Proceedings of the Institute of Acoustics

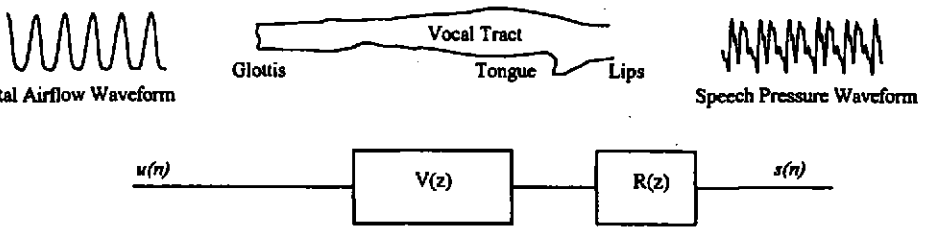
SPEAKER CHARACTERISTICS FROM A GLOTTAL AIRFLOW MODEL USING ROBUST INVERSE FILTERING

D. M. Brookes & D. S. F. Chan

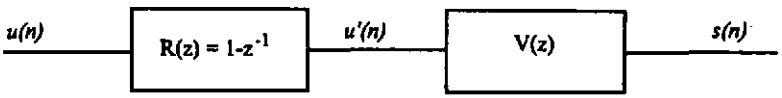
Signal Processing Section, Electrical & Electronic Engineering Dept, Imperial College of Science Technology & Medicine, Exhibition Road, London SW7 2BT

1. INTRODUCTION

The work presented in this paper is based on the discrete-time source-filter model of voiced speech production shown below:



Here the excitation, $u(n)$, is the volume velocity or volume flow rate of air through the glottis, $V(z)$ is a linear filter representing the vocal tract and $R(z)$ is a linear filter representing the lip radiation. The input to $R(z)$ is the volume velocity at the lips while the output, $s(n)$, is the pressure waveform at the microphone. For frequencies below a few kHz, $R(z)$ approximates a differentiator [20] and in this work it is taken to be $(1-z^{-1})$. Providing the characteristics of the vocal tract do not change significantly during the impulse response of $R(z)$, we can interchange $V(z)$ and $R(z)$ without altering the output signal. The input to $V(z)$ is now $u'(n)$, the first difference of the glottal airflow waveform:



The modelling techniques presented in sections 2, 3 and 4 allow parametric models of the waveform $u'(n)$ and the vocal tract filter $V(z)$ to be estimated from the speech signal. The reasons for wishing to do this include the development of improved front ends for speech/speaker recognition systems and of diagnostic tools for speech clinicians. In section 5, we show that the glottal waveform parameters contain information that can be used to discriminate between different speakers.

The APLAWD speech corpus used in this work was recorded as part of the Alvey SPAR project and consists of 10 RP speakers (5 male, 5 female) each speaking 150 one or two word items and 5 phonetically balanced sentences [16]. A sample rate of 20 kHz was used both for the speech and for the simultaneous Laryngograph (or EGG) recordings. Low frequency phase shifts introduced during the recording process were removed by a second order allpass filter whose characteristics were determined from a reference square-wave recording [13]. Waveform fitting procedures are sensitive to low frequency phase shifts and such phase correction is essential.

2. CLOSED PHASE INVERSE FILTERING

In the technique of closed phase inverse filtering, $V(z)$ is initially estimated during the portion of the glottal cycle when the vocal folds are closed (the closed phase). The waveform $u'(n)$ can then be obtained by applying the inverse filter $1/V(z)$ to $s(n)$.

The technique has a long history [19,12,24,23,14]. The distinctive features of the procedure described below are: (1) it is fully automatic rather than being interactive, (2) the speech is not preemphasised, (3) it allows for a DC offset in the speech signal and/or the glottal derivative, (4) consecutive closed phases are combined where necessary to provide sufficient data for analysis, (5) the gain of the inverse filter is clipped at unity, and (6) poorly modelled larynx cycles are reprocessed using the vocal-tract filter from adjacent cycles.

