OBTAINING LIP AND GLOTTAL REFLECTION COEFFICIENTS FROM VOWEL SOUNDS

Huiqun Deng[1], Rabab K. Ward[2], Michael P. Beddoes[2], Douglas O'Shaughnessy[1]

[1]INRS-EMT 800 de la Gauchetiére Ouest, bureau 6900, Montréal H5A 1K6 Canada
[2]Electrical and Computer Engineering Department, University of British Columbia, BC V6T 1Z4, Canada

ABSTRACT

Knowledge about lip and glottal reflection coefficients during phonation is needed to eliminate their distortion effects on the estimates of vocal-tract area functions and glottal waves from vowel sounds. Direct measurements of these coefficients at human mouths are difficult. This paper presents a method for estimating them from vowel sounds. The estimation encounters an ill-defined inverse problem: the number of unknowns is greater than the number of constraints, and non-unique solutions exist for a sound. To overcome this problem, this paper uses a vowel sound produced by a subject whose vocal-tract area function (VTAF) for the sound is known. The estimates of the lip and the glottal reflection coefficients are determined as those that lead to a VTAF solution most similar to the known VTAF for the sound. The lip and the glottal reflection coefficients obtained for /a/ and /i/ are presented.

1. INTRODUCTION

It is known that to obtain accurate estimates of vocal-tract area functions (VTAFs) and glottal waves from vowel sounds, the effects of incomplete glottal closures and frequency-dependent lip reflection coefficients contained in the vocal-tract filter (VTF) estimates must be eliminated [1], [2]. To do so, the parameters of the lip and the glottal reflection coefficients must be known. However, previous knowledge about a glottal reflection coefficient r_g and a lip reflection coefficient r_{lip} is based on some simplified models for glottal impedances and lip radiation impedances. Glottal impedances were derived from a rectangular slit model, and lip radiation impedances were approximated as those of spherical sources, or those of pistons in a sphere or in a baffle [3]. Experiments show that such simplified models cannot lead to satisfactory results in the estimation of VTAFs. More accurate knowledge about r_g and r_{lip} can be obtained by directly measuring r_g and r_{lip} at the glottis and the lip opening of a human speaker. However, such measurements are dangerous and difficult.

This paper presents a signal processing method for estimating r_g and r_{lip} from vowel sounds. As shown later, determining r_g and r_{lip} from a vowel sound needs to solve an underdetermined system of equations about r_g, r_{lip} and the VTAF, and there are non-unique sets of solutions of r_g, r_{lip} and VTAF for a sound. To overcome this problem, this paper obtains r_g and r_{lip} from the vowel sound of a subject whose VTAF for the sound has been obtained using a MRI (magnetic resonance imaging) method. The known vocal-tract area function is used as a guide in determining the estimates of r_g and r_{lip}: the best estimates of r_g and r_{lip} should lead to the VTAF solution that is the most similar to the known VTAF among those determined from other r_g and r_{lip} values. In the next section, the acoustic models for VTF estimates, r_g, r_{lip}, and vowel sound signals are presented. In section 3, the method for obtaining r_g, r_{lip}, and VTAFs from a vowel sound signal is developed. Section 4 presents the results obtained from vowel sounds /a/ and /i/. Section 5 contains conclusions.

2. MODELS

2.1 Glottal-vocal-tract filters and vocal-tract filters

The acoustic model for producing a vowel sound is shown in Fig. 1 [4], where $u_{sc}(t)$ is the equivalent glottal source signal, Z_g is the glottal impedance, $u_g(t)$ is the glottal wave (the total volume velocity at the back end of the vocal tract), $p_1(t)$ is the sound pressure at the back end of the vocal tract, Z_{lip} is the lip radiation impedance, and $u_{lip}(t)$ and $p_{lip}(t)$ are the total volume velocity and sound pressure at the lip opening, respectively. The transfer function from the glottal source to the lip volume velocity is defined as a glottal-vocal-tract filter (GVTF). The transfer function from the glottal wave to the volume velocity at the lip opening is defined as a vocal-tract filter (VTF).

