ANALYSIS AND PERCEPTION OF GLOTTAL PULSE SHAPES

W.A.Ainsworth and D.J.Munden

Department of Communication and Neuroscience
University of Keele, Keele, Staffordshire

INTRODUCTION

Although there have been a number of papers describing techniques for extracting the glottal excitation from the natural speech waveform [1-4], little use has been made of them and there has been no detailed examination of the glottal pulse shape and how it is affected by changes in pitch. The open phase to closed phase glottal ratio is sometimes used as a parameter for synthetic excitation pulses. It would therefore be useful to learn how this parameter varies in natural speech, and how well listeners can perceive changes in it.

The ratio of the glottal open phase to the fundamental period, the open/period ratio or open quotient [5], is a measure of the proportion of one glottal cycle that the glottis remains open and is related to the open phase to closed phase ratio by

$$\text{open quotient} = \frac{\text{open/closed ratio}}{\text{open/closed ratio} + 1} \qquad (1)$$

The analysis and experiments described here are concerned with determining the effect of pitch on the open quotient and how effectivly listeners can percieve changes in this parameter.

Determination of the glottal wave from the acoustic pressure wave presents a problem which usually involves inverse filtering in some form or another. This technique was first used to demonstrate the pulse-like nature of the excitation waveform [6]. The basic idea is to use the natural speech to obtain an estimate of the transfer function for the vocal tract. A series of filters can then be designed with a transfer function which is the inverse of this estimate. Passing the speech through these filters cancels the formant poles, leaving only the excitation waveform. As there is no unique decomposition of the speech signal into its excitation and vocal tract components, the success of the analysis is governed by the accuracy of the constraints and criteria used. A closed phase analysis, as described by Wong et al [3] and Veeneman [4], calculates the inverse filter coefficients during the closed glottis interval. Wong et al describe a computationally expensive algorithm for detecting the points of glottal opening and closure, whereas Veeneman relies on the output from an electroglottographic device to estimate the timing of these events. If the inverse filtering analysis window is chosen during the period that the glottis is completely closed it is assumed that the excitation has no effect on the speech pressure wave and immediately after glottal closure it becomes a freely decaying oscillation which is strictly a function of the vocal tract resonances. The analysis described here is similar to that described by Veeneman but estimates the closed phase directly from the speech pressure wave.

Little is known about how well we perceive changes in the shape of the excitation waveform. In 1970 Rosenberg [7] reported the results of a number of experiments designed to determine if listeners could perceive changes in the glottal pulse shape and if some sounded more natural than others. Using pitch synchronous analysis he extracted the excitation from one sentence and three CVC words. He re-synthesised the same utterances using the natural waveform plus five synthetic pulse shapes varying only in their location and number of slope discontinuities. Listeners showed a preference for speech generated using the naturally derived excitation, followed by the synthetic shapes with similar characteristics. Holmes [8] showed that it was possible, with careful selection of the parameters, to produce natural sounding synthetic speech almost indistinguishable from the natural version from which it was derived. He concluded that the excitation waveform was an important factor in the production of natural sounding synthetic speech.

ESTIMATION OF THE GLOTTAL EXCITATION WAVEFORM

Voiced speech production can be modelled by a linear system. The glottal volume velocity signal g(n) is filtered by the vocal tract transfer function H[Z] to form the lip volume velocity signal v(n), which in turn radiates from the lips to form s(n). Assume that the vocal tract can be modelled by an all pole process, [9,10], given by

$$H[Z] = \frac{1}{1 + \sum_{k=1}^{p} a(k)\, Z^{-k}} \qquad (2)$$

where p/2 equals the number of resonant frequencies of the vocal tract. Assume also that the radiation impedance can be represented by a differencer of the form

$$R[Z] = 1 - Z^{-1} \qquad (3)$$

The speech production model, expressed in difference form, is then given by

$$v(n) = g(n) - \sum_{k=1}^{p} a(k)\, v(n-k) \qquad (4)$$

If the glottal volume velocity, g(n), is replaced by its derivative, the effective driving function q(n), given by

$$Q[Z] = R[Z]\ G[Z] \tag{5}$$

Then voiced speech production can be modelled by the linear system expressed in difference form as

$$s(n) = q(n) - \sum_{k=1}^{p} a(k)\ s(n-k) \tag{6}$$

If only the output s(n) is known it is impossible to determine the glottal volume velocity or the vocal tract transfer function without making further assumptions. In this analysis it is assumed that $q(n) = 0$ during the closed phase and that complete glottal closure occurs for sufficient time during each cycle to ensure that there are enough points with which to perform the analysis. This analysis is restricted to non-nasalised voiced speech as an all pole model of the vocal tract is assumed and it is not possible to separate excitation zeros from vocal tract zeros. If $q(n) = 0$ during the closed phase then

