

Physical Modeling of Voice and Voice Quality

Brad H. Story

Department of Speech and Hearing Sciences
University of Arizona, Tucson, AZ
bstory@u.arizona.edu

Abstract

Physical modeling of the phonatory and vocal tract systems has served as a useful tool to study many aspects of speech production. This paper offers a brief review of two specific types of physically-based models. One for simulating the vibration of the vocal folds and another for representing the vocal tract shape in the form of an area function. While much of the technical detail of these models has been presented elsewhere, the emphasis here will be on providing examples of how physical models may be used for studying *voice quality*.

1. Introduction

Sounds produced by biological systems generally result from the vibration of some type of biomaterial (e.g. tissue). These vibrations create a source of oscillating air pressure(s) that acoustically “encode” information pertaining to the vibratory character of the tissue. In turn, these pressures may propagate through or around a physical structure, taking on information about it that is carried along to a listener’s ear. In speech (voiced portions), this process occurs as the sound generated by the vibrating vocal folds propagates along the distance from larynx to lips, “collecting” information that reflects the characteristics of the current vocal tract shape. Thus, the final output signal contains acoustic features that reveal clues to the generation of the sound at its source as well as the vocal tract structure through which the source sound has traveled. During connected speech, the ongoing motor control commands executed by a speaker cause both the vocal folds and the vocal tract to undergo continuous (or nearly so) structural changes enacted by muscle contractions to produce an ever-evolving signal from which a listener can extract a message.

While the human speech production system is similar enough across people to produce messages of common linguistic content, a speaker’s personal biological endowment of the sizes, shapes, and tissue properties of the speech articulators, as well as their own idiosyncratic use of them, will dictate the actual acoustic signal that is produced. It is these speaker-specific acoustic properties underlying the linguistic message that comprise the *quality* of the voice. Laver[1][2] outlined a convenient framework to describe voice qualities based on the concept of “settings” of the speech organs. These so-called “settings” represent habitual muscle tensions throughout the speech production system that impose a specific pattern of use during speech and consequently a specific voice quality.

The aim of this paper is to briefly discuss the use of two physically-based models of speech production to investigate the notion of voice quality. The first part will discuss a model of vocal fold vibration and the second part a model of vocal tract shape. In each case, the emphasis will be on specific examples that show how common acoustic goals can be achieved even

with significant acoustic variability imposed by various “settings” of the speech articulators.

2. Vocal fold model

The vocal folds are soft tissue structures contained within the cartilaginous framework of the larynx. Their structure is often described by the cover–body concept[3] that suggests they can be roughly divided into two tissue layers with different mechanical properties. The cover layer is comprised of pliable, non-contractile mucosal tissue that is somewhat like a sheath around the body-layer, which consists of muscle fibers (thyroarytenoid) and some ligamentous tissue. Vibration of the vocal folds is initiated, and sustained over time, by the steady air flow and pressure supplied by the lower respiratory system. The pattern of vibration is such that the lateral displacement of the upper (superior) portion of each vocal fold is not in phase with the lower portion. That is, the lower part of the vocal fold leads the upper, creating wave–like motion in the cover from bottom to top. Once in vibration, the vocal folds effectively convert the steady air flow from the lungs into a series of air flow pulses by periodically opening and closing the air space between the vocal folds (called the glottis). The stream of flow pulses provides the sound source for the excitation of the vocal tract resonances in vowels.

Understanding the precise mechanisms responsible for vocal fold vibration has occupied researchers for several decades and this work has spawned a number of useful computational models representing the vocal fold mechanics (e.g. [4][5][6][7]). Chosen as an example for this paper, a three–mass model[8] includes the effect of the cover–body vocal fold structure but also maintains the simplicity of a low–dimensional system, and hence computational efficiency. As shown in a coronal view in Figure 1, this model consists of a “body” mass that is positioned lateral to two cover masses. The cover masses are both connected to the body mass via spring and damper systems that represent stiffness of the cover tissue as well as the effective coupling stiffness between the body and cover. The body mass, in turn, is connected to a rigid lateral boundary with another spring and damper system that account for the effective stiffness of the body which will depend on the level of contraction of the muscle tissue. To account for shear forces in the cover, the two cover masses are coupled to each other with another spring–damper element. A coupling of the equations of motion of this system with simple aerodynamic equations will create a self-oscillating system.

