proposed by Traunmüller (1994) who suggested that speech signals result from allowing articulatory gestures to modulate a “carrier” signal that is phonetically neutral, but descriptive of the “personal quality” of the speaker. In terms of a model of the vocal tract, the neutral shape could be considered the “carrier” that sets the acoustic background on which vowel and consonant perturbations can be imposed.

The first layer of modulation in the model, henceforth referred to as a “modulation tier,” considers that vowels and transitions from one vowel to another can be represented as a superposition of a weighted linear combination of spatial deformation patterns on an underlying neutral vocal tract configuration. These patterns, derived by Principal Component Analysis (PCA), tend to be common across talkers (Story and Titze, 1998; Story, 2005b, 2009). They are configured such that one pattern produces a continuum from /i/-like to /a/-like vocal tract shapes while another represents an /æ/-like to /u/-like continuum, suggesting that they represent a coordination of the speech articulators necessary for vowel production. The results are similar to factor analyses of tongue shape (Harshman, Ladefoged, and Goldstein, 1977; Shirai and Honda, 1977; Nix, Papcun, Hogden, and Zlokarnik, 1996; Zheng, Hasegawa-Johnson, and Pizza, 2003), shaping components used in other articulatory models (Berry, 2003; Maeda, 1990, 1991), as well as analogous research on vocal tract area functions (Mokhtari et al., 2007). Some recent work has also shown that the shape of these patterns can be derived theoretically from calculations of acoustic sensitivity functions of the nonuniform neutral area functions of speakers (Story, 2007). That is, the spatial distribution of acoustic kinetic and potential energy along the tract length, associated with the formants F1 and F2, is predictive of the shaping patterns obtained from analysis of vocal tract data, based on both imaging and kinematic studies. This is important because it suggests that the basic structure of the airway modulation model is not only statistically based, but also acoustically relevant to the spatial aspects of human vocal tracts.

The second tier in the model was motivated, to some degree, by Öhman’s (1966) spectrographic study of vowel-consonant-vowel (VCV) utterances. He concluded that “A VCV utterance of the kind studied here can, accordingly, not be regarded as a linear sequence of three successive gestures,” and proposed that a consonant gesture (constriction) is superimposed on an underlying vowel substrate. The implication is that speech proceeds

as a series of independently controlled vowel-to-vowel transitions, interrupted by superposition of consonant perturbations (Fujimura, 1992). Öhman (1967) subsequently developed a model in which the midsagittal cross-distance (width) of one vowel shape could be interpolated to another over the time course of a vowel-vowel transition. Simultaneously, and in parallel, a consonant constriction function could be activated that varied over the same time course as the vowel component. At each successive point in time, the consonantal function would be superimposed on the modeled vowel substrate to produce a composite tract shape. As compared to X-ray data, the model was successful at describing the time variation of the midsagittal vocal tract shape for a wide range of VCV utterances.

Similarly, Perkell (1969) observed that articulatory structures behave differently depending on whether they are influenced by demands to produce a vowel or a consonant. Gracco (1992) also referred to two general categories of vocal tract movement: “valving” to produce constrictions, and “shaping” to modulate the vocal tract geometry. This paradigm of a functional division of the articulatory system into separate vowel and consonant classes has influenced various control strategies for articulatory and vocal tract models. Nakata and Mitsuoka (1965) and Ichikawa and Nakata (1968) designed a rule-based synthesizer in which consonants were superimposed on a vowel-vowel transition. Similarly, Båvegård (1995) and Carré and Chennoukh (1995) have both reported vocal tract area function models where consonant constrictions could be superimposed on an interpolation of a vowel-to-vowel transition. The primary value of these types of models, regardless of their specific implementation, is that they allow for a concise mathematical representation of the change in vocal tract shape that produces the acoustic characteristics present in the speech signal. In the next section, several examples are presented that demonstrate the multi-tier approach to modeling the time-varying vocal tract shape and corresponding acoustic consequences.

