Publications of Dr. Martin Rothenberg:

Nonlinear inverse filtering technique for estimating the glottal-area waveform

by Martin Rothenberg and Stephen Zahorian

Syracuse University, Syracuse, New York 13210
(Received 1 July 1976; revised 29 December 1976)

Published in the *Journal of the Acoustical Society of America*, Volume 61, Number 4, April 1977

ABSTRACT

Linear time-invariant inverse filtering of the pressure or airflow signal at the mouth during voicing has often been used to provide a reasonable representation of the airflow waveform at the glottis. However, if the glottal impedance, as it varies during the glottal cycle, does not remain much larger than the impedance of the supraglottal vocal tract, the resulting glottal-supraglottal acoustic interaction may make the glottal airflow a poor representation of the motion of the vocal folds. An inverse filter is described here for deriving from oral airflow the Norton equivalent airflow of the glottal source, which is a better estimate of the glottal area than the actual airflow. This inverse filter is nonlinear in that certain filter parameters are varied dynamically, so as to compensate for the time variation of the glottal resistance and inertance. Using a feedback procedure, estimates of these glottal parameters are obtained from the filter output waveform. Results from such nonlinear inverse filtering support the conclusion that the nonlinear glottal-supraglottal interaction may be a significant factor in determining the spectral content of the glottal pulse and for some types of voice, in causing the primary excitation of the vocal tract to occur at the closing of the vocal folds.

PACS numbers: 43.70.Gr, 43.70. Bk

INTRODUCTION

Movements of the vocal folds during voiced speech are often studied through the glottal area waveform. However, because of the relative inaccessibility of the glottis, it would be convenient to be able to estimate glottal area from some measurement made outside of the body. When the supraglottal vocal tract is not overly constricted, the glottal volume velocity can be used as a first approximation to glottal area, and estimates of the glottal-volume-velocity waveform can be derived from the pressure or airflow at the mouth by

the technique of inverse filtering. In linear inverse filtering, for a given vocal-tract configuration the acoustic effect of the supraglottal vocal tract is canceled by a linear, time-invariant filter having a transfer characteristic that is the inverse of that of the supraglottal system. However, we will show that a better approximation to the glottal area function can often be obtained by means of a nonlinear inverse filter which accounts for the nonlinear acoustic interaction between the source and the supraglottal vocal tract.

A lumped-parameter model commonly proposed for the glottal-supraglottal system during voicing is shown in Figure 1. In this model, the subglottal pressure P_{sg} is usually assumed to be constant over the glottal cycle. Although P_{sg} actually has small variations, van den Berg *et al.* estimate these variations to be less than 5% of the average pressure for normal voicing. [1] (Since the amplitude of the subglottal pressure variations tends to be proportional to the glottal pulse height and inversely proportional to lung volume, voicing with an abnormally large glottal pulse and/or with an unusually small lung volume would result in somewhat larger variations.) The glottis is represented by the resistive and inertive parameters $R^*_g(t)$ and $L^*_g(t)$, which vary as functions of the glottal area $A_g(t)$ and glottal volume velocity $u_g(t)$. [1], [2] The resistive parameter, R^*_g, has been shown to consist primarily of a viscous component proportional to A_g^{-3}, and a kinetic component proportional to $u_g A_g^{-2}$, with the kinetic component predominating when the glottal opening is larger than about 1/5 its maximum value during the glottal cycle.

If the acoustic pressure just above the glottis is negligible compared to P_{sg}, there is no glottal-supraglottal acoustic interaction. This condition is sometimes said to occur when the total glottal impedance is large compared to the vocal-tract impedance at the glottis, Z_t in the figure, although the glottal impedance is formally not well defined (except in a small signal sense) due to the nonlinearity (dependence on u_g) of the glottal parameters. Simplified models of the action of the glottis during voicing usually assume that this acoustic interaction between the glottis and the supraglottal vocal tract is negligible. To the extent that this is true, the glottal airflow is determined primarily by the kinetic resistive term, and is approximately proportional to glottal area except at small glottal openings.

However, the acoustic interaction between the glottis and the supraglottal vocal tract cannot always be neglected. This is illustrated in Figure 2. The top waveform shows the volume velocity at the mouth during a vowel [æ] produced by an adult male speaker. It was obtained using the circumferentially vented pneumotachograph mask described previously. [3] The mask had a frequency response flat to about 1000 Hz. The bottom trace shows the volume velocity at the glottis, as obtained by inverse filtering the signal shown in the top trace. The first formant of the vocal tract is excited primarily by the closing of the glottis and the oscillations at the formant frequency die out quickly during the opening phase of the glottis, as energy is lost through the open glottis. The formant energy lost through the open glottis appears in the bottom trace as an oscillation during the rising portion of the glottal wave. Model studies by Flanagan and Landraf [4] indicate that this oscillation does not occur in the actual motion of the vocal folds. This means that the glottal airflow and glottal area functions may be significantly different.

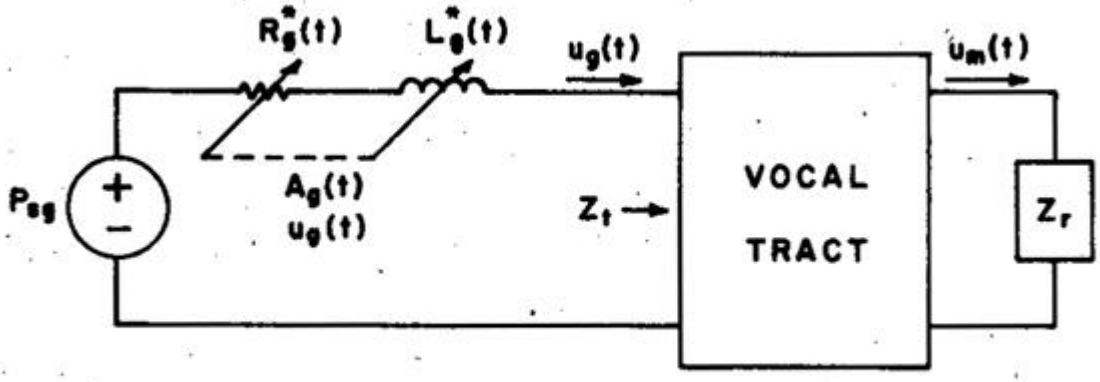

FIG. 1. Model for the glottal–supraglottal system.

