A model for the synthesis of natural sounding vowels

Donald R. Allen and William J. Strong
Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602

(Received 10 October 1983; accepted for publication 11 April 1985)

A model has been developed which is designed to preserve some of the naturalness that is usually lost in speech synthesis. A parametrized function is used to produce an approximation to the cross-sectional area through the glottis. A circuit model of the subglottal and glottal system is used with the supraglottal pressure to generate the glottal volume velocity. The tract used to obtain the supraglottal pressure is represented by its input-impedance impulse response, which can be calculated from the area function of the tract. A convolution of the input-impedance impulse response with the volume velocity determines the supraglottal pressure. The two coupled equations for the volume velocity are solved simultaneously. The output of the model is generated by convolving the resulting glottal volume velocity with the transfer-function impulse response of the tract. This technique preserves the interaction between the glottal flow and the vocal tract, which is usually lost. Comparisons are made between "complete tract loading" and "inductive tract loading." Magnitude spectra of the various pressures and the glottal volume velocity are examined in detail. Effects of varying the glottal parameters are examined for one vowel. Listening tests showed that vowels synthesized with the interaction were preferred as more natural sounding than those without the interactions.

PACS numbers: 43.70.Jt, 43.70.Bk, 43.72.Ja

INTRODUCTION

There have been many attempts to synthesize speech. Most of the techniques used have been motivated by an attempt to reduce the information rate of speech transmission using the redundancy which is inherent in speech. Although the intelligibility has been largely maintained, the synthetic speech is often not very natural sounding. Two main factors cause this lack of naturalness. First, the loading of the vocal tract on the excitation functions is usually neglected. Second, the models do not incorporate talker-specific physical features and are therefore inadequate for synthesizing most voices.

A speech-synthesis system is needed that correctly models the physical processes present in speech production so that the features necessary for a perception of naturalness are present. Before discussing a partial solution, we will briefly review current and past techniques that have been used for speech synthesis.

Many successful speech-synthesis systems are based on formant synthesis. In this technique the formant frequencies, amplitudes, and bandwidths are used to determine either an analog or a digital filter. A source function is used to excite the filter. Any number of functions—an impulse, noise (for unvoiced excitation), a triangular pulse, or other pulse shapes—can be used for the source. This technique reduces the bit rate of speech tremendously and is relatively successful at producing intelligible speech.

Predictive methods for speech synthesis are relatively new but have found wide acceptance and use. With linear prediction, a linear polynomial is used to predict the subsequent values of the speech waveform from previous values of the speech sample. The determination of the coefficients used in the predictive calculation assumes that the speech sample was excited by an impulse. It was soon discovered that very few voices were adequately represented by this technique and efforts were begun to "fix" the synthesis. The most effective method found was to use the error between the predicted wave and the original as the excitation. However, this requires almost as much information as exists in the original waveform so that the bit rate is not reduced. Efforts are now under way at many laboratories to try to encode the error signal so that the bit rate can be reduced while maintaining the advantages of using the error signal as the excitation. Another method that has proven quite successful is the use of a multipulse excitation (Atal and Caspers, 1983).

Both the formant and the predictive methods are noninteracting models of speech. That is, there is no interaction between the glottal-flow excitation function and the vocal-tract filter function. However, it has been found (Fant and Liljencrants, 1979) that on many vowels this interaction can be a significant factor in the naturalness of synthetic speech. The following techniques include at least some of this interaction.

The full interaction between the glottal flow and vocal tract is included by solving the differential equations relating to the flow of air in the vocal tract, including equations to specify the mechanical vibration mechanism of the vocal folds (Flanagan and Landgraf, 1968; Ishizaka and Flanagan, 1972; Titze, 1973–74). With this model it is possible, in principle, to model all voices. However, it is computationally inefficient and one needs to know the area functions of the vocal tract and the physical characteristics of the vocal folds in order to proceed. In most applications, few of these values are easily accessible, making this model hard to use.

Guérin *et al.* (1976) developed a model in which an account was taken of the interaction between the supraglottal

 0001-4966/85/070058-12$00.80

cavities and the glottal air flow. The model demonstrated glottal-flow zeros at vocal-tract impedance maxima.

Fant (1979) developed a semi-interactive model that allows a parametrized volume velocity to be modified through interaction with the vocal tract. The characteristics of the vocal tract are changed by the varying glottal impedance in the model. In particular, the vocal-tract formant bandwidths increase by as much as a factor of three and the vocal-tract formant frequencies typically decrease slightly when the vocal folds are open as opposed to when they are closed. In a study of vowels synthesized by this model, Fant found that, when the interaction was maintained between the vocal folds and the vocal tract, the speech sounded more natural than when there was no interaction.

Instead of parametrizing the volume velocity, Rothenberg (1981) parametrized the glottal conductance. Using a triangular shape for the conductance, he was able to get volume-velocity skewing by loading the conductance with the inertance of the subglottal and supraglottal tracts. A compliance was added to provide the first-formant ripple in the volume velocity.

A further refinement of this method has been made by Ananthapadmanabha and Fant (1982), where the area function is parametrized by Fant's (1979) glottal-wave equations. The full impedance of the glottis is used rather than an approximation to it, and the vocal-tract load is represented with tuned circuits representing the formants. In most of their work, Ananthapadmanabha and Fant used just the first formant as the load. The glottal flow is determined by an iteration of the equation relating to the flow.

The model we propose is very similar to that described by Ananthapadmanabha and Fant in that we have chosen to use a parametrized glottal area function. However, we wanted to be able to represent the full vocal-tract input impedance and yet maintain the interaction between the glottal volume flow and the vocal tract. We have chosen to include the interaction between the glottal-flow source and the vocal tract by calculating and using the time-varying pressure at the input of the tract in deriving the glottal flow. This pressure is the convolution of the volume flow with the vocal-tract impulse response. This choice was motivated in part by Schumacher's (1981) very successful application of input-impedance impulse-response methods to the clarinet. An impulse-response representation of the vocal tract may have an advantage over a circuit representation in that frequency-dependent losses can be represented more accurately. Furthermore, this model for the tract may be more flexible than circuit models because very arbitrary input impedances can be specified. The model incorporates circuit representations of the subglottal tract and the glottis. In the work we report here, we have used the input impedance as calculated from the static area function of a specific vowel to specify the tract load. Vowels are synthesized with this load and compared to vowels synthesized with only an inductive tract load. No provision is made in the model at the present time for the synthesis of consonants or continuous speech.

I. DESCRIPTION OF THE MODEL

We chose to use a parametrized glottal area function in our model rather than modeling the full mechanical motion of the vocal folds. This choice is motivated in part because the vocal folds are relatively massive and their motion, and the resulting glottal area function, should be relatively insensitive to vocal-tract loading (Titze, 1980). Vertical phasing in the vocal fold vibration may also contribute to making the "projected glottal area" relatively insensitive to tract loading. Glottal flow will be sensitive to the glottal-area parametrization and vocal-tract loading via the pressure difference across the glottis. Also, by parametrizing the glottal area we simplify the model, but with some expectation of obtaining variations in the glottal flow seen via inverse filtering.

Of various parametrizations, we have chosen one proposed by Titze (1982, 1984) because it appears to be the most flexible, thus permitting the representation of a wider range of glottal waveforms. In addition, there are only two discontinuities, at the glottal opening and closing, whereas most other parametrizations also have a discontinuity at the glottal-area maximum.

There are five parameters in the Titze area function that are used: (1) maximum area A; (2) period T; (3) open quotient γ, which is the ratio of the open time to total period; (4) speed quotient δ, which controls the symmetry of the waveform; and (5) a slope parameter β, that governs the opening and closing slope. The equation governing the glottal area is

$$A_g(\theta) = A\left[(\theta/\theta_m)^{-\theta_m \cot \theta_m}(\sin\theta/\sin\theta_m)\right]^{\beta}, \quad \theta \leqslant \Pi,$$
$$= 0, \quad \theta \geqslant \Pi, \tag{1}$$

where $\theta = \Pi t/\gamma T$, and $\theta_m = \Pi\delta/(1+\delta)$. Some representative examples of the output possible for this model of the area are given in Titze (1982).