Following acoustic theory [20], the vocal tract filter is taken to be an order-p all-pole filter of the form

$$V(z) = \frac{1}{1 + \sum_{j=1}^{p} a_j z^{-j}}$$

which results in the time-domain recurrence relation

$$s(n) = u'(n) - \sum_{j=1}^{p} a_j s(n-j) + e(n) \qquad (1)$$

where $u'(n)$ is the glottal derivative and $e(n)$ is the model error for sample n. The a_j that minimise the squared error $e(n)^2$ are given by the solution to the normal equations:

$$\Phi \mathbf{a} = -\varphi \quad \text{where} \quad \Phi_{ij} = \sum_n s(n-i)s(n-j) \quad \text{and} \quad \varphi_i = \sum_n s(n)s(n-i) \qquad (2)$$

in which Φ is a $p \times p$ matrix and $\mathbf{a}$ and φ are $p \times 1$ column vectors. In contrast to conventional covariance LPC analysis the summations over n are restricted to values that lie within periods of glottal closure, i.e. when $u'(n)$ is assumed zero. This is a special case of weighted LPC analysis [8].

There are a number of published techniques for identifying the period of glottal closure from the speech waveform [22,24,2,6,18]. In our experience none of these is sufficiently robust for use in automatic inverse filtering so in this work we have used a Laryngograph or EGG [1] instead. This instrument measures the radio-frequency electrical conductance across the larynx and gives an unambiguous indication of glottal closure. We take the closed phase to be the interval between the peak conductance of the Lx signal and the time at which it has fallen by 50% of its amplitude.

From the lossless tube model, the LPC order needed to represent the vocal tract is given by $p = f_s \times 2l / c$ where f_s is the sample frequency, l the vocal tract length and c the speed of sound [20]. To obtain good estimates of the LPC parameters, we would like the summations in (2) above to include at least $2p$ samples: this implies an analysis interval of at least $4l/c$ which for typical vocal tract lengths amounts to about 2 ms. The closed phase of the glottal cycle is frequently shorter than this and so it is necessary to combine the closed phases from two or even three closed cycles [5].

Because the spectrum of the $u'(n)$ waveform falls at roughly -6dB/octave, it is common practice to preemphasise the speech waveform before doing LPC [7]. In closed-phase LPC analysis we have found that such preemphasis often makes the vocal tract coefficients noisier and we have not done this. Unpreemphasised speech may however contain significant DC offsets which vary throughout a recording.

To compensate both for these and for any DC component in the closed phase of $u'(n)$, we modify equation (1) by the addition of a constant term [4]:

$$s(n) = u'(n) - \sum_{i=1}^{p} a_i s(n-i) + e(n) + G$$

Minimising the squared error now gives rise to an augmented set of equations:

$$\begin{pmatrix} N & \mathbf{x}^T \\ \mathbf{x} & \Phi \end{pmatrix} \begin{pmatrix} -G \\ \mathbf{a} \end{pmatrix} = -\begin{pmatrix} x_0 \\ \varphi \end{pmatrix} \qquad (3)$$

where N is the number of values of n included in the summations, $x_i = \sum_n s(n-i)$ and $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix}^T$. A straightforward extension of these equations allows the inclusion of distinct offsets, G_i, for each larynx cycle included in the summations.

Since the mean volume velocities at larynx and lips must be equal the overall gain of the vocal tract filter derived from the LPC analysis is normalised to give unity gain at DC. In addition, any poles on the positive real axis are deleted to eliminate any overall tilt to the spectrum.

Although the forward spectra resulting from LPC analysis provide a good fit to the signal's formant peaks, a much poorer fit is obtained in those portions of the spectrum where little speech energy is present. If the gain of the forward filter is too low in such a frequency region, the gain of the inverse filter will be correspondingly high and the inverse filtered waveform will be noisy. The effect is particularly noticeable in vowels containing formants that have narrow bandwidths or that are widely separated in frequency.