The fundamental frequency (F_0) of vocal fold vibration (which gives rise to the perception of pitch) is dictated primarily by the effective mass and stiffness of the vocal folds. In terms of the three-mass model, this means the actual values of the mass and stiffness elements. However, a speaker controls F_0 largely

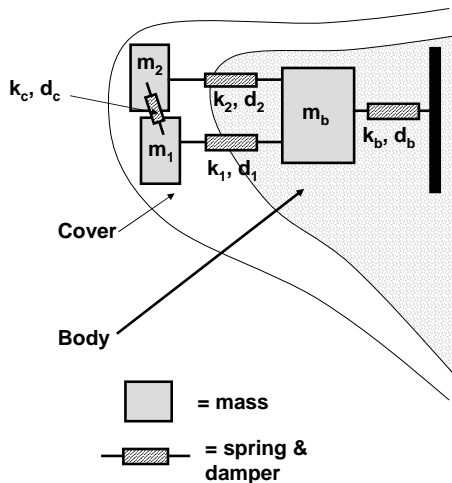


Figure 1: Schematic representation of single vocal fold in the coronal plane. The cover-body structure is indicated by the shaded and white portions and the mass-spring-damper elements representing this structure are also shown.

with two complexes of muscles, the cricothyroid (CT) and thyroarytenoid (TA), that are configured in an agonist/antagonist relationship such that they can rotate the thyroid cartilage relative to the cricoid cartilage (see Figure 2 for a simplified view). In turn, the relative contractions of these muscles influence the strain (length change) of the vocal folds and consequently determines their effective masses and stiffnesses. Thus, for modeling F_0 control that is more relevant to real speakers, a transformation was proposed by [9] that converts normalized muscle activation parameters (a_{TA} , a_{CT}) into the three-mass model parameters of mass and stiffness.

With this transformation, a muscle activation plot for F_0 (Figure 3a) can be generated that shows the level of TA and CT activation on the horizontal and vertical axes, respectively; F_0 contours are indicated within the plot. In other words, each contour line represents combinations of (a_{TA} , a_{CT}) that will produce the same F_0 . Thus, for any given F_0 there are many possible “settings” of a_{TA} and a_{CT} that may be used by a speaker.

The relevance of such a plot to voice quality is that even though each contour line represents the *same* F_0 , the shape of the glottal flow waveform, and hence its distribution of harmonic amplitudes, may be quite different. As an example, the 115 Hz contour line has two large dots (marked A & B), one near the left side of the plot and one near the right side. These two points represent muscle activation pairs of ($a_{TA} = 0.06$, $a_{CT} = 0.44$) and ($a_{TA} = 0.94$, $a_{CT} = 0.15$), respectively. The glottal flows (i.e. the sound source for vowel-like sounds) produced by the three mass model with these two combinations of muscle activations are shown in Figure 3b. Note that the length of each glottal cycle is identical in both cases. However, the activation pair at point B produces a flow pulse that is more heavily skewed in the rightward direction and has a shorter open phase than than point A. The resulting sound produced at point B will perhaps have a more “pressed” quality than the sound produced at point A.

The reason for the difference is that varying the combination of CT and TA activations leads to different phase and am-

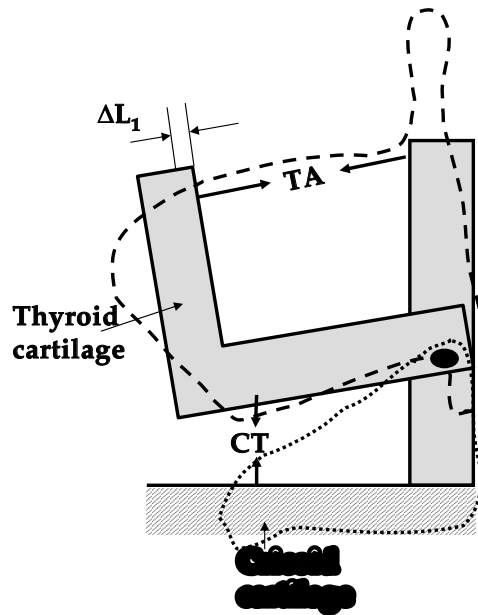


Figure 2: Idealized view of the rotation of the thyroid cartilage relative to the cricoid cartilage in a sagittal (side) view. The directions of the forces produced by the CT and TA muscles are indicated by arrows and the change in length of the vocal folds (strain) is also shown.

plitude relationships between the lower and upper portions of the vocal folds. This creates the possibility of wide variety of vibrational patterns, even though the fundamental frequency of vibration remains the same.