The equivalent circuit for the vocal model is shown in Fig. 1 (Flanagan, 1972). The determination of the glottal resistance, R_g, is made by assuming that the glottal area is in the shape of a long ellipse. The area of the ellipse is approximated by 20 rectangular areas, the total resistance being the resistance of these rectangular areas in parallel. The resistance of each rectangular area (Van den Berg *et al.*, 1957) is determined by Eq. (2):

$$R_g = 12\mu d/lw^3 + 0.875[\rho|U_g(t)|/2(lw)^2], \tag{2}$$

where

$\mu = 0.000186$ dyn-s/cm^2≡viscosity of air,

$\rho = 0.00114$ g/cm^3≡density of air,

$U_g(t)$≡glottal volume velocity,

$l = 0.9$ mm≡length of rectangular area,

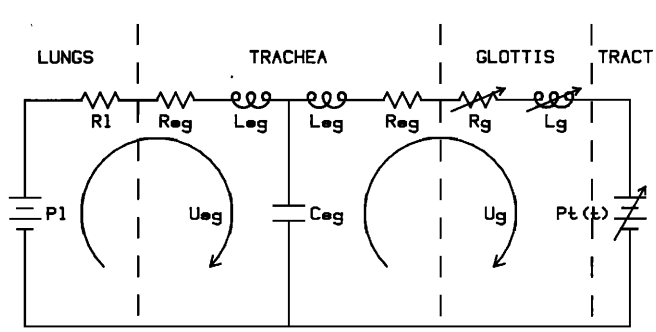

FIG. 1. Electrical circuit analog of the vocal-tract model.

$w = \dfrac{A_g(t)[1-(x/10l)^2]^{1/2}}{10\Pi l} \equiv$ width of rectangular area,

$d = 3$ mm $\equiv$ depth of glottis,

$x \equiv$ position of the rectangular area along the ellipse.

Although the primary inductive loading of the glottal flow comes from the subglottal and supraglottal tracts, the glottal inductance

$$L_g = \rho d / A_g(t) \tag{3}$$

is also important. The differential equations for the glottal flow involve the derivative of $L_g(t)\,U_g(t)$ and can be obtained under the assumption of one-dimensional flow (see Flanagan, 1972).

The rest of the subglottal and glottal system is represented by its equivalent electrical analog. The trachea is represented by a single "T" circuit with a resonance of 500 Hz. Since the damping from the lung coupling is substantial, we assume only one such circuit is needed to allow the subglottal pressure to vary in a realistic manner. Ananthapadmanabha and Fant (1982) showed that only the first resonance of the subglottal system is significant for voice source and effects which supports the use of only one resonance circuit for the trachea. The lungs are represented by a constant pressure source $P_l = 10\,000$ dyn/cm^2 with resistance $R_l = 18\,\Omega$. (See Sec. V, as one of the reviewers noted that this value of lung resistance is too large.)

The volume velocities U_{sg} and U_g, in the circuit in Fig. 1, are represented by the following differential equations:

$$P_l - U_{sg}(t)(R_l + R_{sg}) - L_{sg}\frac{dU_{sg}(t)}{dt} - \frac{1}{C_{sg}}\int [U_{sg}(t) - U_g(t)]\,dt = 0, \tag{4}$$

$$\frac{1}{C_{sg}}\int [U_{sg}(t) - U_g(t)]\,dt - L_{sg}\frac{dU_g(t)}{dt} - U_g(t)[R_{sg} + R_g(t)] - \frac{d\,[L_g(t)U_g(t)]}{dt} - P_t(t) = 0, \tag{5}$$

where

$R_{sg} = S_{sg} l_{sg} (\Pi \rho \mu f_{sg})^{1/2} / 2A_g^2$,

$L_{sg} = \rho l_{sg} / 2A_{sg}$,

$C_{sg} = A_{sg} l_{sg} / \rho c^2$,

l_{sg} = trachea length,

S_{sg} = trachea circumference,

A_{sg} = trachea cross-sectional area,

$f_{sg} = 50$ Hz,

$c = 35\,000$ cm/s.

To solve these equations on a computer, the integrals are represented as sums and the derivatives as differences so that the discrete time values for the two desired volume velocities $U_{sg}(i)$ and $U_g(i)$ become

$$U_{sg}(i) = \left(P_l + U_{sg}(i-1)\frac{L_{sg}}{\Delta T} - \frac{1}{C_{sg}}\sum_{j=0}^{i-1}[U_{sg}(j) - U_g(j)]\Delta T\right) \times \left(R_l + R_{sg} + \frac{L_{sg}}{\Delta T}\right)^{-1}, \tag{6}$$

$$U_g(i) = \left(\frac{1}{C_{sg}}\sum_{j=0}^{i-1}[U_{sg}(j) - U_g(j)]\Delta T + U_g(i-1)\frac{L_{sg} + L_g(i)}{\Delta T} - P_t(i)\right)\left(R_{sg} + R_g(i) + \frac{L_{sg} + L_g(i)}{\Delta T} + \frac{L_g(i) - L_g(i-1)}{\Delta T}\right)^{-1}. \tag{7}$$

The period of the calculation ΔT was set to 0.05 ms.

The supraglottal pressure $P_t(t)$ can be written as the convolution of the vocal-tract, input-impedance impulse response $Z_{in}(t)$, and the glottal volume velocity $U_g(t)$:

$$P_t(t) = U_g(t) * Z_{in}(t). \tag{8}$$

The discrete convolution can be written as a sum of two terms:

$$P_t(i) = \sum_{j=0}^{\infty} U_g(i-j) Z_{in}(j) = U_g(i) Z_{in}(0) + \sum_{j=1}^{\infty} U_g(i-j) Z_{in}(j). \tag{9}$$

The first term involves the current value of the glottal volume velocity and the first value of the input-impedance impulse response; the second term involves only past values of the glottal volume velocity and the other terms of the impulse response of the tract. In this way Eq. (7) may be rewritten to include the tract pressure using the two terms of the convolution sum:

$$U_g(i) = \left(\frac{1}{C_{sg}}\sum [U_{sg}(j) - U_g(j)]\Delta T + U_g(i-1)\frac{L_{sg} + L_g(i)}{\Delta T} - \sum U_g(i-j) Z_{in}(j)\right) \times \left(R_{sg} + R_g(i) + \frac{L_{sg} + L_g(i)}{\Delta T} + \frac{L_g(i) - L_g(i-1)}{\Delta T} + Z_{in}(0)\right)^{-1}. \tag{10}$$

With the inclusion of the convolution terms, the interaction of the tract with the glottal and subglottal systems is included. However, we are free to specify any kind of load we wish by specifying the load's impulse response.

One way to obtain the load is to calculate the input-impedance impulse response of the tract for a given vocal-tract configuration. This is done by first calculating the input impedance of the vocal tract from an area function. The input impedance is calculated via a concatenation of lossy cylindrical sections in a manner similar to that described by Plitnik and Strong (1979). However, cylindrical sections with compliant walls (Ishizaka *et al.*, 1975) were used in the calculations. In addition, the transfer function between the glottal and mouth ends of the tract is determined for later use

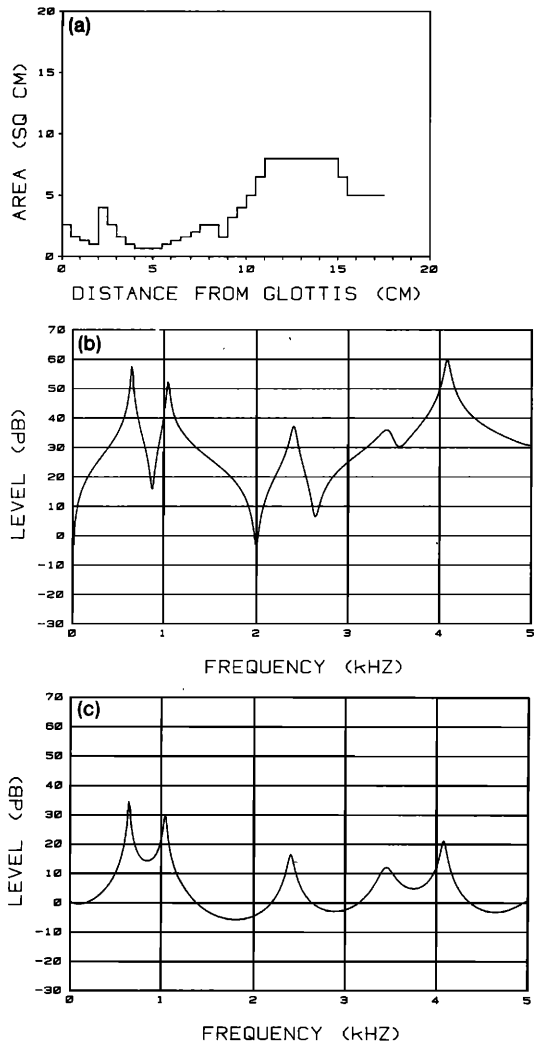

FIG. 2. Russian /a/: (a) area function, (b) input-impedance magnitude, and (c) transfer-function magnitude.