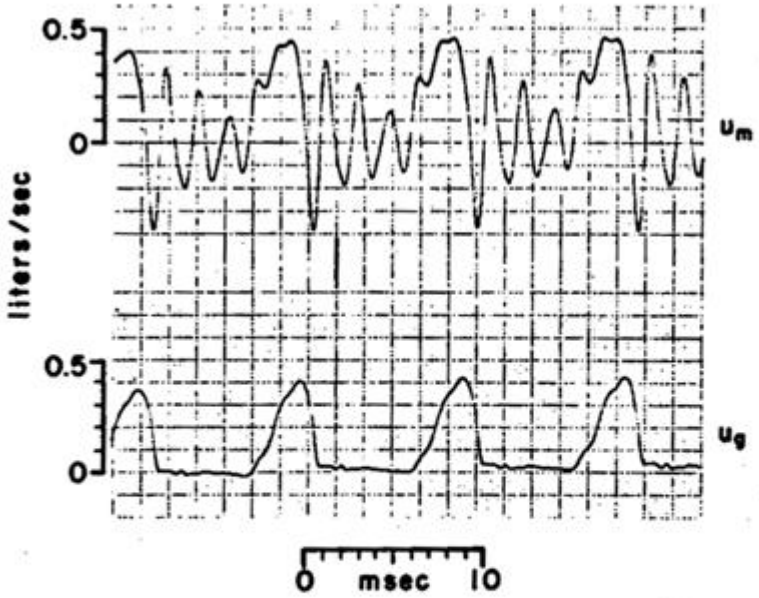

FIG. 2. Oral and glottal volume velocity waveforms illustrating the damping of formant energy during the open phase of the glottal cycle. The u_g trace is delayed about 0.7 msec with respect to the u_m trace due to the inverse filtering procedure used to obtain u_g.

If the supraglottal pressure is not negligible, the perturbation in u_g that results from this pressure is not easy to visualize from the model of Figure 1, since the desired glottal parameters for the incremental calculations are not the nonlinear functions R^*_g and L^*_g, but the time-varying linear differential resistance and inductance, R_g and L_g. R_g is

obtained by partial differentiation of pressure with respect to flow, at the value of area present at that instant, and at the value of flow specified by the model of Fig. 1, with Z_t = 0. Since L^*_g is approximately proportional to A_g^{-1} and independent of u_g, L_g is approximately equal to L^*_g. These parameters can be presented more clearly in the "Norton equivalent" model shown in Figure 3. The source $u_f(t)$ in this model is the flow that would occur at the glottis if the impedance of the supraglottal tract were zero, and, as noted above, is approximately proportional to glottal area. R_g and L_g are the incremental parameters, and the flow through R_g and L_g represents the change in u_g from u_f (and therefore from A_g) caused by the supraglottal pressure. The model can be considered accurate as long as Z_t is fairly small compared to $Z_g=R_g+jwL_g$, or alternatively, if the supraglottal pressure variations are small compared to P_{sg}. The calculations of Flanagan and others indicate that this is the case over a wide range of voice qualities and vowel types. The model becomes less valid for very occluded supraglottal vocal tract configurations and for voicing at high loudness levels, since both these conditions tend to raise the ratio of supraglottal pressure to subglottal pressure.

u_f is related to the airflow at the mouth u_m by a time-varying, nonlinear vocal-tract system, since this system includes R_g and L_g. Since for most vowels, and for the small glottal openings characteristic of normal voicing, it is primarily first formant energy that is lost through the glottis as it opens [2], the variation in the acoustic properties of this nonlinear vocal tract as the glottis opens can be characterized by an increase in the damping of the first formant and a small variation in its frequency. We now consider whether or not it is possible to derive the unloaded glottal flow u_f from the actual airflow at the mouth, u_m, by using some form of time-varying, nonlinear inverse filter, limiting ourselves to the effect of the variation in the frequency and damping of the first formant caused by the variation in glottal impedance.

I. MATHEMATICAL ANALYSIS

A. Linear time-invariant system

The transfer function $H(s)$ of a linear time-invariant system can be defined as the Laplace transform of the impulse response of the system. If the initial conditions are zero, the transform $Y(s)$ of the output $y(t)$ is related to the transform $X(s)$ of the input $x(t)$ by

$$Y(s) = X(s)H(s) \qquad (1)$$

When two systems are cascaded, the overall transfer function is the product of the two individual system transfer functions. Hence, when it is desired that the overall system function be unity for a system of two cascaded blocks, the second block should have a transfer function equal to the reciprocal of that of the first block, and one block is said to be the inverse of the other. Where needed for realizability, the response of an inverse may be modified outside the frequency range of interest, in which case the inverse is approximate, but may be made to be arbitrarily accurate over any given frequency range.

For example, a system with a transfer function with two complex-conjugate poles (roots of the denominator),

$$H(s) = 1 / (as^2 + bs + 1) \quad (2)$$

would have an ideal inverse transfer function

$$H^{-1}(s) = as^2 + bs + 1 \quad (3)$$

where a and b are constants.

However, $H^{-1}(s)$ is unrealizable, since for a realizable inverse system the number of poles must be at least as large as the number of zeros (roots of the numerator), or else the frequency response becomes infinite as f ® ¥. A practical inverse for $H(s)$ might be

$$H^{-1(s) = (as_2} + bs + 1) / (a's^2 + b's + 1) \quad (4)$$

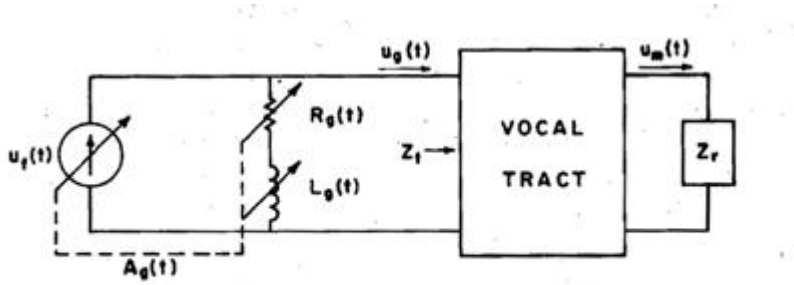

FIG. 3. Norton equivalent model for the glottal source during voicing.

The cascaded system would then have a transfer function with two poles:

$$H_T(s) = H(s)H^{-1}(s) = 1/(a's^2 + b's + 1) \quad (5)$$

The constants a¢ and b¢ that determine the two poles added to the inverse filter would be chosen so as to place the pole frequencies well above the frequency range over which the inverse system is to be accurate. For example, in the inverse filter used in this study to obtain glottal volume velocity by inverse filtering the volume velocity at the mouth, the poles of the inverse were placed above 2000 Hz.

To simplify the equations in the following discussion we assume an inverse filter with an equal number of poles and zeros, although the conclusions are valid for an increased number of poles.

The filter and the practical inverse filter can also be described in differential equation form:

$$a(d^2y(t) / dt^2) + b(dy(t) / dt) + y(t) = x(t) \text{ (filter)} \quad (6)$$

$$a'(d^2z(t) / dt^2) + b'(dz(t) / dt) + z(t) = a(d^2y(t) / dt^2) + b(dy(t) / dt) + y(t) \text{ (inverse filter)} \quad (7)$$

where, as in (1), $x(t)$ and $y(t)$ are the input and output of the filter, while $z(t)$ is the output of the inverse filter.