Demonstration of Formant Frequency Patterns for VCVs

Shown in the first row of Figure 1 is an example of a “neutral” vocal tract shape and its corresponding formant frequencies. The formants are labeled $F1_n$, $F2_n$, and $F3_n$ and plotted with respect to a time axis to represent an utterance that is 0.5 seconds in duration, but has no temporal variation in acoustic characteristics. The second row of the figure demonstrates a relatively slow modulation of the neutral tract shape to produce a transition from an /ɪ/ vowel to an /ɑ/ (Tier 1). The red line in the vocal tract plot indicates the initial shape for the /ɪ/, the blue line is the final shape for the /ɑ/, and the dashed lines represent the temporal change between the initial and final configurations. The formant frequency plot indicates how each of the corresponding three vocal tract resonances (solid gray lines) are displaced about those for the underlying neutral tract (dashed lines) over the time course of the /ɪɑ/ transition.

In the third row of Figure 1, a localized and rapid modulation is superimposed on the underlying /ɪɑ/ transition at the lip end of the vocal tract (Tier 2). The location of the constriction is denoted as $l_c = 1.0$ which is a value normalized relative to the total vocal tract length; i.e., the lip termination is equal to the vocal tract length and any other constriction locations will have values less than 1.0. The cross-sectional area achieved during the modulation is a_c , which in this case is set equal to 0 cm². The corresponding formant plot

shows, along with the neutral tract formants and those of the VV transition, the $F1$, $F2$, and $F3$ frequencies resulting from the superposition of the bilabial occlusion. The formants are deflected away from their vowel-vowel path during the periods of occlusion onset and offset. For this particular location, all three formant frequencies are deflected downward in frequency as indicated by the blue filled regions and minus signs. The break in the formant tracks from approximately 0.15–0.28 seconds is the time period for which the vocal tract was occluded.

Three additional examples are given in Figure 2. In each case, the neutral vocal tract shape and / $\text{I}\alpha$ / modulation are exactly the same as in Figure 1, but the constriction location and cross-sectional area achieved by the constriction are varied. A partial occlusion is demonstrated in the first row of plots. The constriction location and minimum area were set to be $l_c = 0.93$ and $a_c = 0.05 \text{ cm}^2$, respectively, to roughly simulate a lingua-dental consonant. The formant deflection pattern shows that $F1$ and $F2$ decrease in frequency relative to the / $\text{I}\alpha$ / transition as for the bilabial occlusion, but $F3$ is now deflected upward. A full occlusion located at $l_c = 0.86$, the alveolar region, is shown in the second row of the figure. The formant deflection pattern is again different, with $F1$ deflected downward while both $F2$ and $F3$ are deflected upward. In the third row, the occlusion is imposed in the velar region at $l_c = 0.68$ and creates another unique formant deflection pattern. In this case $F1$ and $F3$ are deflected downward during occlusion onset and offset while $F2$ is deflected upward.

The examples presented in Figures 1 and 2 demonstrate how constriction location is related to the temporal pattern of deflection of the first three formants away from the more slowly-varying path dictated by the underlying vowel transition. In fact, if a series of VCV utterances is generated with the model such that the constriction location is incremented by only a small portion of the total tract length from the velar region to the lips, it can be shown that the four deflection patterns depicted in Figures 1 and 2 are maintained. Table 1 gives the ranges of constriction locations l_c that produce specific deflection patterns as indicated by the plus and minus symbols. The velar and alveolar regions specified in the table occupy considerably larger portions of the total vocal tract length than either the dental or lip region. This observation coincides with expectations based on the known anatomy and physiology of the articulators, however, it is noted that these regions have been defined purely on acoustic considerations, not on anatomy or physiology.

