

A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function

Brad H. Story*

Speech and Hearing Sciences, University of Arizona, U.S.A. and National Center for Voice and Speech, WJ Gould Voice Research Center, Denver Center for the Performing Arts, U.S.A.

Ingo R. Titze

National Center for Voice and Speech, WJ Gould Voice Research Center, Denver Center for the Performing Arts, U.S.A. and Department of Speech Pathology and Audiology, University of Iowa, U.S.A.

Received 9th November 2001, and accepted 3rd January 2002

The idea is pursued that voice quality can be partially represented by the underlying shape of a speaker's neutral vocal tract. Using an area function model, which allows direct access to the neutral tract shape, four separate modifications were made to one male speaker's vocal tract. The modifications involve imposing constrictive or expansive effects on the pharyngeal and oral portions of the neutral area function as well as on lip aperture and the epi-laryngeal tube. A single word utterance was first synthesized by superimposing deformation patterns appropriate for the word onto the original neutral tract shape (area function). Then, four additional samples of the word were synthesized using different modified neutral area function each time. The modifications were assessed by comparing $F1$ – $F2$ formant trajectories of the original utterance with those of the modifications. The formant frequencies were observed to shift within the $F1$ – $F2$ plane in directions predictable from simple tube acoustics. However, the modified voice qualities did not preserve the shape of the original $F1$ – $F2$ trajectory. In other words, the modifications did not create a simple linear transformation of formant frequencies even though the "articulatory dynamics" (deformation patterns of the area function) were identical in all cases. These somewhat artificial vocal tract modifications were also compared with formant frequencies extracted from recordings of a speaker attempting to produce the same types of modifications. In general, the speaker's formant trajectories showed some similarities to the synthesized versions. However, the speaker also seemed to grade the "level" of the voice quality that was exerted on the utterance depending on whether the demands of the voice quality were in competition with the linguistic

*E-mail: bstory@u.arizona.edu

demands of a given phonetic segment. Finally, to demonstrate this type of voice quality modification in a broader context, the same procedures were applied to sentence-level speech and results were again shown as $F1$ – $F2$ formant trajectories.

© 2002 Elsevier Science Ltd. All rights reserved.

1. Introduction

The highly variable nature of speech production and the resulting acoustic signal is well known. Much of this acoustic variability falls within the domain of *voice quality*. Laver (1980, p. 1) defined voice quality in a broad sense as “the characteristic auditory coloring of an individual speaker’s voice”, which is a different definition than the more common use of the term to describe a quality generated solely by laryngeal activity. This “auditory coloring” arises from many possible sources, the most obvious of which are organic considerations such as vocal tract length and shape, jaw and tongue size, as well as vocal fold length, mass, and viscosity. To a large degree, such organic sources explain the acoustic differences in the speech of males, females, and children. For example, relative to male speakers, females (and children) have shorter vocal tract lengths (e.g., Goldstein, 1980; Fitch & Giedd, 1999), as well as shorter vocal fold lengths (Titze, 1989), leading to typically higher formant frequencies (e.g., Peterson & Barney, 1952) and higher fundamental phonation frequencies (e.g., Holmberg, Hillman & Perkell, 1988; Titze, 1989), respectively. Furthermore, many speaker-unique qualities arise from the idiosyncratic structure of an individual’s vocal tract and vocal folds.

A second source of variability comes from the way in which the speech production system is habitually used. A nasal voice, for example, is produced by a speaker’s tendency toward keeping the velum in a slightly lowered position, allowing sound to propagate in the nasal passages as well as the main vocal tract. Another example is the “breathy” voice produced by hypo-adduction of the vocal folds or conversely, a “pressed voice” when the vocal folds are hyper-adducted (Titze, 1994). In any of these examples, a learned habitual muscle pattern is likely the cause of the various types of voice quality and may have either pathological or nonpathological origins (Laver, 1980). Influences such as regional dialect, familial tendencies, social and cultural forces, as well as idiosyncratic tendencies may all have a hand in forming the spoken acoustic speech signal (Ladefoged & Broadbent, 1957).

The focus of this study is on the second type of variability. Specifically of interest is the subtle shaping of the vocal tract (at least subtle in comparison with the shape changes created by speech articulation) that speakers may use, either habitually or intentionally, to create a particular voice quality. To describe this type of vocal tract shaping, Laver (1980) proposed a system in which long-term “settings” of the vocal tract, in addition to the biological endowment of the speaker, bias the resulting formant frequency patterns toward a particular type of global timbre. In this system, he defined two broad categories of vocal tract settings: longitudinal and latitudinal. Longitudinal settings describe the state of the long axis of the vocal tract such as larynx height and protrusion/retraction of the lips. The latitudinal settings are “tendencies to maintain a particular constrictive (or

expansive) effect” within some region located along the length of the vocal tract (Laver, 1980, p. 35). Linguistically relevant gestures are then superimposed on these settings resulting in intelligible speech produced against a unique, but neutral, “acoustic background”.

Similarly, in an attempt to explain invariance in speech signals, Traunmüller (1994) has proposed a “modulation theory” in which speech signals are considered to be the result of allowing conventional (common) articulatory gestures to modulate a “carrier” signal. This carrier signal is thought to be phonetically neutral but descriptive of the “personal quality” of the speaker. Thus, embedded within the carrier would be contributions of the biological structure of the vocal tract as well as any vocal tract shaping patterns resulting from habitual muscle tensions. During the production of speech this neutral state of a speaker’s articulatory organs is modulated by linguistically meaningful gestures resulting in an acoustically unique, but linguistically invariant message.

Changes in the carrier signal or the acoustic background seem to occur when speakers transform their speech to satisfy the demands of some immediate circumstance. For example, the acoustic signal resulting from informal conversational speech, “clear” speech, and baby talk have all been shown to differ in terms of formant frequency locations, likely representing an adaptation to the communicative demands of each situation (Lindblom, Brownlee, Davis & Moon, 1992). Similarly, Traunmüller & Eriksson (2000) have observed that several acoustic characteristics, including fundamental frequency and formant frequencies, change significantly when the physical distance over which speech communication takes place is drastically increased. Another obvious example of an active and intentional change in the acoustic background is when actors alter their voice quality to portray a particular character or personality.

Both the ideas of Laver (1980) and Traunmüller (1994) suggest a speech production system that maintains some type of linguistically neutral state containing voice-quality information but also allows for a superposition of linguistically relevant gestures. Thus, the speech signal could be conceived as a convolution of a particular voice-quality and linguistic gestures; or as Traunmüller (1994) suggests, a “modulation” of the two. It is the goal of this paper to present a highly preliminary study of combining various voice-quality characteristics with linguistic gestures in the framework of a recently developed model of the vocal tract area function for vowels. This model is based on recent vocal tract imaging studies (Story, Titze & Hoffman, 1996) and subsequent analysis of the resulting area functions (Story & Titze, 1998).

The model contains three basic parts. The first is a hypothetical neutral state of the vocal tract in the form of an area function. The second and third parts are two separately controlled (scaled) deformation patterns that represent gestural components capable of producing specific vowel shapes when combined with the neutral area function. This 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 area function while leaving the gestural components (i.e., deformation patterns) 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. Operationally, this may be considered analogous to the suggestion by Traunmüller (1994) of a neutral carrier signal modulated by linguistic gestures.

The specific aims of the paper are: (1) to review an area function model based on physiologic imaging data, (2) to synthesize a replica of a single-word voiced utterance (based on an acoustic recording) with the area function model, (3) to transform the synthesized utterance in (1) into the same utterance but with four new voice qualities, and (4) to repeat aims (1) and (2) with sentence-level speech.

2. Area function model

The development of the area function model has been described in Story, Titze & Long (1998). However, to clarify the presentation of the current study, the model is summarized here. A set of vocal tract area functions corresponding to 10 vowels was acquired using MRI (Story *et al.*, 1996) for an adult male speaker. First, each area function was normalized to a length of 17.46 cm and then each area section (there are 44 sections in each area function) was converted into a diameter of a circle having the same area. Thus, area functions, represented by $V(x)$ were converted into “diameter” functions with $d(x) = \sqrt{4V(x)/\pi}$ where x is the distance from the glottis. This transformation was used because it had the effect of expanding small areas and compressing the large ones, leading to a more accurate analysis and subsequent reconstruction of area functions as discussed below.

Next, a principal components analysis (PCA) was used to decompose the collection of diameter functions into a set of orthogonal basis functions and a mean diameter function. Because of their orthogonality and the fact that they can reproduce many possible geometric states of the vocal tract tube, the basis functions have been termed “geometric eigenmodes” or simply “modes” (Story & Titze, 1998). The first four most significant modes individually accounted for variances within the set of diameter functions of 66, 22, 6.1, and 2.2%, respectively. Together, the first two modes accounted for 88% of the total variance suggesting that reasonably accurate reconstruction of the original vowel area functions can be realized by summing the mean diameter functions with a weighted combination of just the first two modes and with the final step of converting the diameter functions into an area function (see Story & Titze (1998) for discussion of the reconstruction accuracy).