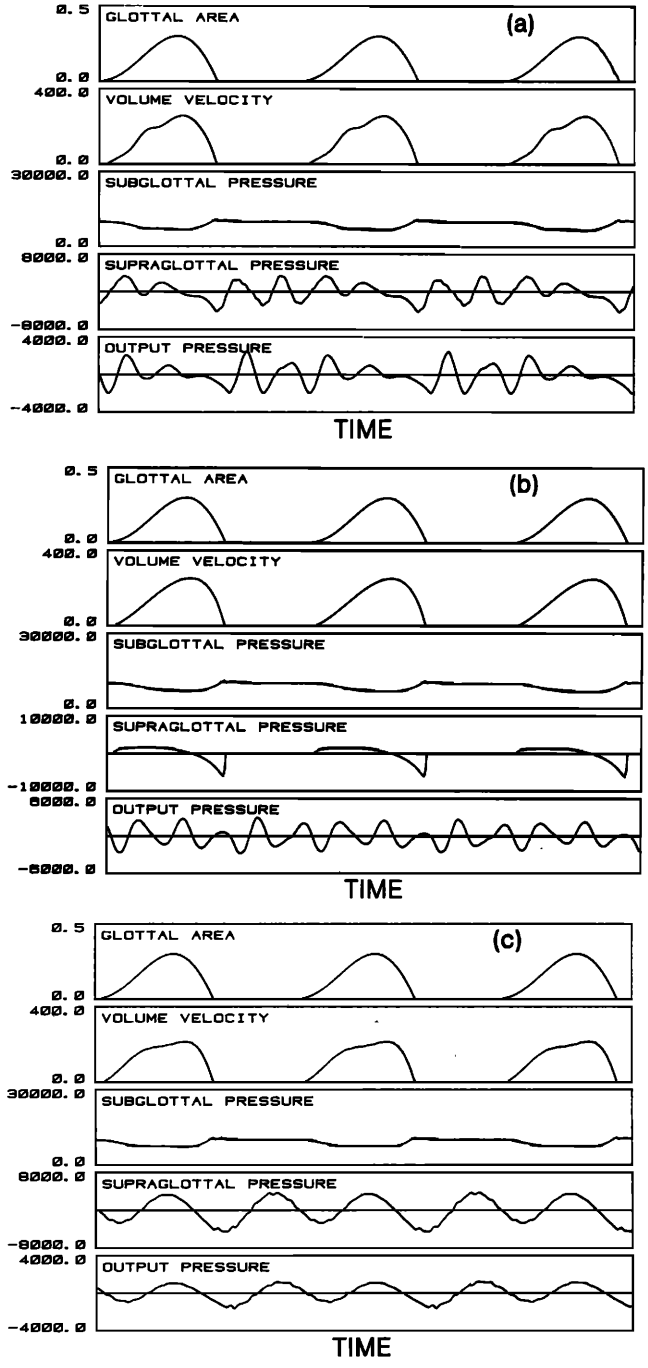

FIG. 3. Synthesized vowel waveforms: (a) tract loaded /a/, (b) inductive loaded /a/, and (c) tract loaded /i/.

in calculating the output. Since we are dealing with a causal system, the real part of the Fourier transform of the input impedance is the impulse response and the imaginary part is zero. Impulse responses calculated from input impedances and transfer functions, had real parts several orders of magnitude larger than their imaginary parts.

The output volume velocity of the model is calculated by taking the convolution of the determined glottal volume velocity with the transfer-function impulse response of the tract. The output pressure is calculated from the output volume velocity through the radiation impedance by Eq. (11):

$$P_r(t) = L_r\left(\frac{dU_0(t)}{dt} - \frac{dP_r(t)}{R_r\,dt}\right), \tag{11}$$

where

$$R_r = 128\rho c/9\Pi^2 A_m,$$

$$L_r = 8\rho/3(\Pi^3 A_m)^{1/2}, \quad A_m = \text{mouth area}.$$

The values of glottal area, subglottal pressure, glottal volume velocity, supraglottal pressure, and output pressure are then stored for further processing and display.

II. GENERATION OF WAVEFORMS AND SPECTRA

Input impedances, transfer functions, and their associated impulse responses were calculated for five Russian vowels (Fant, 1960) and for three area functions of the same American vowel as determined by Atal *et al.* (1978). Only illustrative examples are included here. Figure 2 shows the tract area, the tract input impedance, and the tract transfer function for the Russian /a/. Note the zeros in the input impedance that do not appear in the transfer function.

We chose to explore vowels /a/ and /i/ since they have vocal-tract shapes which have major cavities at opposite ends of the vocal tract; the /a/ near the mouth and the /i/

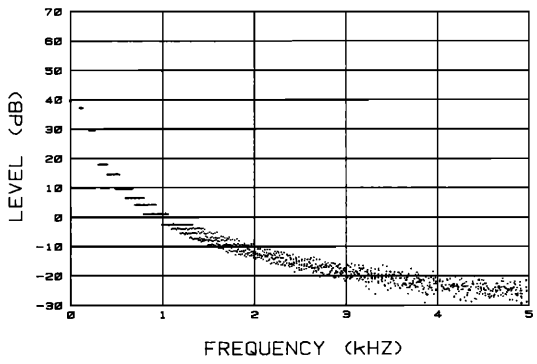

FIG. 4. Glottal area spectra.

near the glottis. In addition, the /a/ has a very high first-formant frequency, while the /i/ has a low first-formant frequency. This should cause a significant difference in the amount of loading that the tract gives to the glottal flow.

In addition to the /a/ and /i/ area functions, we examined two quite different area functions for /u/ given by Atal. The two /u/ area functions have the same transfer function (not shown) out to the third formant, but we wanted to see how their input-impedance functions differed and what effects this might have on the pressure spectra and waveforms. Our results for the transfer functions agree very well with theirs.

The model was used to generate synthetic vowels using the impulse responses described above. The time waveforms for the parametrized glottal area function, glottal volume velocity, subglottal pressure, supraglottal pressure, and output pressure are shown in Fig. 3(a) for the vowel /a/. The volume velocity shows the first-formant ripple that has been seen in inverse-filtering studies. Also, the increased damping that occurs when the glottis is open and which has been discussed by Fant (1979) and by Fant and Ananthapadmanabha (1982) can be seen in the time waveforms of both the supraglottal pressure and the output pressure.

Magnitude spectra were calculated for the area function as shown in Fig. 4. Magnitude spectra were also calculated for subglottal pressure, glottal volume velocity, supraglottal pressure, and output pressure as shown in Fig. 5. The spectra were calculated by running the model at 25 different fundamental frequencies between 100 and 133 Hz. All the glottal parameters except the period were held constant over this range. The spectra were calculated pitch synchronously for each case using the closing instants of the glottis as the period boundaries. The spectra shown are a composite of the spectra for all of the 25 cases. The nominal glottal-area parameter values chosen for all studies were open quotient = 0.6, speed quotient = 2.0, and slope factor = 1.0. For this particular set of parameters, the area-function spectra (Fig. 4) are fairly monotonic. By using such an area function without "major" zeros in the spectra, we can examine the vocal-tract loading without having to interpret the glottal zero effects.