Perceptual Experiments

As an example of using the airway modulation model to facilitate studies of speech perception, Story and Bunton (2010) performed a consonant identification experiment to determine whether listeners are sensitive to the formant deflection patterns associated with the different regions of the vocal tract. The hypothesis was that the boundaries between perceptual categories would be aligned with the constriction locations at which the formant deflection patterns changed. Three VCV continua were generated by superimposing constrictions on the VV contexts / $\text{æ}\alpha$ /, / $\text{æ}\text{i}$ /, and / $\text{æ}\text{u}$ / . The constriction location was incremented from lips to the velar region, and the time-varying vocal tract shape corresponding to each VCV was coupled to a voice source to generate a speech waveform for presentation to listeners.

Because all the constrictions were full occlusions and there was voicing throughout each VCV, the listeners were given a forced choice of /b, d, g/. The results indicated that perceptual categories were aligned with the formant deflection regions, similar to those in Table 1. This outcome is not as obvious as it might seem. The three different VV contexts in the experiment created conditions such that the absolute direction of the formant transitions for a given consonant category (e.g., /d/) might be either upward or downward in frequency. For example, formant patterns are shown in Figure 3 for a constriction imposed on the VV contexts /əi/, and /əu/ at a normalized location of $l_c = 0.86$ (same as middle row of Fig. 2). In the /əi/ context, the F2 transition direction during the release of the constriction (from 0.25–0.45 sec.) is upward, whereas in the /əu/ context it is downward. But in both cases, the deflection pattern of the formants relative to the underlying VV formant contours (dashed lines) is the same. Thus, the perception of a given consonant tended to align with the relative formant deflection patterns rather than the absolute direction of change in formant frequencies.

Demonstration of Kinematic, Aerodynamic, and Acoustic Quantities

To this point, the focus has been on using the airway modulation model to relate the time-varying vocal tract shape to resulting formant frequencies, and to perceptual identification of speech sounds. The model also facilitates understanding how kinematic modulations of the vocal tract and glottis affect the aerodynamic and acoustic signals generated during speech production. As an example, a simulation of the phrase “Happy Birthday” is shown in Figure 4. The time-varying vocal tract shape is plotted in the upper left portion of the figure. It is composed of continuous movement through the vowels $\text{æ} \rightarrow \text{ɪ} \rightarrow \text{ʊ} \rightarrow \text{eɪ}$, which, in turn, is perturbed by constrictions that impose characteristics of the consonants /p, b, d/. A kinematic model of the vocal folds (Titze, 1984; 2006), shown schematically in the lower left part of the figure, generates the time-varying glottal area that results from vibration during voiced segments and the much slower abductory/adductory maneuvers required for unvoiced portions of an utterance. When coupled to the air flows and pressures in the trachea and vocal tract, it produces the glottal flow signal which is the primary sound source transformed into speech by the vocal tract (Titze, 2002).

Glottal area, glottal flow, intra-oral pressure, and radiated acoustic pressure waveforms generated by simulating “Happy Birthday” are plotted on the right side of Figure 4. Note that units for all signals have been omitted to keep the figure visually simple. The glottal area waveform in the lower panel shows three abductory events that are needed to produce the initial /h/ sound, as well as the unvoiced /p/ and /θ/ sounds that occur near 0.25 seconds and 0.7 seconds, respectively. During these events the vibration of vocal folds essentially ceases and eliminates voicing. At other points in time the vocal folds are adducted and vibrating at some specified fundamental frequency. The glottal flow exhibits nearly the same pattern except that its amplitude decreases during the time periods when the vocal tract is occluded by a constriction (e.g., 0.2–0.3 sec., 0.4–0.5 sec., and 0.65–0.85 sec.). During these same periods, the intra-oral pressure increases rapidly during the occlusion and then drops when the constriction is released. The radiated output pressure in the top panel is the final product of acoustic wave propagation through an airway system undergoing continuous modulation of shape at multiple time scales (Liljencrants, 1984; Story, 1995).