B. Time-varying systems

The standard transfer-function representation is generally not possible for linear time-varying systems. [5] However, we can define a transfer function-similar to the transfer function of a linear time-invariant system-and show that it is of value for our application. Consider the filter, analogous to the one previously mentioned, that is described by the follow

$$a(t)(d^2y(t) / dt^2) + b(t)(dy(t) / dt) + y(t) = x(t) \text{ (time-varying filter)} \quad (8)$$

ing differential equation with time-varying coefficients:

If the inverse filter has input y(t), output z(t), and is described by the differential equation

$$a'(d^2z(t) / dt2) + b'(dz(t) / dt) + z(t) = a(t)(d^2y(t) / dt) + b(t)(dy(t) / dt + y(t) \text{ (time-varying inverse filter)} , \quad (9)$$

then

$$a'(d^2z(t) / dt^2) + b'(dz(t) / dt) + z(t) = x(t) , \quad (10)$$

i.e., the overall system consisting of the cascade connection of the filter and inverse filter is described by a linear differential equation with constant coefficients. Once again the transfer function description is applicable and the cascaded system function would be

$$H_T(s) = Z(s) / X(s) = 1/(a's^2 + b's + 1) . \quad (11)$$

Since the composite system is described by a transfer function, we can also formally define a time-varying transfer-function notation to model the time-varying component systems. The transfer function for the time-varying filter has “time-varying poles” and is defined as

$$H(s, t) = 1 / (a(t)s^2 + b(t)s + 1). \quad (12)$$

Similarly, the inverse filter is described as having “time-varying zeros” which cancel the filter “time-varying poles” and fixed poles. In this notation the inverse filter description is

$$H^{-1}(s, t) = (a(t)s^2 + b(t)s + 1)/(a's^2 + b's + 1) . \quad (13)$$

The “transfer functions” given above for the time-varying systems cannot be obtained by taking the Laplace transform of the original differential equations (8) and (9), since $L\{a(t)f(t)\} \neq L\{a(t)\}L\{f(t)\}$. They are obtained formally by taking the transfer functions for the corresponding time-invariant systems and replacing the time-invariant differential equation coefficients which appear in these transfer functions by the corresponding time-varying coefficients. If the coefficients vary slowly with respect to the differential equation variables, these transforms are consistent with our usual notions of Laplace transforms, except that the poles and zeros vary with time.

C. The vocal tract as a time-varying second-order system

In a practical situation, the time variation of the first formant parameters within a glottal period would not be known directly, but would be at best specified only through the variation of the glottal parameters R_g and L_g. If these glottal parameters vary quickly, the relation between them and the formant parameters can be quite complex, however we can get some idea of what the formant variation might be by considering the case of a linear system with slowly varying parameters. According to the calculations of Flanagan [2] and the measurements of Fujimura and Lindquist [6], a finite glottal impedance of a magnitude corresponding to that of the glottal opening during voicing would cause a significant increase in the damping of the first formant, and a small increase in the formant frequency. We derive here expressions for the variation of the formant with R_g and L_g, in a form more convenient for us than the expression derived by Flanagan, using a somewhat different set of assumptions.

Figure 4 shows the glottal source model connected to a one-stage lumped-parameter approximation to the vocal tract. [7] The inertance L includes a component due to radiation impedance, and $u_m(t)$ is the volume velocity at the mouth. The transfer function relating u_t and u_m can be written in the form

$$U_m(s) / U_f(s) = 1 / (s^2LC + s(L / R) + 1 + sL - (R_g + sL_g)) , \quad (14)$$

where the term $sL/(R_g + sL_g)$ represents the effect of R_g and L_g. Though this term makes the system third order, we can obtain a second-order approximation by representing this term by the first three terms in its Taylor series expansion:

$$F(s) = sL / (R_g + sL_g) \approx F(0) + F'(0)s + F''(0)s^2 = 0 + (L / R_g)s - (LL_g / R^2_g)s^2 . \quad (15)$$

Equation (15) will be fairly accurate for frequencies such that wL_g is less than R_g, which, according to Flanagan [2], will be true for frequencies up to about 1000 or 2000 Hz.

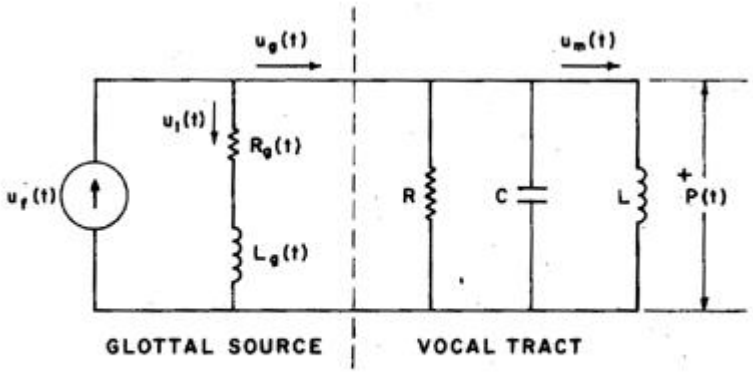

FIG. 4. Simplified electrical network model of glottal source/vocal tract system.

The resulting second-order transfer function is

$$U_m(s) / U_f(s) \approx 1 / (s^2(LC - LL_g / R^2_g) + s(L / R + L / R_g) + 1) . \quad (16)$$

For a denominator of the form $as^2 + bs + 1$, the formant radian frequency is equal to $(1/a)^{1/2}$, where formant frequency is defined to be the magnitude of the complex pole frequency. The damping factor s associated with a formant is the magnitude of the cosine of the angle of the complex pole frequency, and is equal to $(b^2/4a)^{1/2}$. If R_g and L_g are both roughly inversely proportional to glottal area, then according to Equation (16) the formant damping factor will increase with glottal area, since $[L/R + L/R_g]$(coefficient b) increases, while $[LC\text{-}LL_g/R^2_g]$(coefficient a) decreases. The formant radian frequency, as determined by $1/[LC\text{-}(LL_g/R^2_g)]^{1/2}$, will also increase with glottal area. In relating these results to actual formant measurements, however, it should be kept in mind that for a given formant frequency, the frequency of the peak in the spectral analysis of the waveform decreases with an increase in damping factor, and so might actually decrease if the damping increase were sufficiently large relative to the increase in formant frequency. A third measure of formant frequency, the imaginary part of the complex frequency, which occurs in the waveform as the frequency of the ringing in the damped oscillations associated with the formant, is also decreased by an increase in the damping factor, though not as much as is the spectral peak.