The multivaluedness of the spectra in Fig. 5, especially at low frequencies, arises because any given harmonic has a fixed energy level since the Fourier series expansion differs

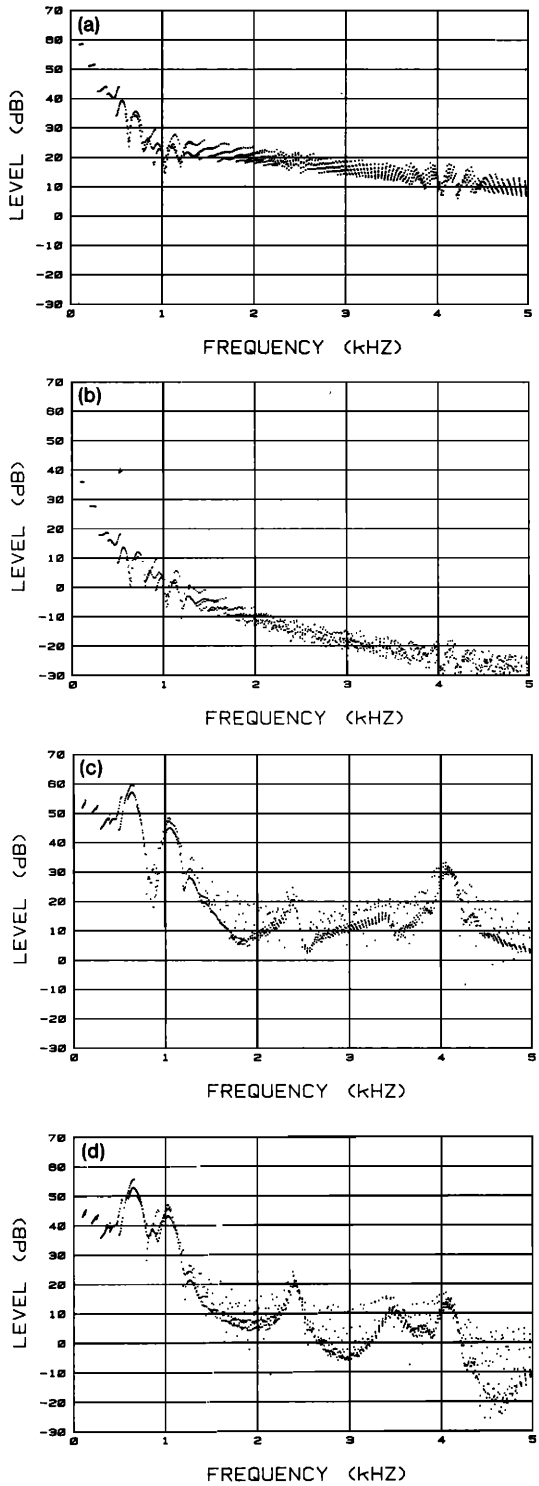

FIG. 5. Spectra for the synthesized /a/: (a) subglottal pressure, (b) volume velocity, (c) supraglottal pressure, and (d) output pressure.

from the Fourier intergral by 1 over the period factor not included here. (Note the horizontal lines on the area-function curve which are generated by the same harmonic at various fundamental frequencies.) Since a higher order harmonic will have less energy at a given frequency than a lower one, another line is created as the pitch is lowered. To offset this effect, the area-function spectra were used as a "normalizing" reference; the area-function spectra were subtracted from the other spectra on a point by point basis. The result of

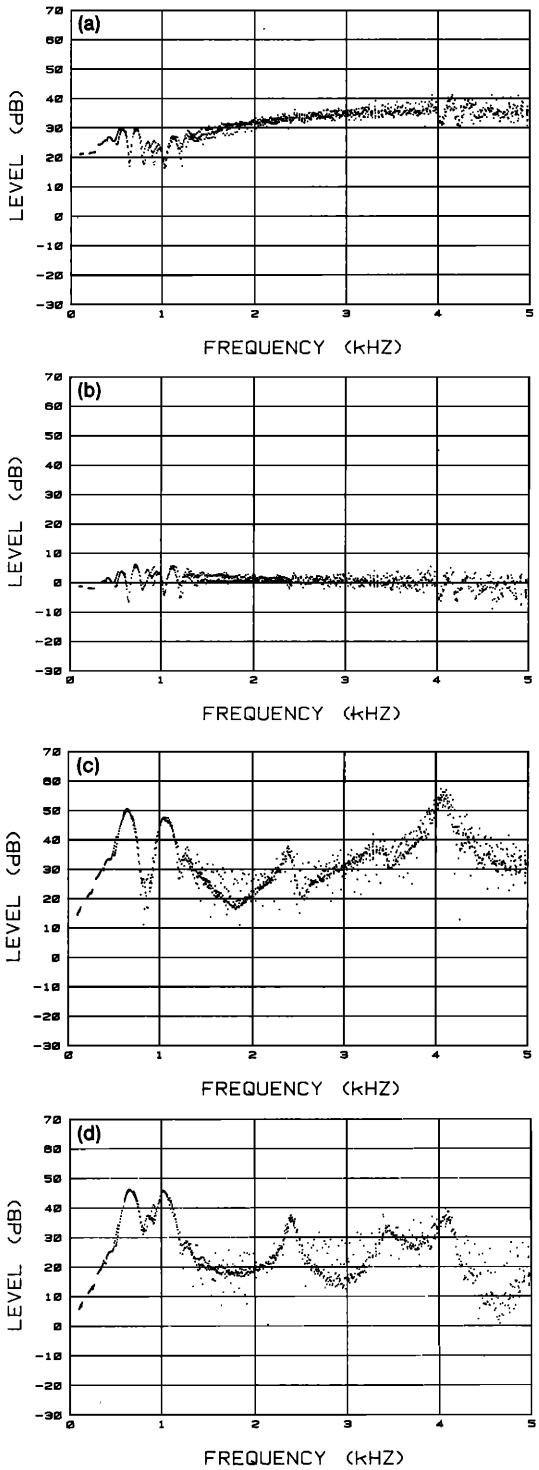

FIG. 6. Normalized spectra for the synthesized /a/: (a) supraglottal-pressure, (b) volume velocity, (c) supraglottal pressure, and (d) output pressure.

area-function "normalization" is seen in Fig. 6 for the vowel /a/. The multivaluedness of the curves is mostly gone, especially at the lower frequencies. Because of the improved appearance, all remaining spectra have been smoothed by the area-function spectra. A disadvantage to the normalized spectra is that the low-frequency energy below the first formant has been "normalized" out. The normalized spectra must be "unnormalized" by adding a smoothed version of

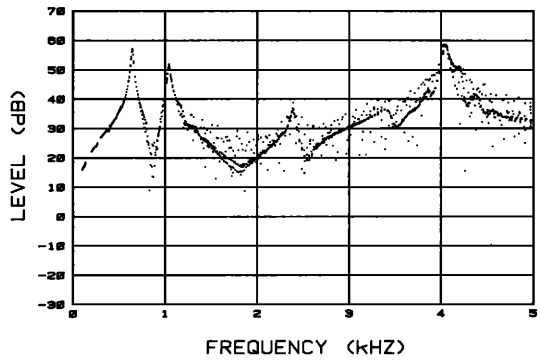

FIG. 7. Supraglottal pressure spectra normalized relative to glottal flow spectra for synthesized /a/.

the area spectra to obtain representative vowel spectra.

The supraglottal pressure normalized relative to the glottal volume velocity is shown in Fig. 7. This should be the input impedance of the vocal tract and as such it provides a means for checking the accuracy of the model computations and the FFT analysis used in the study. It compares quite favorably with Fig. 2(b), especially at lower frequencies. The trend of Fig. 2(b) is clearly seen in Fig. 7. The points scattered around the trend usually occur at frequencies above about 1500 Hz and are probably attributable to normalizing with low level values of the area spectra.

Figure 8 shows spectra for the vowel /a/ for the case when the vocal-tract loading is represented as a pure inductance. The corresponding volume-velcoity waveform [Fig. 3(b)] exhibits the asymmetry characteristic of inductive loading, but does not exhibit the ripple that existed in the complete vocal tract loading case. A comparison of Fig. 8(b) with Fig. 6(b) shows that the variation in the volume-velocity spectra near the poles and zeros of the tract input impedance arises from the tract load. In fact, it is apparent that the first formant ripple arises because the volume velocity has less energy at the formant frequency with a complete vocal tract load than with an inductive load. {Note that the first major zero in the volume-velocity spectra [Fig. 6(b)] occurs at the frequency of the first formant [Fig. 6(d)].} Since the alternating subglottal pressure at the formant frequency has a lower amplitude than does the supraglottal pressure, the tendency is for the flow at the formant frequency to be toward the lungs rather than toward the mouth when the glottis is open. Thus, as the glottis comes open, the formants become more heavily damped because the energy is dissipated toward the lungs.

The subglottal pressure shows much of the same structure in the frequency domain [Fig. 6(a)] as the volume velocity [Fig. 6(b)]. This model does not produce as much structure in the time domain of the subglottal pressure [Fig. 4(a)] as has been reported in the literature from actual measurements (see, for example, Fant, 1982). This indicates that the lung resistance may be too high in our model or that more than one "T" section is needed to represent the subglottal system.

