BRITISH ACOUSTICAL SOCIETY

"SPRING MEETING" at Chelsea College, London, S.W.3 on
Wednesday 25th April / Friday 27th April, 1973.

SPEECH AND HEARING: Session 'B': Speech Analysis and Transmission.

Paper No:

73SHB6

Digital Inverse Filtering of the Speech Waveform.

John Rogers.
Department of Electrical Engineering,
Imperial College, LONDON SW7 2BT.

The method described uses a digital inverse filter to estimate the vocal tract area function and the glottal excitation function for voiced speech. Householder transformation is used to find the inverse filter coefficients which give minimum least squared output during the period of glottal closure. These coefficients uniquely define the area function of the vocal tract model and inverse filtering the speech waveform gives an estimate of the glottal excitation function.

Results are presented for real and synthetic speech, the analysis being carried out in an interactive graphics enviroment on a PDP 15 computer. The results suggest that the area function obtained is insensitive to errors in estimation of the closed glottis period but the deconvolved glottal pulse is very sensitive to errors in this estimation.

INVERSE VOCAL TRACT FILTER.

The vocal tract transfer function for voiced speech is known to be an all pole function, hence the vocal tract can be modelled by a recursive filter. The inverse filter will be a transversal filter whose transfer function $C_n(z)$ is given by

$$C_n(z) = \sum_{i=0}^{n} a_i^{(n)} . z^{-i} \quad , \; a_o = 1$$

where n is the number of filter coefficients.

For voiced speech the glottal excitation function is known to have a closed period, i.e. period of zero input volume velocity. Using this knowledge the coefficients can be found by minimising the energy output of the filter, for speech input, during the closed glottal period. The vector of outputs $\underline{g}$ is given by

$\underline{g} = S\underline{a}$ where S is a matrix of speech samples used and $\underline{a}$ is the required coefficient vector. To avoid the trivial solution $a_i = 0$ for all i, it is necessary to separate out the first column of S which can be done as a_o is assumed to be unity to give

$$\begin{bmatrix} g_o \\ g_1 \\ \cdot \\ \cdot \\ \cdot \\ g_m \end{bmatrix} = a_o \begin{bmatrix} s_o \\ s_1 \\ \cdot \\ \cdot \\ \cdot \\ s_m \end{bmatrix} + \begin{bmatrix} s_{-1} & s_{-2} & \cdots\cdots & s_{-n} \\ s_0 & s_{-1} & \cdots\cdots & s_{-n+1} \\ s_1 & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ s_{m-1} & s_{m-2} & \cdots\cdots & s_{m-n} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_n \end{bmatrix} \qquad 1$$

for m n.

Equation 1 is most conveniently solved for $\underline{a}$ by minimising the least squared output using Householder transformation (1).

Alternative methods of finding the inverse filter or predictor coefficients proposed by Markel (2) and Atal (3) require the calculation of an autocorrelation or an autocovariance matrix which makes these methods computationally less efficient than using Householder transformation.

Once these coefficients have been found the vocal tract area function can be found using Wakita's algorithm (7) which is repeated here as:-

Given the coefficients of an n length filter normalised such that $a_o^{(n)} = 1$ identify $a_n^{(n)} = R_n$ 2

where R_n is the n^{th} junction reflection coefficient. This is related to the area function by

$$R_k = \frac{A_k - A_{k+1}}{A_k + A_{k+1}} \qquad 3$$

where A_k is the area of the k^{th} section of the vocal tract model. The n^{th} section is then removed and replaced by a termination matched to the $(n-1)^{th}$ section. The coefficients of the (n-1) section model are then found from those of the n section model by using

$$a_i^{(n-i)} = \frac{(a_i^{(n)} - R_n \, a_{n-i}^{(n)})}{(1 - R_n^2)} \qquad 4$$

Equations 2 3 and 4 define a recursive algorithm which allows calculation of the vocal tract area function. This algorithm is a direct development of the work of Kinariwala (10).

The glottal excitation function can be estimated by inverse filtering the speech waveform and the pitch can be found by Markel's method (4).

RESULTS FROM SYNTHETIC SPEECH.

Synthetic speech was formed by convolving the vocal tract transfer function, defined by its formant frequencies and bandwidths, with a synthetic glottal pulse consisting of a half period cosine in the opening period, a quarter period cosine in the closing period and zero in the closed period,(see fig. 1) which is shown to be well matched spectrally to real glottal pulses by Stansfield (5).

