MODELLING ENERGY FLOW IN THE VOCAL TRACT WITH APPLICATIONS TO GLOTTAL CLOSURE AND OPENING DETECTION

D. M. Brookes, H. P. Loke
Dept of Electrical & Electronic Engineering, Imperial College,
Exhibition Road, London SW7 2BT, UK.
mike.brookes@ic.ac.uk, http://www.ee.ic.ac.uk/hp/staff/dmb/dmb.html

ABSTRACT

The pitch-synchronous analysis that is used in several areas of speech processing often requires robust detection of the instants of glottal closure and opening. In this paper we derive expressions for the flow of acoustic energy in the lossless-tube model of the vocal tract and show how linear predictive analysis may be used to estimate the waveform of acoustic input power at the glottis. We demonstrate that this signal may be used to identify the instants of glottal closure and opening during voiced speech and contrast it with the LPC residual signal that previous authors have used for this purpose.

1. INTRODUCTION

In voiced speech, the main acoustic excitations of the vocal tract occur at the instants of glottal closure and, to a lesser degree, opening. Determining these instants is important for pitch-synchronous speech analysis techniques such as closed-phase LPC. Such techniques are useful in speech coding, synthesis and prosody manipulation, and in determining speaker characteristics; see references in [7]. Previous attempts to identify these instants have either been based on the LPC residual or have made use of the laryngograph (EGG) signal [7, 5]. In this paper we examine the flow of energy in the lossless-tube model of the vocal tract and propose that the signal representing acoustic input power at the glottis be used to determine the instants of glottal closure and opening.

2. THE LOSSLESS TUBE MODEL

An N^{th}-order lossless-tube model of the vocal tract consists of N concatenated uniform sections having cross-sectional areas A_k, $k = 1, \ldots, N$ [9].

The junction between sections k and $k+1$ is characterized by a reflection coefficient:

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \quad \text{for} \quad k = 0, \ldots, N \tag{1}$$

A_0 and A_{N+1} represent the effective cross-sectional areas of the glottis and of the free space beyond the lips.

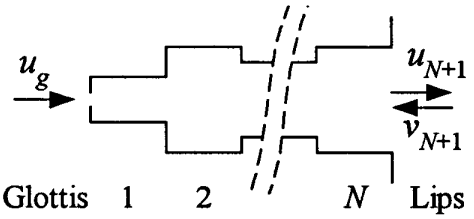


Figure 1. Lossless tube model of the vocal tract

Within each tube section, the acoustic signal is the superposition of two waves travelling in opposite directions at velocity $c \approx 340$ m/s. We define $u_k(t)$ and $v_k(t)$ to be the volume velocities (or volume flow rates) of the two waves at the glottis end of section k as shown in Figure 1 where the arrows associated with each wave indicate the direction both of propagation and of positive flow. At the lips, $u_{N+1}(t)$ is the acoustic output of the vocal tract and $v_{N+1}(t)$ is taken to be zero. Taking z-transforms, we may relate U_k and V_k to U_{N+1} by [9]

$$\begin{pmatrix} U_k \\ V_k \end{pmatrix} = \prod_{j=k}^{N} \frac{z^{1/2}}{1+r_j} \begin{pmatrix} 1 & -r_j \\ -r_j z^{-1} & z^{-1} \end{pmatrix} \times \begin{pmatrix} U_{N+1} \\ 0 \end{pmatrix} \tag{2}$$

and the forward travelling wave at the glottis is given by

$$U_g = \frac{U_1 - r_0 V_1}{1 + r_0} \tag{3}$$

If we neglect the delay and gain factors represented by the term $\frac{z^{1/2}}{1+r_j}$ we can derive from (2) and (3) an N^{th}-order FIR transfer function for the vocal tract

$$\frac{U_g}{U_{N+1}} = 1 + \sum_{j=1}^{N} a_j z^{-j} \tag{4}$$

If we assume that $r_0 = 1$ (or equivalently that $A_0 = 0$) then the remaining N reflection coefficients are uniquely defined by the coefficients a_j.