There is a considerable difference in the supraglottal pressure between the tract loaded [Fig. 6(c)] and the induc-

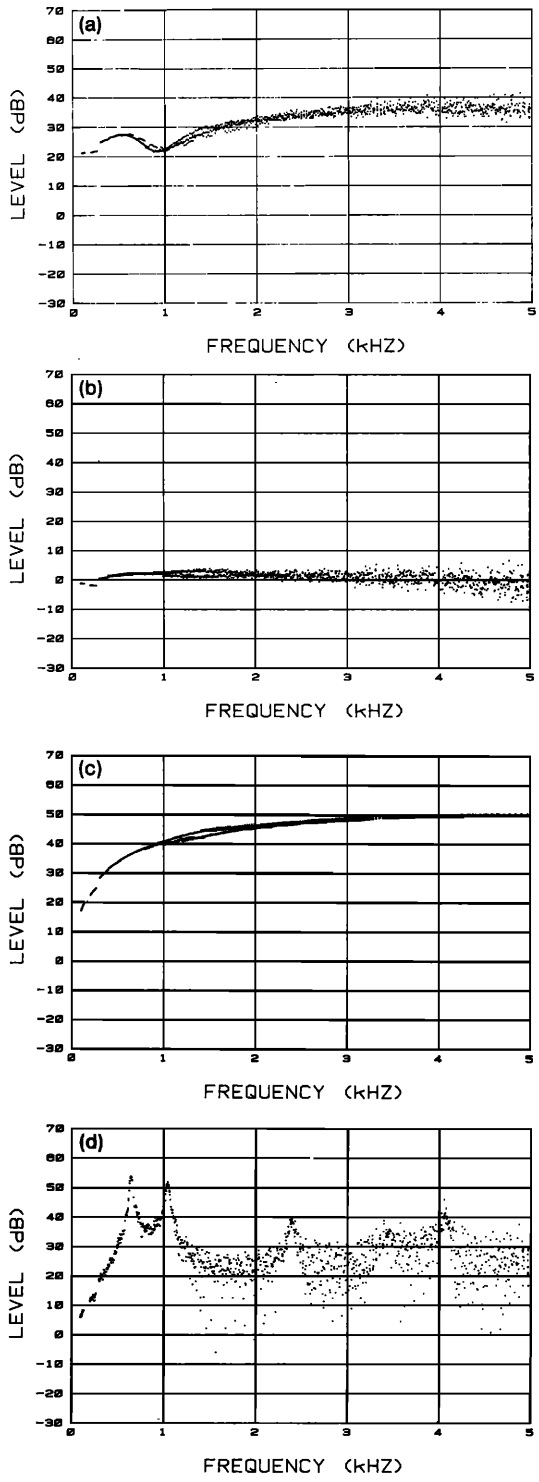

FIG. 8. Normalized spectra for the /a/ synthesized with an inductive load: (a) subglottal pressure; (b) volume velocity, (c) supraglottal pressure, and (d) output-pressure spectra.

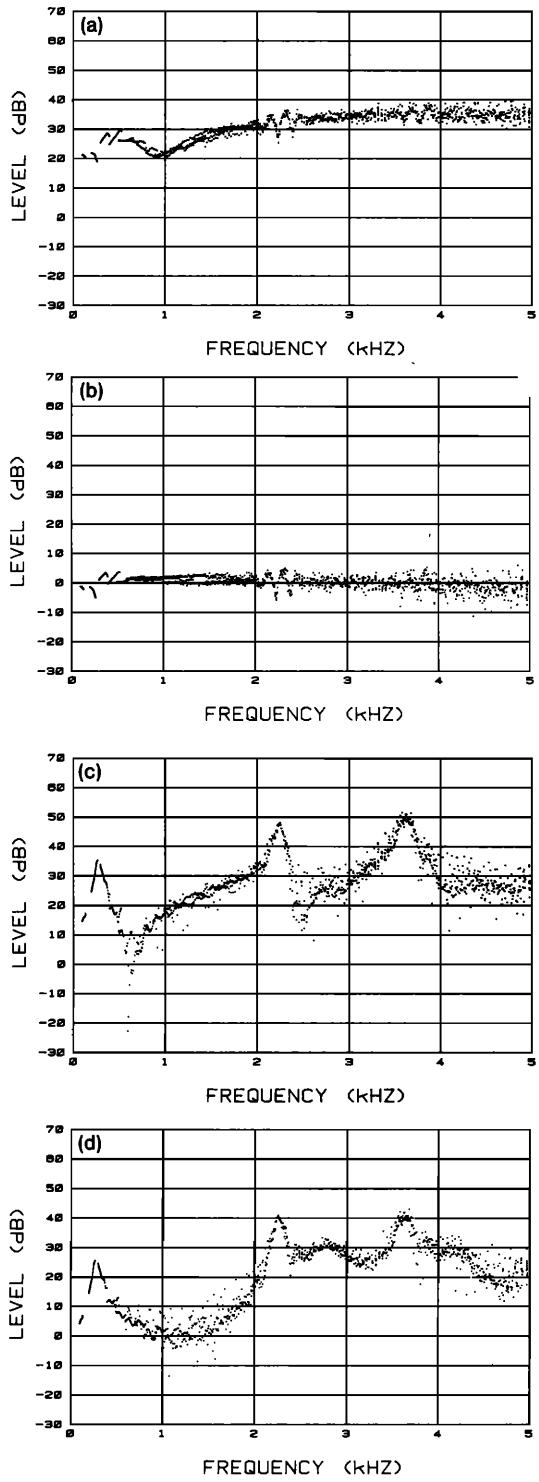

FIG. 9. Normalized spectra for the synthesized /i/: (a) subglottal pressure, (b) volume velocity, (c) supraglottal pressure, and (d) output pressure.

tive loaded [Fig. 8(c)] cases since the supraglottal pressure is highly dependent on the load. In the tract loaded case the supraglottal-pressure spectra are a near copy of the input impedance, which is expected since the volume velocity has a "relatively flat" spectral content. There is, however, an important difference between the input impedance [Fig. 2(b)] and the supraglottal-pressure spectra [Fig. 6(c)]. The peaks in the supraglottal-pressure spectra are decreased in amplitude and broadened relative to those of the input impedance. This is due to the increase in the volume velocity at frequencies both above and below that of the formant and the decrease in the volume velocity at the formant frequency.

A major difference is seen to exist in the radiated spectra between the tract loaded [Fig. 6(d)] and inductive loaded [Fig. 8(d)] cases. The peaks in the output-pressure spectra are flattened and broadened by the introduction of the interac-

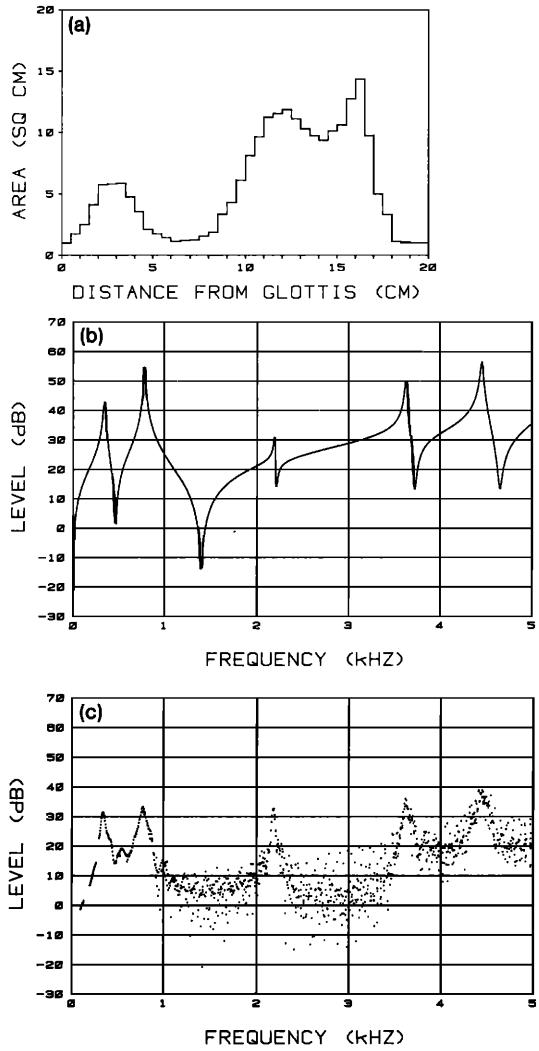

FIG. 10. Atal /u11/: (a) area function, (b) input-impedance magnitude, and (c) output-pressure spectra.