In the discrete-time domain, the vocal tract is modeled as an acoustic tube with M sections each having the same length and a different cross-sectional area. If the vocal-tract length L and the number of sections of the tube model are related as $M=2LF_s/c$, where F_s is the sampling rate of the sound, c is the speed sound, then the Z transform of the GVTF transfer function is [5], [6]:

$$H_{GVTF}(z) \equiv \frac{U_{lip}(z)}{U_{sc}(z)} = \frac{0.5z^{-M/2}(1+r_g)(1+r_{lip})\prod_{m=1}^{M-1}(1+r_m)}{[1, r_g]\begin{bmatrix} 1 & r_1 \\ r_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & r_{M-1} \\ r_{M-1} z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ r_{lip} z^{-1} \end{bmatrix}}, \quad (1)$$

where

$$r_g = (Z_g - \rho c / S_1)/(Z_g + \rho c / S_1) \quad (2)$$

$$r_m = (S_{m+1} - S_m)/(S_{m+1} + S_m) \quad (3)$$

$$r_{lip} = (\rho c / S_M - Z_{lip})/(\rho c / S_M + Z_{lip}) \quad (4)$$

$r_1, \ldots, r_{M-1}$ are the reflection coefficients at each boundary of the tube model of the vocal tract, S_1 is the cross-sectional area at the backend of the vocal tract, and S_M is that near the lips.

The transfer function of a VTF does not contain the effect of an open glottis, and should be estimated from the sound signal

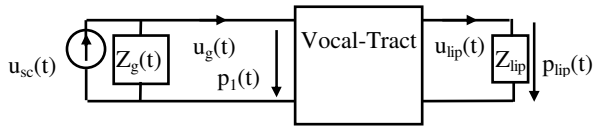

Fig. 1. The acoustic model for vowel sounds.

segments corresponding to closed glottal intervals, i.e., when $Z_g=\infty$. The transfer function of a VTF equals to that of a GVTF corresponding to $r_g=1$:

$$H_{VTF}(z)\equiv\frac{U_{lip}(z)}{U_g(z)}=\frac{z^{-M/2}(1+r_{lip})\prod_{m=1}^{M-1}(1+r_m)}{[1,1]\begin{bmatrix}1 & r_1\\ r_1z^{-1} & z^{-1}\end{bmatrix}\cdots\begin{bmatrix}1 & r_{M-1}\\ r_{M-1}z^{-1} & z^{-1}\end{bmatrix}\begin{bmatrix}1\\ r_{lip}z^{-1}\end{bmatrix}}. \tag{5}$$

2.2 Glottal reflection coefficients

In the discrete-time domain, r_g can be modeled using its Z transform, which can be obtained via the bilinear transform:

$$j\omega=2(1-z^{-1})F_s/(1+z^{-1}), \tag{6}$$

where ω is the angular frequency. Let R_g be the glottal resistance, and L_g be the glottal inductance, then, Eq. (2) becomes:

$$r_g=(R_g+j\omega L_g-\rho c/S_1)/(R_g+j\omega L_g+\rho c/S_1)\cdot \tag{7}$$

Thus, the Z transform of r_g is:

$$\begin{aligned} r_g(z)&=\frac{R_g-\rho c/S_1+L_g2F_s(1-z^{-1})/(1+z^{-1})}{R_g+\rho c/S_1+L_g2F_s(1-z^{-1})/(1+z^{-1})}\\ &=\frac{R_g-\rho c/S_1+L_g2F_s+(R_g-\rho c/S_1-L_g2F_s)z^{-1}}{R_g+\rho c/S_1+L_g2F_s+(R_g+\rho c/S_1-L_g2F_s)z^{-1}}\\ &\equiv K\frac{1+xz^{-1}}{1+yz^{-1}}, \end{aligned} \tag{8}$$

where x and y are parameters of $r_g(z)$. Since r_g at very high frequencies is one, thus K is determined as $K=(1-y)/(1-x)$ in order to have $r_g(z=-1)=1$. In this study, the parameters of $r_g(z)$ are to be estimated from a vowel sound.