To overcome this effect, we limit the peak gain of the inverse filter to unity. This gain-limiting is achieved by taking the fourier transform of the inverse filter impulse response, clipping the resultant magnitude spectrum, taking the inverse fourier transform and then applying a Hamming window. This procedure gives a non-causal filter with the same phase characteristics as the original filter but with a clipped magnitude response.

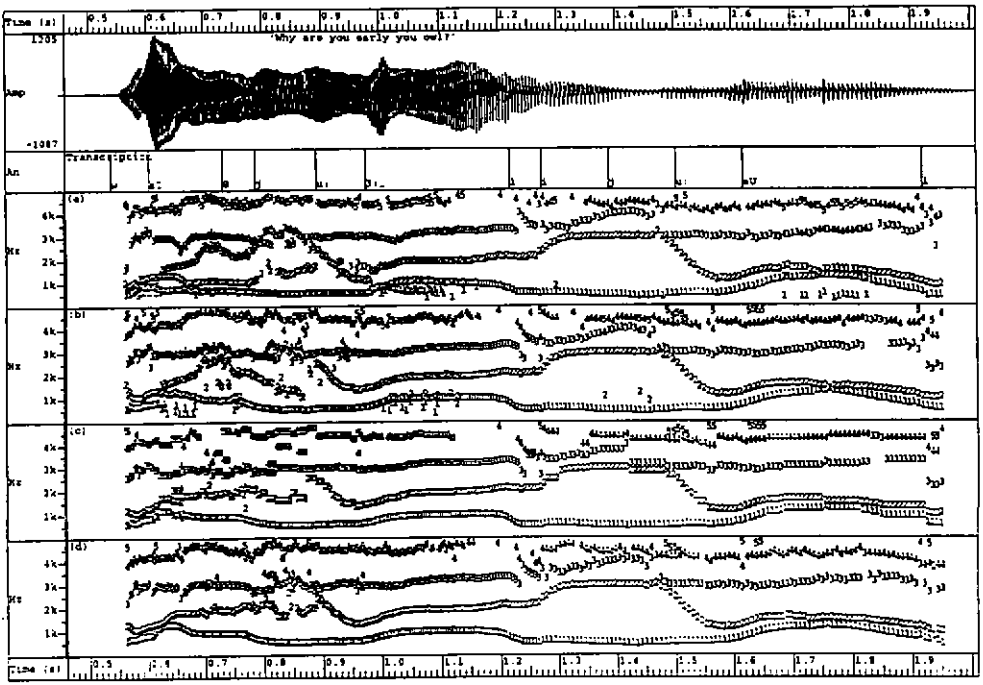
To eliminate the erratic results that sometimes occur when covariance LPC analysis is based on only a small number of data samples, the inverse filtered waveform during the closed phase is compared to the negative exponential shape predicted by the LF and LFCB models of $u'(n)$ described below. The speech signal corresponding to a particular larynx cycle is inverse filtered using the $V(z)$ derived from each of the larynx cycles lying within a window of 30 ms. Whichever of these $V(z)$ yields an inverse filtered waveform that is closest to a negative exponential in a mean square sense is used as the filter for that cycle. The effect of this smoothing technique may be seen in example (c) below.

The graphs overleaf show the sentence *"Why are you early you owl?"* spoken by a female RP speaker with a falling intonation. The formant tracks shown result from (a) whole-cycle pitch synchronous LPC, (b) single-cycle closed-phase LPC, (c) multi-cycle closed-phase LPC with the smoothing technique described above, and (d) single-cycle analysis using the LFCB glottal model described below.