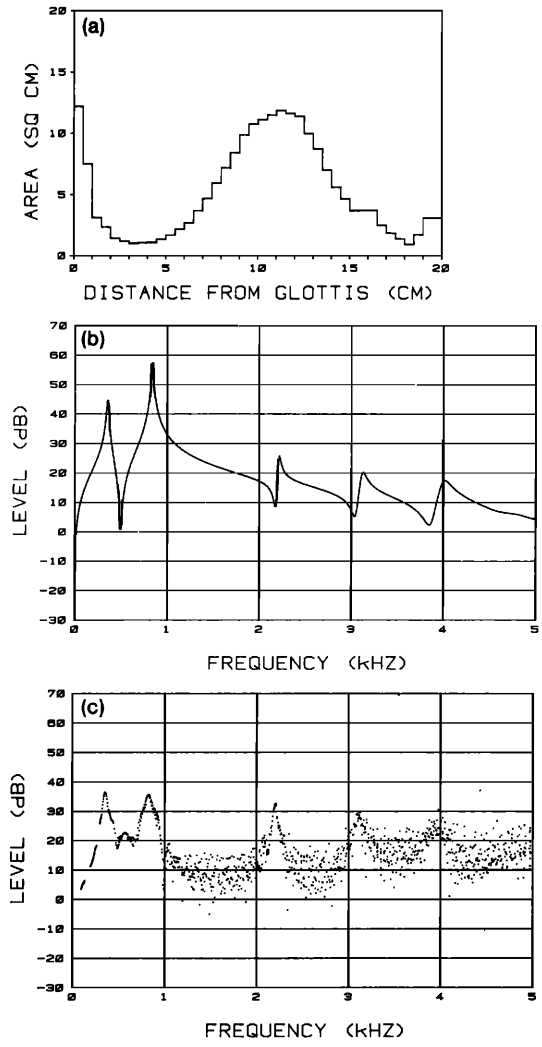

FIG. 11. Atal /u31/: (a) area function, (b) input-impedance magnitude, and (c) output-pressure spectra.

tion between the glottal flow and the vocal tract. The output spectra broadening [Fig. 6(d)] is similar to the supraglottal spectra broadening [Fig. 6(c)] as might be expected since they differ in terms of a zero function only. This has been described by Fant (1979) as arising due to the increased damping which exists when the glottis is open and which is apparent in the output time waveforms. This increased damping can be seen in the time waveforms of Fig. 3(a) but is lacking in the inductive loading case [Fig. 3(b)]. In addition, we see a small peak in the spectra between the first two poles at the location of the input-impedance zero.

Waveforms [Fig. 3(c)] and spectra (Fig. 9) for the vowel /i/ are shown for comparison. The volume-velocity spectra structure near formant frequencies as discussed for /a/ are also apparent here.

Figures 10 and 11 show tract areas, input impedances, and output spectra for the two different /u/-area functions as given by Atal *et al.* (1978). There is a striking difference in the slope of the input impedances [Figs. 10(b) and 11(b)] in the vicinity of the third formant which is reflected in the output spectra. The output spectra are very similar up to the third formant which is the same range over which the input impedances are very similar.

III. VARIATION OF GLOTTAL PARAMETERS

Figures 12–14 illustrate the effects of changing the various glottal parameters. The same /a/ vocal tract discussed above was used, but with a somewhat higher subglottal resonance of 525 Hz. The values of the glottal parameters used for the various configurations studied are shown in Table I. The "standard" values used to generate Fig. 3(a) were used as the middle value in each comparison.

First, the open quotient was varied over three values: 0.2 [Fig. 12(a)], 0.6 [Fig. 3(a)], and 0.9 [Fig. 12(c)]. It is apparent that the damping of the formants increases as the glottis is open longer. This can be seen by comparing the time waveforms of the supraglottal and output pressures as well as the bandwidths of the formants in the supraglottal spectra [Figs. 12(b), 6(c), and 12(d)]. There is always a minimum in the volume-velocity spectra (not shown) at the frequency of the formant. The periodicity of the structure in the glottal volume-velocity spectra (not shown) and the subglottal pressure spectra (not shown) is determined by the open quotient.

Next, the symmetry quotient was varied over three values: 1.0 [Fig. 13(a)], 2.0 [Fig. 3(a)], and 5.0 [Fig. 13(c)]. The more symmetric area function [Fig. 13(a)] introduces major

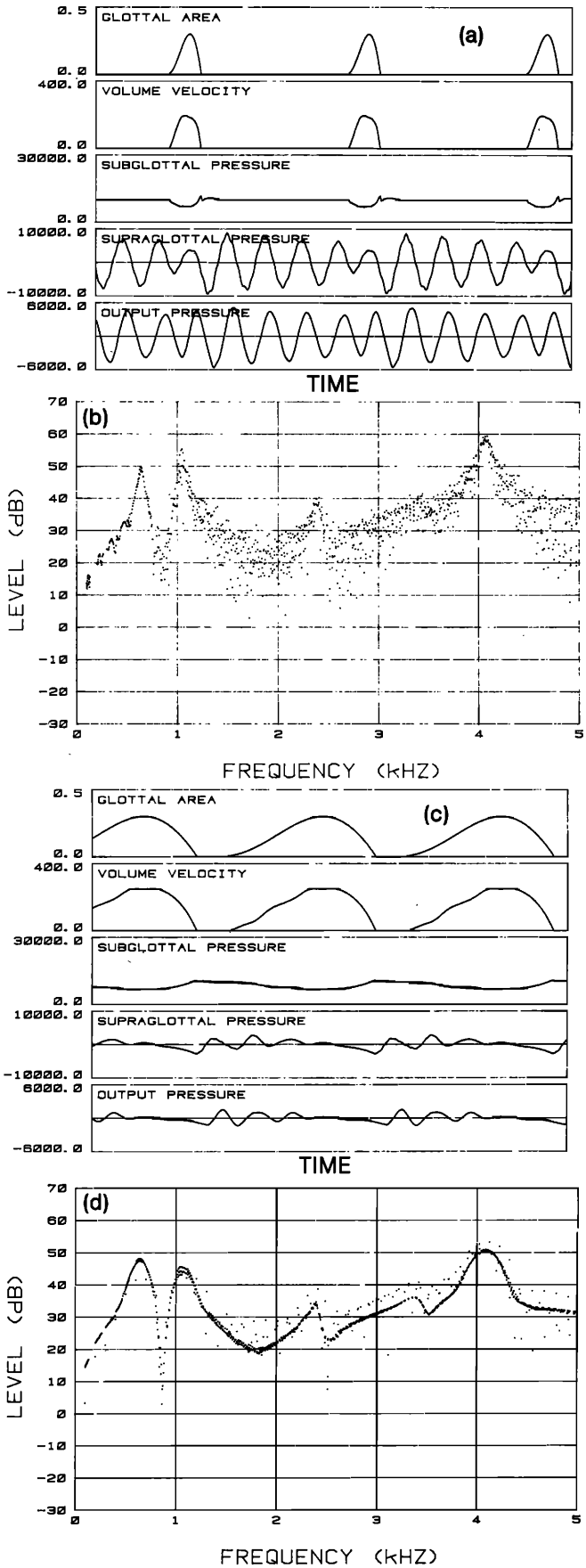

FIG. 12. Modification of glottal parameter open quotient for synthesized /a/: (a) time waveforms for O1 glottal parameters; (b) supraglottal-pressure spectra for O2 glottal parameters.