2.3 Lip reflection coefficients

If the radiation impedance of the lip opening is approximated as that of a spherical sound source with the same area as the lip opening $S_M=\pi a^2$ in free field, then the lip radiation impedance is [3]:

$$Z_{lip}=\frac{\rho c}{S_M}\frac{jka/2}{1+jka/2}, \tag{9}$$

where a is the radius of the lip opening, $k=\omega/c$. Then, Eq. (4) becomes:

$$r_{lip}=1/(1+jka) \tag{10}$$

and the Z transform of r_{lip} obtained via bilinear transform is:

$$r_{lip}(z)=\frac{1+z^{-1}}{1+2aF_s/c+(1-2aF_s/c)z^{-1}}. \tag{11}$$

Considering that lip radiation impedances are more complicated than those of spherical sources, especially at higher frequencies, where the sound wavelength is on the order of head dimensions or less, this paper treats the parameters of $r_{lip}(z)$ as unknowns, and a lip reflection coefficient is modeled using an IIR filter:

$$r_{lip}(z)=\mu(1+\beta z^{-1})/(1+\alpha z^{-1}), \tag{12}$$

where α and β depend on the lip-opening area and are to be determined. Since lip radiation impedances increase with frequency, thus r_{lip} is a stable low-pass filter, i.e., $-1<\alpha<1$, and $\alpha<\beta$. Since $r_{lip}=1$ at $f=0$ Hz, i.e., $r_{lip}(z=1)=1$, thus, $\mu=(1+\alpha)/(1+\beta)\cdot$

2.4 Vowel sound signals

The Z transform of a discrete-time vowel-sound signal $p(n)$ in front of the lips can be modeled as [2] [3]:

$$P(z)=\frac{\rho}{4\pi r}z^{-rF_s/c}(1-z^{-1})U_g(z)H_{VTF}(z), \tag{13}$$

where $P(z)$ is the Z transform of $p(n)$, r is the distance from the lips to the microphone, c is the speed of sound, and ρ is the air density.

3. DERIVING GLOTTAL AND LIP REFLECTION COEFFICIENTS FROM VTF ESTIMATES

It is noted that glottises can hardly be completely closed during phonation. Thus, an $H_{VTF}(z)$ estimate obtained from sounds during closed glottal intervals still contains the effect of incomplete glottal closures, and is in fact equal to an $H_{GVTF}(z)$. The $H_{GVTF}(z)$ corresponding to a glottal area and a VTAF can be modeled by substituting $r_g(z)$ and $r_{lip}(z)$ into Eq. (1):

$$\begin{aligned} H_{GVTF}(z)&=\frac{0.5z^{-M/2}(1+K+(y+Kx)z^{-1})(1+\mu+(a+\beta\mu)z^{-1})\prod_{m=1}^{M-1}(1+r_m)}{[1+yz^{-1},\ K(1+xz^{-1})]\begin{bmatrix}1 & r_1\\ r_1z^{-1} & z^{-1}\end{bmatrix}\cdots\begin{bmatrix}1 & r_{M-1}\\ r_{M-1}z^{-1} & z^{-1}\end{bmatrix}\begin{bmatrix}1+\alpha z^{-1}\\ \mu(1+\beta z^{-1})z^{-1}\end{bmatrix}}\\ &\equiv\frac{0.5z^{-M/2}(1+K+(y+Kx)z^{-1})(1+\mu+(a+\beta\mu)z^{-1})\prod_{m=1}^{M-1}(1+r_m)}{1+\sum_{m=1}^{M+2}b_mz^{-m}}. \end{aligned} \tag{14}$$

As shown in Eq. (14), b_m is a non-linear function of x, y, r_1, ..., r_{M-1}, α, and β, which we denote as $b_m(x, y, r_1,....., r_{M-1}, \alpha, \beta)$. b_m's can be estimated from a sustained vowel sound using the method in [6]. Let the estimated values of b_m's be h_1, ..., h_{M+2}. The goal here is to determine the parameters of $r_g(z)$ and $r_{lip}(z)$ from h_m's such that b_m's determined by Eq. (14) satisfy the following equations:

$$\begin{aligned} &f_1=b_1(x,y,r_1,...,r_{M-1},\alpha,\beta)-h_1=0\\ &f_2=b_2(x,y,r_1,...,r_{M-1},\alpha,\beta)-h_2=0\\ &.....\\ &f_{M+2}=b_{M+2}(x,y,r_1,...,r_{M-1},\alpha,\beta)-h_{M+2}=0. \end{aligned} \tag{15}$$

One might think that r_m's are known from the VTAF measured using MRI, and the above $M+2$ non-linear equations are over-determined about x, y, α, and β. However, experiments have shown that the formant frequencies determined by the given r_m's and the resulting $r_g(z)$ and $r_{lip}(z)$ are higher than the formants observed from the speech sound. This difference can be explained by the fact that the acoustic length of the vocal tract is longer than its geometrical length, since there is end correction caused by the unknown lip radiation reactance. Thus, in this

study, r_m's, and the effective length of the acoustic tube model of the vocal tract, which determines M, are treated as unknowns. The acoustic length of the vocal tract is estimated from the average of lengths determined by the 3rd to 15th formant frequencies using the method in [7].