II. A TIME-VARYING SECOND-ORDER INVERSE FILTER

The inverse filter illustrated in Figure 5, designed and built in our laboratory for use in canceling a complex pole pair of a time-invariant linear system, was converted to a time-varying inverse filter for the vocal tract. The filter is an analog system using standard operational amplifier circuitry. The frequency and damping controls are normally adjusted manually to minimize the formant frequency oscillations during the most-closed portion of the glottal cycle, as determined by observing the output waveform of the filter. $P(s)$ is a third-order polynomial $a's^3 + b's^2 + c's + 1$ such that $1/P(s)$ is a linear-phase low-pass filter with a response time of 0.35 msec. The modified inverse filter is shown in

Figure 6. The differential equation relating the output $z(t)$ of this system to the input $y(t)$ is

$$a'(d^3z(t)/dt^3) + b'(d^2z(t)/dt^2) + c'(dz(t)/dt) + z(t) = [K_1^2 - K_1K_{11}d(t)](d^2y(t)/dt^2) + [K_1K_2 + K_1K_{22}d(t)](dy(t)/dt) + y(t) + E(t)\,, \quad (17)$$

where $E(t)$ is considered to be an error term that becomes small if either the fixed poles are high in frequency (a', b', and c' are small), the time-varying parts of the coefficients are small (K_{ll} and K_{22} are small), or the coefficients vary slowly [the derivative of $d(t)$ is small]. The first condition could presumably be used to make the error term arbitrarily small.

In this system a single variable $d(t)$ is used to vary both damping and frequency. K_{22} controls the amount of change in damping factor caused by $d(t)$, and K_{ll} the amount of frequency deviation, though K_{ll} also has some effect on damping.

Comparison of Eqs. (16) and (17) show that if $E(t)$ *is* sufficiently small and if the ratio of L_g to Rg can be assumed constant, since both vary approximately inversely to area, then with

$$\begin{gathered} K_1 = (LC)^{1/2}\,, \\ K_2 = (1/R)(L/C)^{1/2}\,, \\ K_g = L_g / R_g\,, \\ K_{11} = K_g(L/C)^{1/2}\,, \\ K_{22} = (L/C)^{1/2}\,, \\ \text{and} \\ d(t) = 1 / [R_g(t)]\,, \end{gathered} \quad (18)$$

this inverse filter will be appropriate for the timevarying vocal tract. In actuality, the variations of glottal inertance and resistance are not quite proportional as the glottis opens, however K_{ll} and K_{22} can be trimmed for a best average setting.

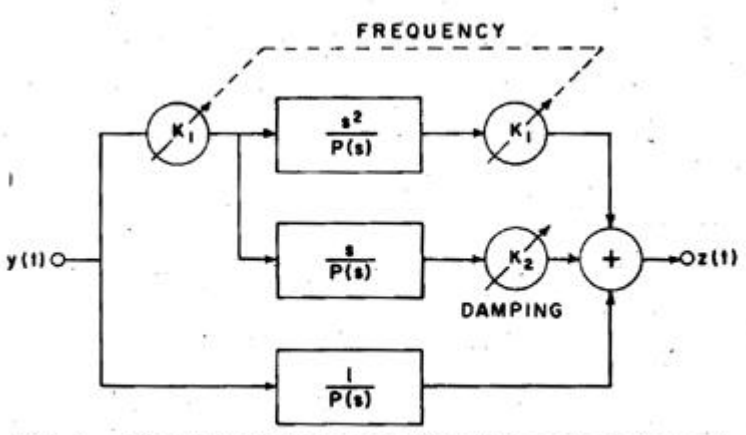

FIG. 5. Time-invariant inverse filter for a single formant.

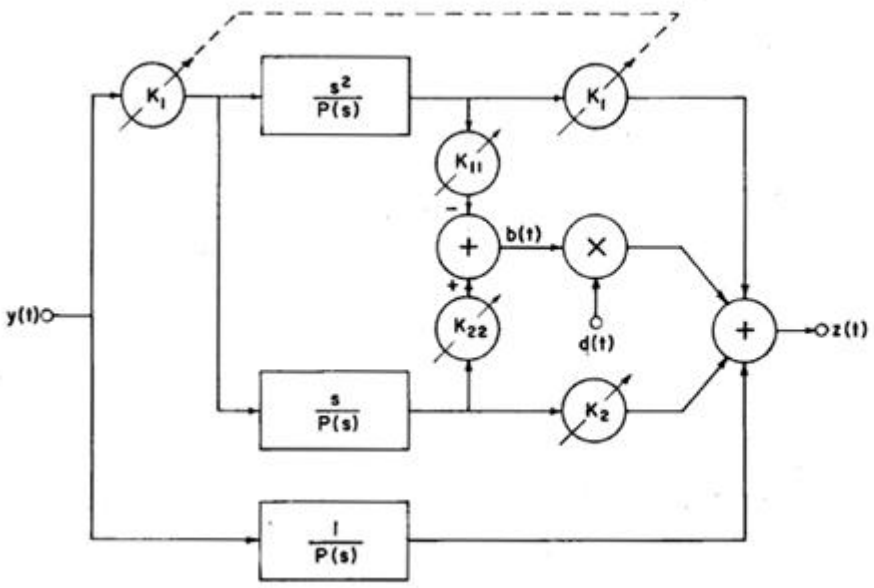

FIG. 6. Time-varying inverse filter.

III. THE USE OF FEEDBACK FOR INVERSE FILTER CONTROL

The practical implementation of a time-varying inverse filter requires the correct control signal $d(t)$ for the time-varying parameters. Since the Norton equivalent glottal flow, $u_f(t)$ is nearly proportional to glottal area, and since the glottal parameters Rg and L_g tend to vary inversely with area (though not entirely linearly), $u_f(t)$ could be used as a first approximation to the control signal. Since $u_f(t)$ is the output of the time-varying inverse filter, we may use a feedback system, with the inverse filter output used as the control signal $d(t)$. Note that the feedback converts the time-varying filter into a nonlinear filter. Therefore this formulation will sometimes be referred to as a nonlinear inverse filter in the following discussion.

When using feedback one naturally asks if the system will be stable and the results consistent with those obtained with a nonfeedback time-varying system [*i.e.*, a system where the inverse filter control signal $d(t)$ is a replica of the glottal parameter]. If, in Figure 6, $1/R\text{g}(t)$ is used as the control parameter $d(t)$, there exists a positive-feedback loop through the multiplier and final summer in Figure 6. The feedback loop will be unstable when the loop gain exceeds unity, which occurs whenever the other input to the multiplier, $b(t)$ in the figure, is more than unity. For vowels with an average or higher than average first formant frequency, it can usually be assumed that the peak amplitude of the formant frequency oscillations at the input is approximately the same as the peak amplitude of the glottal waveform. This property is well illustrated by the example in Figure 8, and can be supported theoretically by observing that at the first formant frequency the termination of the glottal pulse during normal voicing approximates a negative-going step function with an amplitude almost equal to the peak of the glottal pulse. Such a step function would result in a damped oscillation with a peak amplitude exactly that of the step. Carryover of the damped oscillations between glottal pulses is

small because of the increased damping of the first formant during the open phase of the glottal cycle. Making this assumption, it can then be shown that the voltage $b(t)$ will reach unity when either K_{11} is set for an infinite variation of frequency during the glottal pulse, or K_{22} is set for a peak change in the damping factor of 0.5 from its value during glottal closure (with a damping factor of 1.0 representing the critically damped, non-oscillatory case).