Comparing (b) with (a), it is apparent that using only the closed phases results in noisier formant estimates but does avoid the spurious low frequency formants that appear in (a) during /s/ and /aʊ/. Spurious low frequency formants do still appear in (b) during /aʊ/ because the closed phase was too short and the LPC analysis was forced to include some of the following open phase.

Both the spurious formants and the noise are largely eliminated in (c) through the use of multi-cycle analysis and smoothing. In rejecting unreliable analysis frames, the smoothing procedure results in the

same vocal tract filter being used for several successive frames: this is visible as horizontal formant track segments. Graph (d) is discussed in section 4 below.



3. GLOTTAL WAVEFORM MODEL

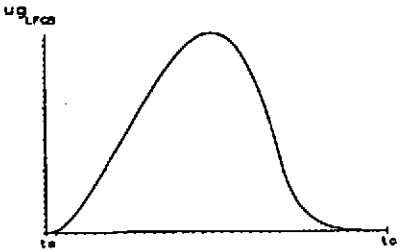
A number of parametric models have been proposed for the waveform $u(t)$ or, more commonly for its derivative $u'(t)$. These models typically divide the waveform into two or three segments and fit some combination of polynomial, trigonometric and exponential functions to each segment. In most models the waveform $u'(t)$ is constrained to be continuous and to integrate to zero over each complete larynx cycle. This last constraint is equivalent to insisting that the glottal flow, $u(t)$, be zero at the start and end of each cycle. The table below lists a number of models and gives for each the number of free parameters: this quantity is the difference between the number of parameters in the model and the number of constraints that must be satisfied. The parameter count does not include t_s and t_e, the start and end times of the cycle.

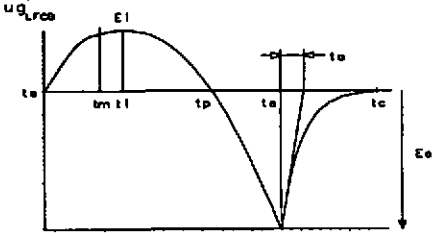
Abbreviation	Parameters – Constraints	Authors
ROS	3 – 0 = 3	Rosenberg [21]
FL	8 – 2 = 6	Fujisaki-Ljungqvist [10]
AD	5 – 0 = 5	Ananthapmanabha [3]
LF	7 – 3 = 4	Liljencrants-Fant [9]
LFCB	8 – 3 = 5	Chan-Brookes

SPEAKER CHARACTERISTICS FROM INVERSE FILTERING

The LFCB model, presented here, is based on the widely used LF model but includes an additional parameter, tm, that controls the skewness of the positive portion of $u'(n)$:

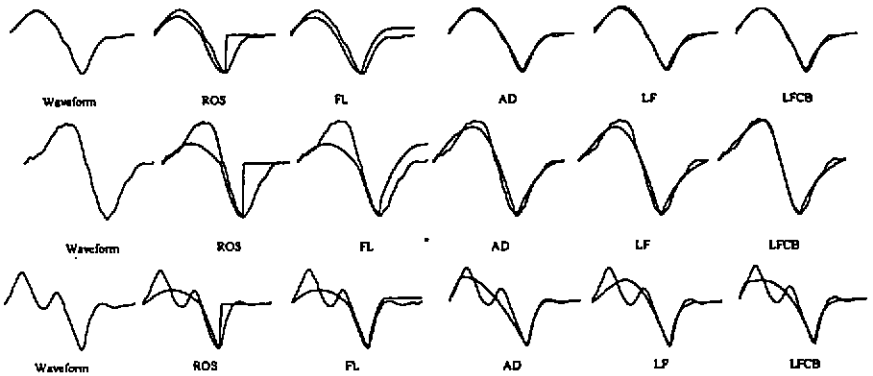
$$u'(t) = \begin{cases} E_0 \exp(\alpha(t-t_s)) \sin\left(\dfrac{\pi(t-t_s)}{2(t_m-t_s)}\right) & t_s < t \le t_m \\ E_0 \exp(\alpha(t-t_s)) \cos\left(\dfrac{\pi(t-t_m)}{2(t_p-t_m)}\right) & t_m < t \le t_e \\ \dfrac{E_e}{\xi t_a}\left(\exp(-\xi(t-t_e)) - \exp(-\xi(t_c-t_e))\right) & t_e < t \le t_c \end{cases}$$