To solve for x, y, r_1,……, r_{M-1}, α, and β, at least M+3 constraints are needed, and one more equation is needed in addition to those in Eq. (15). This study first assumes a candidate value for α, and then solves for x, y, r_1,…, r_{M-1}, and β from Eq. (15). N candidates of α are selected according to Eq. (11):

$$\alpha_i = (1 - 2a_i F_s / c)/(1 + 2a_i F_s / c), \tag{16}$$

where a_i is the radius of an effective lip radiation area. Considering that the effective acoustic radiation area of the lip opening is larger than its geometrical area because of the reflection effects of the head and body on the lip radiation impedance, the radius of a candidate radiation area is determined as:

$$a_i = \sqrt{S_{lip}(1 + \Delta i)/\pi} \qquad i = 1,..., N \tag{17}$$

where S_{lip} is the geometrical area of the lip opening measured using MRI, Δ=0.01 is the step size for searching for the candidate radiation area, and N=600. Then, for each candidate α, this study solves for x, y, r_1,……, r_{M-1}, β from Eq. (15) by applying non-linear optimization:

$$\underset{x,y,\beta,r_1,.....,r_{M-1}}{Min} \sum_{m=1}^{M+2} f_m^{\,2} \tag{18}$$

under the following constraints:

$$\begin{aligned} &-1 < r_m < 1 \\ &\alpha < \beta < 6 \\ &-1 < y < 1 \\ &-1 < x < 1. \end{aligned} \tag{19}$$

The constraint $-1<r_m<1$ is from the physical meaning given by Eq. (3). The constraint $\alpha<\beta$ is due to the fact that $r_{lip}(z)$ is a stable low-pass IIR filter, and its pole is inside the unit circle and is closer to 1 than its zero. The constraint $\beta<6$ is from our experience. The constraint $-1<y<1$ is because the pole of a stable IIR is inside the unit circle. $-1<x<1$ is because the zero of a high-pass IIR filter is larger than the pole, and the angle of $r_g(z)$ at z=1 (zero frequency) is zero according to Eq. (8) since $R_g>\rho c/S_1$.

The non-linear optimization under the above constraints is implemented using the function "*lsqnonlin*" in the Optimization Toolbox in Matlab 7.1. Given an initial value vector, the vector of the objective functions f_m's defined in Eq. (15), and the Jacobin matrix determined by the derivatives of the objective functions, this function finds the solution that minimizes the sum of the squares of the M+2 functions f_m's as defined in Eq. (15). To reach the global optimization, as many as possible different initial values for r_1, ……, r_{M-1}, x, y and β should be given in their feasible ranges. Restricted by computing time, in this paper, 4 sets of initial values for them are used for each candidate α_i. $r_1^{\,0}$ …, $r_{M-1}^{\,0}$ =0; x^0= –0.8, –0.2; y^0=x^0+0.3; β^0=α_i+1, α_i+2, for i=1, …, N. Thus, 4N different sets of solutions of x, y, r_1, ……, r_{M-1}, α and β are obtained from Eq. (15). This study determines the best estimates for x, y, α, and β from the set whose r_m's lead to the VTAF most similar to the known VTAF among other VTAFs determined by other sets of solutions.

The similarity between two VTAFs can be measured using the correlation coefficient between their reflection coefficients. The above idea for determining the best estimate of α is implemented by:

$$\overset{*}{\alpha} = \underset{r_1(\alpha),.....,r_{M-1}(\alpha)}{ArgMax} \frac{\sum_{m=1}^{M'} (r_m(\alpha) R_m)}{\sqrt{\sum_{m=1}^{M'} (r_m(\alpha))^2} \sqrt{\sum_{m=1}^{M'} R_m^{\,2}}}, \tag{20}$$

where $r_m(\alpha)$'s are the solutions obtained from Eq. (15) given a candidate α and a set of the initial values, R_m's are reflection coefficients determined from the VTAF obtained using MRI, and M' is the number of sections determined from the geometrical length of the vocal tract. The estimates of β, x, y are determined from the solution set in which r_m's satisfy Eq. (20).

