Quarterly Progress and Status Report

Voice source dynamics in connected speech

Gobl, C.

journal: STL-QPSR
volume: 29
number: 1
year: 1988
pages: 123-159

http://www.speech.kth.se/qpsr
II. SPEECH PRODUCTION

A. VOICE SOURCE DYNAMICS IN CONNECTED SPEECH

Christer Gobl*

Abstract
Dynamic variations of the voice source in connected speech were studied by means of inverse filtering and waveform parameterization. The speech materials consisted of a few utterances spoken by three adult males and one 10 year-old male. A four parameter source model (the LF model) was utilized to describe the temporal changes of the glottal pulse shape. The properties of this model are briefly explained. Some problems associated with inverse filtering for deriving the volume velocity through the glottis are also discussed.

The results of the analysis provide a general idea of the range of pulse shape variation in normal speech. Different stress environments show major effects on the glottal excitation. Significant changes in the glottal pulse shape are typically found at the onset and, particularly, at the termination of the voice source, and also at many of the boundaries between vowels and consonants. Strong interdependencies are often found between the different parameters of the model, and substantial information about the pulse shape can be inferred directly from the amplitude of the speech waveform. Resynthesized utterances show that the improved source model is more important for the child's voice than for the voices of the adult males. Typical parameter values for different contexts are tabulated and suggestions for developing a preliminary source rule system are presented.

1. Introduction

Most speech synthesizers are based upon the source-filter theory of speech production, for which the source is assumed to be linearly separable from the vocal tract filter. The conventional source of voiced sounds in a formant synthesizer is a sequence of constant shape pulses with a spectral slope of -6 dB per octave, controlled in terms of fundamental frequency and amplitude only. Much more attention has accordingly been paid to filter than to source functions.

Systems for speech synthesis by rule, such as the one developed at KTH by R. Carlson and B. Granström (see Carlson, Granström, & Hunnicutt, 1981 and references therein), have advanced steadily by improving phonetic and prosodic rules. However, the sound generating system (the synthesizer) is basically the same, which means that the conventional source is still used.

* Swedish Telecom Administration (Televerket), Technology Dept., Section for Research and Development, S-123 86 Farsta, Sweden. Graduate student at KTH.
The demand for improved quality and naturalness of synthetic speech is growing with current expectations for a more extended use of text-to-speech systems in information services and handicap aids. Therefore, studies of the human voice source have gained a renewed interest. The present experimental study supplements the considerable amount of theoretical work which has been pursued by others in this field.

In order to convincingly synthesize women's and children's voices and different types of phonation such as breathiness, tense (pressed) voice, loud voice, shouting, etc., it is important to gain a more profound insight into the properties and dynamics of the human voice source.

There are (at least) two different approaches to the problem of improving the voice source. One is to reject the linear source-filter concept of speech production with its inherent limitations. The alternative is an interactive synthesis. (For basic theory and description of interaction, see, for instance, Ananthapadmanabha & Fant, 1982; Fant & Lin, 1987, and references therein. A brief summary of interaction phenomena can be found in Klatt & Klatt, forthcoming.) However, a fully interactive synthesis would be very complex, especially if a complete articulatory coding were adopted. We still lack strategies and rules for controlling articulators rather than formants. A completely new system of rules would have to be developed.

The perceptual importance of source-filter interaction is as yet to a large extent unknown (see, however, Nord, Ananthapadmanabha, & Fant, 1984). Even so, I find it reasonable to believe that the types of source filter interactions that cannot be simulated within the formant synthesizer structure are not likely to be all that important. HiFi-synthesis (using a non-interactive formant synthesizer) carried out at the Dept. of Speech Communication and Music Acoustics and elsewhere, has shown it possible to achieve very close imitations of natural utterances, perceptually almost indistinguishable from the originals (Holmes, 1982; Fant, Gobl, Karlsson, & Lin, 1987a; Klatt, 1987; Klatt & Klatt, forthcoming).

The approach outlined in the present paper is to retain and optimize the source function within the linear source-filter theory. It is important to utilize a parameterized model which is capable of preserving the main shape (disregarding any kind of ripple components due to source-filter interaction) of almost any glottal pulse which is likely to occur. Further, one must study how the glottal pulse shape varies in a linguistic context.

Several different voice source models (e.g., Ananthapadmanabha 1984; Fant, 1979a; 1979b; 1982b; Fant, Liljencrants, & Lin, 1985a; Hedelin, 1984; Klatt & Klatt, forthcoming; Ljungqvist & Fujisaki 1985a; Rosenberg, 1971; Rothenberg, Carlson, Granström, & Lindqvist-Gauffin, 1974) and numerous methods of analysing the glottal airflow (Anan-
thapadmanabha, 1984; Cranen & Boves, 1985; Fant & Sonesson, 1962; Hunt, 1987; Lindqvist-Gauffin, 1965; 1970; Ljungqvist, 1986; Ljungqvist & Fujisaki, 1985b; Rothenberg, 1973; Sondhi, 1975) have been described in the literature. However, these models and these methods have essentially been utilized for studying the voice source under static conditions such as steady-state vowels (Gauffin & Sundberg, 1980; Holmberg, Hillman, & Perkell, 1987; Mårtensson, 1965; Sundberg & Gauffin, 1979, etc.). Relatively little attention has been paid to the dynamic properties of the voice source in connected speech (see however, Ananthapadmanabha, 1984; Fant, 1980; Gobl, 1985). It is the context-dependent temporal variations of the voice source that is the centre of interest in the present paper. The ultimate objective is a set of rules for the variation of the source model parameters according to features like intonation, stress patterns, onset and termination of utterances, phonetic differences, etc. These rules could then be incorporated in text-to-speech synthesis.