$$s(n) = - \sum_{k=1}^{p} a(k)\ s(n-k) \tag{7}$$

where point of glottal closure + p < n < point of glottal opening

The next step is to determine the values of the coefficients a(1)..a(p) which represent the vocal tract transfer function. These are calculated using the covariance method of linear prediction [10], which selects the values of the prediction coefficients such that the total squared error e(n) is a minimum, where e(n) is given by

$$e(n) = s(n) + \sum_{k=1}^{p} a(k)\ s(n-k) \tag{8}$$

After the filter coefficients are obtained the inverse filter is constructed and the speech pressure waveform passed through it to obtain the effective driving function. This can then be integrated to produce the glottal volume velocity waveform.

The primary excitation of a speech waveform occurs at times corresponding to instants of glottal closure. Therefore, the instants of glottal closure can be determined by obtaining the positions of peaks in the digitised speech. The instant of glottal opening is far more difficult to obtain, and in this method no attempt has been made to estimate where glottal opening begins. Instead it is assumed that the glottis will remain closed for at least 3 ms after closure.

The major peaks in the speech waveform are detected by following the waveform until a maximum is found [11]. A marker then begins to decay exponentialy until it crosses the speech waveform, which it then joins and follows until the next maximum is detected. This signal is then differentiated to produce spikes at the points of glottal closure. The whole process is then repeated to remove any spurious spikes. The closed phase analysis window is positioned immediately after the position of one of the spikes and the following 30 points are considered to be within the closed portion of the glottal cycle. The LPC coefficients are calculated and 3 ms of speech immediately following the chosen point of glottal closure is passed through the inverse filter. The roots of the linear prediction coefficient polynomial are calculated. Formant frequencies will be defined by complex pole pairs, any real roots suggest an over specification of the filter order and are therefore removed and the inverse filtering repeated. A filter order of 10 or 12 (5 or 6 formants) and a corresponding window length of 20 or 24 points has been found adequate for isolated vowels. For more complex sounds a higher filter order may be required to resolve closely spaced formants and provide better spectral matching.

THE EFFECTS OF PITCH ON THE OPEN QUOTIENT

Three different male subjects produced a number of samples, varying only in pitch, of steady state /a/ and /3/ vowels. These were digitized, at a rate of 10 KHz, by a CED 1401, interfaced to an IBM PC-XT, using a Revox A77 tape recorder and microphone. The samples were all of 500 ms duration. The speech was bandpass filtered to remove any energy above 5 KHz, and any low frequencies bellow 30 Hz caused by recording noise. The speech was then analysed and the open quotient estimated. The effect of fundamental frequency on the open quotient is shown for the different subjects in Figure 1. A general increase in the open quotient with increasing pitch is observed for two of the subjects while the other shows the opposite trend.

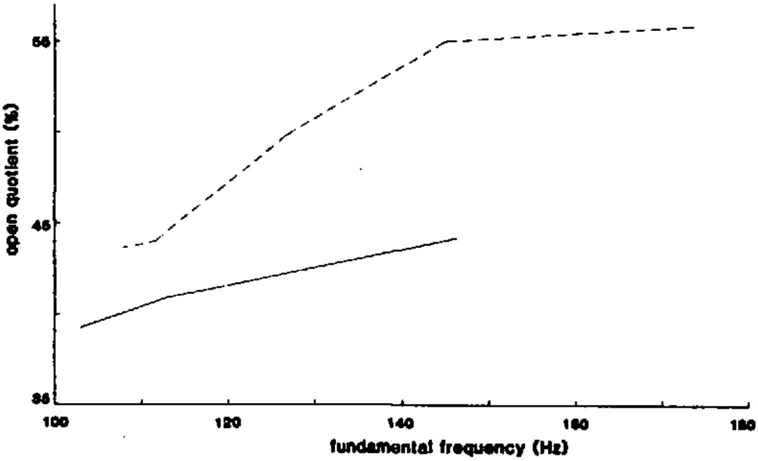

Figure 1(a). Open quotient derived from the speech waveform of /a/ (solid curve) and of /3/ (dashed curve) as a function of fundamental frequency for speaker SD.