3. Vocal tract model

As mentioned in the Introduction, Laver[1][2] proposed a descriptive system for voice qualities based on the notion that people speak under the influence of imposed “settings” on their speech organs. Central to this idea is that a *neutral* reference configuration exists for the vocal tract shape. The neutral configuration doesn’t necessarily imply the “normally used” configuration, but rather one to which all subsequent settings can be referenced. An interpretation of this neutral configuration for the vocal tract would be one that produces an acoustically neutral sound; i.e. equally spaced formant frequencies.

Thus, a “setting” may be considered to be a slight deviation of the vocal tract shape from the purely neutral configuration, and becomes the starting point or “background” on which subsequent articulations are superimposed. Many of the settings proposed in [1] and [2] are divided into the categories of *longitudinal* and *latitudinal*. Longitudinal settings are those in which the deviations from the neutral reference lengthen or shorten some portion of the vocal tract such as larynx lowering/raising or lip protrusion/retraction. Latitudinal settings concern expansions or contractions of cross-sectional area along the vocal tract. An example of a latitudinal setting might be a person who speaks with their tongue predominantly fronted and raised toward the hard palate such that the oral cavity is constricted. Other possible settings involve the velopharynx for nasal or denasal qualities and the phonatory system for source-based qualities such as breathiness, roughness, pressed, etc.

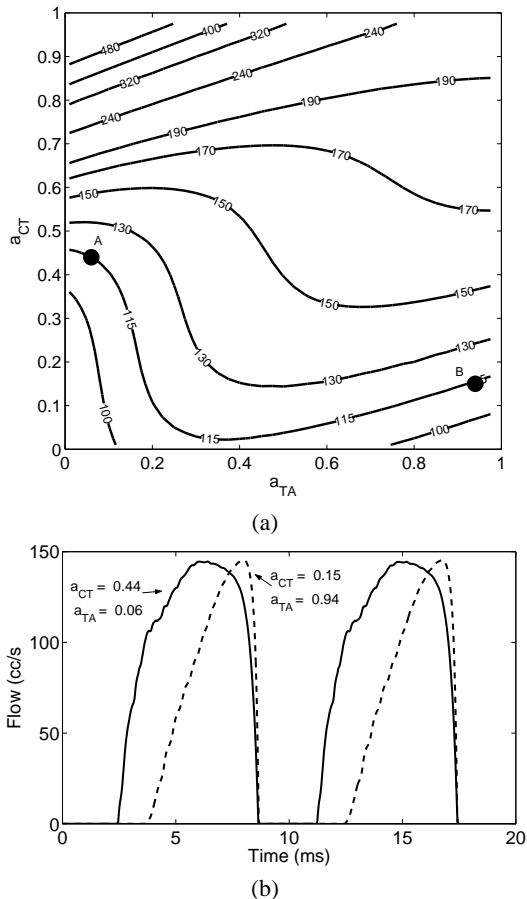


Figure 3: (a) Muscle activation plot showing F_0 contour lines. Each contour represent muscle activation pairs that will produce the same F_0 with the three-mass model. (b) Glottal flow pulses produced by muscle activation pairs corresponding to the two dots in (a). The solid line represents the flow generated at point “A” and the dashed line represents the flow generated at point “B”.

For purposes of modeling, the vocal tract shape is typically represented as a collection of cross-sectional areas extending from the larynx to the lips and is called an *area function*. An example of an area function for a male vowel /a/ is shown in Figure 4; the dashed lines indicate the approximate points of division between various the epilarynx, pharynx, and oral cavities as well as connection points of the piriform side branch and nasal tract. This particular area function was derived from an imaging study [10] in which a speaker produced a series of ten different vowels of American English.

While individual area functions are useful for studying vocal tract acoustics, it is often desirable to have a model from which many vocal tract shapes can be generated (e.g. see [11][12][13]). From analysis of the area functions resulting from the previously cited imaging study[10], a parametric model of the vocal tract shape[14] has been developed that seems to lend itself to modeling the types of settings described above [15]. The following sections will briefly describe this model and then discuss examples of various settings that alter voice quality.