Figs. 1(a) and (b) show the first and second mode shapes, respectively (these have been smoothed with an eighth-order polynomial) and the mean diameter function is shown in Fig. 1(c); each of these is shown as a function of distance from the glottis. The equation for reconstructing any of the 10 vowel area functions is

$$V(x) = \frac{\pi}{4} [\Omega(x) + c_1 \phi_1(x) + c_2 \phi_2(x)]^2 \quad (1)$$

where $\Omega(x)$ is the mean diameter function, $\phi_1(x)$ and $\phi_2(x)$ are the two modes, and x is the distance from the glottis. c_1 and c_2 are the amplitude coefficients (weights) for the first and second modes, respectively, which will reconstruct the original 10 vowel area functions; the numerical values for each vowel are shown in Table I. Note that the first mode creates a front–back asymmetry such that a negative value of c_1 (e.g., the dashed line in Fig. 1(a)) will constrict the front half of the area function and expand the back; a positive value of c_1 will do just the opposite. The second mode has a large influence on the mid-tract region of the area function and at the lips; a

TABLE I. Modal coefficients corresponding to 10 vowels derived from a principal components analysis of 10 vocal tract area functions

Coeff.	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
c_1	-5.02	-2.45	-1.00	0.14	2.61	3.91	3.49	1.76	0.16	-3.59
c_2	1.06	1.07	1.35	2.27	0.16	1.41	-0.49	-1.89	-2.58	-2.37

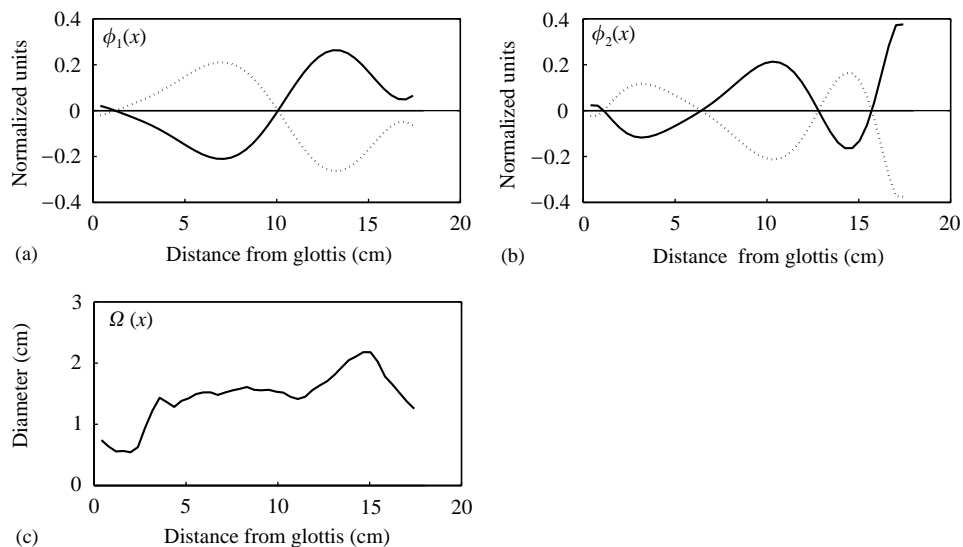


Figure 1. Modal representation of an adult male vocal tract: (a) first mode shape ϕ_1 (dashed line is the reflection of the solid line as for a negative weighting coefficient), (b) second mode shape ϕ_2 , and (c) mean diameter function $\Omega(x)$.

negatively valued c_2 will constrict the mid-tract and the lip opening while a positive c_2 would have the opposite effect.

Initially, the purpose of carrying out the PCA was simply to compress each of the original area functions from a 44 area element representation (area function was specified as 44 cross-sectional areas) into a three-element, two-parameter representation; the three elements being $\Omega(x)$, ϕ_1 , and ϕ_2 , and the two parameters c_1 and c_2 . However, it was found that both the mean area function ($(\pi/4)\Omega^2(x)$) and the two orthogonal modes ϕ_1 and ϕ_2 correspond to systematic, speech-relevant, acoustic characteristics.

In particular, the calculated formant frequencies for the mean area function ($(\pi/4)\Omega^2(x)$) are similar to those of a uniform tube, while its actual shape is not at all uniform. The similarity is apparent in Fig. 2, where frequency response functions for both the mean area function and the uniform tube are shown. Thus, the mean area function appeared to represent a physiologically-based approximation to a neutral vocal tract shape. Henceforth, $\Omega(x)$ will be referred to as the *neutral* diameter function ($(\pi/4)\Omega^2(x)$ is then the *neutral* area function).

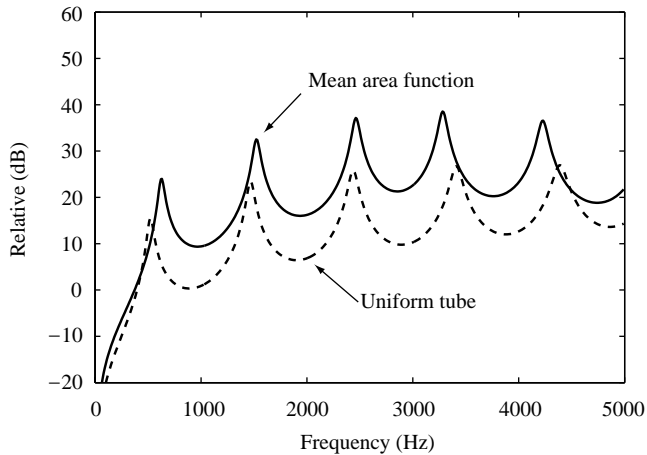


Figure 2. Frequency response functions for the mean area function (solid line) and for a uniform tube of length 17.5 cm and a cross-sectional area of 2 cm^2 (dashed line). Note the similar spacing of the formant frequency peaks.

Furthermore, when either modal coefficient c_1 or c_2 was varied along a continuum from its most negative value to its most positive value (e.g., $c_1 = -5.02$ for vowel /i/ and $c_1 = +3.9$ for /a/) while the other coefficient was held at 0.0, the resulting area functions correspond to *monotonically* changing formant frequencies along this same continuum. Fig. 3(a) shows area functions reconstructed with Equation (1) for the case $(c_1, c_2) = (-5.02, 0.0)$ (solid line) and also $(c_1, c_2) = (+3.90, 0.0)$ (dashed line). These are the extreme cases of the possible area function deformations caused by the first mode ϕ_1 and approximates an [i]-like shape and an [a]-like shape, respectively; the coefficient values are obtained from Table I. The first three formant frequencies ($F1$, $F2$, and $F3$) corresponding to area functions produced along a continuum of c_1 values (while $c_2 = 0$) on the interval $[-5.02, +3.9]$ is shown in Fig. 3(b). At the far left side of the plot, $F1$ and $F2$ are located far apart in frequency as for a typical [i] vowel while at the far right side $F1$ increases and $F2$ decreases to approximate an [a] vowel; $F3$ is nearly unaffected by the variation of c_1 .

Figs. 3(c) and (d) present similar results but for the case, where $c_1 = 0.0$ and c_2 ranges from -2.58 to $+2.27$. The area functions in Fig. 3(c) show that the second mode ϕ_2 creates a continuum from an [o]-like shape (solid line) to something more [æ]-like (dashed line). The $F1$ and $F2$ formant frequencies in Fig. 3(d) are close together at the far left side corresponding to the [o]-like vowel. As c_2 is increased, $F1$ and $F2$ both monotonically increase in frequency but at different rates (i.e., the distance between them is not maintained); $F3$ shows only a slight decrease of about 300 Hz over the course of the c_2 continuum.

The monotonic acoustic characteristics noted above suggest that weighted combinations of the two modes can precisely control the location of $F1$ and $F2$. For $F2$, the two modes exert the opposite effect on its frequency, metaphorically acting like an “antagonist/agonist” pair, while the frequency of $F1$ is similarly influenced by both modes. From Table I it is seen that realistic area functions for 10 vowels are generated by weighted combinations of both the modes. However, if the numerical values of the two modal coefficients are simultaneously varied along their

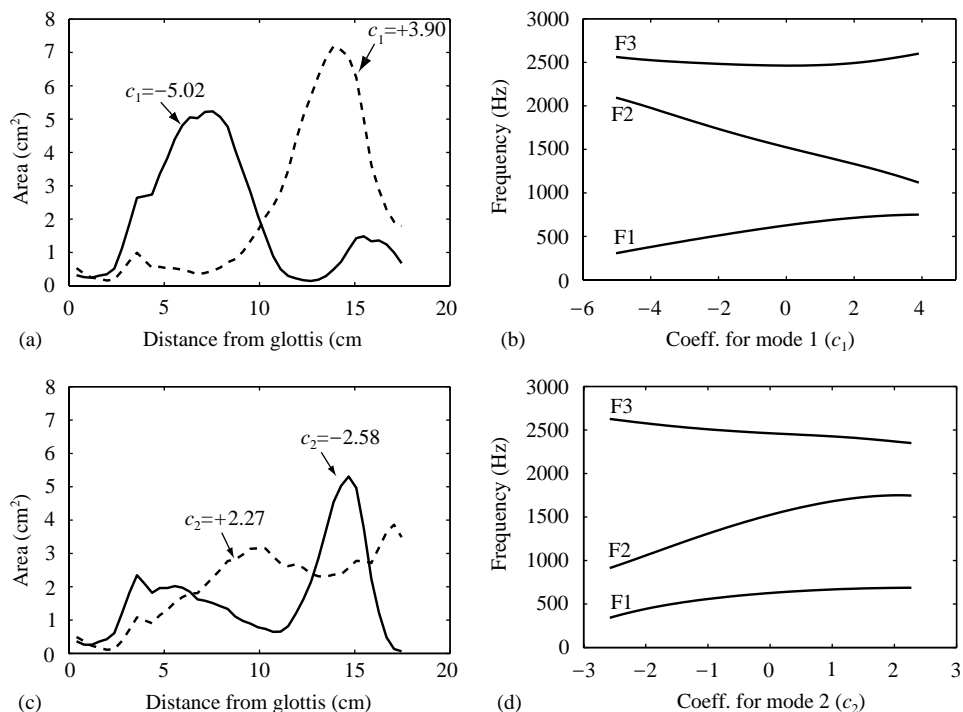


Figure 3. Demonstration of the spatial and acoustic effects of superimposing the modes (ϕ_1 and ϕ_2) on the neutral vocal tract area function: (a) case where $(c_1, c_2) = (-5.02, 0.0)$ (solid) and $(c_1, c_2) = (+3.9, 0.0)$ (dashed), (b) $F1$, $F2$, and $F3$ formant frequencies resulting from a continuum of c_1 values ranging from -5.02 (left side) to $+3.9$ (right side) while $c_2 = 0$; (c) case where $(c_1, c_2) = (0.0, -2.58)$ (solid) and $(c_1, c_2) = (0.0, +2.27)$ (dashed), and (d) $F1$, $F2$, and $F3$ formant frequencies resulting from a continuum of c_2 values ranging from -2.58 (left side) to $+2.27$ (right side) while $c_1 = 0$.

respective continua shown in Fig. 3, thousands of new (hypothetical) vowel shapes can be created that did not exist in the original 10-vowel set. Fig. 4(a) shows an 80×80 mesh of 6400 coefficient pairs (c_1, c_2) based on the coefficient values given in Table I. The solid dots and dark connecting line indicate the coefficient pairs that reconstruct the original 10 vowels. Note that the grid boundaries are based on the maximum and minimum values of c_1 and c_2 . For each coefficient pair in the grid, an area function was created with Equation (1), and then $F1$ and $F2$ frequencies corresponding to each area function were determined by finding the formant peaks in the computed frequency response function (computed with the transmission line approach of Sondhi & Schroeter, 1987). $F1$ and $F2$ are plotted as pairs in the formant mesh shown in Fig. 4(b); the white connected dots represent the formant frequencies of the original 10 vowel area functions.

