Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions (L)

Brad H. Story[a)]

Speech Acoustics Laboratory, Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

(Received 17 May 2005; revised 17 October 2005; accepted 22 November 2005)

A technique for modifying vocal tract area functions is developed by using sum and difference combinations of acoustic sensitivity functions to perturb an initial vocal tract configuration. First, sensitivity functions [e.g., Fant and Pauli, Proc. Speech Comm. Sem. **74**, 1975] are calculated for a given area function, at its specific formant frequencies. The sensitivity functions are then multiplied by scaling coefficients that are determined from the difference between a desired set of formant frequencies and those supported by the current area function. The scaled sensitivity functions are then summed together to generate a perturbation of the area function. This produces a new area function whose associated formant frequencies are closer to the desired values than the previous one. This process is repeated iteratively until the coefficients are equal to zero or are below a threshold value. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2151802]

PACS number(s): 43.70-h, 43.70.Bk [AL] Pages: 715–718

I. INTRODUCTION

The shape of the vocal tract can be approximately represented by an area function; that is, the variation in cross-sectional area as a function of distance from the glottis. A pattern of acoustic resonances can be calculated based on the shape of any given area function, and will indicate the locations of the formant frequencies that contribute to both phonetic and speaker-specific characteristics. It may be of interest to know how particular changes in the formant frequency pattern could be generated by changes to the shape of the area function, and vice versa. For example, how might an area function for the vowel [ɑ] be modified so that the second formant (F2) is increased in frequency while all other formants remain fixed at their original values? Or perhaps the interest may be in altering the area function to generate a particular pattern of the upper formants (e.g., F3–F5) such that a distinct change in sound quality is produced, while F1 and F2 remain fixed.

The purpose of this letter is to present a technique that iteratively adjusts ("tunes") the shape of a given vocal tract area function so that a specific pattern of formant frequencies is produced. The technique is based on perturbing the shape of an initial vocal tract configuration with a summation of scaled acoustic sensitivity functions, such that the formants are systematically displaced toward desired values. Although it is well known that transformation of a set of formant frequencies to a vocal tract area function does not produce a unique solution (e.g., Schroeder, 1967; Mermelstein, 1967; Wakita, 1973; Atal, Chang, Mathews, and Tukey, 1978; Milenkovic, 1984; Sondhi and Resnick, 1983; Sorokin, 1992), the technique described may be useful for generating subtle, hypothetical, modifications to a specific vocal tract shape.

II. ACOUSTIC SENSITIVITY FUNCTIONS

The sensitivity of a particular formant frequency to a change in vocal tract cross-sectional area can be defined as the difference between the kinetic energy (KE) and potential energy (PE) as a function of distance from the glottis, divided by the total energy in the system (Fant and Pauli, 1975). A sensitivity function can be written as

$$S_n(i) = \frac{\mathrm{KE}_n(i) - \mathrm{PE}_n(i)}{\mathrm{TE}_n} \quad n = 1,2,3,\ldots \quad \text{and} \quad i = [1,\ldots,N_{\text{areas}}], \tag{1}$$

where i is the section number (section 1 is just above the glottis and section N_{areas} is at the lips), n is the formant number, and

$$\mathrm{TE}_n = \sum_{i=1}^{N_{\text{areas}}} [\mathrm{KE}_n(i) + \mathrm{PE}_n(i)]. \tag{2}$$

The kinetic and potential energies for each formant frequency are based on the pressure $P_n(i)$ and volume velocity $U_n(i)$ computed for each section of an area function. They are calculated as

$$\mathrm{KE}_n(i) = \frac{1}{2}\frac{\rho l(i)}{a(i)}|U_n(i)|^2 \tag{3}$$

and

$$\mathrm{PE}_n(i) = \frac{1}{2}\frac{a(i)l(i)}{\rho c^2}|P_n(i)|^2, \tag{4}$$

where $a(i)$ and $l(i)$ are the cross-sectional area and length of element i within an area function, respectively, and ρ is the density of air and c is the speed of sound. The area functions used throughout this study contained 44 sections, each with a length of $l(i)=0.396825$ cm. Hence $i=[1,\ldots,44]$ and the actual distance from the glottis for each section is $x = i \cdot l(i)$.