Simulation of speech with a computational model is a useful tool because the signals generated are analogous and comparable to signals that can be obtained in a laboratory or clinic. Thus, a model can provide a system for validation of clinical measurements. In addition, a model allows a researcher control over many aspects of the speech production system that cannot necessarily be controlled by a human talker, thus providing a platform for “what if” types of experimentation. For example, in the case of “Happy Birthday” (or any other simulated word, phrase, or sentence) the timing of the events in the vocal tract or at the glottis could be modified according to a hypothesis regarding a particular speech disorder or speech quality. The vocal tract movements to produce vowels could be attenuated, delayed, or shifted in time relative to the consonantal constrictions, perhaps similar to some dysarthrias. Glottal modulations could be modified to simulate tremor or hypo-adduction during voiced segments. The structure of the vocal folds and vocal tract could be scaled in size based on anatomical measurements to understand the effect of growth during childhood development of speech production. The point is that models can provide unique insight into the human speech production system that compliments observations and measurements obtained from real talkers.

Comparison to Other Models of Speech Production

The airway modulation model described here is part of a class of models that operate on a parametric representation of the vocal tract area function (cf. Stevens & House, 1955; Fant, 1960; Mrayati et al., 1988; Stevens, 1998). Admittedly an abstraction of articulatory reality, the area function is the representation that provides the most direct connection between vocal tract shape and resulting acoustic characteristics. As demonstrated in previous sections, a parametric model of the area function is useful for situations in which precise control of the detailed structure of the vocal tract shape is desired.

In contrast, an articulatory model explicitly represents the structure of individual articulators, such as the tongue, jaw, lips, velum, and larynx, as well as parameters for controlling their positions and movements. Articulatory models often are configured to operate in the midsagittal plane (cf. Lindblom and Sundberg, 1971; Coker, 1976; Mermelstein 1973; Maeda, 1990; Scully, 1990), but more recent developments have included remarkably detailed three-dimensional representations of some aspects of the system (cf. Dang and Honda, 2004; Fels et al., 2006; Birkholz, 2013). Such models are intuitively appealing because of the explicit relation between model parameters and human articulatory structures. Relating articulation to acoustic characteristics with these models, however, usually requires an empirically-based transformation of midsagittal coordinates to the area function domain. Thus, control of the overall vocal tract shape is less direct with an articulatory model than with a model whose parameters explicitly affect the area function.

Typically the control parameters of either area function or articulatory models are obtained by analysis of kinematic, acoustic, or aerodynamic signals collected from real talkers. Once an acceptable simulation of the original speech is produced, parameters can be modified to impose various conditions for experimentation as suggested in the previous section. A different approach, however, is to develop a neural model of speech planning and execution that can control the input parameters of a sound production model. The DIVA model (cf.

Guenther et al., 2006) is a computational neural network that learns to control movements of an articulatory model (Maeda, 1990) through simulated babbling. It includes a feedback component that accounts for somatosensory and auditory information, as well as a feedforward system. Control signals are combined in a model of the motor cortex to plan movements that are executed in the articulatory model. Kroger et al. (2010) have developed a similar neural-based model of speech production that learns to execute movements in a three-dimensional articulatory model (Birkholz, 2013). These types of models provide many new possibilities for simulating motor control of speech production, and a means to test the effects of various types of impairments in the feedback or feedforward control loops. They also provide an interesting computational parallel to the “operators” trained to produce speech with the VODER, or Faber’s learned abilities to talk with his mechanical speaking machine. Perhaps neural-based models coupled with speech simulation systems will eventually uncover some of the “rules” required for the acquisition of speaking skills.