Thus, a mapping between the first two formant frequencies and the two modal coefficients is generated; this is effectively a mapping between formants and area functions that generate those formants. With the exception of the upper border of this mesh, the mapping is one-to-one; that is, one formant pair in Fig. 4(b)

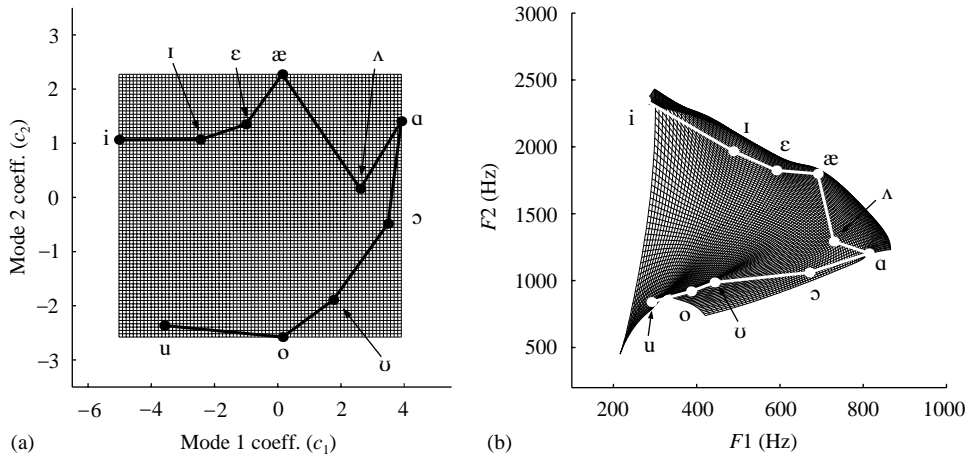


Figure 4. Mapping of (c_1, c_2) coefficient pairs to $F1$ – $F2$ formant pairs, where the solid circles represent coefficients and formant frequencies for the original 10 vowels: (a) 80×80 coefficient grid, and (b) corresponding $F1$ – $F2$ grid.

corresponds to one coefficient pair in Fig. 4(a). It should be noted that this particular version of the mapping made indirect use of two additional modes (modes 3 and 4) to enhance the detail of the reconstructed area functions (this is discussed in Story & Titze, 1998). Theoretically, the relationship between the area function and formant frequencies is many-to-one; that is, many area functions can produce the same formant frequency pattern (Schroeder, 1967; Mermelstein, 1966). Thus, two formants is not sufficient to determine an unambiguous area function. However, the mapping between $F1$ – $F2$ pairs and modal coefficient pairs (c_1, c_2) is physiologically constrained by the fact that the modes and neutral area function were derived from measured vocal tract shapes of a particular speaker. Thus, an $F1$ – $F2$ pair that falls within the mesh shown in Fig. 4(b) corresponds to a unique coefficient pair in Fig. 4(a) and hence to a unique area function. This, of course, does not preclude the possibility that the same $F1$ – $F2$ pair could be generated by an entirely different area function which does not conform to this particular mapping.

Nonetheless, the nearly one-to-one relationship between formant frequencies ($F1$ and $F2$) and articulation (i.e., modal coefficients) offers the possibility of mapping time-varying formant frequencies extracted from recorded natural speech to time-varying modal coefficients. Then, a time-varying area function can be generated by

$$V(x, t) = \frac{\pi}{4} [\Omega(x) + c_1(t)\phi_1(x) + c_2(t)\phi_2(x)]^2 \quad (2)$$

where the mean area function $\Omega(x)$ and the two modes, $\phi_1(x)$ and $\phi_2(x)$, remain unchanged with time but the modal coefficients are time-varying.

2.1. Synthesis of a single-word utterance

In this section, the mapping between $F1$ – $F2$ pairs and modal coefficients is applied to synthesis of a single word. The word ‘‘Iowa’’ was recorded directly onto a hard

disk in a PC via a high-quality sound card (sampling frequency=44 100 Hz). The word was spoken by the male subject from whom the original area functions were obtained. Formant frequencies over the time course of the word were determined by LPC analysis (autocorrelation method) and a peak-picking technique; the analysis was performed over 25 ms windows with a 12.5 ms overlap. The first two formant frequencies are shown as functions of time in Fig. 5(a) (the formants were effectively sampled at 12.5 ms intervals due to the overlap). In Fig. 5(b) the same information is displayed, but as a trajectory in the $F1$ – $F2$ plane superimposed on the formant mesh from Fig. 4. Note that the trajectory remains within the outer boundaries of the formant grid for its entire duration assuring that a corresponding modal coefficient pair exists for each sampled formant pair. Using a simple database lookup technique each formant pair in the trajectory was matched to a modal coefficient pair, generating the coefficient trajectory in Fig. 5(c). Finally, Fig. 5(d)

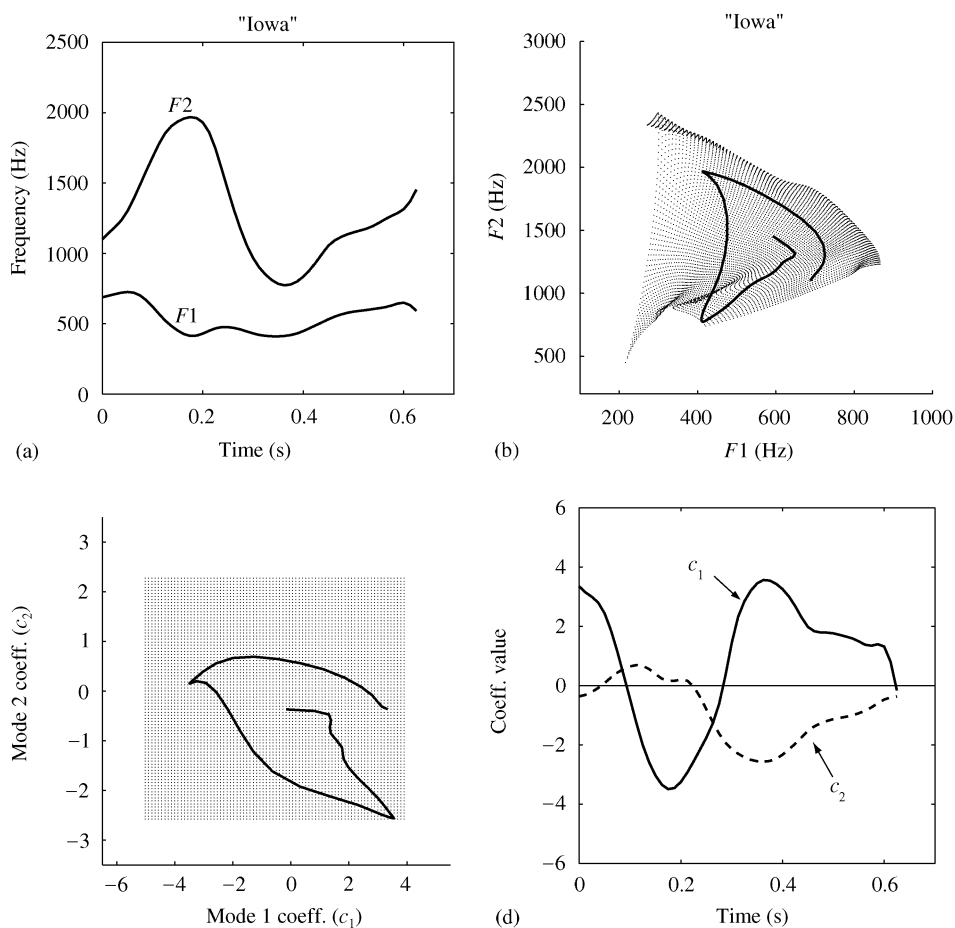


Figure 5. Conversion of formant frequencies $F1$ and $F2$ to modal coefficients c_1 and c_2 for the utterance "Iowa": (a) time-varying $F1$ and $F2$, (b) $F1$ and $F2$ trajectory, (c) coefficient trajectory and (d) time-varying c_1 and c_2 .

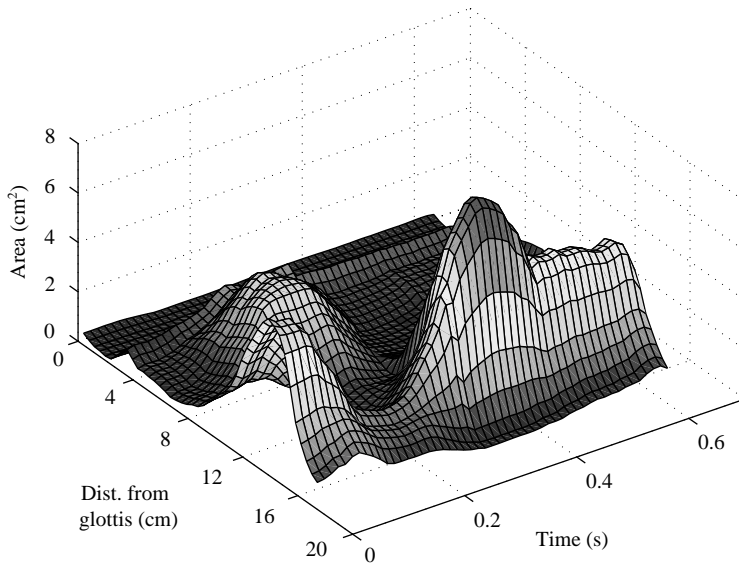


Figure 6. Time-varying area function for “Iowa” constructed from the modal coefficients in Fig. 5(d) using Equation (2).

shows the two modal coefficients (c_1 and c_2) plotted as functions of time. It is noted that the database lookup imposes a slight quantization effect on the resulting time-varying coefficients. Thus, what is shown in Figs. 5(c) and (d) is a mildly smoothed version of the coefficient functions (fourth-order low-pass FIR filter with a cutoff frequency of 10 Hz).