[a)]Electronic mail: bstory@u.arizona.edu

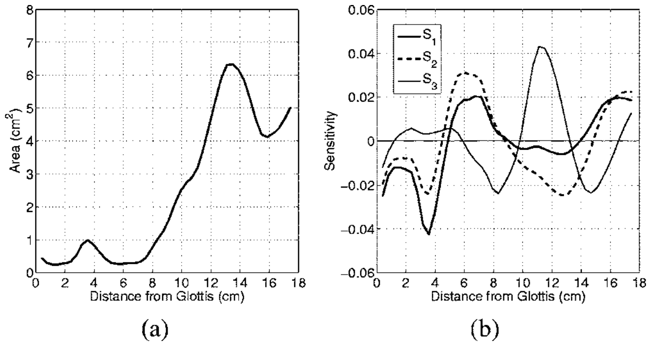

FIG. 1. Sensitivity function calculation for a male [ɑ] vowel based on Story, Titze, and Hoffman (1996). (a) Measured area function, (b) sensitivity functions for F1, F2, and F3.

Calculations of pressures, flows, and frequency response functions for this study were accomplished with a transmission-line type model of the vocal tract (e.g., Sondhi and Schroeter, 1987; Story, Laukkanen, and Titze, 2000) that included energy losses due to yielding walls, viscosity, heat conduction, and acoustic radiation at the lips. This particular implementation did not, however, include any side branches such as the piriform sinuses, sublingual cavities, or nasal passages. While it is recognized that these cavities may significantly affect some formant frequencies (Dang and Honda, 1997; Espy-Wilson, 1992; Makarov and Sorokin, 2004), their omission here does not affect the development of the proposed method, nor would it prevent their inclusion in the future.

As an example, sensitivity functions were calculated for a male [ɑ] vowel based on Story, Titze, and Hoffman (1996) [see Fig. 1(a)] and are shown in Fig. 1(b). Each line extends along the distance from the glottis to lips and indicates the relative sensitivity of the first, second, and third formants (F1, F2, and F3) to a small perturbation of the area function [$\Delta a(i)$]. Mathematically, this can be written as,

$$\frac{\Delta \mathrm{F}_n}{\mathrm{F}_n} = \sum_{i=1}^{N_{\text{areas}}} S_n(i)\frac{\Delta a(i)}{a(i)}, \tag{5}$$

where n is again the formant number. Using S_1 in Fig. 1(b) and Eq. (5) as a guide, it is observed that F1 could be increased by expanding the area in regions along the vocal tract length between 5 cm and 9 cm from the glottis and from 14 cm to the lip termination. F1 could also be increased by constricting the regions between the glottis and 5 cm, as well as between 9 cm and 14 cm. Lowering F1 would require the opposite changes in area within the same regions. For S_2, an increase in F2 could be produced by expanding the regions between 4.5–8.8 cm and 14.8–17.5 cm, and constricting the regions of the area function that extend from 0 to 4.5 cm and 8.8 to 14.8 cm; lowering F2 would require the opposite changes in area. Changes in F3 could be similarly carried out by modifying cross-sectional areas in the positively and negatively valued regions specified by S_3. Although not shown, sensitivity functions corresponding of F4 and F5 were also calculated.

III. AREA FUNCTION PERTURBATION

Whereas changes to an area function that would modify formant frequencies according to the calculated sensitivity functions can be performed manually (e.g., Story, Titze, and Hoffman, 2001), an automated process would be more efficient, and ultimately more useful. The proposed technique consists of superimposing scaled (to affect cross-sectional area) replicas of the sensitivity functions on an area function. Direct superposition of S_1 for any area function would raise F1, whereas its opposite, $-S_1$, would lower it. F2 could be similarly controlled with superposition of a scaled S_2 replica, where $+S_2$ would increase F2 and $-S_2$ would decrease it. Higher frequency formants (e.g., F3, F4, and F5) could also be controlled with superposition of their respective sensitivity functions. Shifting multiple formants simultaneously could be carried out with superposition of the sum of $\pm S_1, \pm S_2, \pm S_3, \ldots, \pm S_n$.

The prediction of formant frequency change based on sensitivity functions is, however, limited to small area changes (approximately <10%). Thus, sensitivity functions need to be recomputed after any small amount of area change, and a new perturbation determined. This can be performed iteratively until arriving at an area function that produces a desired set of formant frequencies. The process is mathematically represented as,

$$a_{k+1}(i) = a_k(i) + \sum_{n=1}^{N_{\text{fmts}}} z_{n_k} S_{n_k}(i)$$
$$i = [1, N_{\text{areas}}] \quad k = [0, N_{\text{iter}}]. \tag{6}$$