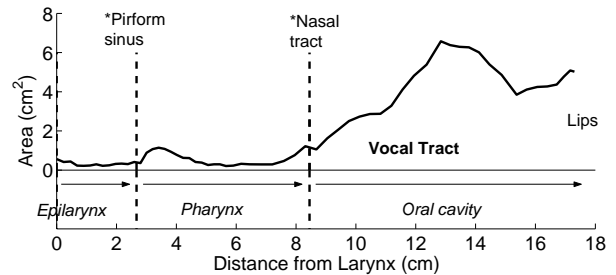


Figure 4: Area function for a male /a/ vowel. The dashed vertical lines show approximate points of division between various sections of the vocal tract as well as connection points of side branches.

3.1. Parametric model of the area function

The model contains three basic parts, each of which was derived from a principal components analysis (PCA) of a ten vowel set of area functions[14]. The first part is a hypothetical neutral state of the vocal tract $\Omega_r(x)$ in the form of a collection of diameters extending from the larynx to the lips (note: eventually the diameters are converted to the more familiar area function of the vocal tract) and x represents the distance from the larynx. The subscript r in this case denotes that this function will serve as the neutral reference configuration.

The second and third parts are two separately controlled (scaled) deformation patterns capable of producing specific vowel shapes when combined with the neutral function. These deformation patterns are denoted as $\phi_1(x)$ and $\phi_2(x)$ and henceforth will be referred to as *modes*. Each mode may be scaled by a coefficient, q_1 for $\phi_1(x)$ and q_2 for $\phi_2(x)$. Thus, a vocal tract *area function* $V(x)$ can be specified as

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

where the squaring operation and scaling factor of $\pi/4$ converts the diameters to areas.

These three functions are shown graphically in Figures 5a–c. When $q_1 = q_2 = 0$, the area function ($\frac{\pi}{4} \Omega_r^2(x)$) produces nearly evenly spaced formant frequencies[14] and thus, is considered to be a “neutral” vocal tract configuration. If the coefficients are increased continuously from their most negative to most positive values (based on the set of vowel area functions from which this model was derived) a continuum of tract shapes is produced that is roughly from /i/ to /a/ for the first mode and /ae/ to /o/ for the second mode. In combination, the coefficient pairs generate a full vowel space as is shown in Figure 6. An 80×80 grid of coefficients is shown in Figure 6a with q_1 on the x-axis and q_2 on the y-axis. For each pair in the grid, an area function was generated with Eqn. 1 and its formant frequencies were calculated, thus allowing a corresponding F1-F2 grid to be formed as is shown in Figure 6b. The point at which $q_1 = q_2 = 0$ is indicated with black dot in the coefficient plane and a white dot in the F1-F2 plane.

Also shown in this figure is a trajectory that produces the vowel utterance /iaui/. Note that the trajectory in the coefficient plane consists of three linear “legs” that progress over time (the time course spans about 1.25 s). By modifying Eqn. (1) to be time dependent [15], a time-varying area function (see Figure 7) can be produced with,

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

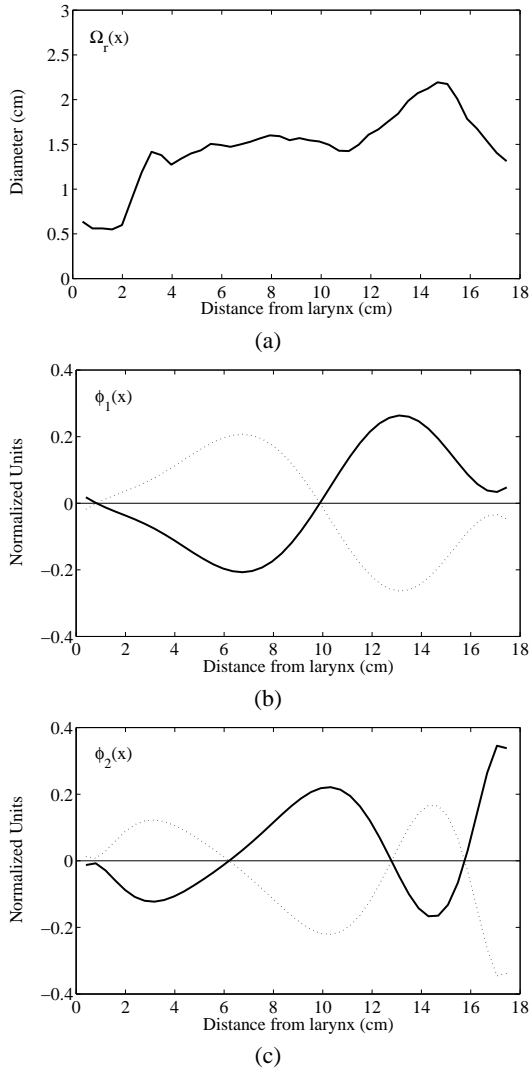


Figure 5: Representation of an adult male vocal tract with “modes” derived from principal components analysis, (a) first mode shape ϕ_1 (dotted line is the reflection of the solid line), (b) second mode shape ϕ_2 , and (c) mean diameter function $\Omega(x)$.

where the only change is that q_1 and q_2 are now functions of time. Calculating the formant frequencies for each area function along this trajectory, the legs in the F1-F2 plane demonstrate the curvature that can result from the nonlinear relation between vocal tract shape and acoustic properties.