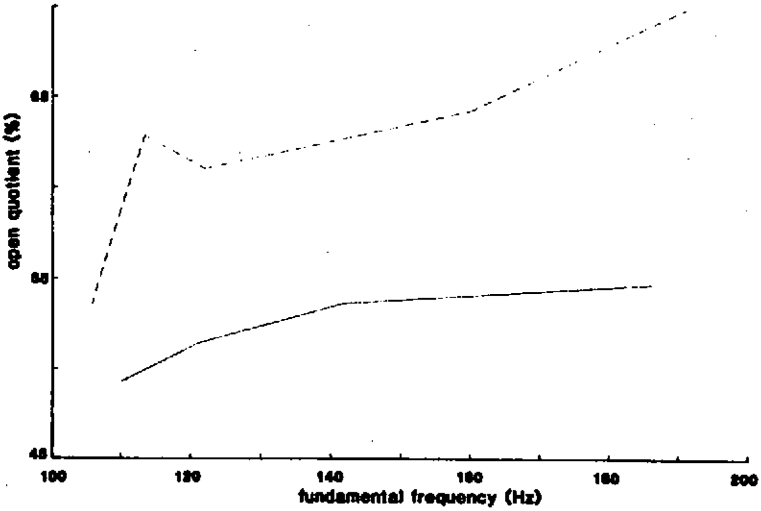

Figure 1(b). Open quotient derived from the speech waveform of /a/ (solid curve) and of /3/ (dashed curve) as a function of fundamental frequency for speaker SM.

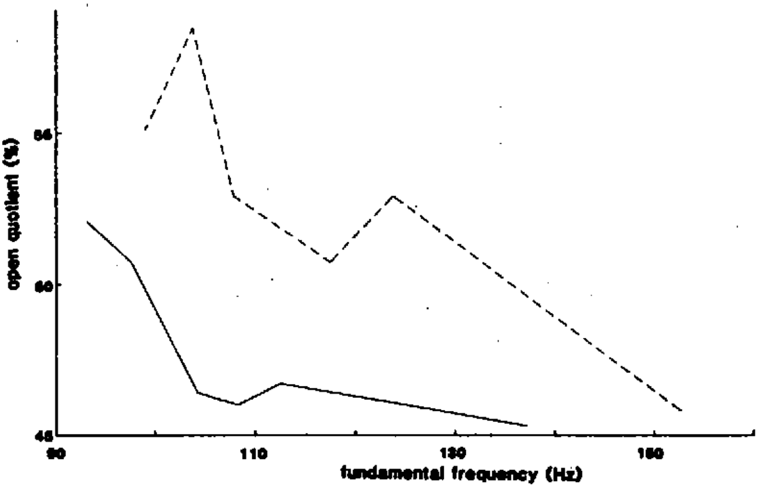

Figure 1(c). Open quotient derived from the speech waveform of /a/ (solid curve) and of /3/ (dashed curve) as a function of fundamental frequency for speaker DJM.

Figure 2 shows a typical effective driving function obtained using this analysis, and the estimated glottal area function which it corresponds to. The non-abrupt closure of the glottis observed in this example is similar to that obtained from many of the samples. In fact, often complete glottal closure may not occur at all during a glottal cycle. This results in the estimate of the vocal tract function being influenced by the excitation, and therefore authentic estimates of the glottal pulse shape are not possible. In practice though there is enough information in the obtained driving function to at least obtain the points of glottal opening and closure if not an acurate estimate of the glottal pulse shape itself. Problems are also encountered when the glottal closed phase is shorter than the analysis window length. This is particularly common at high fundamental frequencies where there are less points per glottal cycle and the closed phase tends to be shorter. This makes the analysis difficult to use on female speakers which is why they have not been included in this study.

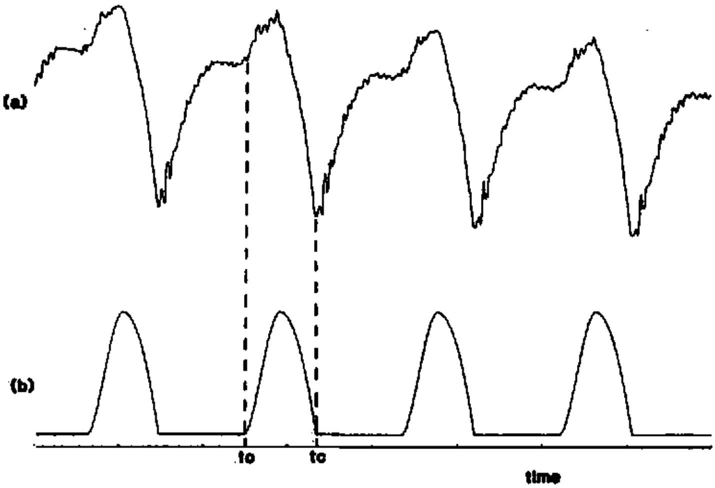

Figure 2. Glottal driving function (a) and glottal area function (b) derived from a speech waveform.