With the initial area function denoted by $a_0(i)$, the $a_k(i)$'s and S_{n_k}'s are vocal tract area functions and sensitivity functions, respectively, at successive iterations. The coefficients z_{n_k} scale the sensitivity functions so the area function perturbation displaces the formant frequencies in the desired direction. At every iteration, the z_{n_k}'s are determined by

$$z_{n_k} = \alpha\left[\frac{\mathcal{F}_n - \mathrm{F}_{n_k}}{\mathrm{F}_{n_k}}\right], \tag{7}$$

where $\mathcal{F}_n$ is a set of target formant frequencies and the F_{n_k}'s are the formants that correspond to the kth area function. α is an additional scale factor that can be used to speed the iterative process and is typically set to $\alpha=10$. The iterations continue until the root of the sum of the squared differences between target formants and those of the kth area function,

$$\delta = \sqrt{\sum (\mathcal{F}_n - \mathrm{F}_{n_k})^2} \tag{8}$$

is less than a desired tolerance value. For this study, the iterations were allowed to proceed until $\delta < 0.1$ Hz.

To protect against cross-sectional areas becoming too small over the course of successive iterations, the superposition is performed logarithmically for those areas within an area function that are less than 1 cm^2 so that Eq. (6) becomes,

$$a_{k+1}(i) = \begin{cases} a_k(i) + \displaystyle\sum_{n=1}^{N_{\text{fmts}}} z_{n_k} S_{n_k}(i) \quad \text{for } a_k(i) > 1 \\ \exp\left(\ln(a_k(i)) + \ln\left(\displaystyle\sum_{n=1}^{N_{\text{fmts}}} z_{n_k} S_{n_k}(i) + 1 \right)\right) \\ \quad \text{for } a_k(i) \leqslant 1 . \end{cases} \tag{9}$$

In addition, a minimum area threshold is set such that,

$$a_{k+1}(i) = \max[a_{k+1}(i), 0.1], \tag{10}$$

where the 0.1 is in units of square centimeters.

It is noted that Carré (2004) proposed a similar iterative technique for modifying the shape of an area function, specifically a uniform tube. His technique was based on using sensitivity functions as deformation patterns, but did not include individual scaling coefficients like the z_{n_k}'s specified here. Instead, each sensitivity function was always scaled with an amplitude of ±1.0, as well as a "deformability function" that constrained specific regions of the area function. Because of the constraints on the scaling coefficients, Carré's method apparently does not allow for the specification of a desired formant pattern, but rather modifies the area function successively with constant coefficients to produce a varying formant contour.

IV. EXAMPLES OF AREA FUNCTION TUNING

To demonstrate the method outlined in the previous section, the area function shown in Fig. 1(a) ([ɑ] vowel) was "tuned" in two different ways. In the first example, the second formant was shifted upward in frequency while all other formants (F1, F3, F4, and F5) were held constant at their original calculated values. For the second example, the upper formants were shifted toward each other so that they formed a cluster in the vicinity of about 3000 Hz [this is a typical frequency range of the "singing formant" (Sundberg, 1974)].

A. Upward shift of F2

With the [ɑ] area function (Fig. 1) serving as the initial vocal tract configuration $a_0(i)$, the corresponding initial (calculated) formant frequencies were,

$$\mathrm{F}_{n_0} = 800, 1136, 2770, 3448, 4240 \text{ Hz}, \quad n = 1,2,3,4,5 \tag{11}$$

and the target set of formant frequencies were set to,

$$\mathcal{F}_n = 800, \mathbf{1400}, 2770, 3448, 4240 \text{ Hz} \quad n = 1,2,3,4,5, \tag{12}$$

where only F2 differs from the original formants. Based on Eq. (7), and with $\alpha = 10$, the initial z coefficients were

$$z_{1_0} = 0.0, \tag{13}$$

$$z_{2_0} = 10\left[\frac{\mathcal{F}_2 - F_{2_0}}{F_{2_0}}\right] = \left[\frac{1400 - 1136}{1136}\right] = 2.32, \tag{14}$$

