

# ESTIMATING VOCAL-TRACT AREA FUNCTIONS FROM VOWEL SOUND SIGNALS OVER CLOSED GLOTTAL PHASES

Huiqun Deng, Rabab K. Ward, Michael P. Beddoes, Murray Hodgson\*

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

## ABSTRACT

Existing methods that estimate the vocal-tract area functions (VTAFs) from vocal-tract filters (VTFs) using speech signals suffer from inadequate elimination of the glottal wave, and the influence of non-ideal vocal-tract boundary conditions. To minimize these effects on the VTF estimation, we present a method that jointly estimates the glottal wave and the VTF corresponding to closed glottal phases. Experimental results show that our method yields better estimates. The VTAFs obtained for /a/ and /i/ each produced by a female and a male subject show that more detailed and accurate VTAFs are obtained using the speech signals corresponding to closed glottal phases.

## 1. INTRODUCTION

Vocal-tract area functions (VTAFs) are needed in many applications, such as speech synthesis, speech recognition, language training, etc. Direct measurements of VTAFs using MRI or X-ray methods are expensive and time consuming. The VTAF, however, can be derived from the vocal-tract filter (VTF), which is estimated from the speech signal [1,2]. There are two methods for deriving the VTAF from the VTF, based on two different assumptions about the vocal-tract boundary conditions. The first method uses **assumption 1** that the glottis is completely closed (or the glottal reflection coefficient  $r_g$  is one), and the lip opening is terminated with some characteristic impedance [1]. The second method uses **assumption 2** that the glottis is terminated with some characteristic impedance, and the lip is terminated with zero impedance (or the lip reflection coefficient  $r_{lip}$  is one) [2].

In applying the above methods, there are some factors that degrade the VTAF estimates obtained from speech signals. Firstly, in estimating the VTF from a speech signal, the unknown glottal wave degrades the estimation of the VTF. In the literature, the compensation for the glottal wave is based on some simplified assumptions, such as zero-flow over the closed glottal phases, or parametric models of the derivative glottal wave. This inadequate compensation for the glottal wave leads to inaccurate estimate of the VTF. Secondly, the assumed vocal-tract boundary conditions cannot be satisfied at all times. The glottis closes and opens periodically during phonation. The lip radiation impedance can only be approximated by zero (or  $r_{lip}$  is

one) at very low frequencies. It is reported that assumption 1 cannot lead to reasonable results, and that reasonable VTAF estimates are obtained based on assumption 2 [2]. However, as shown later, assumption 2 cannot lead to detailed estimates of the VTAFs, and assumption 1 can lead to detailed and reasonable results if the VTF used for deriving the VTAF is estimated over closed glottal phases.

From the above, it is clear that accurate estimation of the VTAF from the speech signal requires that 1) the influence of the glottal wave on the estimation of the VTF be eliminated; and 2) the VTF used correspond to the assumed vocal-tract boundary condition. To overcome these problems, we eliminate the influence of the glottal wave on the VTF estimation through their joint estimation without using assumptions about the glottal wave [5]. Furthermore, we estimate the VTF over closed glottal phases, in order to carry out the VTAF estimation based on assumption 1. The reason why we prefer assumption 1 over assumption 2 is shown in the following section.

## 2. ESTIMATING THE VTAF FROM THE VTF

In this section, we review and compare the two methods [1,2]. The vocal tract can be modeled as a multi-sectional cylindrical tube, with each section having the same length and a different cross-sectional area [3], as shown in Figure 1, where  $S_m$  is the cross-sectional area of section  $m$ . The cross-sectional area of the tube is regarded as the VTAF. The VTF is represented by the transfer function from the glottal volume velocity  $U_g$  (i.e., the glottal wave) to the lip volume velocity  $U_{lip}$ . The VTF is time varying even if the shape of the vocal tract is fixed for a sustained vowel sound, because the glottal impedance is time varying during phonation [5]. The VTAF can be represented using the reflection coefficients  $r_m$  of the tube:

$$r_m = (S_m - S_{m+1}) / (S_m + S_{m+1}) \quad m=1, 2, \dots, M-1. \quad (1)$$

For a vowel sound, the reflection coefficients of the tube model of the vocal tract can be obtained from the all-pole VTF if the vocal-tract boundary condition is known. Under assumption 2, the reflection coefficient of the multi-sectional tube can be obtained from [2]:

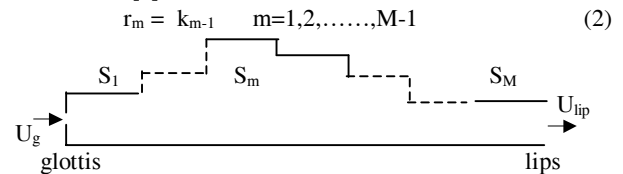
$$r_m = k_{m-1} \quad m=1, 2, \dots, M-1 \quad (2)$$


Fig. 1. The tube model of the vocal tract.

where  $m$  is the sectional number which increases from the lip end to the glottal end, and  $k_m$  is determined by the autocorrelation coefficient of the impulse response of the VTF. Under assumption 1, the reflection coefficient of the tube can be obtained from [1,2,6]:

$$r_m = -k_{m-1} \quad m=1,2,\dots,M-1 \quad (3)$$

where  $k_m$  is the same as in (2), but  $m$  increases from the glottal end to the lip end.

The two methods work under different boundary conditions; thus their capabilities for measuring detailed structures of the VTAF are different. Computer simulations [6] show that the VTAF estimation based on assumption 1 requires the glottal reflection coefficient  $r_g$  to be one (i.e., the glottis to be completely closed) and the lip reflection coefficient  $r_{lip}$  to be constant. This means that the VTAF estimation based on assumption 1 requires the VTF be estimated when the glottis is closed, and allows the signal to cover a wide frequency range (given  $r_g=1$  the influence of the frequency-dependent  $r_{lip}$  can be eliminated by post-processing). On the other hand, VTAF estimation based on assumption 2 requires  $r_{lip}$  to be one (i.e., the lip radiation impedance to be zero) and  $r_g$  to be constant, meaning that the speech signal should be in a very low frequency range, and the glottal area should be constant.

The frequency range of the signal used determines the resolution of the VTAF estimate. It is known that, in deriving the VTAF from the VTF, the section number of the tube model should be  $M=2LF_s/c$ , where  $F_s$  is the sampling rate of the speech signal,  $L$  is the vocal-tract length, and  $c$  is the sound speed [2]. Thus, to obtain more detailed structure of the VTAF, a higher sampling rate is needed. However, if a speech signal contains only low frequency components, higher than Nyquist sampling rate cannot help provide more information about the vocal tract. Since the VTAF estimation based on assumption 2 requires the signal to be limited to the low frequency range, this assumption cannot result in a detailed structure of the VTAF. For example, the glottal ends of the VTAF estimates appear unreasonably wider than actual shapes [2]. In contrast, the VTAF estimation based on assumption 1 allows the signal to have a wide frequency range, and thus it can obtain a more detailed structure of the VTAF, if the VTF corresponds to closed glottal phases. In the next section, we present the method for estimating the VTF over closed glottal phases.

### 3. ESTIMATING THE VTF OVER CLOSED GLOTTAL PHASES

In the literature, the VTF corresponding to closed glottal phases (VTF<sub>cgp</sub>) is estimated under the assumption that the glottal wave during closed glottal phases is always zero. This assumption is not always true. For some voices there are no completely closed glottal phases. We consider a non-zero glottal wave, while estimating the VTF over closed glottal phases to eliminate its influence on the VTF estimate.

At a sampling rate of  $F_s=Mc/(2L)$ , the Z transform of the vowel sound signal at the microphone is [4,5]:

$$P_{mic}(z) = K \frac{0.5(1+r_g)(1+r_{lip}) \prod_{m=1}^{M-1} (1+r_m) z^{-M/2-\Delta} (1-z^{-1}) U_g(z)}{1 - \sum_{m=1}^M a_m z^{-m}} \quad (4)$$

where  $a_1, \dots, a_M$  are the coefficients of the all-pole VTF,  $K=\rho/4\pi r$ ,  $r$  is the distance from the lips to the microphone,  $\rho$  is the air density,  $\Delta=rF_s/c$ . It is convenient not to consider the delay  $\Delta$ , by defining  $P(z) \equiv z^\Delta P_{mic}(z)$ , and

$$P(z) = K \frac{0.5(1+r_g)(1+r_{lip}) \prod_{m=1}^{M-1} (1+r_m) z^{-M/2} (1-z^{-1}) U_g(z)}{1 - \sum_{m=1}^M a_m z^{-m}} \quad (5)$$