Using the time-varying modal coefficients as input to Equation (2), a time-varying vocal tract area function was generated and is displayed as a 3-D plot in Fig. 6, where areas are shown varying spatially from the glottis to the lips in the x - z plane and temporally along the y -axis. To synthesize the word “Iowa” with this time-varying area function, a wave-reflection type of vocal tract model was used to compute the output sound pressure (Liljencrants, 1985; Story, 1995). In this model, each cross-sectional area in the area function at a given time point defines transmission and reflection coefficients which are used to calculate the acoustic wave propagation along the vocal tract length and radiation at the lip terminal. The model is sampled at a frequency of 44 100 Hz, however, the time-varying area function (Fig. 6) is sampled at only 80 Hz (0.0125 ms interval). To generate an area function for each time sample of the wave propagation, a linear interpolation was used between the consecutive area function samples that make up the time-varying area function in Fig. 6. The synthesis also included a glottal flow pulse model to generate the voice source (Titze, Mapes & Story, 1994). The fundamental frequency (F_0) and amplitude of the glottal flow pulse were varied according to F_0 and amplitude contours extracted from the original recorded utterance. The F_0 extraction was based on low-pass filtering and zero-crossing detection, while the amplitude contour was derived from the envelope of the negative going peaks in the acoustic signal; this envelope was then inverted to be positive going and arbitrarily scaled to have a maximum peak value of 200 cm³/s (similar to Fant, 1993). Other parameters

of the voice source included the open quotient (Q_o) and the skewing quotient (Q_s) which were maintained at constant values for all synthesized samples.

Spectrograms of the natural (recorded) and synthesized versions of “Iowa” are shown in Fig. 7. As expected, the dynamics of the first two formants are nearly identical in the natural and synthesized samples while there are some differences in the upper formants. It is again emphasized, however, that the time-varying area function in Fig. 6 does not necessarily represent the actual area functions used by the speaker, especially in terms of minute details. However, because the area function model is based on this same speaker, it is at least reasonable to expect the mapping to produce similar area functions. Informal listening tests verify a close match in the perceptual characteristics of the two samples but with the largest perceptual difference seeming to be in the voice source (this is entirely expected since only F_0 and amplitude contours were varied in the voice source model). Audio samples of original and synthesized versions of “Iowa” are available as “Supplementary material” for this article on IDEAL at www.idealibrary.com. The synthesized versions of the voice quality modifications discussed in the following sections can also be found at this website.

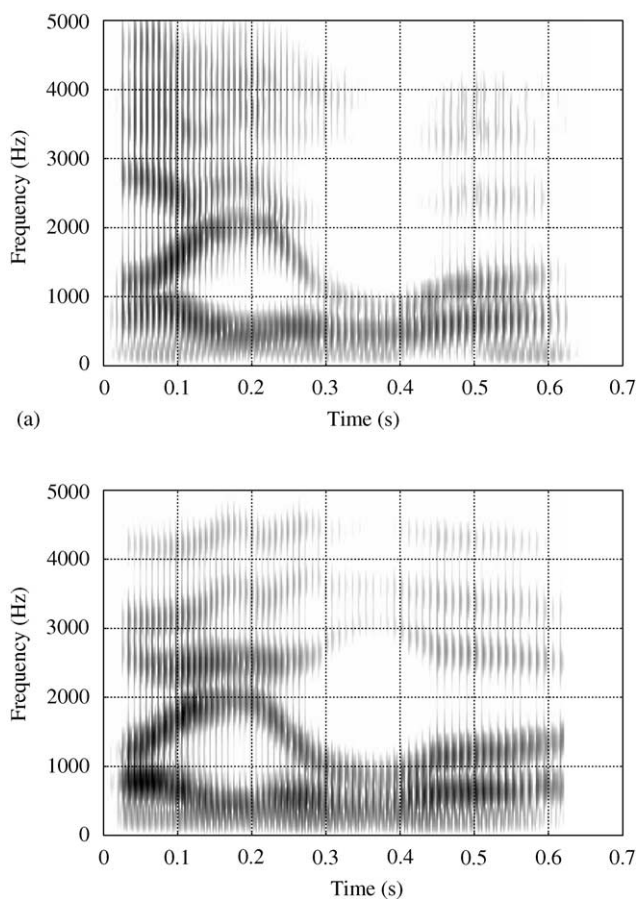


Figure 7. Spectrograms for “Iowa”: (a) natural (recorded) speech, and (b) synthesized version based on the acoustic to articulatory mapping.

3. Transformation of voice quality

In the framework of the area function model specified by Equation (1), an acoustic “background” for a speaker can be considered to be embodied in the acoustic characteristics of the neutral vocal tract. The linguistic gestures are enacted by superimposing weighted combinations of the two modes onto the neutral vocal tract. It is hypothesized that this acoustic “background” can be selectively altered by modifying the neutral area function, hence producing a new voice quality. The modifications can be considered to be new vocal tract “settings” upon which the weighted mode combinations will be superimposed just as they were for the original case. This transformation is perhaps analogous to imposing a modification on neutral carrier signal of Traunmüller (1994), while leaving the modulations unchanged and nearly equivalent to altering the “settings” proposed by Laver (1980).

For this study, four modifications, designed to create distinctly different voice qualities, were imposed on the neutral area function. All four modifications impose changes only to cross-sectional areas; the vocal tract length is held constant. The first two closely correspond to *latitudinal* settings of Laver (1980) *latitudinal settings of palatization and pharyngealization* and primarily affect the middle portions of the vocal tract. Palatization consists primarily of a tendency to impose a constrictive effect in the region of the palate. However, because of physiological constraints, such as the conservation of tongue volume, a constrictive effect in the oral cavity also typically implies an expansion in the pharynx (Laver, 1980, pp. 45–47). Conversely, pharyngealization is realized by imposing a constrictive effect in the middle pharynx and, again due to physiologic constraints, some expansion of the oral cavity.

The modifications to create the palatized and pharyngealized settings were first imposed on the neutral *area* function ($(\pi/4)\Omega^2(x)$) and then converted back into a *diameter* function suitable for Equation (1). The solid line shown in Fig. 8 was used as a scaling function to convert the neutral area function into one that is palatized; i.e., the modified neutral area function is the product of the scaling function in Fig. 8 and the original neutral area function. The scaling function is 1.0 at both the glottal and lip ends meaning that the cross-sectional area in these regions is left unchanged. In the middle portion of the vocal tract, a sine function produced values >1.0 (indicating an area expansion) in the range 3.5–10 cm from the glottis and values <1.0 (indicating an area constriction) at a distance of 10–16 cm from the glottis. For pharyngealization, the same scaling function was used except the sine portion in the middle was shifted in phase by 180° (see the dashed line in Fig. 8). In either case, the maximum increase or decrease in area is 50%. This technique for generating the two settings is obviously artificial in the sense that perfect symmetry is assumed in the form of the sine function. Actual articulation would likely be something other than symmetric, but at this preliminary stage this issue is not a concern; the primary goal here is to establish the method for modifying voice quality.

Figs. 9(a) and (c) show the original and modified neutral area functions for both palatization and pharyngealization, respectively. Frequency response functions (formant spectra) corresponding to each area function are plotted in Figs. 9(b) and (d). Relative to the original neutral area function, the palatization setting causes

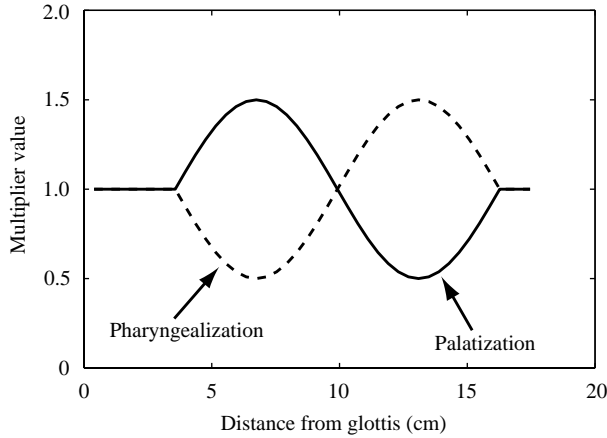


Figure 8. Multiplier functions used to transform the neutral area function to a voice quality that is either palatized (solid line) or pharyngealized (dashed line).

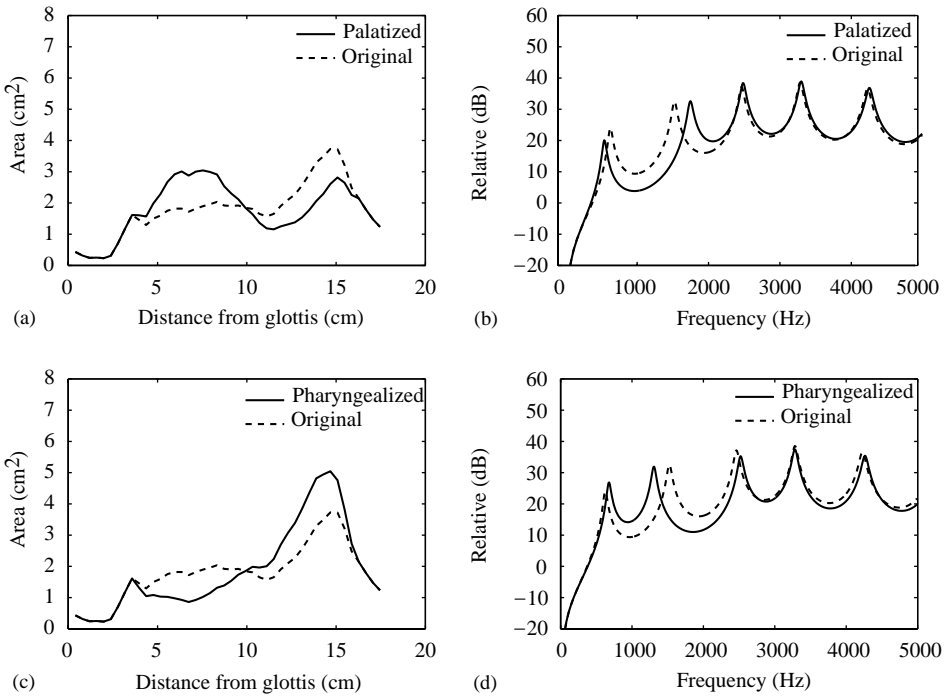


Figure 9. Neutral area functions and corresponding frequency response functions for original and modified voice qualities: (a) original (dashed line) and palatized (solid line) neutral area functions, (b) frequency response functions for the area functions in (a), (c) original (dashed line) and pharyngealized (solid line) neutral area functions, and (d) frequency response functions for the area functions in (c).