Given the estimates of $r_{lip}(z)$ and r_1,……, r_{M-1}, the $H_{VTF}(z)$ can be constructed according to Eq. (5). As a byproduct, the derivative of the glottal wave (DGW) can then be obtained by filtering the vowel sound using $1/H_{VTF}(z)$ according to Eq. (13).

4. RESULTS AND DISCUSSION

The sustained vowel sounds were produced in a sound-treated room by a male subject in a supine position, whose VTAFs have been obtained using the MRI method. The sounds were received using an AKG 300B microphone and recorded via Kay CSL 4400 to a disk. The signals were digitized at 16 bits/sample and a rate of F_s=44.1 kHz. Signal analysis shows that, for unknown reasons, the signals contain background noise, which increases as frequency decreases from about 3000 to 10 Hz. Before the estimations, each signal is decomposed into 8 layers of "Meyer" wavelet coefficients using the Matlab Wavelet Toolbox. Then, the wavelet coefficients in the layer of 0-86 Hz, which is below the fundamental frequency of the vowel sounds (about 120 Hz), are set to zeros. The inverse wavelet transform of the processed wavelet coefficients is the de-noised speech signal, keeping the speech signal and the higher frequency components of the noise unchanged. Some segments of the de-noised signals for /a/ and /i/ and the results obtained are shown in (a)–(f) of Figs. 2 and 3.

The value of M is estimated as 46 for /a/, and 43 for /i/. One set of estimates of b_m's for each sound can be obtained from two signal segments corresponding to two adjacent closed glottal intervals using the method in [6]. The closed glottal intervals were first estimated as marked using zero lines in Figs. 2(a) and 3(a). 304 sets of estimates of b_m's were obtained for /a/ and 137 sets for /i/. Due to the remaining background noise in the sound signal and the change in loudness and pitch of the "sustained" vowel sound, the estimates of b_m's vary for different analysis segment pairs. The b_m estimates obtained from different analysis segment pairs are shown by the dots in the m^{th} vertical line in Fig. 2(b) for /a/, and Fig. 3(b) for /i/. The value of h_m in Eq. (15) is the average of the b_m estimates and is marked by a circle, and the standard deviation of b_m estimates is marked by the cross in the m^{th} vertical line. "Pure" VTFs, which do not contain the effects of incomplete glottal closures, were then constructed according to Eq. (5) using the estimates of $r_{lip}(z)$ and r_1, …, r_{M-1} obtained using the method in section 3. The frequency responses of VTF estimates and the "pure" VTFs are plotted in solid lines and broken lines in Fig. 2(c) for /a/, and in Fig. 3(c) for /i/. The frequency responses of the estimated $r_{lip}(z)$ and $r_g(z)$ are shown using solid and broken lines in Fig. 2(d) for /a/ and Fig. 3(d) for

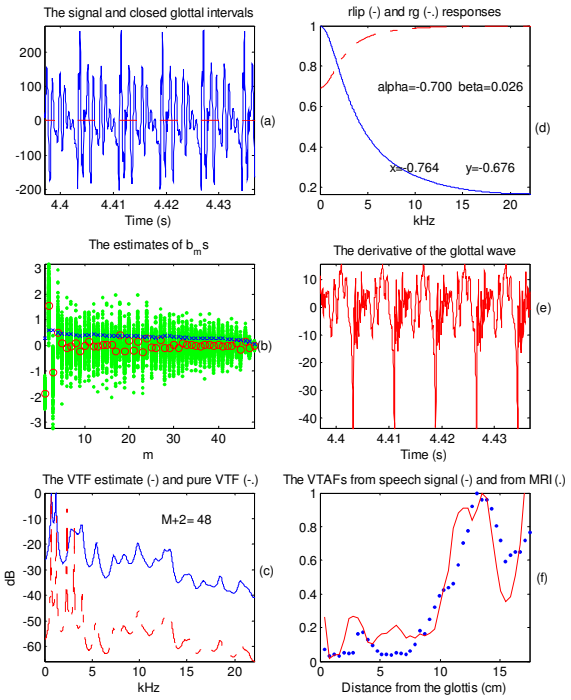

Fig. 2. The speech signal of /a/ and the estimates obtained.