3.2. Voice quality modifications based on “settings”

The separation of the vocal tract area function into three parts allows for vocal tract variations related to voice quality to be exclusively imposed on the neutral configuration $\Omega_r(x)$ while leaving the deformation patterns ($\phi_1(x)$ & $\phi_2(x)$) unchanged. Thus, the “neutral” state of the vocal tract can be considered to contain particular “settings” that create the acoustic background for a speaker while invariant gestures are superimposed on it with the two deformation patterns to create the speech signal.

As a demonstration, some work reported previously [15] concerning settings will be reiterated. The first example setting that will be imposed on the neutral configuration is one

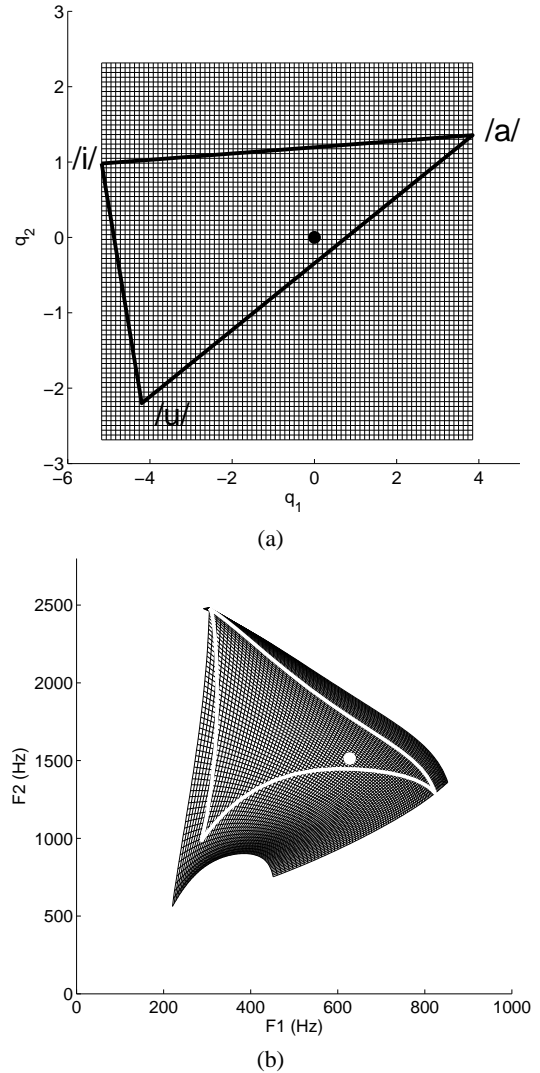


Figure 6: Mapping of 6400 (q_1, q_2) coefficient pairs to the same number of F1-F2 formant pairs. (a) coefficient grid, (b) corresponding F1-F2 grid. Also shown in each figure is a trajectory that courses through the vowels /aui/.

in which the pharyngeal space is compressed and the oral cavity is somewhat expanded[15]. This roughly corresponds to a “pharyngealized” voice [1]. Figure 8a shows the original neutral diameter function $\Omega_r(x)$ along with a modified version that will be referred to as $\Omega_{pha}(x)$. Now, Eqn. 2 can be modified to use $\Omega_{pha}(x)$ instead of $\Omega_r(x)$,

$$V(x, t) = \frac{\pi}{4} [\Omega_{pha}(x) + q_1(t)\phi_1(x) + q_2(t)\phi_2(x)]^2 \quad (3)$$

and the /aui/ trajectory recomputed in the F1-F2 plane. Note, that the coefficient trajectory used here is identical to that shown previously in Figure 6a.