Thus we see that the damping control is the more critical in determining stability. Under the above conditions, if the damping factor in the closed-glottis state is about 0.05 (a typical value when a mask is used), an open-glottis damping peak of up to about 0.55 should be possible with no instability. This was approximately what was found experimentally, when processing speechlike synthesized waveforms and noting the setting of K_{22} at which instability began. When the feedback was in the stable range, the filter output was very similar to that obtained with a timevarying inverse filter not using feedback. Some typical results are shown in Figure 8.

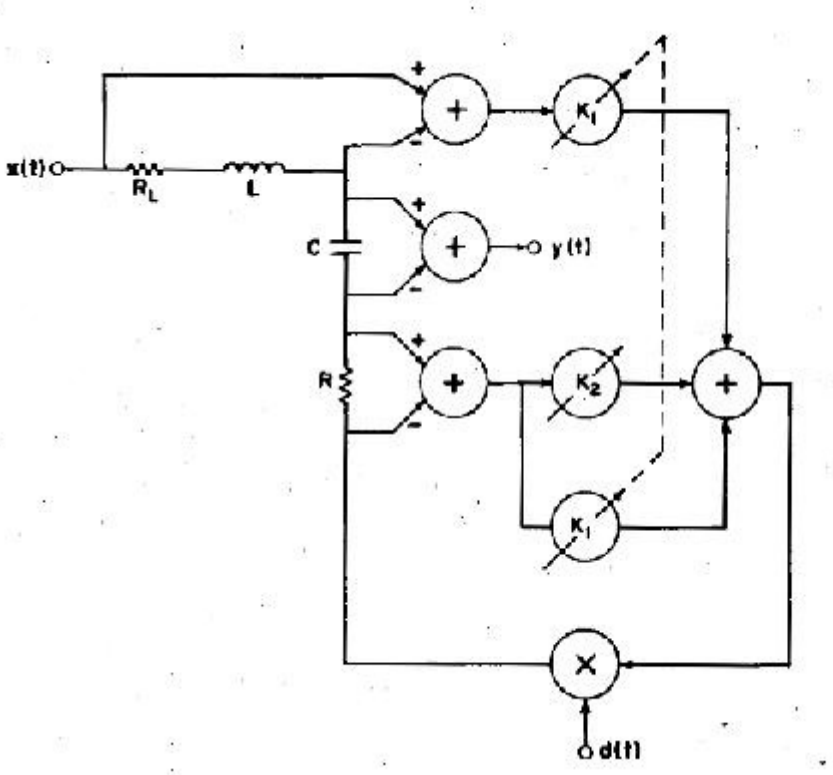

FIG. 7. Time-varying filter.

IV. EXPERIMENTAL RESULTS

A. Constructed waveforms

The time-varying filter circuit shown in Figure 7 was constructed to simulate a pair of complex-conjugate “time-varying poles”. For this system the differential equation relating the input $x(t)$ and the output $y(t)$ is

$$\{LC[1 - K_1 d(t)]\}(d^2y / dt^2) + \{R_L + R[1 + K_2 d(t)]\}(dy(t) / dt) + y(t) = x(t) \qquad (19)$$

Comparison of the above equation with Eq. (16) shows that this filter is of the same form as the vocal tract if the ratio of $L_g(t)$ to $R_g(t)$ is constant.

The filter/inverse filter combination was tested with constructed waveforms similar to the glottal volume velocity during voicing, using a repetition rate of 180 Hz and a pulse width of approximately 3.5 msec. The photos in Figure 8 show the output of the inverse filter when there is no compensation for the time-varying effects [$d(t)$ in the inverse filter equal to zero] and when there is compensation for these effects, using a nonlinear inverse filter [with $d(t)$ for the inverse obtained from the inverse filter output]. The magnitudes of the variations in formant frequency and damping during the simulated glottal pulse, as set by K_1 and K_2 of Eq. (19), were chosen to be slightly larger than the variations to be expected in real speech. Df in the figure is the peak change (increase) in the formant frequency during the glottal pulse. Our vocal-tract filter was calibrated in damping for percent decay/cycle, m, where m varies from zero (undamped) to 100% (critically damped). This measure is related to damping factor d by the following functions:

$$m = 100\{1 - exp[-2\pi\sigma / (1 - \sigma^2)^{1/2}]\} , \qquad (20)$$

$$\sigma = (-lm(1 - (.01)m) / \{(2\pi)^2 + [ln(1 - (.01)m)]^2\}^{1/2} . \qquad (21)$$

In Figure 8 Dm is the peak change in the decay/cycle during the glottal pulse. For the case of the nonlinear inverse filter (using feedback) the constants K_{11} and K_{22} were set to yield the same peak values of frequency and damping variation as in the vocal-tract filter.

It can be seen in Figure 8 that the output of the cascaded system consisting of the time-varying vocal-tract filter and nonlinear time-varying inverse filter closely resembled the input. Furthermore, with a linear time-invariant inverse filter, the formant frequency oscillations during the glottal opening phase resemble those observed in glottal volume velocity waveforms obtained by time-invariant inverse filtering of oral airflow.

B. Actual speech

Figure 9 shows the system used to test nonlinear inverse filtering with real speech. A section of the vowel [æ] from the center of the nonsense syllable [*b* æ *p*] was used for the experiment. With $d(t) = 0$ (a time-invariant filter), the inverse filter was first adjusted for the best representation of the glottal airflow signal by minimizing the formant frequency oscillations during the interval of glottal closure. [3] The resultant settings were $F_1 = 560$ Hz at 25% decay/cycle and $F_2 = 1030$ Hz at 25% decay/cycle. The first formant filter was then made time varying, and the constants K_{11} and K_{22} determining the frequency and damping changes during glottal openings were adjusted to give the best-appearing glottal airflow signal, as judged by the elimination of the formant frequency oscillations during

the opening phase of the glottis. In the result shown in Figure 10, F_1 varies from 560 to 640 Hz, and the damping varies from 25% to 85% decay/cycle. The baseline clamping circuit modified $d(t)$ slightly to remove any small off set from zero during the interval of glottal closure.

The range of damping and f requency variation required, though reasonable in terms of the acoustic properties of the vocal tract, cannot be considered to have any absolute significance. It should be kept in mind that our assumption of a variation of frequency and damping proportional to the output waveform was crude in light of the nonlinearity of the actual time-varying parameters.