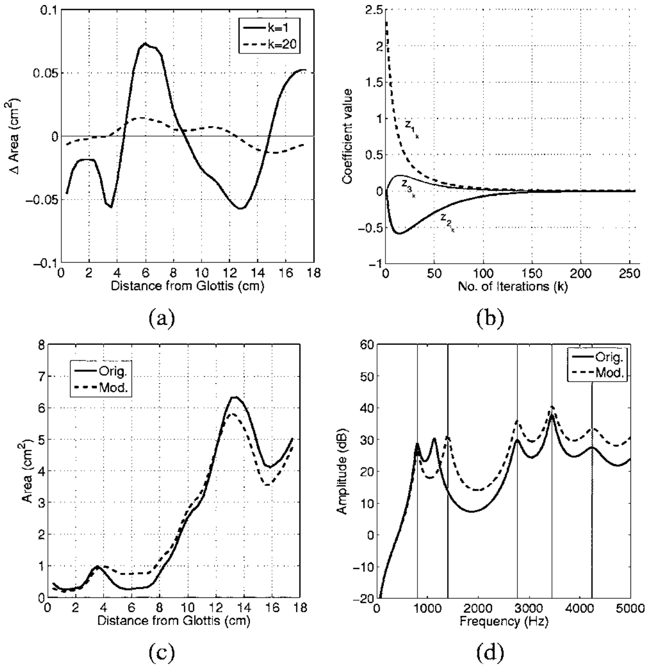

FIG. 2. Example of area function tuning to produce an upward shift of F2 based on the [ɑ] vowel in Fig. 1(a). (a) Area function perturbations at the first (solid) and the twentieth (dashed) iterations. (b) Values of the z for the first three coefficients as they converge toward 0.0 over 257 iterations. (c) Initial (solid) and modified (dashed) area functions. (d) Frequency response functions of the initial (solid) and final (dashed) area functions; the vertical lines represent the target formant frequencies.

$$z_{3_0} = z_{4_0} = z_{5_0} = 0.0. \tag{15}$$

The perturbation pattern imposed on the initial area function at the first iteration is simply $2.32 S_2$ and is plotted as the solid line in Fig. 2(a). Note that the y-axis is shown as "Δ area cm^2" to indicate the cross-sectional area that is added to, or subtracted from, the area function and represents $\Delta a(i)$ in Eq. (5). The shape of the perturbation will gradually evolve over the course of the iterations to reflect the progressive change in sensitivity of the modified area function. As an example, the perturbation at the 20th iteration is shown as the dashed line.

After 257 iterations, $\delta < 0.1$ Hz and the z coefficients corresponding to each formant have converged toward zero as shown for the first three formants in Fig. 2(b); the z coefficients for F4 and F5 similarly converge but are not shown in order to preserve the clarity of the figure. The original and final area functions are shown in Fig. 2(c), where it can be seen that the tuning process has generated an expansion of the pharyngeal part of the vocal tract, and a reduction of the cross-sectional areas in the oral cavity and just above the glottis. The corresponding frequency response functions are plotted in Fig. 2(d) and indicate that the second formant frequency produced by the modified area function is precisely matched to the target frequency of 1400 Hz. The other formants were successfully maintained at their original values.

B. Clustering of F3, F4, and F5

A second example of area function tuning is shown in Fig. 3. The initial area function was the same [ɑ] vowel used in the first example. As before, the initial calculated formants were,

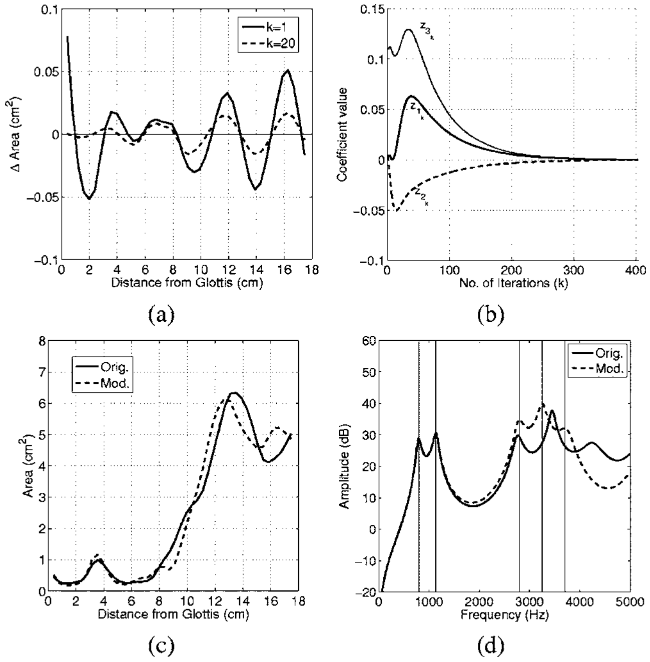

FIG. 3. Example of area function tuning to compress the distance between F3, F4, and F5, while maintaining F1 and F2 at their original values. (a) Area function perturbations at the first (solid) and the twentieth (dashed) iterations. (b) Values of the z coefficients as they converge toward 0.0 over 403 iterations. (c) Initial (solid) and modified (dashed) area functions. (d) Frequency response functions of the initial (solid) and final (dashed) area functions; the vertical lines represent the target formant frequencies.

$$\mathrm{F}_{n_0} = 800, 1136, 2770, 3448, 4240 \text{ Hz}, \quad n = 1,2,3,4,5 \tag{16}$$

but the target values were set to be,

$$\mathcal{F}_n = 800, 1136, \mathbf{2800}, \mathbf{3250}, \mathbf{3700} \text{ Hz}, \quad n = 1,2,3,4,5, \tag{17}$$

which specifies that the distance between the upper formants (F3, F4, and F5) will be decreased.