$F1$ to be shifted downward in frequency and $F2$ to be shifted upward; the formants above $F2$ are largely unaffected. Not surprisingly, the formant frequency shifts for the pharyngealization setting are just the opposite, that is, $F1$ is shifted upward while $F2$ is shifted downward. Again, the upper formants are mostly unaffected, with the exception that $F3$ is shifted slightly upward in frequency relative to the original neutral area function. The direction of frequency shift of $F1$ and $F2$ in the modifications is predictable from simple acoustic tube considerations. For the palatized setting, the neutral area function is forced to become slightly more [i]-like than the original; hence the shift of $F1$ down and $F2$ up in frequency. The pharyngealized setting does the opposite and moves the neutral area function toward a more [a]-like shape which shifts $F1$ up and $F2$ down in frequency.

The third and fourth modifications consist of altering the extreme ends of the area function while leaving the middle portion of vocal tract unchanged. Specifically, the third setting is a constriction of the lip aperture and an expansion of the epi-laryngeal tube located just above the glottis. A scaling function was again used to impose this setting on the neutral area function, however, this time a partial Gaussian pulse was used to generate the “>1.0” and “<1.0” portions of the function. The result, shown as a solid line in Fig. 10, is a scaling function that will expand cross-sectional areas in the region just above the glottis (0–4 cm) and constrict cross-sectional areas at and near the lips (13.5–17.5 cm). Note that the middle portion is set to 1.0 indicating no change in cross-sectional areas. The fourth setting is the opposite of the third, that is, expansion of the lip aperture and constriction of the epi-laryngeal tube. The same scaling function is used for the fourth setting except that it has been reflected about the 1.0 line extending from the glottis to lips (see the dashed line in Fig. 10).

Original and modified neutral area functions are shown in Figs. 11(a) and (c) for the third and fourth settings of the vocal tract. Additionally, the corresponding frequency response functions are given in Figs. 11(b) and (d). It is noted that the lip constriction/epi-laryngeal expansion setting shifts all formants down in frequency

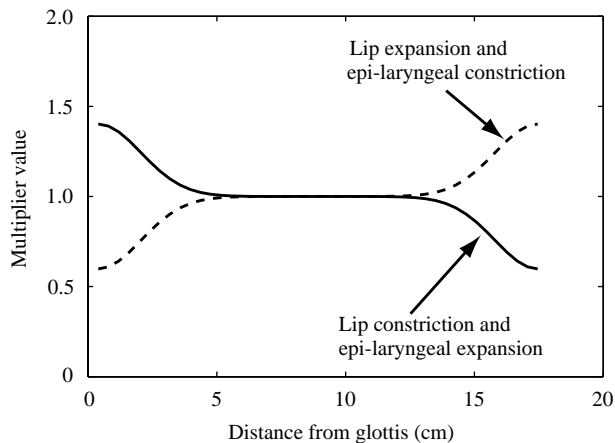


Figure 10. Multiplier functions used to transform the neutral area function to a voice quality represented by either a lip constriction/epi-larynx expansion (solid line) or a lip expansion/epi-larynx constriction (dashed line).

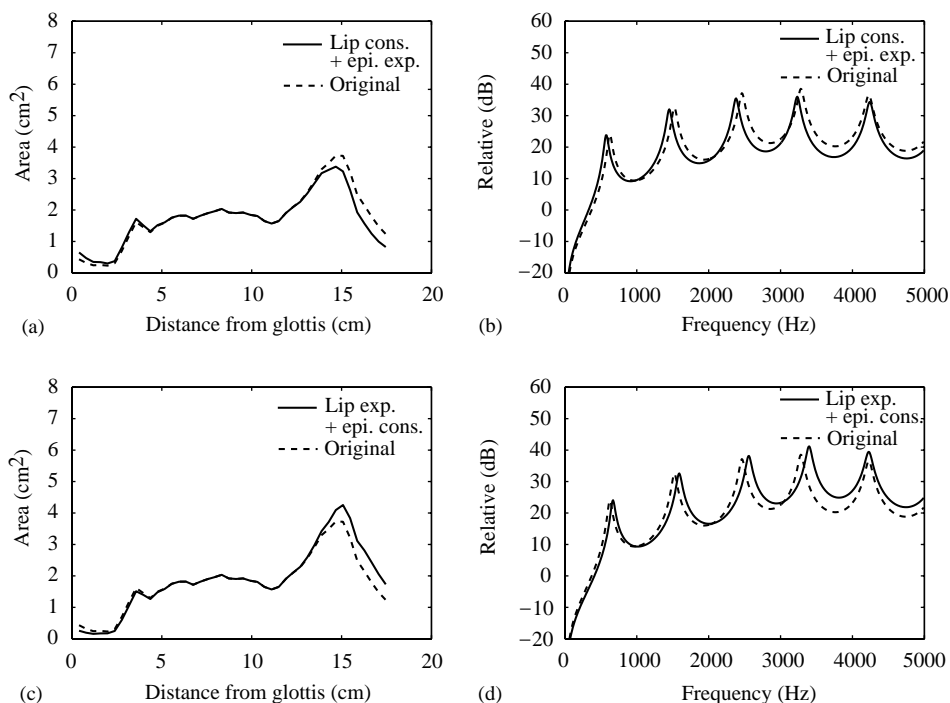


Figure 11. Neutral area functions and corresponding frequency response functions for original and modified voice qualities: (a) original (dashed line) and lip constriction/epi-larynx expansion (solid line) neutral area functions, (b) frequency response functions for the area functions in (a), (c) original (dashed line) and lip expansion/epi-larynx constriction (solid line) neutral area functions, (d) frequency response functions for the area functions in (c).

while the lip expansion/epi-laryngeal constriction setting does the opposite, that is, all formants are shifted upward. Relative to the original, these formants are predictable from simple closed–open tube acoustics. Any time the open end is narrowed and/or the closed end is expanded, all of the formant frequencies will be shifted downward; an upward shift of all formants occurs when the open end is expanded and/or the closed end is narrowed.

3.1. Single-word voiced utterance

The time-varying area function generated in Section 2.1 for “Iowa” can now be modified by exchanging the original neutral diameter function $\Omega(x)$ with one of the four modified versions. Thus, a modified time-varying area function can be constructed as

$$V_m(x, t) = \frac{\pi}{4} [\Omega_m(x) + c_1(t)\phi_1(x) + c_2(t)\phi_2(x)]^2 \quad (3)$$

where the subscript “*m*” refers to one of the four modifications discussed in the previous section. Note that the only change in Equation (3) from Equation (2) is the

new (modified) neutral shape $\Omega_m(x)$. The time-varying coefficients $c_1(t)$ and $c_2(t)$ are identical to those used to produce the synthesis shown in Fig. 7(b).

With Equation (3), four new time-varying area functions for “Iowa” (similar to Fig. 6) were generated for each of the modified voice qualities. Using a transmission-line approach (Sondhi & Schroeter, 1987), F_1 and F_2 formant frequencies were calculated directly from the area functions over the time course of each modified utterance and then plotted against each other as a trajectory. In addition, these area functions were combined with the same glottal flow pulse signal as used previously (i.e., with the same F_0 and amplitude contours) to generate four synthesized versions of “Iowa” representing the voice-quality modifications.

The F_1 – F_2 trajectories resulting from the four modifications are shown in Fig. 12 relative to that of the original utterance; the starting and ending points are denoted by a solid circle and an open circle, respectively. For the palatized voice quality (Fig. 12(a)), the entire trajectory is shifted toward the upper left corner of the F_1 – F_2 plane, meaning that F_1 is lowered and F_2 is raised. These are the same directions of formant frequency shifts observed for the palatized neutral area function in Fig. 9(b). A trajectory shift in the opposite direction occurs for the pharyngealized (Fig. 12(b)) vocal tract. That is, F_1 is increased while F_2 is decreased, resulting in a downward diagonal shift of the trajectory toward the lower right corner of the plane. Again, the direction of formant shifts is the same as those calculated for the pharyngealized neutral area function. The F_1 – F_2 trajectories for the two lip/glottal modifications are shown in Figs. 12(c) and (d). In the case of the lip constriction/epi-laryngeal expansion (Fig. 12(c)), both F_1 and F_2 are decreased resulting in a diagonal shift downward toward the lower left corner of the plane. F_1 and F_2 are both increased in frequency for the opposite case of lip constriction/epi-laryngeal expansion (Fig. 12(d)).

The observed shifts in the F_1 – F_2 trajectories were, for the most part, predictable based on the calculated frequency response functions of the modified neutral area functions (Fig. 11). In addition, perturbations of a simple closed–open tube would indicate the same directions of formant frequency shifts. However, a careful examination of Fig. 12 indicates that the shifted F_1 – F_2 trajectories for the modifications are not simple linear transformations of the original formant frequencies. Instead, the characteristic shape of the path taken through the F_1 – F_2 space under each modification is somewhat distorted relative to the original. For example, a section of the palatized F_1 – F_2 trajectory (Fig. 12(a)) which extends from the extreme upper left hand point (low F_1 , high F_2) to the extreme lower left hand point (low F_1 , low F_2), shows slightly less concavity in the F_1 dimension than does the same section of the original trajectory. The pharyngealized trajectory (Fig. 12(b)) seems to have been “stretched” in the diagonal direction of increasing F_1 and F_2 , but compressed in the opposite direction of decreasing F_1 and increasing F_2 . The main distortion effect in the lip constriction/epi-larynx expansion trajectory is that the second section (from the most [i]-like point to the most [o]-like) is an increase in the concavity relative to the original. For the lip expansion/epi-larynx, this same section of the trajectory shows less concavity.

