Estimating the Vocal-Tract Area Function and the Derivative of the Glottal Wave from a Speech Signal

*Huiqun Deng, Michael Beddoes, Rabab Ward, Murray Hodgson**

Electrical and Computer Engineering Department, *Mechanical Engineering Department
The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
huid@ece.ubc.ca, mikeb@ece.ubc.ca, rababw@icics.ubc.ca, hodgson@mech.ubc.ca

Abstract

We present a new method for estimating the vocal-tract area functions from speech signals. First, we point out and correct a long-standing sign error in some literature related to the derivation of the acoustic reflection coefficients of the vocal tract from a speech signal. Next, to eliminate the influence of the glottal wave on the estimation of the vocal-tract filter, we estimate the vocal-tract filter and the derivative of the glottal wave simultaneously from a speech signal. From the vocal-tract filter obtained, we derive the vocal-tract area function. Our improvements to existing methods can be seen from the vocal-tract area functions obtained for vowel sounds /ɑ/ and /i/, each produced by a female and a male subject. They are comparable with those obtained using the magnetic resonance imaging method. The derivatives of the glottal waves for these sounds are also presented, and they show very detailed structures.

1. Introduction

There are two signal-processing approaches for obtaining vocal-tract area functions from speech signals. The first derives the vocal-tract area function from formant frequencies estimated from a speech signal. This is a one-to-many problem; i.e., for one set of formant frequencies there are an infinite number of vocal-tract area functions corresponding to this set [1]. The other approach derives the vocal-tract area function directly from the acoustic speech waveform. This is a one-to-one problem under some conditions stated below. We take the second approach.

It is well known that the vocal tract acts as an acoustic filter to the glottal wave. The vocal tract can be modeled as a concatenation of M cylindrical acoustic tubes, each with equal length and a different cross-sectional area [2]. The reflection coefficients at each boundary of two adjacent sections determine the frequency response of the vocal-tract filter, and can be derived from the vocal-tract filter [3,4]. From the reflection coefficients, the relative vocal-tract area function can be obtained. Going from the speech waveform to the vocal-tract area function is a one-to-one mapping if, 1) the boundary condition of the vocal tract is known, 2) the vocal-tract is modeled as a concatenation of cylindrical tubes with equal length, and 3) the relationship $M=2LF_s/c$ is satisfied, where F_s is the sampling rate of the speech signal, L is the vocal-tract length, M is the number of sections of the tube model, and c is the sound speed.

There are two different assumptions about the vocal-tract boundary conditions made in deriving the vocal-tract area function from a speech signal. Boundary condition 1 means that the lip opening is terminated with zero acoustic impedance, and the glottal end is terminated with a characteristic impedance [3]. Boundary condition 2 means that the glottis is completely closed, and the lip opening is terminated with a characteristic impedance [3,4]. Atal and Hanauer [4] showed that if the vocal tract satisfies boundary condition 2, the reflection coefficients of the tube model can be derived from the coefficients of the vocal-tract filter obtained using LPC (Linear Prediction Coding). Wakita [3] showed that if the vocal tract satisfies boundary condition 1, the reflection coefficients of the tube model are related to k_m, where k_m is defined in Equation (11) in his paper to obtain the coefficients of the vocal-tract filter. Wakita also reported that the area functions estimated based on boundary condition 2 are not reasonable. However, it has also been reported that Wakita's method gives unreasonable results in many cases [5].

The two methods [3, 4] attempt to estimate vocal-tract area functions from speech waveforms. In applying these methods, however, we found three problems. First, we found a sign error in [3] and [4]. It appears in the equation of continuity of volume velocity, and in the equation of continuity of sound pressure. Second, according to the fact that the glottal waves for different sounds and subjects are different, we found that the influence of the glottal source on the estimation of the vocal-tract filter cannot be properly eliminated if the compensation for the glottal wave is based on certain assumptions. One such assumption is that the glottal wave has a spectrum with –12dB/oct. slope. Third, according to the fact that the glottis opens and closes alternatively during phonation, we found that the vocal-tract boundary condition is ambiguous or time varying over each pitch period. In section 2, we point out and correct the sign error. In section 3, we present our method for estimating the vocal-tract filter (VTF) and the derivative of the glottal wave (DGW). In section 4, we present the vocal-tract area functions, and the DGWs obtained for vowel sounds /ɑ/ and /i/ produced by a female and a male subject.