The time domain equivalent of  $P(z)$  is  $p(n)=p_{mic}(n+\Delta)$ , referred to as the sound pressure at the lips, in this paper. The numerator of (5) can be viewed as the weighted delayed derivative of the glottal wave. Let  $u_g'(n-M/2)$  denote the time domain equivalent of the whole numerator in Eq. (5), the vowel sound signal  $p(n)$  in the time domain can be expressed as:

$$p(n) = u_g'(n-M/2) + a_1 p(n-1) + \dots + a_M p(n-M) \quad (6)$$

To estimate the VTF<sub>cgp</sub>, we choose the samples of  $p(n)$  in the interval  $[n_0, \dots, n_0+N-1]$  corresponding to the closed glottal phase within a period of the signal. Let the samples of the weighted delayed derivative of the glottal wave corresponding to these samples be  $u_g'(n_0-M/2), \dots, u_g'(n_0+N-1-M/2)$ . Then, the coefficients of the VTF<sub>cgp</sub> and the samples of  $p(n)$  are related as  $p_0 = U_g' + P_0 A$ :

$$\begin{bmatrix} p(n_0) \\ \vdots \\ p(n_0+N-1) \end{bmatrix}_{p_0} = \begin{bmatrix} u_g'(n_0-M/2) \\ \vdots \\ u_g'(n_0+N-1-M/2) \end{bmatrix}_{U_g'} + \begin{bmatrix} p(n_0-1) & \dots & p(n_0-M) \\ \vdots & \ddots & \vdots \\ p(n_0+N-2) & \dots & p(n_0+N-1-M) \end{bmatrix}_{p_0} \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix}_A \quad (7)$$

In Eq. (7),  $U_g'$  and  $A$  are to be estimated,  $P_0$  and  $p_0$  are the measurements obtained corresponding to one interval. For a sustained vowel sound, the pitch and the vocal tract are time-invariant. Thus, we approximate the derivative of the glottal wave corresponding to the intervals that have the same length and the same relative position as  $p_0$  in different periods using the same  $U_g'$ . To obtain the estimates of  $U_g'$  and  $A$ , more measurements are then needed. We use 6 sets of measurements taken from 6 closed glottal phase intervals in 6 periods:  $[n_0, \dots, n_0+N-1]$ ,  $[n_0+T, \dots, n_0+N-1+T]$ ,  $\dots$ ,  $[n_0+5T, \dots, n_0+N-1+5T]$ , where  $T$  is the period of the sustained vowel sound signal, and  $T>N$ . The derivatives of the glottal waves corresponding to these intervals are  $U_g'$ . For natural speech signals, samples do not repeat over pitch periods exactly, and thus they provide different measurement points  $(P_i, p_i)$ ,  $i=0, \dots, 5$ . Combining Eq. (7) with the other 5 equations corresponding to these intervals, an over-determined equation relating  $U_g'$ ,  $A$  and samples of  $p(n)$  is obtained:

$$\begin{bmatrix} 1 & 0 & 0 & p(n_0-1) & \dots & p(n_0-M) \\ 0 & 1 & \dots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & p(n_0+N-2) & \dots & p(n_0+N-1-M) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & p(n_0+5T-1) & \dots & p(n_0+5T-M) \\ 0 & 1 & \dots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & p(n_0+5T+N-2) & \dots & p(n_0+5T+N-1-M) \end{bmatrix}_{\hat{Q}} \begin{bmatrix} u_g'(n_0-M/2) \\ \vdots \\ u_g'(n_0+N-1-M/2) \\ a_1 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} p(n_0) \\ \vdots \\ p(n_0+N-1) \\ \vdots \\ p(n_0+5T) \\ \vdots \\ p(n_0+5T+N-1) \end{bmatrix}_{\hat{S}} \quad (8)$$

The column vectors in  $Q$  are linearly independent. Thus,  $U'_g$  and  $A$  corresponding to these intervals can be separated and obtained from the least squares error solution [7]:

$$\begin{bmatrix} U'_g \\ A \end{bmatrix} \approx (Q^T Q)^{-1} Q^T S \quad (9)$$

$U'_g$  and  $A$  can be viewed as optimum solutions of the delayed derivative glottal waveform, and of the coefficients of the VTF, respectively, corresponding to these intervals.