This is subject to the same three constraints as the LF model namely: $u'(t)$ is continuous at t_e, $u'(t)$ integrates to 0 over a cycle, and $u'(t_e)=E_e$. If t_m is set midway between t_s and t_e, the LFCB model is identical to the LF model. The new model will therefore always fit experimental data at least as well as the LF model since it contains the latter as a special case.

The examples below show the result of fitting each of these models to three typical waveforms obtained from inverse filtering: In each case, the model parameters have been chosen to minimise the mean square error subject to an exact fit at the waveform's negative peak. The graphs show the target waveform superimposed on the best-fit output from each model.



With only three free parameters, the ROS model gives a poor fit in each case. The FL model, which consists of four polynomial segments, also gives a relatively poor fit despite having 6 free parameters. The remaining three models all give an excellent fit to the first waveform whose positive portion has a

symmetric rise and fall. The positive portion of the second waveform is much less symmetrical and the AD and LF models are unable to fit well. The additional parameter in the LFCB model results in a much improved fit.

Occasionally, the inverse filtering procedure yields an estimate of $u'(n)$ that is flat during the closed phase but that has significant ripple during the open phase. This is illustrated in the third example above and none of the models provides a good fit. The reasons for the open-phase ripple are unclear. It is commonly accepted that the reduction in source impedance at the open larynx causes the formant bandwidths and frequencies to change slightly and that the inverse filter is therefore unable to cancel the formants completely [15]. We have found however that such open phase ripple does not always occur; it is not clear why the formants are in some cases cancelled out almost perfectly.

The models listed above were tested on the inverse filtered output from the fully voiced sentence *'Why are you early you owl'* uttered by a male and a female speaker. In the table below, the error is given relative to the energy in the inverse filtered waveform.

	ROS	FL	AD	LF	LFCB
Male	-4.4dB	-5.9dB	-10.8dB	-11.1dB	-12.1dB
Female	-4.9dB	-6.8dB	-11.1dB	-10.8dB	-12.3dB

4. VOCAL TRACT REESTIMATION

Once an estimate of the glottal waveform has been obtained, this can be used to reestimate the vocal tract filter. For each larynx cycle, an input waveform is generated from the extracted glottal parameters using a high sample rate. This is then filtered and downsampled to provide the assumed input to the vocal tract, $u'(n)$. The LPC analysis procedure is now repeated but the normal equations include an additional term:

$$\begin{pmatrix} N & \mathbf{x}^T \\ \mathbf{x} & \Phi \end{pmatrix}\begin{pmatrix} -G \\ \mathbf{a} \end{pmatrix} = -\begin{pmatrix} x_0 \\ \varphi \end{pmatrix} + \begin{pmatrix} Y \\ \mathbf{y} \end{pmatrix} \quad \text{where} \quad Y = \sum_n u'(n) \quad \text{and} \quad y_i = \sum_n u'(n)s(n-i) \qquad (4)$$

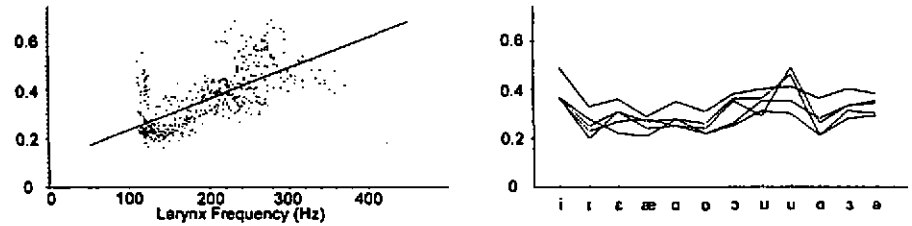
The summations over n now include the entire larynx cycle rather than being restricted to the closed phase; this results in much smoother formant estimates as can be seen in example (d) of the formant tracks above. It has been pointed out in [17] that knowledge of the vocal tract input waveform allows the straightforward estimation of a vocal tract filter containing both poles and zeros. This was not done in this work as the test sentences did not contain nasal sounds.