Summary

There is much history of research in speech production on relating articulatory movement and vocal tract shape to acoustic characteristics, such as formant frequencies. Similarly, there is a long history in speech perception research of generating stimuli for experiments by directly manipulating acoustic characteristics with a speech synthesizer. There have been only limited studies, however, in which the stimuli for a perceptual experiment have been generated with a simulation of the speech production process, where the formant frequencies and other acoustic characteristics contained in each stimulus result directly from the resonant structure of the vocal tract, and not from idealizations of formant frequency patterns. This has been for good reason. Articulation-based synthesis, while invaluable for understanding the sound producing mechanisms of speech, has typically required large amounts of user interaction and parameter tuning to produce even a few short utterances, and these are often of a quality that circumvents their use in perceptual tests. Yet listening experiments based on simulated speech are important to pursue because they can answer questions about the perceptual relevance of various types of kinematic and structural variations of the trachea, larynx, vocal tract, and nasal passages. A model that simplifies speech production to a multi-tier system of airway modulations can serve as research tool to relate structure, movement, sound, and perception.

Acknowledgments

The research was supported by NIH R01 DC011275 and NSF BCS-1145011.

References

- Berry, J. Ph.D. Dissertation. University of Wisconsin-Madison; 2003. Second formant effects of low dimensional models of speaking rate change in articulatory synthesis.
- Birkholz P. Modeling consonant-vowel coarticulation for articulatory speech synthesis. PloS one. 2013; 8(4):e60603. [PubMed: 23613734]
- Båvegård, M. Proceedings Eurospeech 95. Madrid: Spain; 1995. Introducing a parametric consonantal model to the articulatory speech synthesizer.
- Brewster, D. Letters on Natural Magic. Piccadilly, London: Chatto and Windus; 1883. p. 267-271.

- Carré R, Chennoukh S. Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures. *J. Phonetics*. 1995; 23:231–241.
- Coker CH. A model of articulatory dynamics and control. *Proc. IEEE*. 1976; 64(4):452–460.
- Dang J, Honda K. Construction and control of a physiological articulatory model. *J. Acoust. Soc. Am*. 2004; 115:853–870. [PubMed: 15000197]
- Dudley H, Riesz RR, Watkins SSA. A synthetic speaker. *J. Franklin Inst*. 1939; 227(6):739–764.
- Dudley H. The carrier nature of speech. *Bell Sys. Tech. J*. 1940; 19(4):495–515.
- Dudley H, Tarnoczy TH. The speaking machine of Wolfgang von Kempelen. *J. Acoust. Soc. Am*. 1950; 22(2):151–166.
- Fant, G. *Acoustic Theory of Speech Production*. The Hague: Mouton; 1960.
- Fels, S.; Vogt, F.; Van Den Doel, K.; Lloyd, J.; Stavness, I.; Vatikiotis-Bateson, E. *International Seminar on Speech Production*. Brazil: Ubatuba; 2006. Artisynth: A biomechanical simulation platform for the vocal tract and upper airway.
- Fujimura O. Phonology and phonetics - A syllable based model of articulatory organization. *J. Acoust. Soc. Jpn*. 1992; 13:39–48.
- Gracco V. Characteristics of Speech as a Motor Control System. Haskins Laboratories Status Report on Speech Research. 1992; SR-109/110:13–26.
- Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*. 2006; 96(3):280–301. [PubMed: 16040108]
- Harshman R, Ladefoged P, Goldstein L. Factor analysis of tongue shapes. *J. Acoust. Soc. Am*. 1977; 62(3):693–707. [PubMed: 903511]
- Ichikawa, A.; Nakata, K. Speech synthesis by rule. In: Kohasi, Y., editor. *Reports of the 6th Intl. Cong. on Acoust*. Tokyo: International Council of Scientific Unions; 1968. p. 171-1744.
- Kröger BJ, Birkholz P, Lowit A, Neuschaefer-Rube C. Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production. *Speech motor control: New developments in basic and applied research*. 2010:23–36.
- Liljencrants, J. DS Dissertation. Stockholm, Sweden: Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech.; 1985. *Speech Synthesis with a Reflection-Type Line Analog*.
- Lindblöm B, Sundberg J. Acoustical consequences of lip, tongue, jaw, and larynx movement. *J. Acoust. Soc. Am*. 1971; 4(2):1166–1179. [PubMed: 5117649]
- Lindsay D. Talking head. *American Heritage Invention & Technology*. 1997; 13(1):57–63.
- Maeda, S. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In: Hardcastle, WL.; Marcha, A., editors. *Speech production and speech modeling*. Dordrecht: Kluwer Academic Publishers; 1990. p. 131-149.
- Maeda S. On articulatory and acoustic variabilities. *J. Phon*. 1991; 19:321–331.
- Mattingly, IG. Speech synthesis for phonetic and phonological models. In: Sebeok, TA., editor. *Current Trends in Linguistics*. The Hague: Mouton; 1974. p. 2451-2487.
- Mermelstein P. Determination of the vocal-tract shape from measured formant frequencies. *J. Acoust. Soc. Am*. 1967; 41(5):1283–1294. [PubMed: 6074791]
- Mermelstein P. Articulatory model for the study of speech production. *J. Acoust. Soc. Am*. 1973; 53(4):1070–1082. [PubMed: 4697807]
- Mokhtari P, Kitamura T, Takemoto H, Honda K. Principal components of vocal tract area functions and inversion of vowels by linear regression of cepstrum coefficients. *J. Phonetics*. 2007; 35:20–39.
- Mrayati M, Carré R, Guérin B. Distinctive regions and modes: A new theory of speech production. *Speech Comm*. 1988; 7:257–286.
- Nakata K, Mitsuoaka T. Phonemic transformation and control aspects of synthesis of connected speech. *J. Radio Res. Labs*. 1965; 12:171–186.
- Nix DA, Papcun G, Hogden J, Zlokarnik I. Two cross-linguistic factors underlying tongue shapes for vowels. *J. Acoust. Soc. Am*. 1996; 99(6):3707–3717. [PubMed: 8655802]
- Öhman SEG. Coarticulation in VCV utterances: Spectrographic measurements. *J. Acoust. Soc. Am*. 1966; 39:151–168. [PubMed: 5904529]