2. Correction of the sign error

In speech synthesis, the vocal tract can be modeled as a lossless, M-section cylindrical tube, as shown in Figure 1. $u_g(t)$ is the glottal wave, $u^+_m(t)$ and $u^-_m(t)$ are the positive-going and negative-going volume velocities at position x_m and at time t, respectively; the arrows indicate the transmission directions of the positive-going and of the negative-going sound waves (not the assumed directions of the volume velocities), respectively.

At two sides of each boundary, continuity of volume velocity must be satisfied:

$$u^+_{m+1}(t-D) + u^-_{m+1}(t+D) = u^+_m(t) + u^-_m(t) \qquad (1)$$

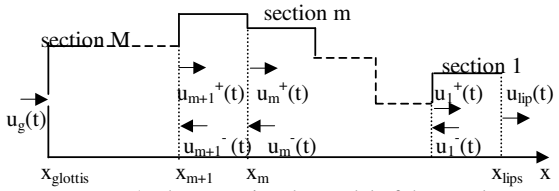

Figure 1. The acoustic tube model of the vocal tract.

and continuity of sound pressure must be satisfied:

$$\{u_{m+1}^+(t-D)-u_{m+1}^-(t+D)\}\rho c/S_{m+1}=\{u_m^+(t)-u_m^-(t)\}\rho c/S_m \quad (2)$$

where D is the sound-wave transmission time in each section, ρc is the characteristic impedance of the air, S_m and S_{m+1} are the m^{th} and $m+1^{th}$ cross-sectional areas, respectively. The volume velocity reflection coefficient at the boundary of the m^{th} and $m+1^{th}$ sections can be shown to be:

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

Combining Equations (1), (2) and (3), we get:

$$u_m^+(t)=(1+r_m)u_{m+1}^+(t-D)-r_m u_m^-(t) \quad (4)$$

$$u_{m+1}^-(t+D)=r_m u_{m+1}^+(t-D)+(1-r_m)u_m^-(t) \quad (5)$$

In addition to the reflections at each boundary of two adjacent sections, there are reflections from the lip opening and the backend of the vocal tract. The lip reflection coefficient r_{lip}, and the glottal reflection coefficient r_g are given in [6]. The lip volume velocity is related to the sound pressure in front of the lips by a delayed derivative factor [7]. Combining the relations in Equations (4) and (5), for discrete-time signals, we relate the glottal wave to the sound pressure in front of the lips by the signal flow diagram shown in Figure 2. $U_g(z)$, $U_m^+(z)$, $U_m^-(z)$ and P(z) are the Z transforms of $u_g(t)$, $u_m^+(t)$, $u_m^-(t)$, and of the sound pressure in front of the lips p(t), $K(1-z^{-1})z^{-\Delta}$ is the transfer function from the lip volume velocity to the sound pressure at the microphone, where $\Delta=rF_s/c$, r is the distance between the lips and the microphone, F_s is the sampling rate of the signal, c is the sound speed, and $K=\rho/4\pi r$. It should be noted that, in Equations (1) and (2), the assumed directions for $u_m^+(t)$ and $u_m^-(t)$, are the same. This is the convention used in acoustic textbooks [8]. The total volume velocity at time t at position x_m is the sum of $u_m^+(t)$ and $u_m^-(t)$. The negative-going sound pressure is the product of $u_m^-(t)$ and the negative of the characteristic impedance [8].

In Wakita's paper [3], the equations of continuity of volume velocity and of sound pressure are incorrect:

$$u_{m+1}^+(t-D)-u_{m+1}^-(t+D)=u_m^+(t)-u_m^-(t) \quad (a)$$

$$\{u_{m+1}^+(t-D)+u_{m+1}^-(t+D)\}\rho c/S_{m+1}=\{u_m^+(t)+u_m^-(t)\}\rho c/S_m \quad (b)$$

Some people would think that the assumed directions of $u_m^+(t)$ and $u_m^-(t)$ are different in [3], and it is a matter of convention, not a matter of fault.