We note that some authors number the tube sections in the reverse order; the assumption $r_0 = 1$ is then equivalent to taking our $A_{N+1} = \infty$. Symmetries in (2) and (3) mean that the a_j are unaffected by reversing the order of the reflection coefficients.

3. ACOUSTIC ENERGY FLOW

The acoustic pressure at the glottis end of section k of the vocal tract is given by

$$p_k(t) = \frac{\rho c}{A_k}\left(u_k(t) + v_k(t)\right) \tag{5}$$

where ρ is the density of air. The corresponding acoustic energy density is given by [8]

$$q_k(t) = \tfrac{1}{2}\rho\left(\frac{u_k(t) - v_k(t)}{A_k}\right)^2 + \tfrac{1}{2}\frac{p_k^2(t)}{\rho c^2} \tag{6}$$

in which the first term represents the kinetic energy density and the second the potential energy density. Substituting (5) into (6) gives

$$q_k(t) = \frac{\rho}{A_k^2}\left(u_k^2(t) + v_k^2(t)\right) \tag{7}$$

which may be interpreted as the sum of two energy density components associated respectively with the forward and reverse waves $u_k(t)$ and $v_k(t)$ and travelling with them at a speed c.

At the lips, we have $v_{N+1}(t) \equiv 0$ so (7) reduces to

$$q_{N+1}(t) = \frac{\rho}{A_{N+1}^2} \times u_{N+1}^2(t) \tag{8}$$

and the acoustic output power from the vocal tract is given by

$$\begin{aligned} w_{N+1}(t) &= q_{N+1}(t) \times cA_{N+1} \\ &= \frac{\rho c}{A_{N+1}} \times u_{N+1}^2(t) \end{aligned} \tag{9}$$

We can obtain a similar expression for the nett acoustic power entering the vocal tract at the glottis, $w_1(t)$, by subtracting the forward and reverse components of power flow at the glottis end of the first tube section

$$w_1(t) = \frac{\rho c}{A_{N+1}} \times \frac{A_{N+1}}{A_1}\left(u_1^2(t) - v_1^2(t)\right) \tag{10}$$

This expression may be decomposed as a product of pressure and nett volume velocity:

$$w_1(t) = \frac{\rho c}{A_{N+1}} \prod_{j=1}^{N} \frac{1 + r_j}{1 - r_j} \times \left(u_1(t) + v_1(t)\right) \times \left(u_1(t) - v_1(t)\right) \tag{11}$$

Using (2) we can derive FIR filters that allow the pressure and volume velocity terms of (11) to be determined from $u_{N+1}(t)$.

The leading scale factor in (9) and (11) depends on A_{N+1}: the effective area of the free-space beyond the lips. The value of A_{N+1} depends on A_{lip} and cannot be deduced from the speech signal. In the results presented below, we neglect the scale factor when calculating $w_1(t)$.

4. EXPERIMENTAL RESULTS

4.1 Data Processing

In the following examples, speech signals taken from the APLAWD database [6] are sampled at 20 kHz and processed using covariance LPC with non-overlapping 15 ms frames, preemphasis, and a filter order of 22. The LPC coefficients are converted to reflection coefficients using the "step-down" procedure [2, 9] and thence via (2) into the filters required to evaluate the terms of (11). The lip volume velocity, $u_{N+1}(t)$, has been estimated from the speech signal, $s(t)$, using the approximation (based on [9])

$$\frac{U_{N+1}}{S} \approx \frac{1 - z^{-1}}{\left(1 - \alpha z^{-1}\right)^2} \tag{12}$$

where

$$\alpha = \exp\left(-2\pi \times 50\ \text{Hz} / 20\ \text{kHz}\right) \tag{13}$$

This filter has a low-pass response but its gain falls at frequencies below 50 Hz to avoid amplifying any DC offsets within the speech signal. MATLAB software for these processing steps is available in [2]. Phase distortions introduced by the recording apparatus have been corrected using the approach of [4].