#### 4. LOCATING THE CLOSED GLOTTAL PHASES

The signal samples used in  $Q$  and  $S$  of Eq. (8) are crucial in obtaining an accurate  $VTF_{cgp}$  estimate. The glottis opens and closes periodically during phonation, and only these signal samples that do not contain the influence of the open glottis should be used in Eq. (8). Traditionally, glottal phases are detected using EGG (electroglottograph) signals. In our research, glottal phases are detected using the glottal waveform estimated from the vowel sound signal recorded at a distance from the lips,  $p_{mic}(n)$ . From a sustained vowel sound, we can obtain the weighted delayed derivative of the glottal wave signal  $u'_g(n-M/2-\Delta)$  [5]. The glottal waveform  $u_g(n-M/2-\Delta)$  can be obtained by integrating  $u'_g(n-M/2-\Delta)$ . The zero line of  $u_g(n-M/2-\Delta)$  cannot be recovered from the sound pressure signal due to the derivative factor in the transfer function from the lip volume velocity to the sound pressure at the microphone (see Eq. (4)). Because actual glottal volume velocity cannot be negative, the zero line of  $u_g(n-M/2-\Delta)$  is set to be the minimum value of  $u_g(n-M/2-\Delta)$ .

The closed glottal phases are determined according to the amplitude of the glottal wave. It is noted that for incomplete glottal closure, the glottal wave can hardly stay at zero. In this case, closed glottal phases are determined when the glottal wave amplitude is below half peak value. Assume over the interval  $[n_1, n_1+1, \dots, n_1+N_1-1]$ , the amplitude of  $u_g(n-M/2-\Delta)$  is below the level, then in the interval  $[n_1-M/2-\Delta, \dots, n_1+N_1-1-M/2-\Delta]$ , the amplitude of  $u_g(n)$  is below this level. The interval  $[n_1-M/2-\Delta, \dots, n_1+N_1-1-M/2-\Delta]$  is taken as the closed glottal phase, and is denoted as  $[n_{close}, \dots, n_{open}]$ , where  $n_{close} = n_1-M/2-\Delta$ ,  $n_{open} = n_1+N_1-1-M/2-\Delta$ .

The samples of  $p_{mic}(n)$  that do not contain the influence of the open glottis are used to construct Eq. (8) and estimate the  $VTF_{cgp}$ . This section shows how to locate these samples. At the instant  $n=n_{close}+3M/2$ , (i.e.,  $3M/2$  sampling periods after the closure instant), the sound pressure at the lips is  $p(n_{close}+3M/2) = u'_g(n_{close}+M) + a_1 p(n_{close}+3M/2-1) + \dots + a_M p(n_{close}+M/2)$ , with the last term leaving from the glottis at  $n_{close}$  and arriving at the lips at  $n_{close}+M/2$ , because the time for the sound wave to propagate through the vocal tract is  $M/2$  sampling periods (recall  $M=2LF/c$ ). While, at the instant  $n_{open}+M/2$ , the sound pressure at the lips is  $p(n_{open}+M/2) = u'_g(n_{open}) + a_1 p(n_{open}+M/2-1) + \dots + a_M p(n_{open}-M/2)$ , with the first term leaving from the glottis at  $n_{open}$  and arriving at the lips at  $n_{open}+M/2$ . Therefore, the  $p_0$  needed for estimating the  $VTF_{cgp}$  in Eq. (7) should be  $p_0 = [p(n_{close}+3M/2), \dots, p(n_{open}+M/2)]^T$ , and the length of  $p_0$  is  $N=N_1-M$ . Since  $n_{close} = n_1-M/2-\Delta$ ,  $p(n-\Delta) = p_{mic}(n)$ , then  $p_0 = [p_{mic}(n_1+M), \dots, p_{mic}(n_1+N_1-1)]^T$ . Other intervals for making the measurements are

determined by shifting multiple pitch periods relative to  $p_0$  interval. Constructing Eq. (8) with  $p(n) = p_{mic}(n)$ ,  $n_0 = n_1+M$ , and  $N = N_1-M$ , then  $a_1, \dots, a_M$ , can be solved.

A reasonable estimate of the  $VTF_{cgp}$  should be stable, and should not have sharp formants at high frequencies. The lip radiation damps the high frequency resonance of the VTF. Due to the noise and calculation error, the resulting  $VTF_{cgp}$  may be unstable, or may contain sharp formants at high frequencies (meaning unreasonably small damping). In this case, one needs to adjust the starting point  $n_1$  and the length of the analysis intervals  $N_1$ , or to modify the poles of the  $VTF_{cgp}$  estimate (not yet in this study), to get a better estimate of the  $VTF_{cgp}$  and hence a reasonable estimate of the VTAF.