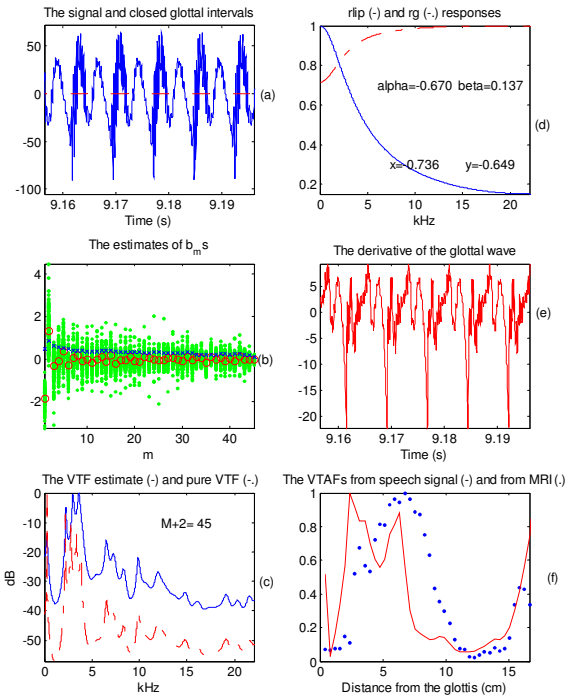

Fig. 3. The speech signal of /i/ and the estimates obtained.

/i/. The r_{lip} obtained for /a/ exhibits sharper low-pass frequency response than that for /i/. This is because of that the lip radiation impedance is a function of the product of frequency and the lip-opening area, and that the lip opening for /a/ is larger than that for /i/. The obtained derivatives of the glottal waves are shown in Figs. 2(e) and 3(e) for /a/ and /i/, respectively. The VTAFs obtained from MRI [8], and the first M' sections of VTAFs obtained from the vowel sounds are plotted using dotted and solid lines in Figs. 2(f) and 3(f). More accurate estimates can be obtained if the step size for the candidate α is finer, the recordings are less noisy and more sustained, and more initial values in the feasible ranges are given.

5. CONCLUSION

This paper presents a signal processing method for determining the parameters of the glottal and the lip reflection coefficients from vowel sounds, given the vocal-tract area functions. The relationship between the pole of the lip reflection coefficient and the lip opening area can be obtained by analyzing more vowel sounds with different lip openings in future. Such a relationship allows one to predict the pole of the lip reflection coefficient from the lip opening area, and to eliminate the effects of incomplete glottal closures and the frequency-dependent lip reflection coefficient on the VTAF and the glottal wave estimated from a vowel sound using the method shown above.

ACKNOWLEDGMENT

The authors thank professor B. H. Story in the University of Arizona for providing us with the speech signals for this paper.

REFERENCES

[1] H. Deng, R. Ward, M. Beddoes, M. Hodgson, "Effects of Glottal and Lip Boundary Conditions on Vocal-Tract Area Function Estimates from Speech Signals," Proceedings IEEE ICASSP, Philadelphia, USA, Mar. 2005, vol. I, pp. 901-904.

[2] H. Deng, R. Ward, M. Beddoes, "Glottal Waves via Inverse Filtering of Vowel Sounds," Proceedings 27th Annual International Conference of IEEE Engineering in Medicine and Biological Society, Sep. 1-4, 2005, Shanghai, China.

[3] J. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, New York, 1972.

[4] T. V. Ananthapadmanabha and G. Fant, "Calculation of True Glottal Flow and Its Components," *Speech Communication,* Vol. 1 pp. 167-184, 1982.

[5] L. R., Rabiner, and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, New Jersey, 1978.

[6] H. Deng, R. Ward, M. Beddoes, M. Hodgson, "A New Method for Obtaining Accurate Estimates of Vocal-tract Filters and Glottal Waves from Vowel Sounds," *IEEE Trans. on Speech and Audio Processing,*2006 (in press).

[7] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. on Acoust., Speech, Signal Processing*, Vol. ASSP-25, No. 2, pp.183-192, April 1977.

[8] B. H. Story, Titze, I. R. and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoustic Soc. Am.*, Vol. 100, No. 1, pp. 537-554, July 1996.