VOCAL TRACT AREA ESTIMATION - EXTENDING THE WAKITA INVERSE FILTER

Róin Scáife
Dept. of Electronic Engineering, UCD, IRELAND
(now at Dept. of Electronic Engineering, DCU, Dublin 9, IRELAND)
EARN: "SCAIFER@VAX1.NIHED.IE"

ABSTRACT

This paper describes an algorithm for identifying the parameters of a Vocal Tract model from a record of the lip-pressure/glottal-flow impulse response. The basis of the identification algorithm is a formal inverse to the reference Vocal Tract network.

The Vocal Tract model is in the form of a digital network, topologically equivalent to the normal lossless acoustic tube model, but with frequency-dependent terminations. The topological equivalence suggests an extension of the Burg [1] algorithm for identification of a lattice network production model. This extended algorithm is described. To assess the performance of the algorithm, it was tested on impulse response records generated from the six Russian vowels of Fant[2]. The results are compared with those of Strube[3].

INTRODUCTION

The Kelly-Lochbaum Vocal Tract model [4], consisting of a chain of cylindrical tubes of equal length, is the basis of a very large class of speech production models. The most widespread version ignores the original (Kelly-Lochbaum) distributed propagation losses, and models all energy dissipation mechanisms by a constant reflection coefficient at each end (glottis and lips) of the model. This acoustical configuration can be exactly simulated by an all-pole digital filter.

In Wakita's formulation [5], the reflection coefficient at the lip end of the model is arbitrarily assigned a value of 1.0, (an acoustical "short-circuit"). The cross-sectional areas of the individual tube sections (the "Area Function"), may then be estimated from the autocorrelation of the response, at the lip end of the tube, to an impulsive or noisy flow at the glottis end.

This model is equivalent to a linear prediction filter, and is therefore often called the "LPC" model. It is important to note that an all-pole filter is still obtained with values of the lip reflection coefficient less than 1.0; the assumption of complete reflection is only necessary to allow direct estimation of the area function via LPC analysis.

The all-pole vocal-tract model is not, however, a realistic model of speech production. It captures only general features of the real vocal tract, and generally incorrectly estimates formant bandwidths. The attractiveness of the Wakita model for estimating the area function lies in its simplicity and computational economy.

This paper reports an investigation of the feasibility of extending Wakita's "inverse filter" approach by replacing the fixed glottal and lip terminations by frequency-varying networks which model the actual loss mechanisms more accurately.

REFERENCE VOCAL TRACT MODEL

Fig. (1a) shows the production model used. \( U_G \) is the volume velocity at the glottis, and \( U_L \) the resulting flow at the lips. When \( r_L \) and \( r_G \) become constant multipliers, this reduces to the all-pole, constant-termination network. It is
easily shown that, whatever the nature of $t_k$ and $r_k$, the network of fig. (1b) is (except for a constant delay) the exact inverse of fig. (1a). That is, if the production model has a transfer function $H_0$, then the inverse filter has a transfer function $x^2/H_0(Q > 0)$.

In the model under discussion, frequency-dependent networks were used for $t_k$ and $r_k$, to improve the modelling of some of the energy-loss mechanisms in a real vocal tract.

The radiation load at the lips is modelled by a one-pole/one-zero impedance due to Laine [6]:

$$Z_r(z) = Z_{LIP} \frac{a'(1-z^{-1})}{(1+bz^{-1})} \tag{1}$$

where $Z_{LIP}$ is the characteristic impedance of a tube of cross-sectional area equal to the lip area. This is an optimal (in the mean-square sense) approximation to the "piston-in-a-square" model of Morse [7]. It is a better match than the usual bilinear-transformation of a parallel resistor-inductor circuit [8], and yet requires only the same computational effort. In the original [6] form, $a$ and $b$ were approximated by linear functions of the square root of the lip area. In the new model, the sampling rate and wide range of lip areas give unacceptable errors with such an approximation, so linear interpolation from a pre-computed table is used instead.

In the reference model, a composite termination at the glottal end is used to model: the source resistance of the larynx, a lumped approximation to the vocal tract wall vibration impedance, and a general "leakage" term to account for the extra dissipation needed to yield realistic formant bandwidths. The source impedance and additional losses are approximated by a constant resistance of 110 (cgs) $\Omega$. The wall load is modelled as a series R-L-C shunt across this glottal impedance. The component values of the lumped wall impedance model are those of Liljencrants [9,11]. (This lumped model is intended to be located at about 4 cm above the larynx [10]; the precise effects of moving the lumped wall load are not yet certain.) A modification of Liljencrants' [11] pole-zero model, (originally used for a distributed wall impedance) is used to represent all the combined impedances. This version has one extra pole, and moves the Liljencrants pole at $z^1$ towards the origin.