V. SOME CONCLUSIONS

The waveforms from actual speech in Figure 10, as also the waveforms in Figure 8 derived from constructed signals, show that the nonlinear time-varying filter can produce a glottal pulse that is slightly more symmetrical (having rising and falling phases more nearly equal) than does the linear filter. Since the linear filter, if properly adjusted, produces the actual glottal airflow, and the nonlinear filter in theory produces a waveform more related to the glottal area function, it can be inferred that the acoustic interaction between the glottis and the supraglottal vocal tract is at least partially responsible for the often-noted skewing of the glottal airflow waveform, which causes the primary excitation of the supraglottal vocal tract to occur during the glottal closing phase. A similar relation between glottal area and glottal airflow can be seen in computer simulations of the vocal tract which take into account the nonlinear glottal-supraglottal interaction. [8-10] In fact, high-speed motion pictures of the vocal folds have sometimes shown a tendency for the vocal cords to open faster than they close, which would result in an asymmetry in the area waveform opposite to that usually noted in the airflow waveform. [11] It is also interesting to compare the effect of glottal-supraglottal interaction in the human vocal tract with the conclusion of Backus and Hundley [12] that the mechanism primarily responsible for harmonic generation in the trumpet is the relationship between the input impedance of the trumpet and the time-varying impedance of the player’s lips during a cycle.

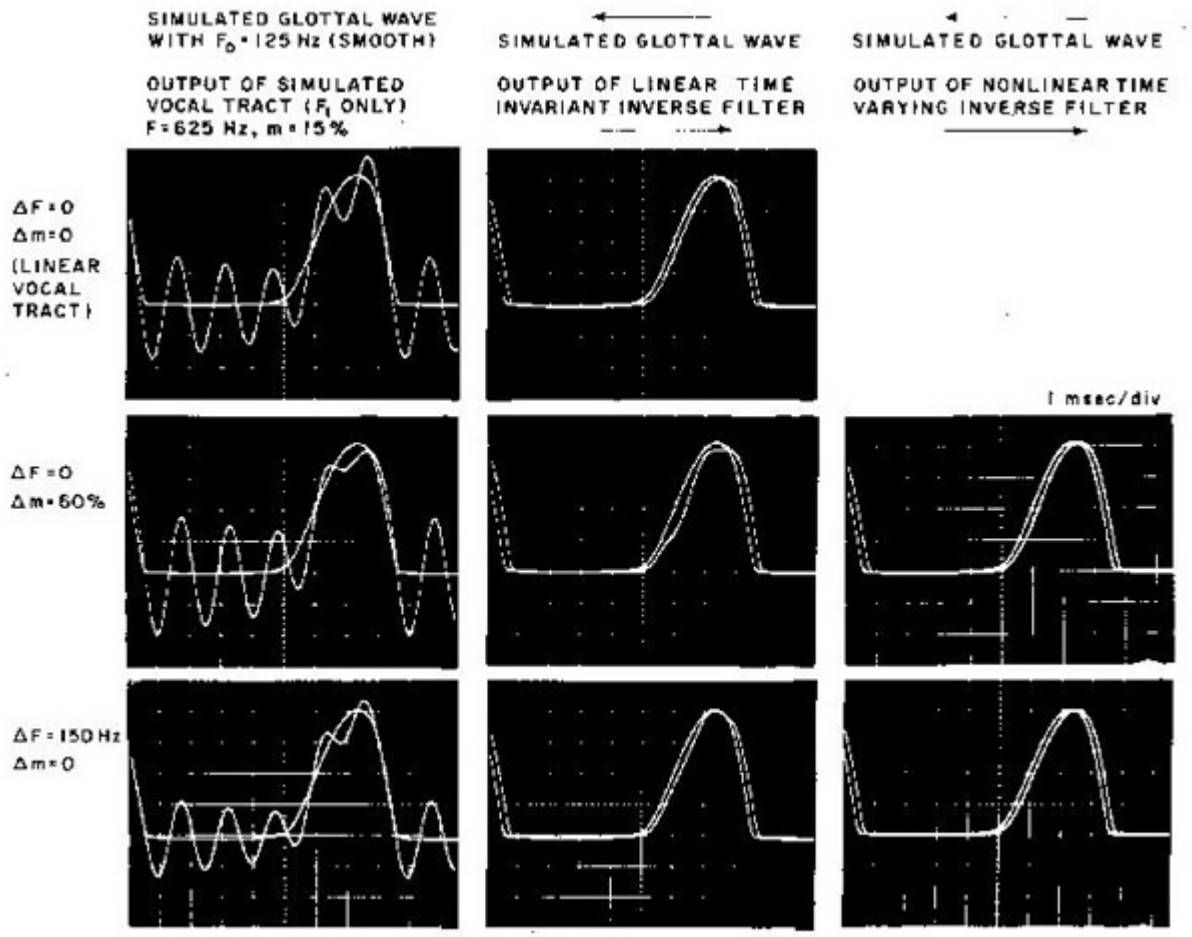

FIG. 8. Inverse filtering results with simulated glottal waveforms.

Though glottal-supraglottal interaction may be typically small during speech at low or moderate amplitudes, it is possible that this effect may be emphasized in some voice modes. For example, nonlinear interaction may be used by trained singers to increase the strength of the higher harmonies. In the operatic or “ring” mode, it is known that the energy near the third formant is much stronger than in ordinary speech, presumably to allow the singer’s voice to carry over the lower frequencies of the sounds of the orchestra. [13] In glottal airflow waveforms obtained by inverse filtering from two male trained singers singing either *F* (174.6 Hz) or *G* (196 Hz) we have commonly noted glottal pulses in which most of the decay of airflow occurred within about 1/2 msec. Such rapid changes would not be expected to occur in glottal area due to the mass of the vocal folds, and do not seem to have been reported in high-speed photographs of the vocal folds. It might be inferred that the fast closing phase in the flow waveform is generated by some nonlinear effect, and that the trained singer may be able to emphasize this effect.

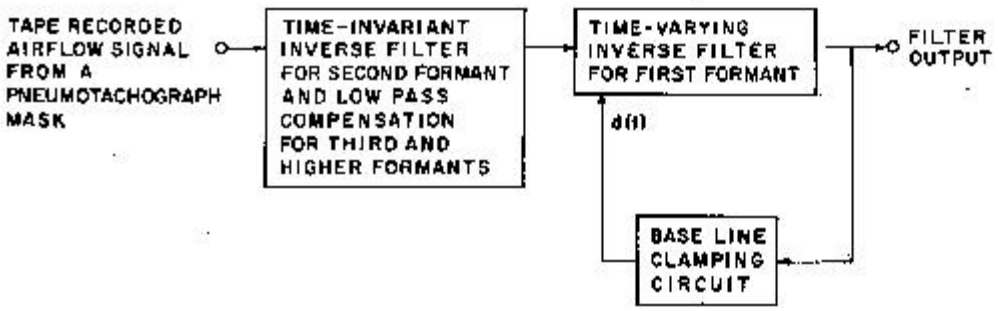

FIG. 9. System for nonlinear inverse filtering of real speech.