#### 5. RESULTS AND DISCUSSION

The vowel sound signals produced by a female and a male subject are sampled at 48 kHz and 16 bits/sample using a DAT (digital tape recorder) in an anechoic room. The digital speech signal is then fed to the computer from the DAT via a sound card. The reflection coefficients of the tube model are derived from the  $VTF_{cgp}$  using Eq. (3). The VTAF is obtained from the reflection coefficients of the tube model given  $S_1=1$ :

$$S_{m+1} = S_m(1-r_m)/(1+r_m) \quad m=1,2,\dots,M-1 \quad (10)$$

where  $m$  increases from the glottis to the lips. For the female subject,  $M=42$  for /a/, and  $M=40$  for /i/; for the male subject,  $M=51$  for /a/, and  $M=46$  for /i/. The sectional length is  $c/(2F_s) = 340/(2 \times 48000) = 3.5$  mm.

The recorded speech signals  $p_{mic}(n)$ , the estimates of  $u'_g(n-M/2-\Delta)$ ,  $u_g(n-M/2-\Delta)$  and the VTAF for /a/ and /i/ each by a female and a male subject are shown in Figures 2-5. For the female subject, the length of the pitch period (samples) is  $T=161$ , the length of closed glottal phase (samples)  $N=80$ , for /a/;  $T=159$ ,  $N=70$ , for /i/. For the male subject,  $T=325$ ,  $N=70$ , for /a/;  $T=264$ ,  $N=70$ , for /i/. It is noted that the glottal waves obtained are different for different vowel sounds produced by the same subject, because of the interaction between the glottal source and the vocal tract. For the sounds /a/ by the two subjects, the glottal waves stay near zero for some time, indicating that incompletely closed glottal phases exist. For the sounds /i/, the glottal waves hardly stay near zero, meaning that there are no obvious closed glottal phases. In this case, the intervals when the amplitude of the glottal wave is below half the peak value are taken as closed glottal phases. The intervals corresponding to  $p_0, p_1, \dots, p_5$  are marked with solid lines.

In this paper, only plane waves are considered to contribute to the speech signal, and higher mode waves are ignored. In spite of this assumption, the VTAFs estimated from the vowel sounds /a/ and /i/ over closed glottal phases are comparable with the MRI [8] results. The discrepancy between the VTAF estimated in this paper and those measured using MRI is due to individual differences, the frequency-dependent lip reflection coefficient, the incomplete glottal reflection (caused by incomplete glottal closure), and the noise effect. The lip reflection coefficient acts as a low pass filter to the waves reflected from the lips into the vocal tract. More accurate VTAF estimates can be obtained by eliminating the influences of the frequency dependent lip reflection coefficient and of the glottal loss contained in the  $VTF_{cgp}$ . This is our future research.

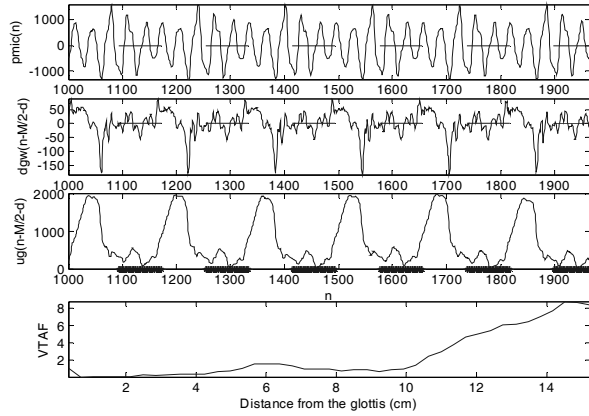


Fig. 2.  $p_{mic}(n)$ , the estimates of  $u_g'(n-M/2-\Delta)$ ,  $u_g(n-M/2-\Delta)$ , and VTAF of /a/ produced by a female subject.

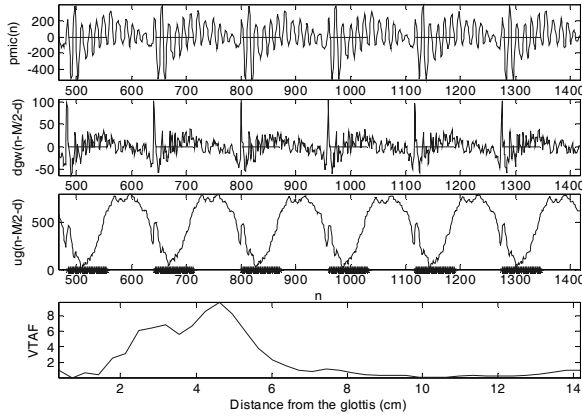


Fig. 3.  $p_{mic}(n)$ , the estimates of  $u_g'(n-M/2-\Delta)$ ,  $u_g(n-M/2-\Delta)$ , and VTAF of /i/ produced by a female subject.