PERCEPTION OF THE GLOTTAL OPEN QUOTIENT

A parallel formant synthesiser, as described by Quarmby and Holmes [12], was used for these experiments. The open quotient is effectively under dynamic control using the mark/space ratio parameter, which is directly proportional over the range of interest. The excitation waveform is stored as a set of 12 sample values, which is sufficient to avoid a loss of naturalness, and represents both the glottal area waveform and its second derivative which provides the voiced excitation waveform. The duration of the excitation pulse is determined by the current fundamental frequency and the mark/space ratio. The glottal area waveform is similar to that shown by Rosenberg [7] to produce the most natural sounding speech, with one sharp slope discontinuity at the point of glottal closure.

Experiment 1

The first experiment was performed in order to determine how great a change in the open quotient is required before there is a detectable difference in the perceived speech output. Four long vowels /i,a,u,3/ with a constant duration of 500ms and fundamental frequency 100Hz were synthesised using excitation pulses with 9 different values of open quotient ranging from 30% to 90%, increasing in steps of 7.5%. This more than covers the range observed in natural speech, usually between 30% and 70%, [13]. Five subjects were presented with all 81 possible pairs of stimuli for each vowel and asked to judge them same or different. This was repeated five times. It was found that the greater the difference in the open quotient between the two stimuli, the more likley they were to be judged dissimilar. The results showed that a 7% difference in the open quotient is required before there is a better than chance probability of hearing a difference in the stimuli.

Experiment 2

Experiment 1 was repeated with one vowel, /a/, for three different fundamental frequencies, 60 Hz, 100Hz and 200Hz. The results showed that a change in fundamental frequency has very little effect on the listeners ability to discriminate between speech produced using excitations differing only in their open quotient. The results of an Analysis of Variance test showed that the effect of pitch is not significant.

Once again it was found that a 7% change in the open quotient was required before there was any perceptual difference in the synthetic speech output. This Figure is independant of both the vowel and the fundamental frequency. Small changes in the open quotient can be detected which suggests that this may be an important parameter in producing good quality synthetic speech. Most synthetic glottal pulse shapes have this ratio as one of their parameters, yet it often remains constant. This is partly due to the fact that it is often not known what the value of this ratio should be, or how and by how much it should vary during an utterance. Further analysis of natural speech excitation should be performed to investigate these properties.

CONCLUSION

An inverse filtering analysis system for extracting the excitation waveform from digitised speech has been developed. Like other methods it relies on there being a phase during each glottal cycle where the glottis remains firmly closed for long enough, in order that a stretch of speech uninfluenced by the excitation can be selected. The resulting waveforms obtained suggest this is not always the case but estimates of the opening and closing, if closure occurs, have been obtained. Experiments have also been performed to measure the perceptual sensitivity to changes in this parameter. A greater knowledge of how this ratio varies during continuous speech may help in achieving more natural synthetic speech.

Acknowledgements
This work was supported by a SERC CASE award and was carried out in collaboration with the IBM U.K. Scientific Centre, Winchester.

REFERENCES

[1] J. Lindqvist, 'The voice source studied by means of inverse filtering', STL-QPSR, Vol.1, 3-9, (1970).

[2] M. Rothenberg, 'A new inverse filtering technique for deriving the glottal air flow waveform during voicing', J.A.S.A., Vol. 53, 1632-1645, (1973).

[3] D.Y. Wong, J.D. Markel and A.H. Gray jr, 'Least squares glottal inverse filtering from the acoustic speech waveform', IEEE Trans., Vol. ASSP-33, no.4, 350-355, (1979).

[4] D.E. Veeneman and S.L. Bement, 'Automatic glottal inverse filtering from speech and electroglottographic signals', IEEE Trans., Vol. ASSP-33, no.2, 369-377, (1985).

[5] G. Fant, 'Preliminaries to analysis of the human voice source', STL-QPSR, Vol.4, 1-27, (1982).

[6] R.L. Miller, 'Nature of the vocal cord wave', J.A.S.A., Vol.31, 667-677, (1959).

[7] A.E. Rosenberg, 'Effect of glottal pulse shape on the quality of natural vowels', J.A.S.A., Vol.49, 583-590, (1971).

[8] J.N. Holmes, 'The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer', IEEE Trans., Vol. AU-21, no.3, 298-305, (1973).

[9] G. Fant, 'Acoustic Theory of Speech Production', Gravenhage, The Netherlands: Mouton, (1960).

[10] J.D. Markel and A.H. Gray jr, 'Linear Prediction of Speech', New York: Springer-Verlag, (1976).

[11] O. Gruenz jr and L.O. Scott, 'Extraction and portrayal of pitch of speech sounds', J.A.S.A., No.21, 487-495, (1949).

[12] D.J. Quarmby and J.N. Holmes, 'Implementation of a parallel-formant speech synthesiser using a single-chip programmable signal processor', IEE Proc., Vol.131, Pt.F, No.6, 563-569, (1984).

[13] W.J. Hess, 'Pitch Determination of Speech Signals', Berlin: Springer-Verlag, (1983).