- Öhman SEG. Numerical model of coarticulation. *J. Acoust. Soc. Am.* 1967; 41:310–320. [PubMed: 6040806]
- Perkell, JS. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge, MA: The MIT Press; 1969.
- Pieraccini, R. *The Voice in the Machine*. MIT Press; Cambridge, MA: 2012.
- Rothenberg, M.; Dorman, KW.; Rumm, JC.; Theerman, PH. *The Papers of Joseph Henry: Volume 6, 1846 letter from Joseph Henry to Henry M. Alexander*. Washington, DC: Smithsonian Institution Press; 1992. p. 359-364.
- Scully, C. Articulatory synthesis. In: Hardcastle, WJ.; Marchal, A., editors. *Speech Production and Speech Modeling*. Dordrecht: Kluwer Academic Publishers; 1990. p. 151-186.
- Schroeder MR. Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am.* 1967; 41(4):1002–1010. [PubMed: 6046539]
- Shirai, K.; Honda, M. Estimation of articulatory motion. In: Sawashima, M.; Cooper, F., editors. *Dynamic Aspects of Speech Production*. Univ. of Tokyo Press; 1977. p. 279-302.
- Stevens KN, House AS. Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Am.* 1955; 27(3):484–493.
- Stevens, KN. *Acoustic Phonetics*. Cambridge, MA: MIT Press; 1998.
- Story BH, Titze IR. Parameterization of vocal tract area functions by empirical orthogonal modes. *J. Phonetics*. 1998; 26(3):223–260.
- Story BH, Titze IR, Hoffman EA. The relationship of vocal tract shape to three voice qualities. *J. Acoust. Soc. Am.* 2001; 109:1651–1667. [PubMed: 11325134]
- Story BH, Titze IR. A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function. *J. Phonetics*. 2002; 30:485–509.
- Story, BH. Ph. D. Dissertation. University of Iowa; 1995. Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract.
- Story BH. A parametric model of the vocal tract area function for vowel and consonant simulation. *J. Acoust. Soc. Am.* 2005a; 117(5):3231–3254. [PubMed: 15957790]
- Story BH. Synergistic modes of vocal tract articulation for American English vowels. *J. Acoust. Soc. Am.* 2005b; 118(6):3834–3859. [PubMed: 16419828]
- Story BH. A comparison of vocal tract perturbation patterns based on statistical and acoustic considerations. *J. Acoust. Soc. Am.* 2007; 122(4):EL107–EL114. [PubMed: 17902738]
- Story BH. Vocal tract modes based on multiple area function sets from one speaker. *J. Acoust. Soc. Am.* 2009; 125(4):EL141–EL147. [PubMed: 19354352]
- Story BH, Bunton K. Relation of vocal tract shape, formant transitions, and stop consonant identification. *J. Spch. Lang. Hear. Res.* 2010; 53:1514–1528.
- Titze IR. Parameterization of the glottal area, glottal flow, and vocal fold contact area. *J. Acoust. Soc. Am.* 1984; 75:570–580. [PubMed: 6699296]
- Titze IR. Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *J. Acoust. Soc. Am.* 2002; 111:367–376. [PubMed: 11831809]
- Titze IR. *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech. 2006:197–214.
- Trautman H. Conventional, biological and environmental factors in speech communication: A modulation theory. *Phonetica*. 1994; 51:170–183. [PubMed: 8052672]
- Zheng Y, Hasegawa-Johnson M, Pizsa S. Analysis of the three-dimensional tongue shape using a three-factor analysis model. *J. Acoust. Soc. Am.* 2003; 113(1):478–486. [PubMed: 12558285]