It is well known that for some speech signals, the Normal equations can become ill-conditioned and that the resultant LPC filter will be unreliable when used for inverse filtering. A number of authors have suggested techniques for improving the robustness of LPC-based inverse filtering [3, 10].

4.2 Processed Waveforms

Figure 2 shows four waveforms for the vowel /a/ from a male speaker: (i) the speech signal, (ii) the laryngograph signal, (iii) the glottal input power, $w_1(t)$, and (iv) the

LPC residual (which approximates the time-derivative of $u_g(t)$).

The laryngograph signal, (ii), measures the high frequency conductance of the larynx and is at a maximum during the glottal closed phase [1]. The laryngograph signal is also known as the electroglottograph (EGG) and is inverted by some authors so that it is approximately in phase with glottal airflow [5]. In all the figures, the laryngograph signals have been delayed to compensate for the larynx-to-microphone propagation time.

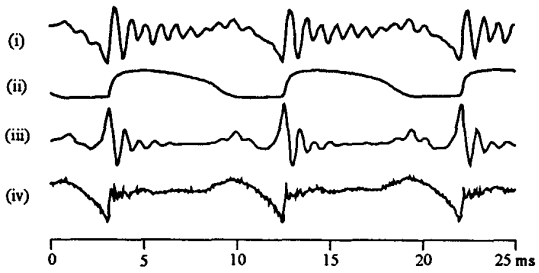


Figure 2. (i) Speech, (ii) Laryngograph, (iii) Glottal input power and (iv) LPC residual waveforms for the vowel /a/, male speaker.

For the example of Figure 2 it is possible to determine glottal opening and closure instants from either the glottal input power or the LPC residual waveforms although the excitation due to glottal opening is more clearly defined in the former. An advantage of using the input power waveform is that because it is a quadratic function of the speech signal, it is unaffected by an inversion of the input. It follows therefore that all vocal tract excitations will result in positive peaks.

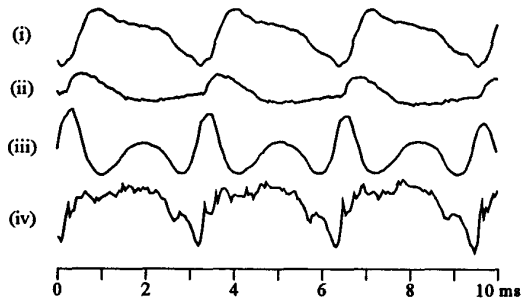


Figure 3. (i) Speech, (ii) Laryngograph, (iii) Glottal input power and (iv) LPC residual waveforms for the vowel /ɔ/, female speaker.

Figure 3 shows a vowel from a female speaker with a much higher larynx frequency. In this example, the excitation due to glottal opening is barely visible in the LPC residual waveform and a robust algorithm for locating it automatically would be very difficult. In contrast, the excitation is quite distinct in the input power waveform.

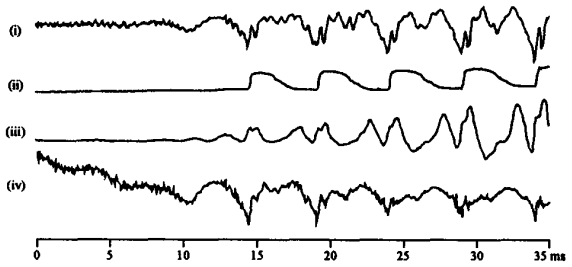


Figure 4. (i) Speech, (ii) Laryngograph, (iii) Glottal input power and (iv) LPC residual waveforms for the unvoiced-voiced transition in /θri/, male speaker.

Fig. 4 shows the same waveforms for a male speaker during a voiceless-to-voiced transition. The glottal input power waveform clearly identifies the onset of voicing and shows a peak for both closure and opening in each larynx cycle. As in the previous example, the glottal opening excitations are far less distinct in the LPC residual signal and even some of the closure excitations are poorly defined.

The processing has been applied to a large number of speech files and the outputs visually inspected. Although the algorithm is generally very robust, it does on rare occasions give anomalous results.