## 6. CONCLUSION

Detailed and reasonable VTAF estimates are obtained from speech signals corresponding to closed glottal phases. The closed glottal phases are determined using glottal wave estimates obtained from vowel sounds only. Over the closed glottal phases, the non-zero glottal wave is jointly estimated with the  $VTF_{cgp}$ , which is used for deriving the VTAF. Comparing the resulting VTAFs with those obtained using MRI methods implies that our method yields reasonable and better estimates than others. Improvements can be made after further eliminating the influences of the frequency-dependent lip reflection coefficient and of the glottal loss in the  $VTF_{cgp}$ .

## 7. REFERENCES

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

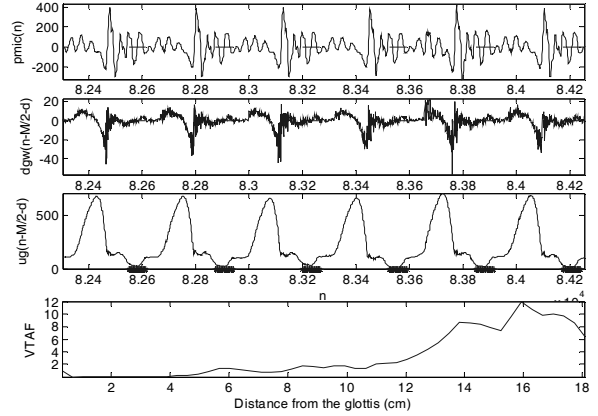


Fig. 4.  $p_{mic}(n)$ , the estimates of  $u_g'(n-M/2-\Delta)$ ,  $u_g(n-M/2-\Delta)$ , and VTAF of /a/ produced by a male subject.

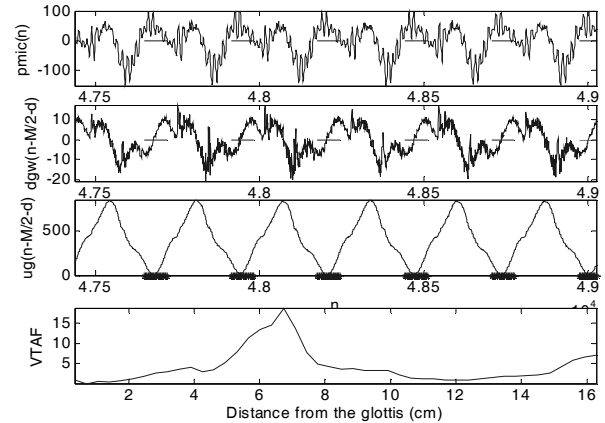


Fig. 5.  $p_{mic}(n)$ , the estimates of  $u_g'(n-M/2-\Delta)$ ,  $u_g(n-M/2-\Delta)$ , and VTAF of /i/ produced by a male subject.

Inverse Filtering of Acoustic Speech Waveforms," *IEEE Trans. Audio Electroacoust.* Vol. AU-21: 417-427, 1973.

- [3] J. L. Jr. Kelly, and C. C. Lochbaum, "Speech Synthesis," *Fourth International Congress on Acoustics*, Copenhagen, p. G42, Aug. 21-28, 1962.
- [4] Rabiner, L. R., and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, New Jersey, 1978.
- [5] H. Deng, M. P. Beddoes, R. K. Ward and M. Hodgson, "Estimating the Glottal Waveform and the Vocal-Tract Filter from a Vowel Sound," *IEEE PacRim Conf. on Comm., Comp. and Sig. Proc.*, pp. 297-300, Aug. 2003.
- [6] H. Deng, M. P. Beddoes, R. K. Ward, and M. Hodgson, "Estimating the Vocal-Tract Area Function from a Speech Signal," *Proceeding of Canadian Acoustic Week*, pp.40-41, Edmonton, Canada, Oct., 2003.
- [7] Hogben, L., *Elementary Linear Algebra*, West Publishing Company, MN, 1987.
- [8] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal Tract Area Functions from Magnetic Resonance Imaging", *J. Acoustic Soc. Am.*, Vol. 100, No. 1, July, 1996.