Base Structure:

Neutral vocal tract shape sets the acoustic “background”

Tier 1:

Global modulations of the neutral vocal tract shape perturb the formants over a “long” time scale

Tier 2:

Local perturbation superimposed on underlying vocal tract shape - deflects the formants away from their path over a “short” time scale

Figure 1.

Demonstration of the airway modulation model. The middle column shows vocal tract shapes and in the right column are the time-varying formant frequencies associated with the vocal tract shapes. The lower rightmost panel show the formant deflection pattern for a bilabial constriction superimposed on the underlying /ɪɑ/ vowel transition.

Figure 2.

Three additional examples of vocal tract shape change representative of VCVs, and the corresponding time-varying formant frequencies. Note that the formant deflection pattern is unique for each of the three cases.

Figure 3.

Time-varying formant frequencies of VCVs produced by superimposing the same constriction on two different VV contexts. The normalized constriction location along the vocal tract length was $l_c = 0.86$ (alveolar region) and the VV contexts were /əi/ and /əu/. In both plots, the solid lines are formants for the VCV and the dashed lines are those for the VV alone. The colored portions indicate the period of time during which the constriction affects the vocal tract shape and perturbs the formants upward (red) or downward (blue) relative to the formants of the VV. (a) /əi/ context where the absolute direction of F2 change in the latter half of the VCV is upward. (b) /əu/ context where the absolute direction of F2 change in the latter half of the VCV is downward. In both (a) and (b), the formant deflection pattern relative to the underlying VV is the same.

Figure 4.

Simulation of the phrase "Happy Birthday." The time-varying vocal tract is shown in the upper left, and the kinematic model of the vocal folds used as the voice source is in the lower left corner. Waveforms of glottal area, glottal flow, oral pressure, and radiated output pressure are plotted on the right side of the figure, respectively, from bottom to top. The light gray windows indicate the approximate temporal extent of the consonantal constrictions.

Table 1

Regions of normalized vocal tract length in which the formant deflection patterns are maintained. In particular, a constriction within any of the four regions shown will result in a formant deflection pattern given by the plus and minus symbols.

	Region			
l_c	~ Velar 0.65 → 0.75	~ Alveolar 0.77 → 0.89	~ Dental 0.91 → 0.93	~ Lip 0.95 → 1.0
$F3$	–	+	+	–
$F2$	+	+	–	–
$F1$	–	–	–	–