These distortions originate in the fact that the modified neutral area functions alter the degree of constriction over the course of the utterance. For example, in the palatized modification, any part of a vocal tract deformation gesture that constricts the front portion of the vocal tract will be enhanced as a more severe constriction

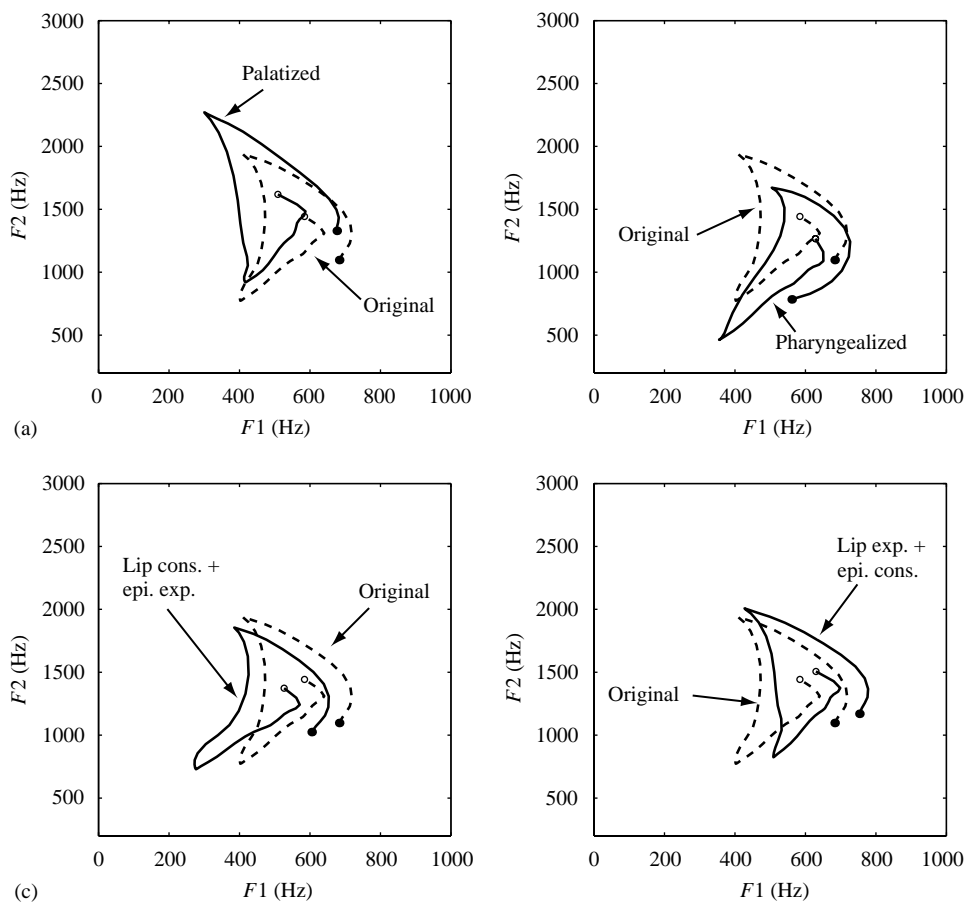


Figure 12. F_1 – F_2 formant trajectories for “Iowa” comparing four neutral area function modifications (solid lines) to the original (dotted line): (a) palatized, (b) pharyngealized, (c) constricted lip aperture + expanded epi-larynx tube, and (d) expanded lip aperture + constricted epi-larynx tube.

than the original. Likewise, the degree of constriction in the pharynx will be reduced under the influence of palatization. The opposite is true for the pharyngealized modification, where the constrictions in the pharynx will be made more severe than in the original, and those in the oral cavity will be reduced in degree (a lesser constriction). The two lip/eppi-larynx cases would undergo similar increases and decreases in the degree of constriction, but at the lip and glottal ends. As is well known, a nonlinear relationship between constriction area and formant frequencies exists (Schroeder, 1966; Stevens, 1989) such that small changes in small areas generally cause much larger formant frequency changes than similar changes in large cross-sectional areas. Thus, as the vocal tract changes shape to produce the “Iowa” utterance, the formant frequencies at any given point in time will depend on the degrees of constriction imposed by the particular voice quality. Since these constriction degrees differ between voice qualities, the trajectory paths would be

expected to differ, as has been observed. It should be noted, however, that these $F1$ – $F2$ trajectory modifications do not reflect any dynamical change in the articulation since the vocal tract movement is entirely governed by the time-varying modal coefficients (which have not been changed).

3.1.1. Comparison to voice-quality modifications of real speech

For comparison to the synthesized modifications of “Iowa”, additional recordings were obtained (from the same speaker as before) representing the four voice qualities discussed in the previous section. The speaker was relatively adept at producing these voice qualities but no assessment was made as to the level of competence. For each recording, the formant frequencies for $F1$ and $F2$ were determined over the time course of the utterance with LPC analysis and a peak picking technique as described above.

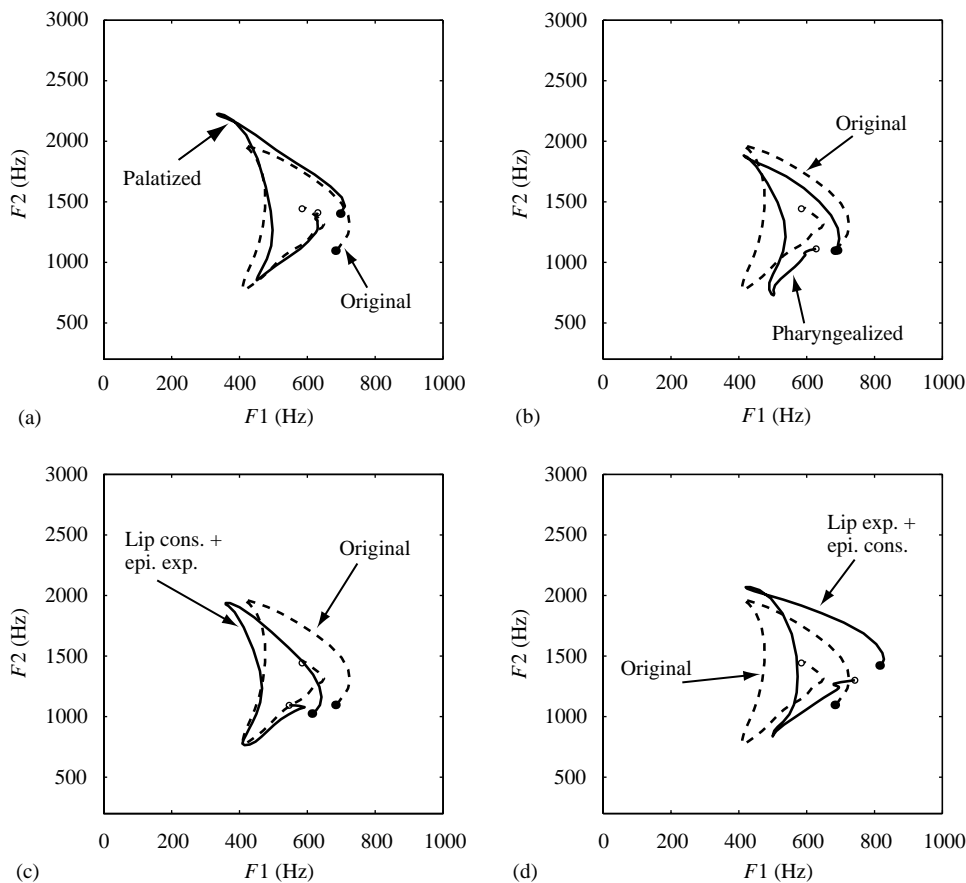


Figure 13. $F1$ – $F2$ formant trajectories for “Iowa” derived from audio recordings of a speaker attempting to produce each of the four voice qualities. Each trajectory (solid lines) is compared with the original (dotted line) as in Fig. 12: (a) ~palatized, (b) ~pharyngealized, (c) ~constricted lip aperture + expanded epi-larynx tube, and (d) ~expanded lip aperture + constricted epi-larynx tube.

The $F1$ – $F2$ trajectories for each voice quality are shown, relative to the original trajectory, in Fig. 13; the arrangement of the graphs is identical to Fig. 12. For the palatized quality (Fig. 13(a)), much of the trajectory is nearly the same as the original except in the upper left-hand portion. It is only in this portion of the trajectory that the frequency of $F2$ increases and that of $F1$ decreases, as was observed for the simulated palatized quality (see Fig. 12(a)). The reason for much of the trajectory remaining nearly unchanged may be that the speaker only imposed the modification during the most /i/-like portion. This would not be surprising since a palatized modification would be most easily maintained during production of a front vowel; i.e., the setting would not compete with the linguistic demands.

The $F1$ – $F2$ trajectory for the pharyngealized quality (Fig. 12(b)) shows an overall decrease in $F2$, but an increase in $F1$ only at the most [o]-like portion (i.e., at about $(F1, F2) = (500, 700)$ Hz. Recall that the synthesized version of this quality showed a decrease in $F2$ and an increase in $F1$ over the course of the entire utterance. Again, the explanation for this is probably that the speaker more easily maintained the pharyngeal constrictive effect during the production of vowel shapes with mid-tract constrictions; i.e., such vowels would not compete with the pharyngeal quality.

In the case of the lip constriction/epi-larynx expansion (Fig. 12(c)), the trajectory shows that both $F1$ and $F2$ were decreased everywhere except in the region where [o]-like vowel shapes would be in use. Apparently, in this region, the lip-constrictive effect has already been maximally used to produce the rounded vowels, thus, the voice-quality modification does not add any further change. The opposite case of lip expansion/epi-larynx constriction (Fig. 12(d)) indicates an increase in both $F1$ and $F2$ formant frequencies, similar to the changes observed for the corresponding synthesized version (see Fig. 12(d)). The expansive effect imposed at the lips for this case does not compete with any other articulatory maneuvers required to produce the word “*Iowa*”, thus allowing the formant changes to be roughly the same (at least in direction) as those predicted by the synthesis version.