We believe that two different conventions are acceptable only if they lead to the same result. If we follow Wakita's convention, the reflection coefficient r_m is linked by [3]:

$$r_m=k_{m-1} \quad m=1,\ldots\ldots,M \quad (c)$$

where k_m is defined in Equation (11) in [3]. On the other hand, if we follow the convention that the assumed directions of $u_m^+(x,t)$ and of $u_m^-(x,t)$ are the same, as used in acoustic textbooks, we obtain different reflection coefficients:

$$r_m=-k_{m-1} \quad m=1,\ldots\ldots,M \quad (6)$$

We believe that only the convention that satisfies physical principles is correct. In modeling the vocal-tract filter, the linear superposition principle must be satisfied. At the left side of the lip boundary, two components of the sound pressure $p^+(x_{lip},t)$ and $p^-(x_{lip},t)$ drive the air from the left side of the lip boundary. $p^+(x_{lip},t)$ produces $u^+(x_{lip},t)$, and $p^-(x_{lip},t)$ produces $u^-(x_{lip},t)$. If we follow the convention that the assumed directions of $u^+(x_{lip},t)$ and of $u^-(x_{lip},t)$ are the same, the lip volume velocity is $u^+(x_{lip},t)+u^-(x_{lip},t)$, and the linear superposition principle is satisfied. But, if we follow Wakita's convention that the assumed directions of $u_m^+(t)$ and of $u_m^-(t)$ are different, the lip volume velocity is $u^+(x_{lip},t)-u^-(x_{lip},t)$, and the linear superposition principle is violated. Therefore, Wakita's convention is incorrect. We should follow the convention used in Equations (1) and (2). The correctness of our correction can also be seen from the vocal-tract area functions obtained in section 4.

3. Estimating the vocal-tract filter and the glottal wave

Accurate estimation of the vocal-tract area function requires accurate estimation of the VTF. In this section, we estimate the VTF and the DGW simultaneously without assuming the nature of the shape of the glottal wave. The relationship between the sound pressure in front of the lips and the glottal wave is used in the estimation.

At a sampling rate of $F_s=cM/(2L)$, the Z transform of the time delay factor D in Figure 2 is $z^{-1/2}$, and the Z transform of the vowel sound signal can be represented as [6]:

$$P(z)=K\frac{0.5(1+r_g)(1+r_{lip})\prod_{m=1}^{M-1}(1+r_m)z^{-M/2-\Delta}(1-z^{-1})U_g(z)}{1-\sum_{m=1}^{M}a_m z^{-m}} \quad (7)$$

$z^{-M/2-\Delta}(1-z^{-1})U_g(z)$ can be viewed as the delayed DGW. If the frequency dependence of r_g and r_{lip} are negligible, the numerator of Equation (7) is proportional to the delayed DGW. Therefore, the vowel sound signal can be viewed as the convolution of the delayed DGW and an all-pole filter.

Assuming the length of one period of the speech signal is N, we represent the sampled waveform of the delayed DGW as:

$$u'_g(n-M/2-\Delta)=b_0\delta(n)+.....+b_{N-1}\delta(n-N+1) \quad (8)$$

According to Equation (7), the time-domain sound pressure sample at time n can be represented by:

$$p(n)=b_0\delta(n)+b_1\delta(n-1)+.....+b_{N-1}\delta(n-N+1)+a_1p(n-1)+a_2p(n-2)+.....+a_M p(n-M) \quad (9)$$

where the b_i's contain the product of coefficients in the numerator in (7), p(n), p(n-1),……, p(n-M) are known samples of the speech signal, b_0, b_1,……, b_{N-1} are the unknown samples of the DGW, and a_1, a_2, ……, a_M are unknown coefficients of the VTF (the all-pole part of Equation (7)). Mathematically, to solve for N+M unknown variables, at least N+M independent equations are needed. To have enough equations relating these unknown variables and the known speech signal samples, we used the same DGW waveform $\{b_0, b_1, b_2, \ldots\ldots, b_{N-1}\}$ in both the current pitch period and the following pitch period, assuming the DGW is periodic for a sustained vowel sound. We construct N+M+Q linear equations relating a_m's and b_i's to the speech signal samples p(n), n=0, 1, ……N+M+Q-1, as below:

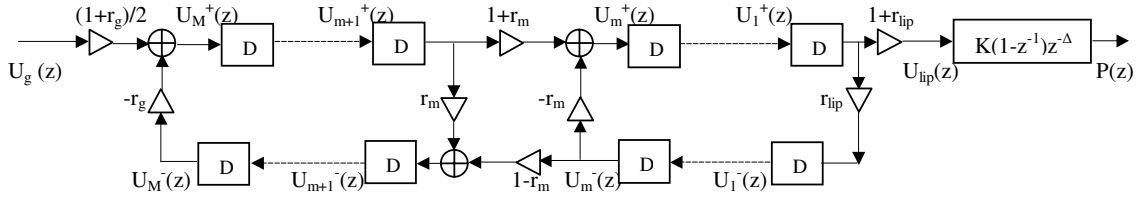

Figure 2. The signal flow diagram of the vowel sound production system.

$$\begin{bmatrix} 1 & 0 & . & . & 0 & p(-1) & . & . & p(-M) \\ 0 & 1 & 0 & . & 0 & p(0) & . & . & p(-M+1) \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & . & . & 1 & p(N-2) & . & . & p(N-M-1) \\ 1 & 0 & . & . & 0 & p(N-1) & . & . & p(N-M) \\ . & . & . & . & . & . & . & . & . \\ 0 & . & 1 & . & 0 & p(N+M+Q-2) & . & . & p(N+Q-1) \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ . \\ b_{N-1} \\ a_1 \\ . \\ a_M \end{bmatrix} = \begin{bmatrix} p(0) \\ p(1) \\ . \\ p(N-1) \\ p(N) \\ . \\ p(N+M+Q-1) \end{bmatrix} \quad (10)$$

There are N+M+Q rows and M+N columns in the first matrix. Mathematically, solving Equation (10) is a least squares problem. Larger Q can reduce the effect of noise on the solution. More periods of the DGW waveform can be obtained by making N equal to the corresponding length in which multiple pitch periods are contained. From our experience, the starting edge of the analysis window, which we refer to as n=0, should be selected not close to an amplitude peak of the speech signal. Otherwise, the resulting vocal-tract filter may be unstable. As can be seen from Equation (10), the waveform of the DGW over one or more periods, and the coefficients of the VTF can be solved simultaneously from the over-determined linear equation constructed using speech signal samples of the sustained vowel sound. The waveform of the glottal wave can be obtained by integrating the waveform of the DGW using the filter $1/(1-z^{-1})$.

4. Results and discussion

After obtaining the coefficients a_m's of the VTF using (10), we obtain the impulse response of the all-pole VTF. Then, we obtain k_m from the impulse response of the all-pole VTF [3]. We obtain the reflection coefficients using Equation (6). The relative vocal-tract area function can be obtained iteratively using:

$$S_{m+1} = S_m \frac{1 - r_m}{1 + r_m} \qquad m = 1,....., M-1 \quad (11)$$

where S_1=1, and m is the section number of the tube model. The numbering of the sections depends on the boundary condition of the vocal tract. If the vocal tract is assumed to satisfy boundary condition 1, m increases from the lips to the glottis [3]. While, if the vocal tract is assumed to satisfy boundary condition 2, m increases from the glottis to the lips [4]. We found that if we number the sections of the tube model from the lips to the glottis and use Equations (6) and (11), the resulting area function is unreasonable. However, if we number the sections from the glottis to the lips, the resulting area functions are reasonable. This means boundary condition 2 is closer to the actual boundary conditions of the vocal tract than boundary condition 1.