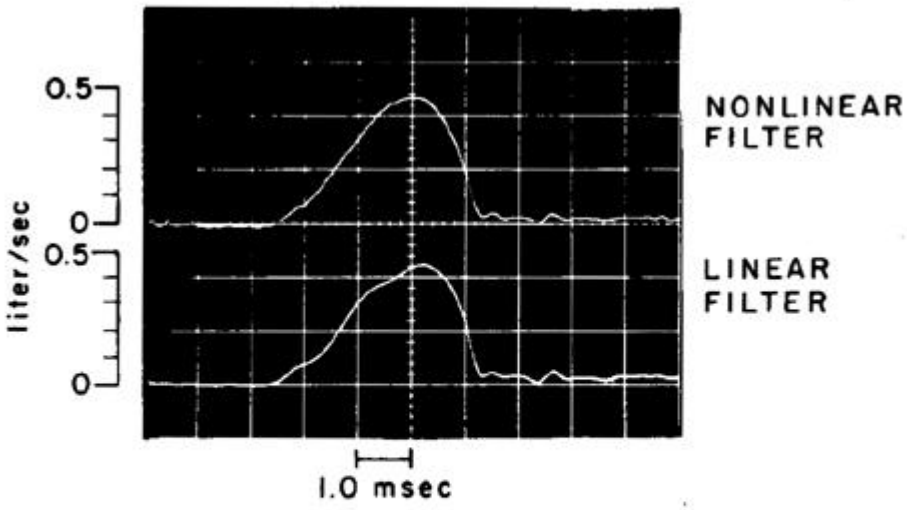

FIG. 10. Nonlinear inverse filtering of a spoken vowel [æ]. The top trace shows the output of the time-varying nonlinear inverse filter. The bottom trace shows the output of the time-invariant inverse filter.

If we approximate the supraglottal load impedance by a simple first-order system, *i.e.*, in the lumped-parameter electrical analogy, an *RL* or an RC circuit with one *L* or one *C*, then the skewing of the glottal airflow pulse to the right occurs when the load is inertive in phase. This condition holds for frequencies below the first formant, and therefore a strictly inertive load would be a good approximation for vowels with a high first formant, such as [a]. To illustrate this type of nonlinear interaction, Figure 11 shows the result of a simulation of a time-varying resistive glottis $R_g(t)$ and an inertive supraglottal load L_t. if P_{sg} is assumed to be constant, the Norton equivalent glottal flow $u_f(t)$ can be considered to be approximately proportional to the glottal conductance $1/R_g$ and to the glottal area waveform. The function $u_f(t)$ was represented by a (symmetrical) half-wave rectified sinusoid. The time-varying *RL* network was represented by an analog simulation of the differential equation of the network:

$$L_t(du_g / dt) + R_g(t)u_g = P_{sg} . \quad (22)$$

The simulated glottal flow u_g is related to u_f by the differential equation

$$u_f(t) = (L_t / R_g(t))(du_g(t) / dt) + u_g(t) . \quad (23)$$

It can be seen that the pulse in u_g is skewed to the right, with an accelerated closing phase. The parameters L_t and R_g were selected empirically to produce a degree of skewing of u_g similar to that occurring in actual glottal airflow waveforms.

A primary limitation in the method used here for adjusting the time-varying parameters of the nonlinear inverse filter is that it depends on the presence of a feature in the waveform definitely attributable to the glottal-supraglottal acoustic interaction (formant frequency oscillations during the opening phase of the glottis). If a better mathematical model were available for the dependence of formant damping on the nonlinearly filtered airflow, the setting of the time-varying parameters could be less *ad hoc.* A less satisfactory alternative would be to use settings for the time-varying parameters derived from one vocalization for other vocalizations in which the optimum setting is not as clear.

A limitation on the accuracy of the feedback method for deriving the control parameters of a time-varying inverse filter is the assumption that the Norton equivalent glottal airflow u_f is proportional to the glottal conductance. The most tenuous part of this representation is probably the assumption of constant subglottal pressure over the glottal cycle. Even a simple low-frequency approximation for the subglottal system might be useful in deriving a better approximation to glottal conductance for the control parameter $d(t)$ in Figure 9.

The feedback method, as used, also ignores the component of u_f due to the air displaced by the vocal folds as they vibrate. [3] If we assume that the volume of air ejected from between the vocal folds as they move is proportional to the change in glottal area A_g, and that the area is proportional to the glottat conductance $G_g = 1/R_g$, it follows that the volume velocity due to this component is proportional to $-dG_g/dt$. If we also ignore the glottal inertance present at small glottal openings, the Norton equivalent glottal flow without the displaced air is given by

$$u_f(t) = P_{sg}G_g(t) , \quad (24)$$

and the total glottal flow u'_f is

$$u'_f(t) = u_f(t) - K'_d(dG_g(t) / dt) = u_f(t) - K_d(du_f(t) / dt) \quad (25)$$

where K'_d depends on the magnitude of the displaced air component, and $K_d = K'_d / P_{sg}$.

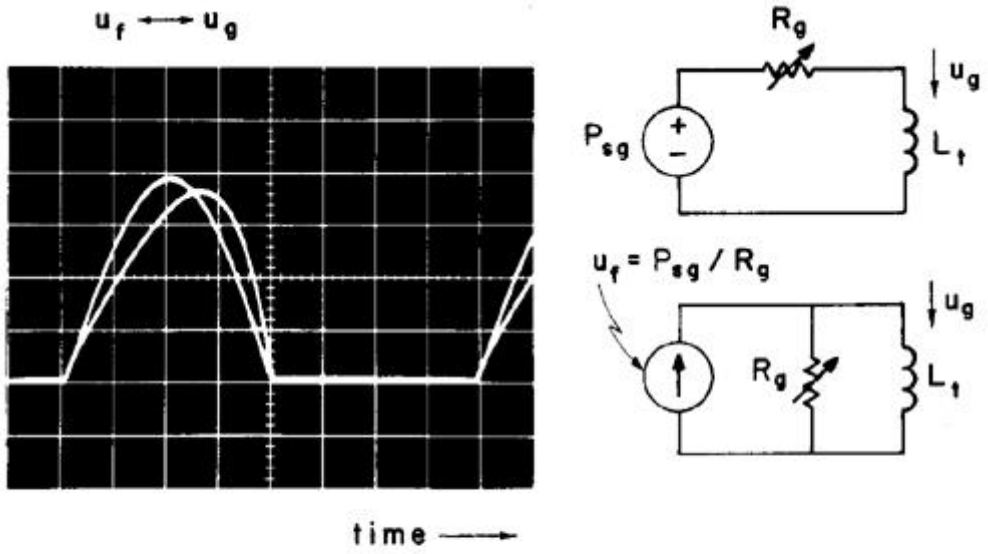

FIG. 11. Change in the glottal volume velocity waveform caused by the interaction of a time varying glottal resistance and an inertive supraglottal vocal tract loading. At the right are the Thevenin and Norton equivalent models for the system used in the simulation.