Figure 5 shows a segment of vowel in which the analysis frame boundaries have been marked. The speech waveform appears to be stationary and the LPC residual has clearly identified all glottal closures. Despite the evident success of the LPC analysis; the glottal input power waveform is a poor indicator of glottal activity. In the first frame, the glottal opening excitation is masked by the first formant oscillations while in the second and third frames several glottal closures are completely absent.

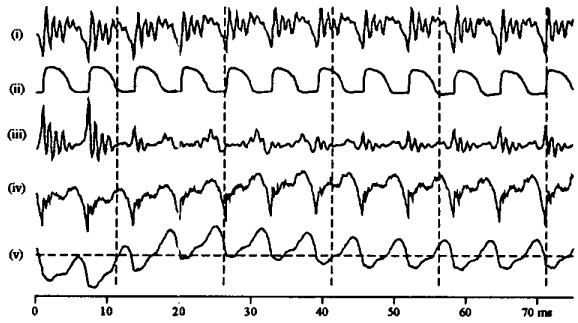


Figure 5. (i) Speech, (ii) Laryngograph, (iii) Glottal input power,(iv) LPC residual and (v) nett glottal flow waveforms for the vowel /a/, male speaker.

On closer inspection, the problems arise because of a transient shift in the DC level of the nett glottal flow waveform, $u_1(t) - v_1(t)$, which is shown as trace (v) in Figure 5 together with a horizontal line indicating zero. The nett glottal flow is one of the components in (11) and should be zero during the glottal closed phase; this is approximately true in the final two frames of the figure but manifestly false during the first three. It appears that a low frequency component of the original speech signal, barely visible in trace (i), is amplified by the filter used to generate $u_{N+1}(t)$ and that this causes the problems visible in the figure.

5. SUMMARY

This paper has presented a procedure for estimating the waveform of the acoustic power supplied to the vocal tract during speech. The procedure is based on LPC analysis and is unaffected by an inversion of the speech signal. It has been shown that the glottal input power signal may be used to find the instants of glottal opening and closure and that it indicates these events more clearly than does the LPC residual signal.

The calculation of the glottal input power appears to be generally robust but is sensitive to low frequency noise in the speech signal. We are currently seeking ways to improve the robustness of the modelling and are developing a dynamic programming algorithm to select automatically the peaks in the input power waveform that correspond to glottal opening and closure excitations.

6. REFERENCES

[1] Abberton E. R. M., Howard D. M. and Fourcin A. J., "Laryngographic assessment of normal voice: a tutorial", *Clinical Linguistics and Phonetics*, 3:281-296, 1989.

[2] Brookes D. M., *VOICEBOX: speech processing toolbox for MATLAB*, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[3] Brookes D. M. and Chan,D. S. F., "Speaker Characteristics from a Glottal Airflow Model using Robust Inverse Filtering", *Proc Inst of Acoustics*, **16**(5):501-508, 1994.

[4] Hunt M.J., "Automatic Correction of Low-Frequency Phase Distortion in Analogue Magnetic Recordings", Acoustics Letters, 2:6-10, 1978.

[5] Krishnamurthy A. K. and Childers D. G., "Two-channel speech analysis". *IEEE Trans. Acoust. Speech & Signal Processing*, **34**(4):730-743, 1986.

[6] Lindsey G., Breen A. and Nevard S., "SPAR's Archivable Actual-word Databases", *Internal Report*, University College London, Jun 1987

[7] Ma C., Kamp Y. and Willems L. F., "A Frobenius Norm Approach to Glottal Closure Detection". *IEEE Trans, on Speech and Audio Processing*, **2**(2):258-265, 1994.

[8] Morse P. M. and Ingard K. U., *Theoretical Acoustics*, Princeton University Press, 1986, ISBN 0-691-08425-4.

[9] Rabiner L. R. and Schafer R. W., *Digital Processing of Speech Signals*, Prentice Hall, 1978, ISBN 0-13-213603-1.

[10] Ramachandran R. P., Zilovic M. S. and Mammone R. J., "A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification", *IEEE Trans on Speech and Audio Processing*, **3**(2):117-125, 1995.