The speech signals used are sustained vowel sounds, and are recorded at sampling rate F_s=48kHz. The number of sections of the multi-sectional tube model should equal M = $2LF_s$ /c. The sound speed is approximately c=340 m/s. The lengths of the vocal tracts are different for different sounds and subjects [11]. For the female subject, the vocal-tract length L is assumed to be 14.9 cm (/ɑ/) and 14.2 cm (/i/); while for the male subject L= 17.7 cm (/ɑ/) and 16.3 cm (/i/). Hence, M=42 (/ɑ/), M=40 (/i/) for the female subject's tube model, and M=50 (/ɑ/), M=46 (/i/) for the male subject's tube model. The pitch periods are 161 samples (/ɑ/), 158 samples (/i/) for the female subject, and 325 samples (/ɑ/), 264 samples (/i/) for the male subject. In the estimation, five pitch periods of speech signal samples are used. The frequency responses of the VTFs, the vocal-tract area functions and the DGWs for the vowel sounds /ɑ/, and /i/ by female and male subjects are shown in Figures 3-6.

The DGWs obtained contain more detailed structure than the characteristics of DGW described in [9]. The sharp negative peaks correspond to the closure instants of the glottis. Positive values indicate the glottis is opening, and negative values indicate the glottis is closing. Where the values of the DGW are close to zero corresponds to the closed phase of the glottis. The duration of the closed phase is shorter for higher pitch voices. Also, from the DGWs obtained, we found that, in closed phases of the glottis, the DGWs are not zero. This may be because that the vocal cords provide non-zero volume velocity input to the vocal tract, or that the glottis is not completely closed. The differences between the DGWs for /ɑ/ and /i/ by the same subject can be explained by the interactions between the glottal source and the vocal tract [10]. The differences for the same vowel by different subjects are due to individual differences.

The resonance frequencies of the VTFs obtained are consistent to those in [12]. The vocal-tract area functions obtained for the male subject are comparable with those obtained for a male subject using the magnetic resonance imaging (MRI) method [11]. The differences between the vocal-tract area functions obtained using our method and those obtained using the MRI method are due to individual differences and the glottal loss. In estimating the VTF, the vocal tract is modeled as a lossless tube that satisfies boundary condition 2. However, in reality, the glottis opens and closes alternatively in each pitch period, and introduces non-infinite glottal impedance. The VTFs obtained from one or several pitch periods contain the influence of the glottal impedance, which damps the bandwidth of the first resonance, as can be seen from the frequency responses of the VTFs. Consequently, the derived vocal-tract area function can be degraded. More accurate estimation of the vocal-tract area function needs more accurate estimates of VTFs, in which not only the influence of the glottal wave, but also that of the glottal impedance is eliminated.

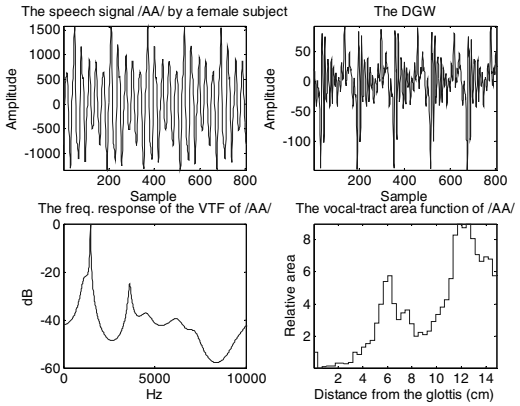

Figure 3. The estimates for /ɑ / of a female subject.

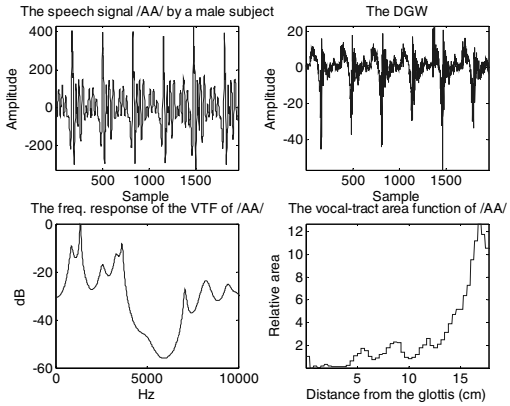

Figure 5. The estimates for /ɑ/ of a male subject.