If the glottal waveform and LPC parameters are used to resynthesise speech, the result is of high quality and both the spectra and waveforms of vowels are reproduced well

5. SPEAKER CHARACTERISATION

From our model of the glottal waveform, we have derived four dimensionless parameters: ①: $\left|\frac{E_i}{E_e}\right|$, the ratio of the peak rise and fall slopes of u(t); ②: $\frac{t_e - t_s}{t_c - t_s}$, the fraction of the cycle corresponding to the open phase; ③: $\frac{t_e - t_p}{t_p - t_s}$, the ratio of fall to rise time; ④: $\frac{t_a}{t_c - t_s}$, the vocal fold closure time as a fraction of the cycle. These quantities are strongly correlated with larynx frequency as can be seen from the scatter diagram below which shows parameter ① versus frequency for a single

female RP speaker for the vowel /ɔ/. The rightmost graph shows mean values of parameter Φ for several vowels for five female RP speakers. The other three parameters showed broadly similar behaviour.



To evaluate the potential of these glottal waveform parameters for text-independent speaker identification, we used a radial basis function network to discriminate between 10 RP speakers: 5 male and 5 female. For each speaker, the net was trained using vowels extracted from a single repetition of a phonetically balanced set of 44 isolated words. Recognition used three repetitions by each speaker of each of five sentences. Of the 35 distinct words in the sentences, 10 were also contained in the training vocabulary. Unvoiced segments within the sentences were ignored.

The 10 inputs to the network were the four glottal waveform parameters defined above, the frequencies and bandwidths of the first two formants, the larynx frequency and the speech energy. The radial basis function centres were chosen at random from the training data points with an equal number from each of the ten speakers; the total number of centres was varied from 100 to 1000. The table below shows the recognition results obtained when each larynx cycle was considered individually and when the results were summed over one and three sentences. For comparison, the table also gives in parentheses the results obtained when the four glottal parameters were omitted from the training and recognition.

Total RBF centres	100	300	500	800	1000
Per Lx cycle	33% (22%)	38% (20%)	41% (19%)	44% (15%)	43% (16%)
Per sentence	70%	74%	79%	82%	79%
Per 3 sentences	70%	84%	80%	92%	90%

6. CONCLUSIONS

This paper has presented a procedure for estimating the glottal volume velocity waveform from the speech signal using inverse filtering. The procedure includes a number of features that improve its robustness and obviate the need for interactive intervention. A new parametric model for the glottal flow waveform is also presented and this is shown to give an improved fit to experimental data. The parameters extracted from the glottal waveform fitting procedure have been shown to contain information that is useful for text-independent speaker identification.

The most significant drawback of the inverse filtering procedure described above is the need for a Laryngograph signal to identify the closed phase of each larynx cycle. We have yet to find an algorithm for doing this from the speech signal that is sufficiently robust. The practice of limiting the inverse filter gain to unity is effective but somewhat ad-hoc. A more rigorous approach would be to use H$^\infty$ techniques to provide an optimum estimate of $u'(n)$ in the presence of measurement noise and modelling errors [11].