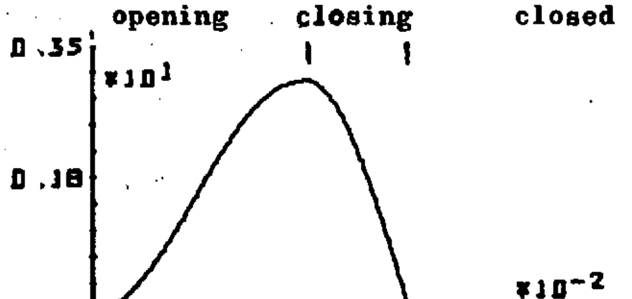
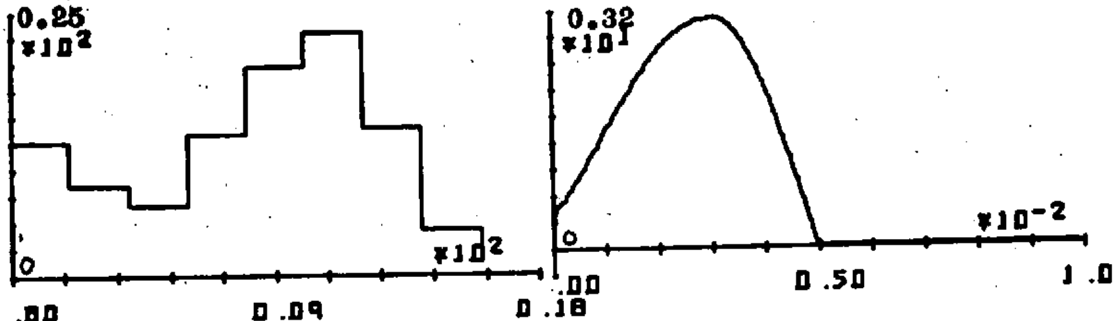


Fig. 1 Synthetic glottal pulse.

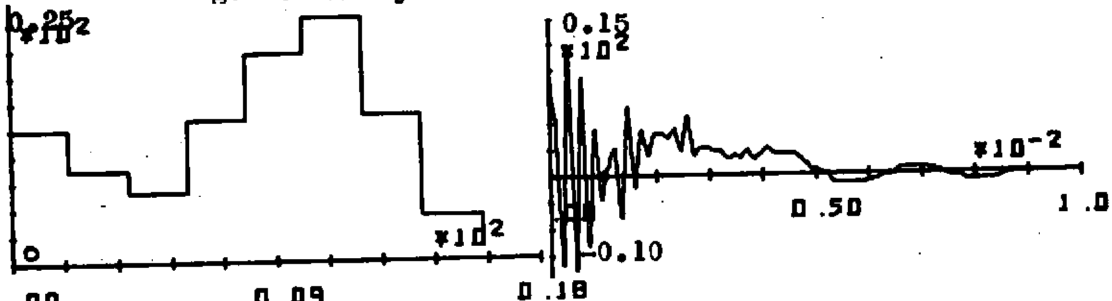
For this type of synthetic speech the glottal pulse is recovered perfectly when the order of the inverse filter is correctly chosen, i.e. equal to twice the number of poles used in the vocal tract, and the analysis is performed over any number of samples entirely in the closed period equal to or greater than the order of the filter. Typical results are shown in fig. 2.

The effect of moving the analysis window to include part of the open period was investigated. It was found that the glottal pulse could no longer be perfectly recovered, but that providing the analysis was carried out entirely during the closed or closing period a very good estimate of the area function is still obtained, see fig. 3. If however the analysis is carried out entirely during the opening period the area function found is vastly different from that obtained by analysis in the closed period, and in some cases the analysis fails, see fig. 4.



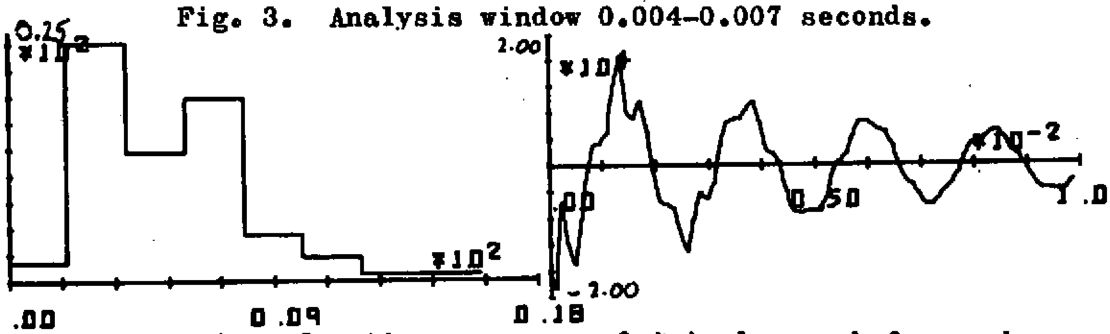
Area function. Recovered glottal pulse.

Fig. 2. Analysis window 0.006-0.01 seconds.



Area function. Recovered glottal pulse.

Fig. 3. Analysis window 0.004-0.007 seconds.



Area function. Original speech for each case.

Fig. 4. Analysis window 0.000-0.003 seconds.