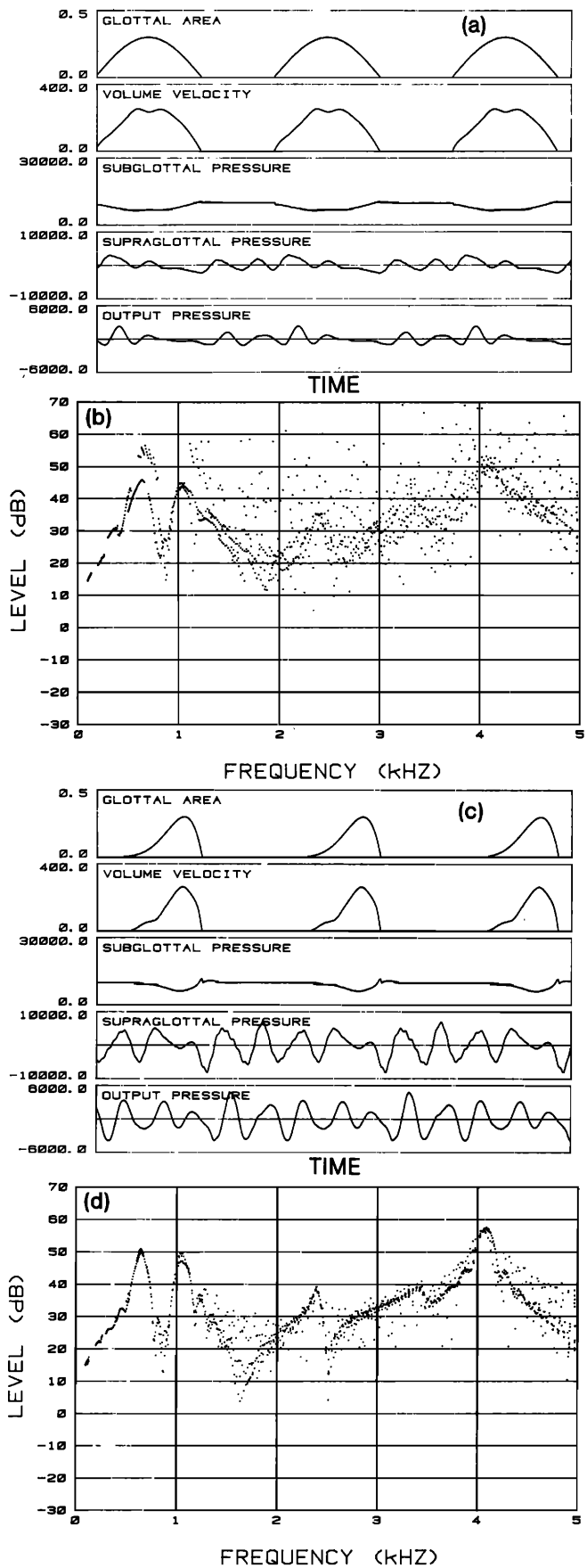

FIG. 13. Modifications to glottal parameter symmetry quotient for synthesized /a/: (a) time waveforms for S1 glottal parameters, (b) supraglottal-pressure spectra for S1 glottal parameters, (c) time waveforms for S2 glottal parameters, and (d) supraglottal-pressure spectra for S2 glottal parameters.

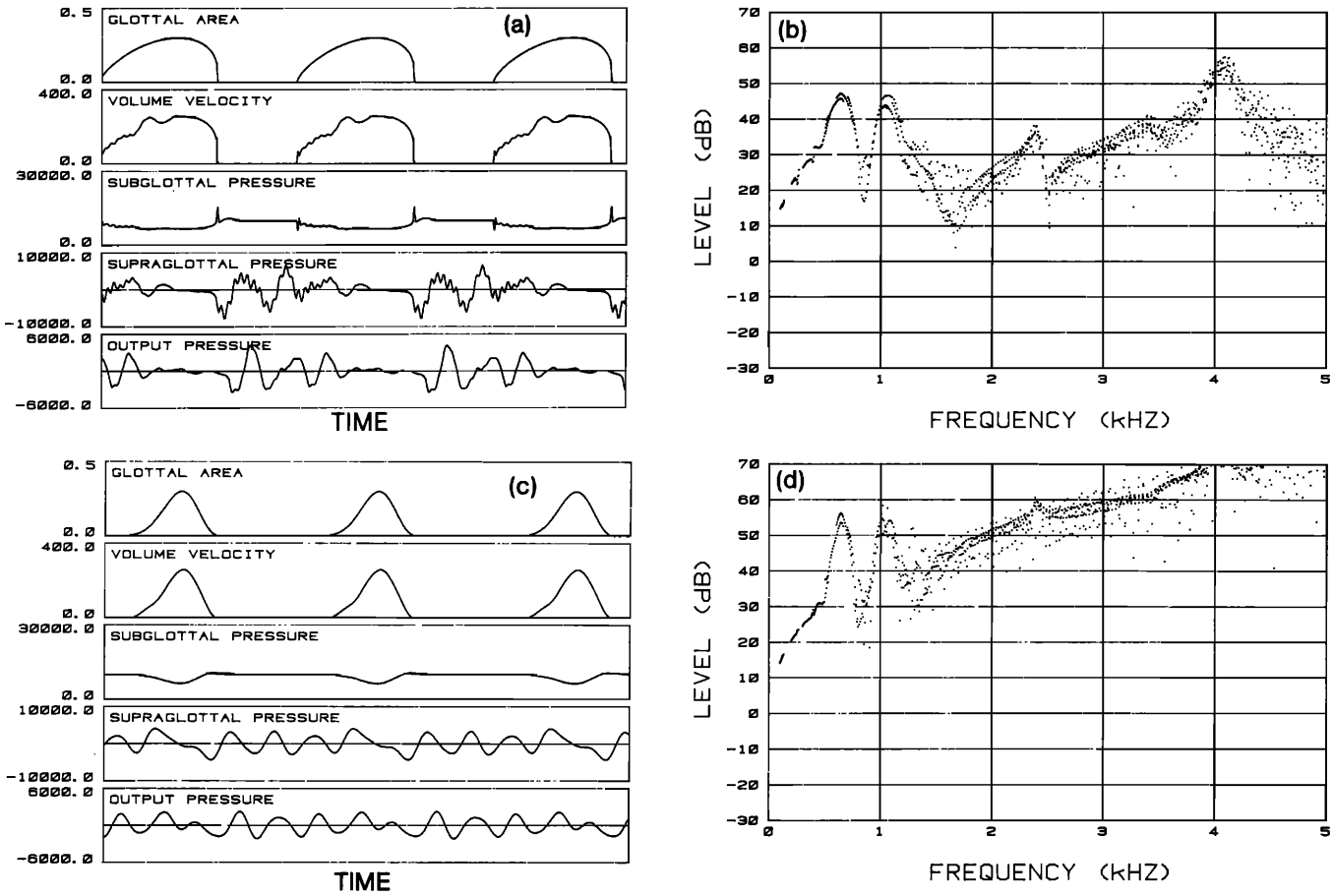

FIG. 14. Modifications to glottal parameter β for synthesized /a/: (a) time waveforms for B1 glottal parameters, (b) supraglottal-pressure spectra for B1 glottal parameters, (c) time waveforms for B2 glottal parameters, and (d) supraglottal-pressure spectra for B2 glottal parameters. High-frequency boost in the spectra arises from "calculation noise."

zeros in the area-function spectra (not shown). Although the overall formant structure of the output remains about the same, the zeros cause considerable "noise" in the supraglottal-pressure spectra when normalized relative to the area-function spectra. There is more high-frequency energy in the more asymmetric area function [Fig. 13(c)] than in either of the other two as might be expected because of its more pulse-like character. The formants display different bandwidths in each case because of the different glottal impedances that exist for different glottal-area values. [The open quotients in Figs. 13(c) and 14(c) appear to be smaller than the specified value of 0.6. This is due to "interaction" of the symmetry and "abruptness" parameters with the open quotient parameter in the model. The "effective" open quotient is made smaller because of this interaction in each of the two cases.]

Finally, the slope factor β which controls the abruptness of opening and closing was varied through the values: 0.3 [Fig. 14(a)], 1.0 [Fig. 3(a)], and 3.0 [Fig. 14(c)]. The spectral balance of the glottal area function (not shown) is significantly changed from one case to the next. The anomalous boost of high frequencies in the supraglottal pressure spectra of Fig. 14(d) can be accounted for by reference to the extremely low values of the area spectra (not shown). All spectra are weighted by the area-function spectra and for this case, "calculation noise" predominates at high frequencies. Changing β also changes the damping of the formants by varying the losses through the glottis. The larger integrated area of Fig. 14(a) produces greater damping than in Figs. 3, 6, or 14(d) as can be seen in both the time waveforms and the amplitude spectra.

TABLE I. Glottal parameters used to study the effect of varying the glottal parameters. The open quotient (O1 and O2), symmetry quotient (S1 and S2), and slope parameter (B1 and B2) were varied over three values each with standard (ST) used as the middle value in each case.

Run name	ST	O1	O2	S1	S2	B1	B2
Open quotient	0.6	0.2	0.9	0.6	0.6	0.6	0.6
Symmetry quotient	2.0	2.0	2.0	1.0	5.0	2.0	2.0
Slope factor	1.0	1.0	1.0	1.0	1.0	0.3	3.0

IV. LISTENING TEST

A listening test was performed to see if the interacting synthetic vowels sounded more natural than the noninteracting synthetic vowels for the five Russian vowels and one of the Atal vowels. All vowels were synthesized with a slight random variation of the period so that the period was not constant over the whole token. A 1-s segment of the steady-state vowels was constructed with a simple cosine time weighting at the beginning and the end. In addition, the levels for all vowels were set to the same peak value. Inductive loaded versus tract loaded versions of the same vowel were paired. The one heard first was chosen at random. After low-

TABLE II. Listener responses from comparison of tract loaded (first column) and inductive loaded (second column) synthetic vowels. The number indicates how many times that particular version was chosen as more natural.