The perturbation pattern at the first iteration is shown in Fig. 3(a), and the converging z coefficients for F1, F2, and F3 are plotted in Fig. 3(b). In this case, 403 iterations were required so that $\delta < 0.1$ Hz. Shown in Fig. 3(c) are the initial and modified area functions. Relative to the initial configuration, the modifications consist of slight constrictions and expansions along the entire vocal tract length. The most apparent changes are in the oral portion of the area function, but the small changes between the glottis and about 8 cm may have also contributed significantly to the acoustic changes. The frequency response functions in Fig. 3(d) indicate that the target formant frequencies are achieved with the modified area function. F1 and F2 are the same as in the initial configuration, but F3, F4, and F5 were moved toward each other, and their combined effect produces a somewhat enhanced amplitude in the 2800–3800 Hz range. It can be noted parenthetically that the cluster of F3, F4, and F5 was produced without any major changes to the length and area of the epilaryngeal portion of the area function. This contrasts with the typical articulatory interpretation of the singing formant (Sundberg, 1974).

V. DISCUSSION

A technique has been proposed that allows area functions to be modified so that their formant frequencies match a set of targets. In both of the examples, the technique was used to successfully modify an original area function, reducing the difference between a set of original and target formant frequencies to nearly zero. An inherent limitation of the technique, however, is that it is not possible to know if the resulting modifications to an initial area function are those that would actually be produced by a human speaker. Hence, the modifications can only be considered hypothetical. Nonetheless, the ability to systematically perturb an area function such that a desired formant pattern is produced may be a useful tool to help understand how relatively subtle changes in cross-sectional area can produce significant acoustic changes.

ACKNOWLEDGMENTS

This study was supported by Grant No. R01 DC04789 from the NIDCD.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting-sorting technique," J. Acoust. Soc. Am. **63**, 1535–1555.

Carré, R. (**2004**). "From an acoustic tube to speech production," Speech Commun. **42**, 227–240.

Dang, J., and Honda, K. (**1997**). "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101**(1), 456–465.

Espy-Wilson, C. Y. (**1992**). "Acoustic measures for linguistic features distinguishing the semivowels/w j r l/ in American English," J. Acoust. Soc. Am. **92**, 736–757.

Fant, G., and Pauli, S. (**1975**). "Spatial characteristics of vocal tract resonance modes," in Proceedings of the Speech Comm. Sem. 74., Stockholm, Sweden, August 1–3, pp. 121–132.

Makarov, I. S., and Sorokin, V. N. (**2004**). "Resonances of a branched vocal tract with compliant walls," Acoust. Phys. **50**(3), 323–330.

Mermelstein, P. (**1967**). "Determination of the vocal-tract shape from measured formant frequencies," J. Acoust. Soc. Am. **41**(5), 1283–1294.

Milenkovic, P. (**1984**). "Vocal tract area functions from two-point acoustic measurements with formant frequency constraints," IEEE Trans. Acoust., Speech, Signal Process., **ASSP-32**(6), 1122–1135.

Schroeder, M. R. (**1967**). "Determination of the geometry of the human vocal tract by acoustic measurements," J. Acoust. Soc. Am. **41**(4), 1002–1010.

Sondhi, M. M., and Resnick, J. R. (**1983**). "The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis," J. Acoust. Soc. Am. **73**(3), 985–1002.

Sondhi, M. M., and Schroeter, J. (**1987**). "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**(7), 955–967.

Sorokin, V. N. (**1992**). "Determination of vocal tract shape for vowels," Speech Commun. **11**, 71–85.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**(1), 537–554.

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

Story, B. H., Titze, I. R., and Hoffman, E. A. (**2001**). "The relationship of vocal tract shape to three voice qualities," J. Acoust. Soc. Am. **109**, 1651–1667.

Sundberg, J. (**1974**). "Articulatory interpretation of the singing formant," J. Acoust. Soc. Am. **55**, 838–844.

Wakita, H. (**1973**). "Direct estimation of vocal tract shape by inverse filtering of acoustic speech waveforms," IEEE Trans. Audio Electroacoust. **AU-21**(5), 417–427.