The resulting network, used as a synthesizer, matches well the formant frequencies and bandwidths of the six Fant Russian vowels, as calculated in [10]. In particular, the shifts (relative to the "LPC" model) in the first formant due to the wall vibration impedance are accurately reproduced. Table (1) shows values of the first 4 formant frequencies and their bandwidths for the Fant [2] vowels /a/ and /u/; according to Badin and Fant [10], as calculated by the new model, and using an all-pole network with $u_9 = 0.9$ and $u_7 = 0.7$.

**ESTIMATING THE NEW MODEL**

To fit a $N$-parameter model to a particular signal, we must adjust the parameters to minimise some measure of the error of the approximation. In general, this requires one of the many available N-dimensional search algorithms, which iteratively approach the minimum. In a few special cases, (e.g. all-pole or all-zero models) the optimal parameters may be obtained non-iteratively as a transformation of the signal. For example, the Burg algorithm [1] sequentially estimates the reflection coefficients of a lattice (all-pole) model one at a time. In the notation of fig. (1b), given the forward and backward error sequences $f_k$ and $b_{k+1}$, the reflection coefficient $r_{k+1}$ is chosen to minimise the sum of the mean-square values of $f_k$ and $b_k$.

Assuming that a record $M$ samples long of the signal is available, this is achieved by computing:

$$r_{k+1} = \frac{\sum_{i=k+1}^{M} \left( \frac{f(i) * b(i-1)}{\Sigma_{i=1+k}^{M} \left( \Sigma_{i=1+k}^{M} \left( f(i) + b(i-1) \right) \right)} \right)}{1-k-1} \tag{2}$$

The primary motivation of this study was the possible extension of the Burg algorithm to the estimation of a (constrained) pole-zero (ARMA) model.

It is important to note that this problem does not need a general ARMA estimation technique. Although the glottal termination adds two zeros and two poles to the all-pole model, the locations of all four singularities are determined by a single variable parameter, the area of the pharynx tube. Similarly, the lip termination pole and zero are not independent, but are both determined by the lip area.

Unlike many area function estimation methods, the present algorithm works from a record of the lip flow due to an impulsive flow at the larynx. This is the inverse Fourier Transform of the frequency response of the vocal tract. Most all-pole methods assume that only the magnitude frequency response is measurable, and therefore use the autocorrelation function of the lip flow. To measure the complex frequency response from the lip flow requires either knowledge of, or an assumption about, the glottal flow. The most common solution is to assume that the magnitude spectrum of the glottal flow falls off at -12 dB/octave. The present model, however, is intended eventually to be used in a joint estimation of both the area function and a parametric model of the glottal flow during voiced speech, (see [12]-[14]). Such a model, it is hoped, will yield fairly reliable information about the phase spectrum of the glottal flow.

A disadvantage of the LPC approach to area function estimation is that the shape is determined only to within an unknown scaling factor. This algorithm to be described actually estimates the lip area, and thus gives an absolutely scaled area function. This is possible because the reflection coefficient at the lips in this model is a function of the lip area.

The strategy of the estimator is to use a golden-mean

<table>
<thead>
<tr>
<th>F1 (B1)</th>
<th>F2 (B2)</th>
<th>F3 (B3)</th>
<th>F4 (Hz) (B4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Badin-</td>
<td>Fant</td>
<td>Fant</td>
<td>Saclfe</td>
</tr>
<tr>
<td>766</td>
<td>1127</td>
<td>2473</td>
<td>3629</td>
</tr>
<tr>
<td>vowel /a/</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F1 (B1)</th>
<th>F2 (B2)</th>
<th>F3 (B3)</th>
<th>F4 (Hz) (B4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Badin-</td>
<td>Fant</td>
<td>Fant</td>
<td>Saclfe</td>
</tr>
<tr>
<td>313</td>
<td>633</td>
<td>2385</td>
<td>3797</td>
</tr>
<tr>
<td>vowel /a/</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table (1) Formant data for vowels /a/ and /u/; (see text for details), * means bandwidth too wide to be accurately measured.
search of lip area over the range 0.0 - 20.0 cm² to determine the configuration which minimises the mean-squared output from the inverse filter. For each value of lip area, \( g(n) \) and \( b_m(n) \) are computed as shown in fig.(1b), and the reflection coefficients \( r_k \) between tubes are calculated according to (2) above. The output of the inverse filter is then calculated as shown in fig.(1b). The tube areas are then computed from the \( r_k \) as in the Wakita method, except that the estimated lip area now provides an absolute reference for scaling the area functions. In order to prevent utterly unrealistic area function estimates, the sum-of-squares of the inverse filter output is weighted by the following factor:

\[
W = 1 + 2 \cdot \left( \frac{V_{\text{EST}} - V_{\text{AVE}}}{V_{\text{EST}}} \right)^2
\]

where \( V_{\text{EST}} \) is the volume of the vocal tract estimated by the model, and \( V_{\text{AVE}} = 92 \) cm³ is the average volume of the six Fant [2] Russian vowels.

The configuration chosen is the one which minimises \( W \) times the mean-square value of \( U(n) \). This is a loose constraint, which appears not to significantly bias the area function estimates.

In its present form, the estimator does not estimate the glottal reflection coefficient \( r_G \), but fixes it at the value corresponding to a pharynx tube area of 2.7 cm².

Tests with Fant Data

Records of the lip-flow/glottal-flow impulse response ratio of the reference model were calculated, using the area functions of the six Russian vowels of Fant [2]. (A sampling rate of 35300 Hz was used). These records were submitted to the estimator. The log-areas of each of the estimated area functions is shown in fig.(2), together with the log-area function of the original Fant data. It is clear that, except in the case of the vowel "y", the general features of the Fant vowels are reproduced, but that there are significant errors in such important parameters as point of constriction, lip area, and especially just above the glottis end of the vocal tract. (It should be remembered that the Burg algorithm is designed for estimating stationary random signals; the impulse response of the vocal tract is far from stationary).

To show the way in which the optimisation proceeds, fig.(3) is a graph, versus assumed lip area, of the sum-of-squares at the output of the estimated inverse filter, (weighted by the factor \( W \) in eqn.(3)), for the six vowels. In each case the location of the minimum is, in fact, dominated by the dependence of the inverse filter output on the assumed lip area; the volume weighting factor \( W \) does not act substantially to bias the estimated vocal tract volume. Convergence takes 6 to 12 passes of inverse filter evaluation.

Another version of the estimator, without a wall-load model, was also tried. This was used to estimate area functions from data generated by the full reference model: (see next section).

Comparison with Strube Data

To assess the performance of the estimator relative to a global optimising estimator, the area function estimates were compared to those computed by Strube’s constrained “cosine-series” model [3]. Fig.(4) shows the errors in the estimated log-areas (relative to the Fant data) of the new estimator (a) with, and (b) without the wall-load model, (c) of the Strube estimates, and (d) of a classical “LPC” model, terminated with \( r_G = 0.9 \) and \( r_L = 0.7 \). (In case (d), the same minimisation criterion was used as in cases (a) and (b)).

It is apparent from curves (a) and (b) that the presence or absence of the wall-load model makes no great difference to the quality of the estimate: this is quite surprising, in view of the effect of the wall-load model on the location of the first formant in the reference model.

In general, the new estimator seems to do substantially better than the Strube cosine-series model. Again, this is surprising; one would expect the Strube (global) estimator, to do better.

Rather than replicate the iterative procedure used by Strube, the new estimation algorithm was also used to estimate the reference model of Strube. This differs from the present model in using the Flanagan [7] parallel resistor-inductor lip radiation model, and a series resistor-inductor glottal termination to model the wall-vibration load. Fig.(5) shows the log-area errors (relative to the Fant data) of Strube’s [3] estimated area functions, and of the new estimator using the Strube acoustic model. The implicit criteria in the two estimates are obviously very different.

Summary

A simple extension of the all-pole vocal tract model is described, which matches quite accurately the formant frequencies and bandwidths computed from the Fant area functions [10]. The new vocal-tract model has an easily computed formal inverse network.

An extension to the Burg algorithm [1] is proposed to
estimate an otherwise all-pole lattice network with frequency-dependent blocks at each end. The algorithm has been applied to the estimation of the vocal tract model described above. Although the proposed composite glottal termination considerably improves the fidelity of the frequency response of the reference model, its inclusion in the inverse filter does not significantly improve the estimated area function.

ACKNOWLEDGEMENTS

This work was largely carried out while the author was on sabbatical leave at KTH Stockholm, and Dublin City University. The author is very grateful to Profs. B. Granström (KTH) and C. Mc Corkell (Dublin City University), and to his colleagues at University College Galway for their support.

REFERENCES