Both the previous F1-F2 trajectory (based on the neutral reference configuration) and the new “pharyngealized” version are shown in Figure 8b. In addition, the F1-F2 pair corresponding to the $q_1 = q_2 = 0$ point is also indicated by the black dots; the arrow points from the original neutral configuration to the new version. It is observed that this new setting has shifted the

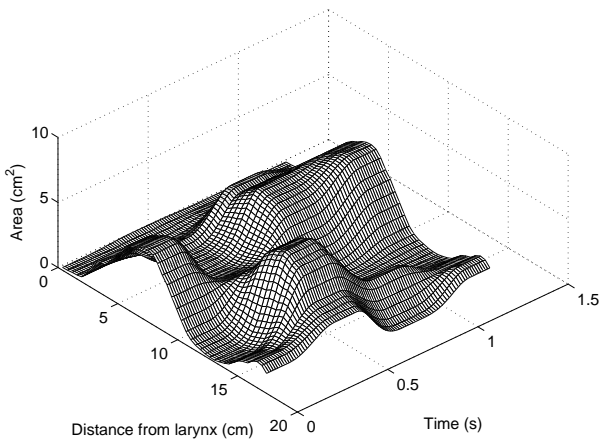


Figure 7: Time-varying area function generated with the coefficient trajectory in Fig. 6a and Eqn. 2.

first formant frequencies in a generally upward direction while the second formants are shifted downward; this shift is also indicated by the $q_1 = q_2 = 0$ points. However, the formants are not simply shifted in a linear translation, as the shape of F1-F2 trajectory has also been distorted relative to that produced with the neutral reference configuration. Thus, a setting not only shifts the “operating point” (i.e. a translation within the vowel space) but also changes the dynamic properties of a formant trajectories, even though the articulatory dynamics (coefficient trajectories) remained unchanged.

A second example is derived from [16] where vocal tract area functions were measured for speakers producing vowels under the voice quality conditions of normal, yawny, and twangy. The twangy quality for the speakers in this study was found to be largely produced by setting the lip end of the vocal tract to be predominantly more open than in the neutral reference. The original neutral diameter function $\Omega_r(x)$ along with a modified version $\Omega_{twg}(x)$ is shown in Figure 9a. As before, Eqn. 2 can be modified to use $\Omega_{twg}(x)$ instead of $\Omega_r(x)$,

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

and again the /iaui/ coefficient trajectory generates the time-varying area function, but with the “twangy” setting imposed. The resulting F1-F2 trajectory is shown in Figure 9b along with the original; also shown are the points at which $q_1 = q_2 = 0$. For most points along the trajectory, both F1 and F2 have increased relative to the neutral reference. However, at the left side of the trajectory F1 is observed to slightly decrease. Also significant for this setting is that while the /u/ is in the appropriate position along the trajectory, it has an F2 value that is inappropriately high. For a real speaker this setting would likely be released during the production of the /u/ vowel, only to be reset for production of vowels where a high F2 is acceptable.

4. Possible voice quality “settings” from audio recordings

The database provided for this conference contains many variations of voice quality, all spoken with the same sentence by one person (“She has left for a great party today.”). In an attempt to

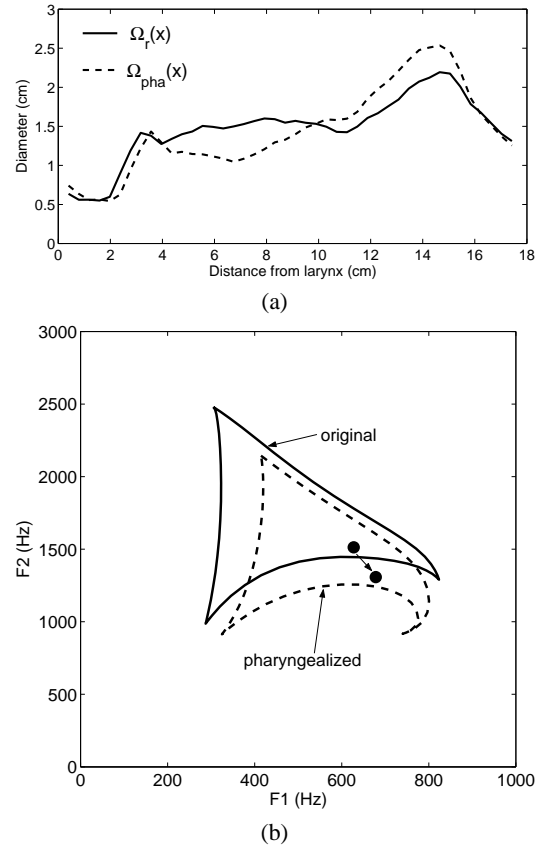
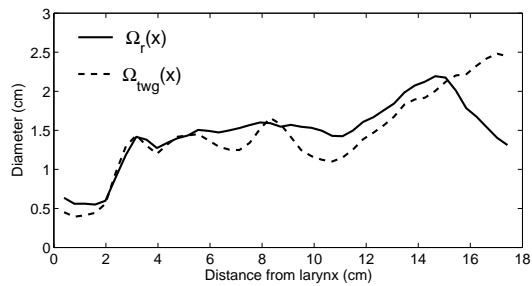