3.2. Sentence-level speech

The area function model used in this study is, at this point, strictly valid only for vowels. Hence, the formant-to-coefficient mapping (Fig. 4) cannot be expected to determine occluded or partially occluded vocal tract states. This precludes its direct use for synthesizing and modifying sentence-level speech. However, the same procedures used previously for the single-word utterance can be applied to the vowel and vowel-like portions of a sentence-level utterance. This allows for at least a preliminary demonstration of the voice quality modifications in a broader context than a single word.

The sentence “*I need to catch the eight o'clock flight to New York*” was recorded from the same speaker that participated in the previous parts of this study. In addition to the audio recording, an electroglottograph (EGG) signal was also recorded to allow for simple detection of voiced and unvoiced portions of the sentence. Within the voiced portions, an LPC analysis was used to determine the $F1$ and $F2$ formant frequencies. Then, each time-varying formant ($F1$ and $F2$) was artificially extended across consonantal parts with a parabolic function to form a smooth vowel-like formant contour over the entire sentence. This step preserves the

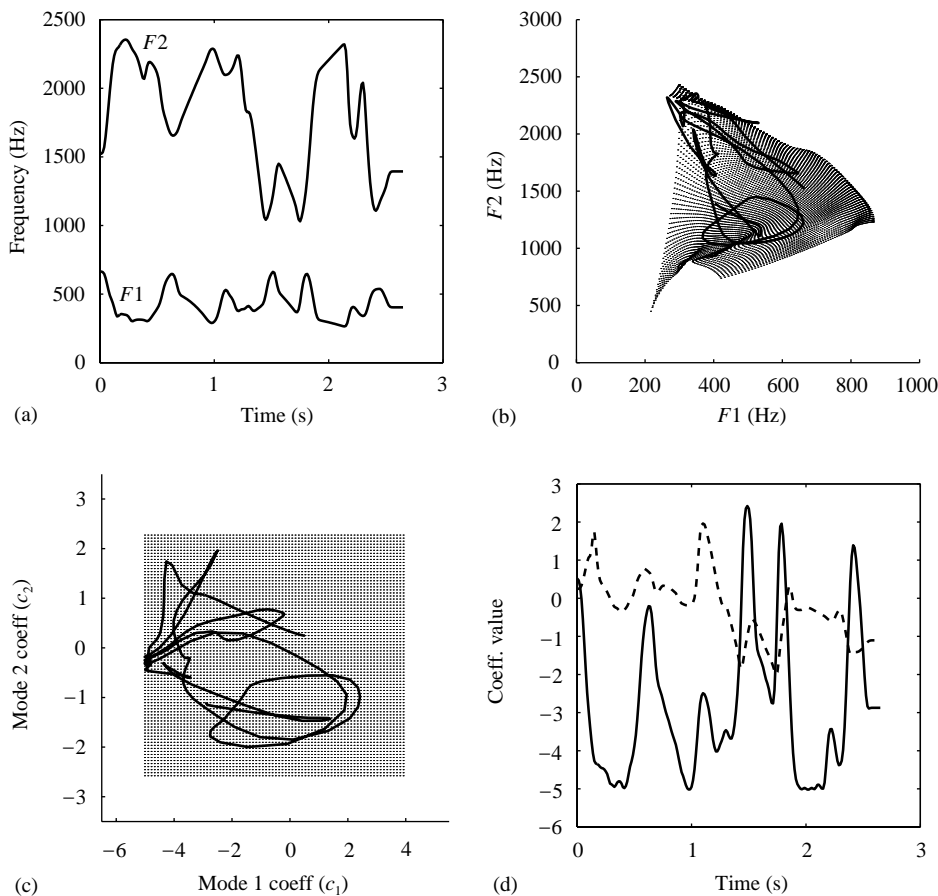


Figure 14. Conversion of formant frequencies $F1$ and $F2$ into modal coefficients c_1 and c_2 for the sentence “*I need to catch the eight o’clock flight to New York.*”: (a) time-varying $F1$ and $F2$, (b) $F1$ and $F2$ trajectory, (c) coefficient trajectory and (d) time-varying c_1 and c_2 .

timing of the sentence and creates a vowel formant “substrate” that can be mapped to modal coefficients. Fig. 14(a) shows these modified formant tracks for $F1$ and $F2$.

Next, the mapping shown in Fig. 4 was used to transform the $F1$ and $F2$ formant contours into time-varying modal coefficients. This process is demonstrated in Fig. 14, as it was for the single-word utterance in Fig. 5. Fig. 14(b) shows the $F1$ – $F2$ trajectory superimposed on the formant mesh of Fig. 4 while the corresponding modal coefficient trajectory is given in Fig. 14(c). Finally, the time-varying modal coefficients are displayed in Fig. 14(d). So, once again, time-varying formants are transformed into time-varying modal coefficients. Subsequently, the time-varying area function was generated with Equation (2) to create the vowel “substrate” of the sentence and is shown in Fig. 15(a); the substrate is the time-varying area function for the sentence without consonantal constrictions.

As an attempt to enhance the synthesis quality, highly approximated consonant constrictions were superimposed on the vowel substrate with an overlay (multiplier)

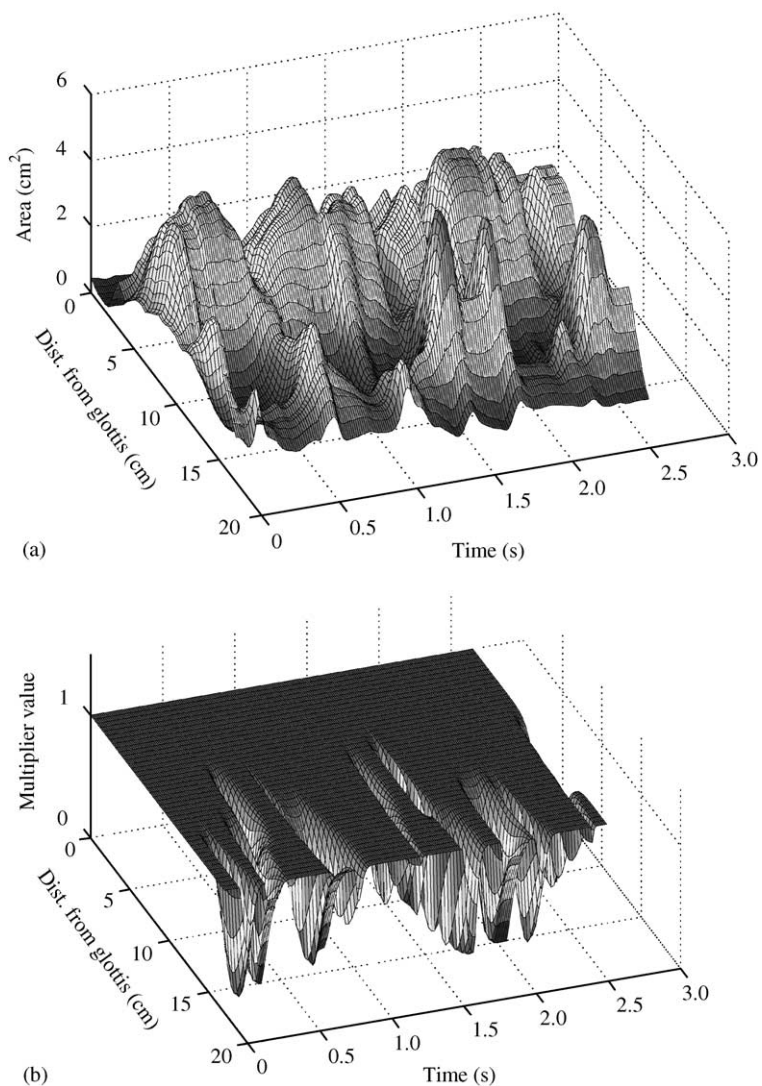


Figure 15. Demonstration of how the time-varying area function for the sentence was created: (a) Time-varying area function for the "vowel substrate" of the sentence. The final time-varying area function used for synthesis is the product of the consonant overlay and the vowel substrate. (b) Consonant overlay function which has a value of 1.0 everywhere except in the region of a consonantal constriction where it may be allowed to be 0.0.

function as was described by Story *et al.* (1998). The overlay is simply a time-varying multiplier that has a value of 1.0 along the entire length of the vocal tract except in the region of a consonant constriction. A constriction is specified in terms of its location (distance from the glottis) and the length of the vocal tract over which it can act. The timing of these consonant constrictions is, at this point, still performed manually. Fig. 15(b) shows the time-varying consonant overlay function

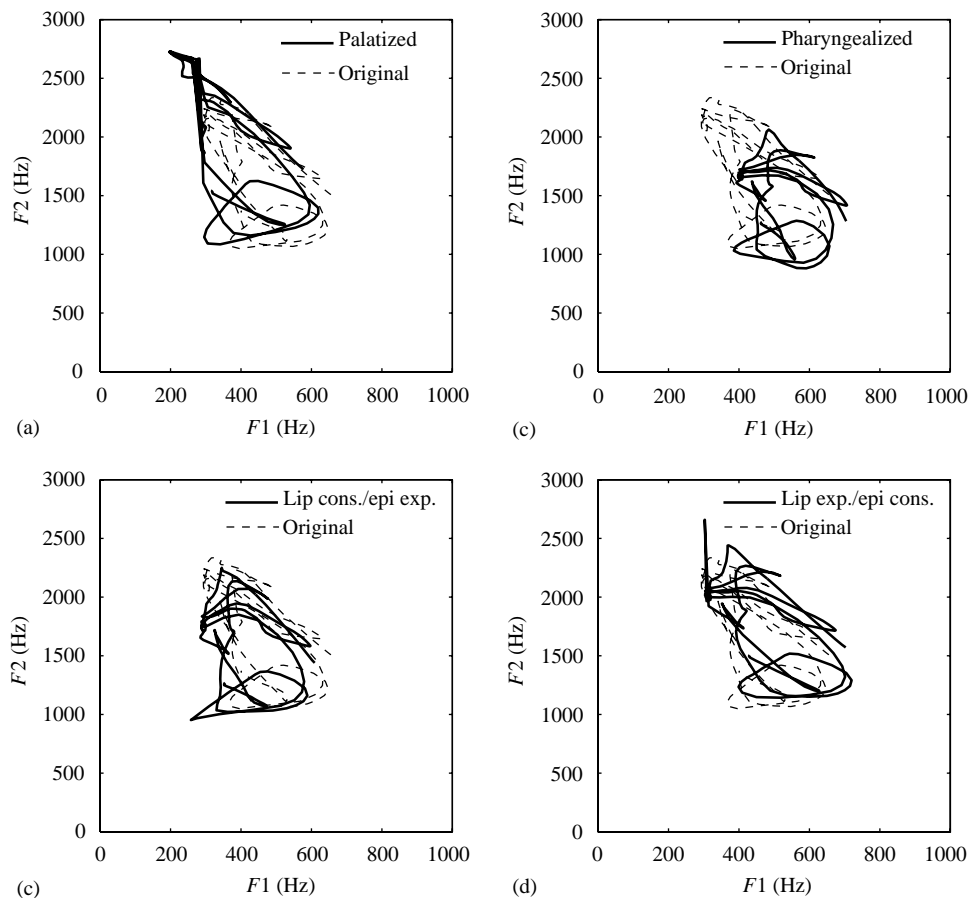


Figure 16. *F1–F2* formant trajectories for “I need to catch the eight o'clock flight to New York” comparing four neutral area function modifications (solid lines) to the original (dotted line): (a) palatized, (b) pharyngealized, (c) constricted lip aperture + expanded epi-larynx tube, and (d) expanded lip aperture + constricted epi-larynx tube.