7. REFERENCES

[1] E R ABBERTON, D M HOWARD & A J FOURCIN, "Laryngographic assessment of normal voice: a tutorial", *Clinical Linguistics & Phonetics*, 3 pp281-296 (1989)

[2] T V ANANTHAPADMANABHA & B YEGNANARAYANA, "Epoch Extraction from Linear Prediction Residual for Identification and Closed Glottis Interval", *IEEE Trans ASSP* 27 pp309-319 (1979)

[3] T V ANANTHAPADMANABHA, "Acoustic Analysis of Voice Source Dynamics", *STL-QPSR*, 2-3 pp1-24 (1984)

[4] M J BEROUTI, D G CHILDERS & A PAIGE, "Glottal area versus glottal volume velocity" *IEEE ICASSP*, 1 pp33-36 (1977)

[5] D S F CHAN & D M BROOKES "Variability of Excitation Parameters from Robust Closed Phase Glottal Inverse Filtering" *European Conf on Speech Communication & Technology*, pp33.1.1 - 33.1.4 (1989)

[6] Y M CHENG & D O'SHAUGHNESSY, "Automatic and Reliable Estimation of Glottal Closure Instant and Period", *IEEE Trans ASSP*, 37 pp1805-1815 (1989)

[7] J R DELLER, "Some notes on closed phase glottal inverse filtering" *IEEE Trans ASSP*, 29 pp917-919 (1981)

[8] J R DELLER, J G PROAKIS & J H L HANSEN, *Discrete-Time Processing of Speech Signals*, ISBN 0-02-328301-7, Macmillan (1993)

[9] G FANT, J LILJENCRANTS & Q LIN, "A Four-Parameter Model of Glottal Flow", *STL-QPSR*, 4 pp1-13 (1985)

[10] H FUJISAKI & M LJUNGQVIST, "Estimation of Voice Source and Vocal Tract Parameters Based on ARMA Analysis and a Model for the Glottal Source Waveform" *IEEE ICASSP*, 2 pp637-640 (1987)

[11] B HENDEL, "The use of Game Theory in Robust Control and Filtering", *PhD Thesis*, Imperial College (1993)

[12] J N HOLMES, "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter", *Proc IV Internat Congr on Acoustics*, pp227-230 (1962)

[13] M J HUNT, "Automatic Correction of low-frequency phase distortion in analogue magnetic recordings", *Acoustics Letters*,2 pp6-10 (1978)

[14] A K KRISHNAMURTHY, "Two channel analysis for formant tracking and inverse filtering" *IEEE ICASSP* 3 pp36.3.1-36.3.4 (1984)

[15] A K KRISHNAMURTHY & D G CHILDERS, "Two channel speech analysis" *IEEE ASSP* 34 pp730-743 (1986)

[16] G LINDSEY, A BREEN & S NEVARD, "SPAR's Archivable Actual-word Databases", *Internal Report*, Dept of Phonetics & Linguistics, University College London, June 1987

[17] A P LOBO & W A AINSWORTH, "Evaluation of a Glottal ARMA Model of Speech Production", *IEEE ICASSP*, 2 pp13-16 (1992)

[18] C MA, Y KAMP & L F WILLEMS, "A Frobenius Norm Approach to Glottal Closure Detection from the Speech Signal", *IEEE Trans SAP*, 2 pp258-265 (1994)

[19] R L MILLER, "Nature of the vocal cord wave", *J Acoust Soc Am*, 31 pp667-677 (1959)

[20] L R RABINER & R W SCHAFER, *Digital Processing of Speech Signals*, ISBN 0-13-213603, Prentice-Hall (1978)

[21] A E ROSENBERG, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", *J Acoust Soc Am*, 49 pp583-590 (1971)

[22] H W STRUBE, "Determination of the instant of glottal closures from the speech wave", *J Acoust Soc Am*, 56 pp1625-1629 (1974)

[23] D E VEENEMAN & S L BeMENT, "Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals" *IEEE Trans ASSP*, 33 pp369-376 (1985)

[24] D J WONG, J D MARKEL & A H GRAY Jr, "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform" *IEEE Trans ASSP*, 27 pp353-362 (1979)