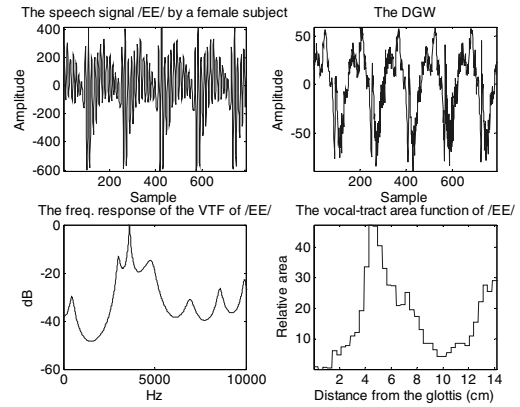

Figure 4. The estimates for /i/ of a female subject.

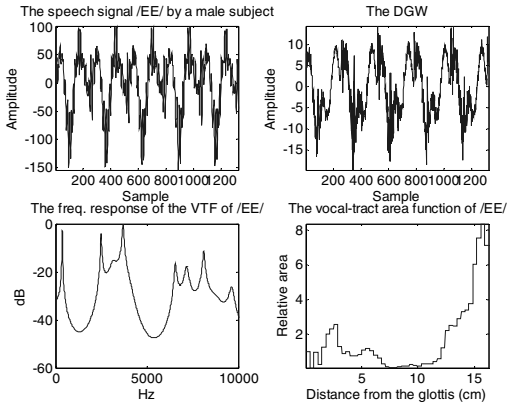

Figure 6. The estimates for /i/ of a male subject.

5. Conclusions

We made improvements to existing methods in three aspects. First, we corrected the long-standing sign error in the vocal-tract reflection coefficients given in the literature [3,4]. Second, we eliminated the influence of the glottal wave on the estimation of the vocal-tract filter. Third, assuming the vocal tract satisfies boundary condition 2, we derived the vocal-tract area functions from the vocal-tract filters. The vocal-tract area functions we obtained for the vowel sounds /ɑ/ and /i/ are comparable with those obtained with the MRI method. This means that our improvements to and correction of existing methods are effective. We believe that our method can be further improved by eliminating the influence of the glottal loss. This will be our future research.

6. References

[1] Mermelstein, P., "Determination of Vocal Tract Shape from Measured Formant Frequencies", *J. Acoust. Soc. Amer., Vol. 41, 1967, p 1283-1294.*

[2] Kelly, J. L. Jr., and Lochbaum, C. C., "Speech Synthesis", *Fourth International Congress on Acoustics, Copenhagen, p. G42, Aug. 21-28, 1962.*

[3] Wakita, H., "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", *IEEE Trans. Audio Electroacoust. Vol.AU-21:417-427, 1973.*

[4] Atal, B. S. and Hanauer, L. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Amer., Vol. 50, Number 2 (part 2), 1971, p 637-655.*

[5] Ray, G. C., "Determination of the Area Function of Individual Vocal Tract From Average for Sustained Vowels," *Proceedings RC IEEE-EMBS & 14th BMESI, India, p 2/78 -2/79, 1995.*

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

[7] Flanagan, J. L., *Speech Analysis Synthesis and Perception,* Springer-Verlag, 1972.

[8] Fahy, F., *Foundations of Engineering Acoustics*, Academic Press, London, 2001.

[9] Quatieri, T. F., *Discrete-time Speech Signal Processing,* Prentice Hall, New Jersey, 2001.

[10] Childers, D. G. and Wong, C., "Measuring and Modeling Vocal Source-Tract Interaction", *IEEE Trans. Biomedical Engineering, Vol. 41: 663-671, 1994.*

[11] Story, B. H., and Titze, R., "Vocal tract area functions from magnetic resonance imaging", *J. Acoust. Soc. Amer., Vol. 100,1995, p 537-554.*

[12] Deller, J.R Jr., Proakis, J. G. and Hansen, J. H., *Discrete-time Processing of Speech Signals*, Prentice-Hall, 1993.