which is multiplied by the vowel substrate in Fig. 15(a) to produce the final set of area functions that are used for synthesizing the sentence. This description of the consonantal model and how it is combined with the vowel substrate set of area functions is admittedly short and incomplete. However, a full description of this particular aspect of the model is beyond the scope of this article since it is the voice-quality modifications that are the focus.

What is shown in Fig. 15 constitutes the set of area functions for an “original” synthesis of the sentence; i.e., no modification. In addition, four other sets of area functions were created by using the modified neutral diameter functions from Section 3 (in Equation (3)); the consonant overlay was exactly the same for all the four modified versions.

The sentence was synthesized 5 times: first with the product of the area function and consonant overlay shown in Fig. 15, then with the four modified versions.

Again, a glottal flow pulse model (Titze, 1994) was used as the voice source whose fundamental frequency was driven by the F_0 contour determined from the sentence. It should be noted that these syntheses were generated without the turbulent noise generation characteristic of plosive and fricative consonants.

As was done for the single-word utterance, $F1$ – $F2$ trajectories over the time course of the sentence were determined directly from the vowel substrate area function sets for each voice quality modification (without the consonant overlay). The trajectories are shown relative to the original in Fig. 16. In general, each trajectory for the four modified voice qualities shows formant frequency shifts in the same directions as were observed for the single-word utterance. That is, the palatized version exhibits a decrease in $F1$ and an increase in $F2$, the pharyngealized quality decreases $F2$ but raises $F1$, the lip constriction/epi-larynx expansion decreases the frequency of both $F1$ and $F2$, and finally the lip expansion/epi-larynx constriction increases both $F1$ and $F2$. It is also apparent that considerable distortions of the formant trajectories occur in the transformation from the original into the modified versions. The reasons for such distortions can be again be assumed to be due to the change in constriction degree that each modification can produce.

4. Discussion

As a preliminary step toward understanding voice quality, this study has shown that the idea of separating articulatory gestures from an underlying neutral vowel shape may provide a useful paradigm for investigating voice quality independent of the speech movements themselves. The results showed that modification of the neutral vocal tract area function does bias (shift) the formant structure of a given utterance in a generally consistent and predictable manner when the same gestural deformations are applied. The observed formant shifts are, for the most part, explained by simple tube acoustic considerations. The modification for palatization biases the tract configuration in the direction of an [i] vowel, whereas the pharyngealization creates a tract shape bias directed toward an [a] vowel. For the two other cases, it is well known that a widened area at the lip end and a constricted area at the glottal end will move all formants upward in frequency while the opposite area change in both regions will lower all formants (Fant, 1960).

The distortion of the formant trajectories across the four modifications, relative to the original, is interesting from the point of view that the exact same articulation pattern (as specified by the time-varying coefficients in Fig. 4(d)) was used in each of the four cases. Thus, the dynamic characteristics of the formant trajectories do not necessarily reflect the underlying articulatory dynamics. This occurs as a result of the nonlinear relationship between the area function (articulation) and acoustic characteristics (e.g., Stevens, 1989). For example, in the case of palatization, the oral cavity will experience a much greater constriction (equivalently, a much smaller area) during the [i]-like portion of the *Iowa* utterance than in any of the other cases. Similarly, the pharyngealized modification would force the middle pharynx to be more highly constricted for the [a]-like portions of the utterance. This suggests that, in addition to predictable formant shifts, various voice qualities may produce different formant dynamics even though the articulatory dynamics producing the speech are the same. Apparently, the common articulatory dynamics could only be

deciphered with knowledge of the underlying neutral tract shape and its acoustic characteristics.

The formant trajectories of the same four voice qualities produced by a real speaker showed many similarities to the synthesized versions. However, there were also significant differences that seemed to indicate that some portions of the utterance were more or less affected by the modifications than others. For example, the palatization quality was most apparent during the portion of the word "Iowa" that required fronted vowel shapes. Conversely, the portion of the trajectory that corresponds to vowel shapes with mid-tract constrictions was significantly more affected by the pharyngealized modification. What the acoustic recordings show is that a speaker may impose a particular voice quality only when articulatory conditions permit it rather than uniformly across the utterance. For the synthesized versions, the modifications were strictly imposed throughout the entire utterance; hence, their effects would necessarily be expressed in some way at all points in time.

The voice-quality modifications investigated in this paper can be considered to represent only "latitudinal" changes in the neutral vocal tract. That is, only cross-sectional areas were increased or decreased in the neutral area function. There were no "longitudinal" modifications along the axis of the vocal tract such as lip rounding/spreading, larynx lowering/raising, or an overall vocal tract length change. Presumably in natural speech, both latitudinal and longitudinal elements could occur simultaneously or wax and wane over the course of an utterance. Thus, future studies need to include the effects of both expansive and constrictive effects of areas as well as various types of tract length changes. In addition, analysis of natural speech samples representing a variety of voice qualities and speakers is needed in order to determine the relevance of this work to natural speech. Finally, there is a need to perform formal perceptual experiments to assess and verify particular modifications.

This study was supported by grants NIH R01 DC02532-05 and NIH R01 DC04789-01 from the National Institutes on Deafness and other Communication Disorders (NIDCD). The authors also acknowledge two reviewers whose comments have helped to improve upon an earlier version of this manuscript.

References

- Fant, G. (1960) *The acoustic theory of speech production*. The Hague: Mouton.
- Fant, G. (1993) Some problems in voice source analysis, *Speech Communication*, **13**, 7–22.
- Fitch, W. T. & Giedd, J. (1999) Morphology and development of the human vocal tract: a study using magnetic resonance imaging, *Journal of the Acoustical Society of America*, **106**, 1511–1522.
- Goldstein, U. G. (1980) *An articulatory model for the vocal tracts of growing children*, Doctoral Dissertation, Department of Electrical Engineering and Computer Science, M.I.T.
- Holmberg, E. B., Hillman, R. E. & Perkell, J. S. (1988) Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice, *Journal of the Acoustical Society of America*, **84**, 511–529.
- Ladefoged, P. & Broadbent, D. E. (1957) Information conveyed by vowels, *Journal of the Acoustical Society of America*, **29**, 98–104.
- Laver, J. (1980) *The phonetic description of voice quality*. Cambridge: Cambridge University Press.
- Liljencrants, J. (1985) *Speech synthesis with a reflection-type line analog*. DS Dissertation, Department of Speech Comm. and Music Acous., Royal Institute of Technology, Stockholm, Sweden.
- Lindblom, B., Brownlee, S., Davis, B. & Moon, S.-J. (1992) Speech transforms, *Speech Communication*, **11**, 357–368.
- Mermelstein, P. (1966) Determination of the vocal-tract shape from measured formant frequencies, *Journal of the Acoustical Society of America*, **41**(5), 1283–1294.

- Peterson, G. E. & Barney, H. L. (1952) Control methods used in a study of the vowels, *Journal of the Acoustical Society of America*, **24**(2), 175–184.
- Schroeder, M. R. (1967) Determination of the geometry of the human vocal tract by acoustic measurements, *Journal of the Acoustical Society of America*, **41**(4), 1002–1010.
- Sondhi, M. M. & Schroeter, J. (1987) A hybrid time-frequency domain articulatory speech synthesizer, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-35**(7), 955–967.
- Stevens, K. N. (1989) On the quantal nature of speech, *Journal of Phonetics*, **17**, 3–45.
- Story, B. H. (1995) *Physiologically-based speech simulation with an enhanced wave-reflection model of the vocal tract*, PhD dissertation, University of Iowa.
- Story, B. H., Titze, I. R. & Hoffman, E. A. (1996) Vocal tract area functions from magnetic resonance imaging, *Journal of the Acoustical Society of America*, **100**(1), 537–554.
- Story, B. H. & Titze, I. R. (1998) Parameterization of vocal tract area functions by empirical orthogonal modes, *Journal of Phonetics*, **26**(3), 223–260.
- Story, B. H., Titze, I. R. & Long, R. (1998) Synthesis of sentence-level speech based on measured vocal tract area functions. In *Proceedings of the ICA/ASA Joint Meeting*, Seattle, WA, 20–26 June, pp. 2663–2664.
- Titze, I. R. (1989) Physiologic and acoustic differences between male and female voices, *Journal of the Acoustical Society of America*, **85**, 1699–1707.
- Titze, I. R. (1994) *Principles of voice production*, Englewood Cliffs, NJ: Prentice-Hall.
- Titze, I. R., Mapes, S. & Story, B. H. (1994) Acoustics of the tenor high voice, *Journal of the Acoustical Society of America*, **95**(2), 1133–1142.
- Trautmüller, H. (1994) Conventional, biological and environmental factors in speech communication: a modulation theory, *Phonetica*, **51**, 170–183.
- Trautmüller, H. & Eriksson, A. (2000) Acoustic effects of variation in vocal effort by men, women, and children, *Journal of the Acoustical Society of America*, **107**, 3438–3451.