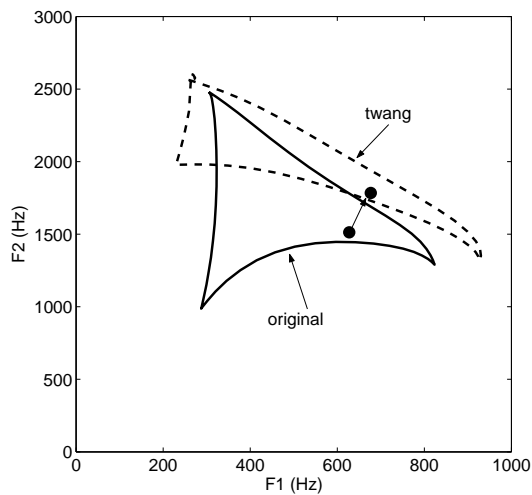
Figure 8: Simulation of a “pharyngealized” setting. (a) Modification of the neutral reference configuration such that the pharynx becomes slightly constricted and the oral cavity slightly expanded. (b) Formant trajectory for /iaui/ resulting from the pharyngealized setting. The solid dots represent the $q_1 = q_2 = 0$ points for both settings.

connect this data with the voice quality modeling in the previous section, three samples were chosen for analysis. They were BrianNormal5.wav, a normal speaking voice; BrianClos1.wav, a somewhat closed vocal tract, possibly with clenched teeth; and BrianSmil5.wav, where the lips were kept in a “smiling” position to whatever degree was possible. For each sample, all unvoiced portions were removed and then a formant analysis was performed on all remaining voiced sections using Burg’s method as incorporated into the Praat speech analysis program [17]. The resulting formant trajectories were subsequently filtered to remove spurious peaks and the mean F1 and F2 were then calculated as a rough measure of the center of gravity of the trajectory (in the F1-F2 plane). The goal was to use the “normal” sample as the reference setting and then compare the other two samples relative to the normal.

Figure 10a shows the mean F1, F2 points and the outer boundary of each F1-F2 trajectory for both the BrianNormal5 and BrianClos1 samples. Overall, the BrianClos1 trajectory appears to have shifted F1 upward in frequency and possibly F2 downward; this is also indicated by the direction of shift in the center of gravity point. However, in the upper portion of the trajectory the shift in F2 appears to be upward, but a check of the formant data showed that this portion of the trajectory resulted from just short period of time at the beginning of the utterance,



(a)



(b)

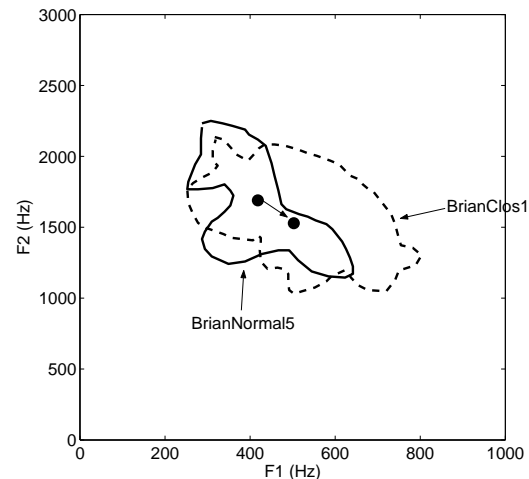
Figure 9: Simulation of a “twangy” setting. (a) Modification of the neutral reference configuration such that the lip end of the vocal tract becomes expanded. (b) Formant trajectory for /iaui/ resulting from the twangy setting. The solid dots represent the $q_1 = q_2 = 0$ points for both settings.

hence the center of gravity point is an accurate indicator of the shift direction. Based on the overall formant shifts, this voice quality may share some similarities with the simulation of the “pharyngealized” voice quality in the previous section. In that case, F1 was shifted upward and F2 downward. This suggests that with clenched teeth, perhaps there is some tendency to retract the tongue and consequently compress the pharynx.