RESULTS FROM REAL SPEECH.

The input for the real speech work was obtained from a capacitor microphone which measures the pressure at a distance from the lips. It has been shown Flanagan (6) that this pressure is approximately given by differentiating the volume velocity at the lips. This means that the speech waveform should be integrated before applying analysis to obtain realistic glottal waveforms. As the contributions due to excitation are specifically excluded from the filter it is not necessary to apply a spectral weighting to account for them as in some methods (7). The results from real speech support those from synthetic speech in the following ways;

1). Realistic area functions can be obtained which are consistent providing the analysis is carried out on the correct portion of the speech waveform.

2). The deconvolved glottal pulse is far more sensitive to changes in the analysis interval than is the area function which is remarkably stable to such changes.

At present the author determines the period of glottal closure experimentally in an interactive graphics enviroment. However, experience suggests that the spikes which are obtained by inverse filtering the speech waveform directly, i.e. without differentiation, give a good estimate of the open glottis period and could provide the basis of an automatic inverse filter program. It is also important to remember that as was pointed out by Holmes (8) reliance can only be placed on the accuracy of the deconvolved glottal pulses if the input speech is obtained from an anechoic chamber and faithfully reproduced, i.e. without phase distortion. No such data is at present available to the author.

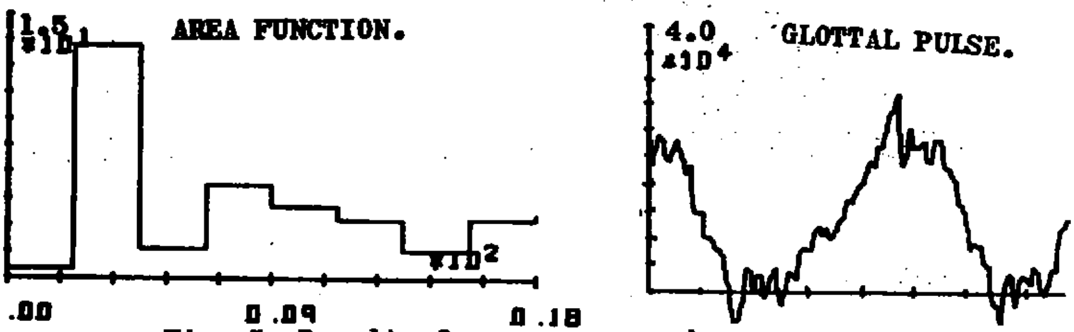


Fig. 5. Results from real speech.

CONCLUSIONS.

A method has been developed which uses Householder transformation and an algorithm consisting of two recursions to give the vocal tract area function. It is the author's belief that more reliable area functions would be obtained if more realistic lip conditions were imposed and losses were included in the analysis. The lip loading can be better approximated if the area of the lips is known. To this end, and to avoid the area normilasation at present necessary, an electronic lip reader based on a television camera has been developed at Imperial College. The losses in the vocal tract can be approximated rather badly at best and one method of including these losses which uses the area of the vocal tract calculated at each stage to modify the transfer function of the remaining sections to avoid accumulative errors is at present under investigation.

The main experimental observations reported here are the facts that the area function is insensitive to errors of the types likely to be encountered in articulatory analysis; this supports earlier work of the author (9), and that the deconvolved glottal pulse is very sensitive to these types of errors. It is hoped now to investigate the usefulness of this method as a feedback loop in teaching deaf people to talk.

REFERENCES.

1). G.H.Golub.1965. Numerische Mathematik 7 (pp206-216)
Numerical methods for solving linear least squares problems.
2). J.D.Markel. 1971. S.C.R.L. monograph No. 7 (pp18-21)
Formant trajectory estimation from a linear least squares inverse filter formulation.
3). B.S.Atal. 1970. J.A.S.A. Vol. 47. No. 1. (pp652-653)
Speech analysis and synthesis by linear prediction of the speech wave.
4). J.D.Markel. 1972. IEEE. AU20 No. 5. (pp. 367-378).
The SIFT algorithm for fundamental frequency estimation.
5). E.V.Stansfield. 1971. Ph.D. Thesis, University of London.
An articulatory model for speech recognition.
6). J.L.Flanagan. 1972. Springer Verlag.
Speech analysis synthesis and perception.
7). H.Wakita. 1972. S.C.R.L. monograph No. 9
Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods.
8). J.N.Holmes. 1962. Paper G13. @ 4th International acoustics congress. An investigation of the volume velocity at the larynx during speech by means of an inverse filter.
9). R.E.Bogner and J.Rogers. 1972.Paper J5 IEEE conference on speech. Determination of the vocal tract area function from a pole description of the vocal tract.
10).B.K.Kinariwala. 1966. B.S.T.J. Vol. 45(pp 638).
Theory of cascaded structures:Lossless transmission lines.