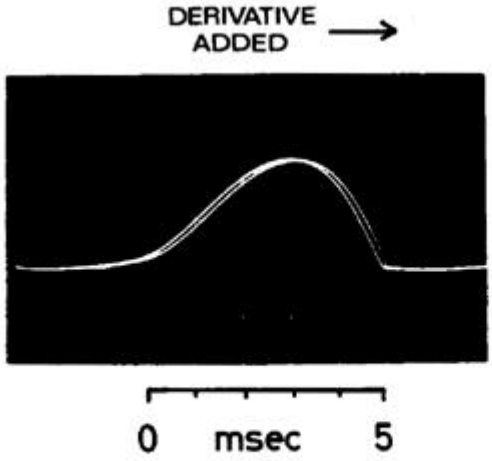

FIG. 12. Effect on a simulated glottal airflow pulse u_t of adding a term $-K(du_t/dt)$ to simulate the air displaced by the motions of the vocal folds.

Figure 12 shows the result of simulating Eq. (25) electronically using a typical pulse waveshape, with $K_d = 1.8 \times 10^{-4}$. These pulse waveshapes have been used to operate the OVE III speech synthesizer at the Royal Institute of Technology, Stockholm, Sweden, and the value of K_d in the figure was the one that seemed to produce the most natural sounding vowels in informal listening tests. [14] The acceleration of the closing phase caused by subtracting the derivative can be seen to skew the glottal pulse to the right, and

cause a sharper termination of the pulse. The derivative term increases the amplitude of energy above the (+ 3 dB) break frequency w = $1/K_d$. The value of K_d used for the photograph would result in a break frequency of about 900 Hz.

A theoretical estimate of K_d can be obtained from the volume displaced during a glottal closing (or opening) movement and the change in glottal airflow during that same movement. If t_1 and t_2 mark the beginning and end of the glottal closing period, then we can integrate the displaced air term in Eq. (25) to obtain the total volume of air displaced,

$$V_d = K_d(du_f(t) / dt)dt \,|_{t1}^{t2} = K_d(duf / dt)dt \,|_{t1}^{t2} = K_d[u_f(t_2) - u_f(t_1)] \; . \qquad (26)$$

Solving for K_d, we obtain

$$K_d = V_d / (u_f(t_2) - u_f(t_1)) = \text{(displaced volume during a glottal closing)/(change in } u_f \text{ during glotal closing)} \qquad (27)$$

Assuming glottal walls 1.8 cm long by 0.3 cm high, for an area of 0.54 cm^2, and a maximum separation of the vocal folds during the glottal cycle of 0.1 cm when averaged over the length of the glottis, the volume of air displaced in a glottal closing would be 0.054 ml (Ref. 3). Assuming further that the displaced air flows equally into the pharynx and into the trachea, and that the total change in u_f as the glottis closes is 400 ml/ sec, as in the waveform in Figure 2, Eq. (27) yields an estimate for K_d of 0.68×10^{-4} sec. Though this figure is a factor of 2.6 from the crudely obtained “best-sounding” value, it does offer some support for the experimental results.

The transfer function associated with Eq. (25), relating u'_f to u_f, is

$$H_d(s) = (1 - K_d s) \; , \qquad (28)$$

with an inverse

$$H_d^{-1}(s) = 1 / (1 - K_d s) \; . \qquad (29)$$

This inverse transform, unfortunately, is unstable. However, it may be possible to find an acceptable stable approximation if K_d is not too large. In nonlinear inverse filtering, an inverse filter for displaced glottal air would be inserted within the feedback loop for the control parameter $d(t)$ so as to make it a better approximation to glottal conductance.

REFERENCES

1. J. van den Berg, J. T. Zantema, and P. Doornenbal, Jr. , “On the Air Resistance and the Bernoulli Effect of the Human Larynx”, J. Acoust. Soc. Am. 29, 626-631 (1957).

2. J. L. Flanagan, “Some Properties of the Glottal Sound Source”, J. Speech Hear. Res. 1, 99-116 (1958).

3. M. Rothenberg, “A New Inverse-Filtering Technique for Deriving the Glottal Airflow Waveform During Voicing”, J. Acoust. Soc. Am. 53, 1632-1645 (1973).

4. J. L. Flanagan and L. L. Landgraf, “Self Oscillating Source for Vocal Tract Synthesizers”, IEEE Trans. Audio Electroacoust. 16, 57-64 (1968).

5. A. D’Angelo, *Linear Time-Varying Systems: Analysis and Synthesis* (Allyn and Bacon, Boston, 1970).

6. O. Fujimura and J. Lindqvist, “Sweep Tone Measurements of the Vocal Tract”, J. Acoust. Soc. Am. 49, 541-558 (1971).

7. G. Fant, *Acoustic Theory of Speech Production* (Mouton, ’S-Gravenhage, 1960).

8. J. L. Flanagan, “Source-System Interaction in the Vocal Tract”, in *Sound Production in Man,* Ann. N. Y. Acad. Sc I. 155, 9-17 (1968).

9. K. Ishizaka and J. L. Flanagan, “Synthesis of Voice Sounds from a Two-Mass Model of the Vocal Cords”, Bell Syst. Tech. J. 51(6), 1233-1268 (1972). In the two-rnass model of the vocal folds, some of the differences in the glottal air flow and the glottal-area waveforms may be due to phase differences between oscillations at the upper and lower edges of the vocal folds.

10. B. Guérin, M. Mryati, and B. Carré, “A Voice Source Taking Account of Coupling with the Supraglottal Cavities”, *Conference Record of the 1976 IEEE International Co-nference on Acoustics, Speech, and Signal Processing,* 12-14 April, 1976, Philadelphia, PA (Canterbury, Rome, NY, 1976), pp. 47 -58.

11. R. Timcke, H. von Leden, and P. Moore, “Laryngeal Vibrations: Measurements of the Glottic Wave. Part I. The Normal Vibratory Cycle”, Arch. Otolaryngol. 68, 1-19 (1958).

12. J. Backus and T. C. Hundley, “Harmonic Generation in the Trumpet”, J. Acoust. Soc. Am. 49, 509-519 (1971).

13. J. Sundberg, “Articulatory Interpretation of the Singing Formant”, J. Acoust. Soc. Am. 55, 838-844 (1974).

14. M. Rothenberg, E. Carlson, B. Granström, and J. Lindqvist-Gauffin, “A Three-Parameter Voice Source for Speech Synthesis”, *Speech Communictation,* edited by G. Fant (Almqvist & Wikrell, Stockholm, 1975), Vol. 2.