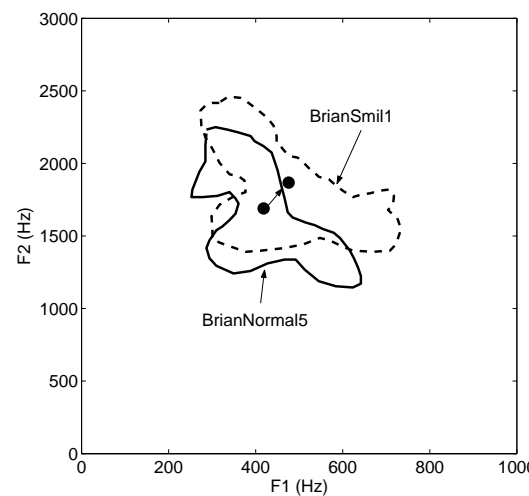
A similar analysis of BrianSmil1 is given in Figure 10b. In this case, the shift of the center of gravity point is upward in frequency for both F1 and F2 and the outline of the trajectory boundary would indicate this trend as well. This is the expected direction of formant shifts for a setting that maintains a predominantly wide lip termination. Also, based on these formant shifts, this quality is comparable to the simulation of the twang quality in the previous section. Again, this not surprising since the twang quality consisted primarily of an expansion at the lips.

5. Conclusion

Presented in this paper have been examples of physically-based models that were configured to simulate various aspects of voice quality. While still early in their development, such models seem to offer a potentially useful tool for investigating qualities of the voice and may eventually help to improve systems for recognition and synthesis of speech.



(a)



(b)

Figure 10: Analysis audio recordings in the conference database. (a) Comparison of F1-F2 formant trajectories of BrianNormal5 with BrianClos1. (b) Comparison of F1-F2 formant trajectories of BrianNormal5 with BrianSmil1.

6. Acknowledgements

This work was supported by NIH R01-DC04789.

7. References

- [1] Laver, J., The Phonetic Description of Voice Quality, Cambridge University Press, 1980.
- [2] Laver, J., Principles of Phonetics, Cambridge University Press, 1994.
- [3] Hirano, M., Morphological structure of the vocal cord as a vibrator and its variations. Folia Phoniat., 26: 89–94, 1974.
- [4] Ishizaka, K., and Flanagan, J.L., Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell Sys. Tech. J., 512:1233–1268, 1972.
- [5] Alipour, F., Berry, D.A., and Titze, I.R., A finite-element model of vocal fold vibration. J. Acoust. Soc. Am., 108(6):3003–3012, 2000.

- [6] Titze, I.R., and Talkin, D.T., A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *J. Acoust. Soc. Am.*, 66:60–74, 1979.
- [7] Titze, I.R., The human vocal cords: A mathematical model, part I. *Phonetica*, 28:129–170, 1973.
- [8] Story, B.H., and Titze, I.R., Voice simulation with a body-cover model of the vocal folds, *J. Acoust. Soc. Amer.*, 97(2):1249-1260, 1995.
- [9] Titze, I.R. and Story, B.H., Rules for controlling low-dimensional vocal fold models with muscle activation, *J. Acoust. Soc. Amer.*, 112(3):1064-1076.
- [10] Story, B.H., Titze, I.R., & Hoffman, E.A., Vocal tract area functions from magnetic resonance imaging, *Journal of the Acoustical Society of America*, 100(1):537-554, 1996
- [11] Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W., Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting-sorting technique, *J. Acoust. Soc. Amer.*, 63(5):1535-1555, 1978.
- [12] Stevens, K. N., and House, A. S., Development of a quantitative description of vowel articulation, *J. Acoust. Soc. Amer.*, 27(3): 484-493, 1955.
- [13] Mermelstein, P., Articulatory model for the study of speech production, *J. Acoust. Soc. Amer.*, 53(4):1070-1082, 1973.
- [14] Story, B.H., & Titze, I.R., Parameterization of vocal tract area functions by empirical orthogonal modes, *J. Phonetics*, 26(3): 223-260, 1998.
- [15] Story, B.H., & Titze, I.R., A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function, *J. Phonetics*, 30:485:509, 2002.
- [16] Story, B.H., & Titze, I.R., The relationship of vocal tract shape to three voice qualities, *J. Acoust. Soc. Amer.*, 109:1651–1667.
- [17] Boersma, P., Praat Speech Analysis Software, www.praat.org, 2003.