Listeners	/a/	/e/	/i/	/o/	/u/	/u11/
L1	0 2	2 0	2 0	2 0	2 0	2 0
L2	2 0	1 1	2 0	0 2	2 0	2 0
L3	0 2	2 0	2 0	2 0	2 0	2 0
L4	1 1	1 1	1 1	1 1	2 0	2 0
L5	0 2	0 2	1 1	1 1	2 0	2 0
L6	0 2	1 1	2 0	2 0	2 0	2 0
L7	1 1	2 0	2 0	1 1	2 0	2 0
Totals	4 10	9 5	12 2	9 5	14 0	14 0

pass filtering at 5 kHz, the tokens were presented over earphones to listeners who were asked to select which member of a pair sounded more natural. A listener was allowed to hear any given pair as many times as desired before making a decision. During the test the subject responded to all six pairs once and then responded to a different randomization of the same six pairs again.

The results of the listening test appear in Table II. The numbers represent the number of times the listener selected that token during both listening events. The first column for each vowel is the tract case and the second column is the inductive case. For the three vowels /i/, /u/, and Atal *et al.*'s (1978) /u11/, the choice was decisively in favor of the interacting case. Interestingly, only the /a/ was not perceived as more natural sounding in the tract case than in the inductive case. (The result for /a/ might be partly ascribed to an overestimate of vocal tract interaction discussed below.)

Table III shows the results of a preference test composed of the three vowels generated from Atal's three /u/ area functions. In each case a tract load was used. /u31/ and /u51/ were preferred as being more natural than /u11/. The sharp resonances between 4 and 5 kHz that exist in /u11/ and that do not exist in /u31/ and /u51/ may be the reason that it was not preferred since more high-frequency energy was produced. There is also a preference for /u51/ over /u31/.

V. COMMENTS

Several overall observations can be made about the model and its implementation. The most obvious effects of vocal-tract interaction are the increase in formant bandwidths and the appearance of some ripple in the glottal flow spectra. Spectral normalization is used to reduce the multivaluedness of various spectra and to enhance interaction effects. The normalization significantly alters the main effects of the source waveform and the overall spectral balance in the spectra. The normalization procedure also results in "fuzzy" spectra and anomalous frequency boosts, particularly when the chosen glottal area parameters give extreme glottal area functions not representative of speech.

TABLE III. Preference test for naturalness between the three pairs of /u/ area functions each of which give the same transfer function.

	/11/ /31/	/11/ /51/	/31/ /51/
L1	0 2	0 2	0 2
L2	1 1	0 2	1 1
L3	0 2	0 2	0 2
L4	2 0	2 0	2 0
L5	1 1	0 2	0 2
L6	0 2	0 2	0 2
L7	0 2	0 2	2 0
Totals	4 10	2 12	5 9

From our examination of the Russian vowels /a/ and /i/ (and with the foregoing cautions in mind), we see that the input impedance of the tract plays an important role in an interacting model. From the data we have presented and that which has been presented by others, it is clear that the implementation of a glottal-flow, vocal-tract interacting model improves the perceived naturalness of synthetic vowels. Seven of the eight vowels we examined were more natural sounding when an interacting model was used. In addition to the listening test, the waveforms which are produced by the model are more realistic than the waveforms resulting with an inductive load.

There are likely other factors which contribute to the naturalness of speech. For example, we varied the period in a random and arbitrary way. The importance of such things in relation to the glottal-flow, vocal-tract interaction needs to be assessed.

The magnitude spectra of the volume velocity indicate that for /a/, the glottal flow interacts more strongly with the vocal tract than it does for /i/, but the listener responses indicated a preference for the inductive loaded /a/ rather than the tract loaded /a/. The opposite was true for /i/, where a very strong preference was shown for the tract loaded /i/. The reason for this discrepancy is not clear. Comparison of real speech with the equivalent synthetic speech from the model would be a fairer test of the naturalness of the synthetic speech. Before this can be done, a reasonable analysis system must be developed.

The lack of agreement between our model and measurements of the subglottal-pressure waveform would indicate that our description of the subglottal system is inadequate. As noted earlier, one of the reviewers pointed out that our value of lung resistance is too large, resulting in values of glottal flow that are too small. This results in values of glottal resistance that are too small because the glottal resistance is approximately proportional to flow. In this case, the vocal-tract interaction is likely overestimated because it is related to the input impedance of the vocal tract relative to the glottal impedance (which is underestimated). However, since the vocal tract interaction is generally of greater significance than the subglottal interaction, the results presented should be qualitatively correct, even though not quantitatively realistic. Clearly, further understanding of the role of the subglottal system should be obtained by systematically changing the tracheal parameters and by varying the lung pressure and lung resistance. In addition, the effect of using multiple "T" circuits for the trachea should be examined.

There are several other things which still need to be done. The perceptual significance of the volume velocity

produced peaks needs to be assessed. A careful study using a number of different glottal parameters would give a better understanding of their physical significance.

To understand better the interaction occurring at the glottis, input impedances should be constructed that have either a single peak or a single zero, the frequency of which is varied so that a simpler model for the calculation of the input-impedance impulse response can be determined.

Although we have used the area function of the tract to determine the input impedance in the work described, it may be possible to imply the input-impedance function by other means. Furthermore, it may be possible to use an interaction that includes only the first part (in the frequency domain) of the input impedance to arrive at an impulse response. This should be so because more energy exists at low frequencies and also because low first formants contribute more to loading than high first formants.

ACKNOWLEDGMENTS

We are indebted to the reviewers for raising important questions and making helpful suggestions that have aided us in the clarification of the method and some of its results and deficiencies.

Ananthapadmanabha, T. V., and Fant, G. (1982). "Calculation of true glottal flow and its components," STL-QPSR 1, 1–30.

Atal, B. S, and Caspers, B. E. (1983). "Periodic repetition of multipulse excitation," J. Acoust. Soc. Am. Suppl. 1 74, S51.

Atal, B. S, Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am. 63, 1535–1555.

Fant, G. (1982). "Preliminaries to analysis of the human voice source," STL-QPSR 4, 1–47.

Fant, G. (1979). "Glottal source and excitation analysis," STL-QPSR 1, 85–107.

Fant, G., and Ananthapadmanabha, T. V. (1982). "Truncation and superposition," STL-QPSR 2-3, 1–17.

Fant, G., and Liljencrants, J. (1979). "Perception of vowels with truncated interperiod decay envelopes," STL-QPSR 1, 79–84.

Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands), p. 115.

Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, New York).

Flanagan, J. L., and Landgraf, L. L. (1968). "Self-oscillating source for vocal-tract synthesizers," IEEE Trans. Audio Electroacoust. AU-16, 57–64.

Guérin, B., Mrayati, M., and Carré, R. (1976). "A voice source taking account of coupling with the supraglottal cavities," IEEE ASSP Conf. Rec., 47–50.

Ishizaka, K., French, J. C., and Flanagan, J. L. (1975). "Direct determination of vocal tract wall impedance," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 370–373.

Ishizaka, K., and Flanagan, J. L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J. 51, 1233–1268.

Plitnik, G. R., and Strong. W. J. (1979). "Numerical method for calculating input impedances of the oboe," J. Acoust. Soc. Am. 65, 816–825.

Rothenberg, M. (1981). "An interactive model for the synthesis of voice sounds," STL-QPSR 4, 1–17.

Schumacher, R. T. (1981). "AB INITIO calculations of the oscillations of a clarinet," Acustica 48, 71–85.

Titze, I. R. (1984). "Parametrization of the glottal area, glottal flow, and vocal fold contact areas," J. Acoust. Soc. Am. 75, 570–580.

Titze, I. R. (1982). "Synthesis of sung vowels using a time domain approach," in *Transcripts of the Eleventh Symposium: Care of the Professional Voice*, edited by V. L. Lawrence (The Voice Foundation, New York), pp. 90–98.

Titze, I. R. (1980). "Comments on the myoelastic-aerodynamic theory of phonation," J. Speech Hear. Res. 23, 495–510.

Titze, I. R. (1973–74). "The human vocal cords: A mathematical model," Phonetica 28, 129–170 (part 1); 29, 1–21 (part 2).

Van den Berg, Jw., Zantema, J. T., and Doorneball, P. (1957). "On the air resistance and the Bernoulli effect of the human larynx," J. Acoust. Soc. Am. 29, 626–631.