2. The Voice Source Model

The model that was utilized for simulating the volume velocity flow is a four parameter model, named the LF model. Its properties have been fully described in "A four-parameter model of glottal flow" (Fant & al., 1985a).

The four parameters are used to model the differentiated flow rather than the real glottal flow. A differentiation of the source function emphasizes the higher frequencies of the source spectrum (+6 dB/octave) and therefore allows a more precise spectral match in this frequency region. The differentiated flow is commonly used in speech synthesis, and includes the effect of radiation at the lips.

The model consists of two parts (cf. Fig. 1). The first part is an exponentially growing sinusoid to which three of the four parameters of the model pertain. This segment is

\[
\frac{du(t)}{dt} = E(t) = E_0 \cdot e^{at} \sin \omega_0 t
\]

\[t_0 \leq t \leq t_e\]

modelling the flow from glottal opening until the main excitation occurs (the moment of maximum discontinuity in the glottal airflow function, which normally coincides with the moment of maximum negative flow derivative). As opposed to most other models of glottal flow, the LF model is a continuous function until the main excitation, and therefore does
not introduce additional excitations. In comparison, the Fant model (Fant, 1979a; 1979b) is composed of two different segments; a rising branch up to maximum flow and a falling branch down to complete closure. The discontinuity between the two segments introduces a secondary weak excitation at the flow peak.

Fig. 1. The LF model. A four parameter voice source model.

The three parameters pertaining to the first segment of the LF model are:

(1) $E_0$ which is merely a scale factor.

(2) $a = -B \pi$ where $B$ is the "negative bandwidth" of the exponentially growing amplitude.

(3) $\omega_g = 2\pi F_g$ where $F_g = 1/2t_p$ and $t_p$ is the rising-time (the time from glottal opening to maximum flow).
The second part of the model is an exponential segment that allows a residual flow (dynamic leakage) after the main discontinuity, at time $t_e$, when the vocal folds close. The segment used for this "return phase" is

$$E(t) = -\frac{E_e}{\epsilon t_a} \left[ e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)} \right]$$

$t_e \leq t \leq t_c$

where $t_a$ is the fourth parameter of the model. $t_a$ is the time constant of the exponential curve and is determined by the projection on the time axis of the derivative at time $t_e$.

$\epsilon$ can iteratively be determined from

$$\epsilon t_a = 1 - e^{-\epsilon(t_c-t_e)}$$

and for small values of $t_a$, $\epsilon$ is approximately equal to $1/t_a$. $E_e$ is the negative amplitude of the excitation-spike and $t_c$ is the moment when complete closure is reached.

The effect of the return phase on the source spectrum is, due to its exponential waveshape, approximately a first order low-pass filter with a cutoff frequency $F_a = 1/(2\pi t_a)$. This means; the longer the return phase, the lower the cutoff frequency, and the larger the high-frequency reduction. The attenuation, $\Delta L_a(f)$, in decibels at frequency $f$ (as compared to a constant spectral slope of -6 dB/oct) is approximately:

$$\Delta L_a(f) = 10 \log_{10}(1 + f^2/F_a^2)$$

By convention $t_c = t_o$, the time of glottal opening for the forthcoming pulse period. This implies that the model lacks a closed phase. In practice this is no drawback; for normal (small) values of $t_a$, the exponential curve will fit closely to the zero line, providing, for all extents and purposes, a closed phase. The lesser number of parameters makes the implementation of the model simpler.

Apart from the four parameters, there is a requirement of area balance,
which keeps the zero flow line from drifting.

One of the properties of this model is the possibility of achieving a smooth and gradual change from a sharp excitation to a perfect sinusoid ($\alpha = 0$).

For the analysis, the parameters $E_i$, $r_g$, $r_k$, and $r_o$ were used, as these are more convenient and more closely related to different properties of the voice source (cf. Fig. 2).

*Fig. 2.* $U(t) =$ amplitude of true glottal flow as a function of time.  
$E(t) =$ amplitude of differentiated glottal flow as a function of time.  
$U_p =$ peak flow.  
$E_i =$ maximum positive rate of change in the flow function (the inflexion point).  
$E_e =$ negative level of the flow-derivative at maximum flow-discontinuity (the excitation).  
t_i, t_p, and t_e = time points of $E_i$, $U_p$, and $E_e$, respectively.  
t_o = time of glottal opening.  
t_c = time of complete (or maximum) closure.  
Note: $t_i$, $t_p$, and $t_e$ are also used to denote the durations $t_i - t_o$, $t_p - t_o$, and $t_e - t_o$, where $t_o$ is defined as equal to zero.  
t_n = duration of $t_e - t_p$.  
For explanation of $t_a$, see text.
The model parameters $E_0, a, \omega_q$, and $t_a$ can then, from the analysis parameters, be calculated by using an iterative method. Alternatively, other quantities such as open quotient defined as $O_q = t_e / T_o$, the cut-off frequency of the return phase, $F_a = 1/(2\pi t_a)$, etc., may be used as analysis parameters. Such alternative parameters will be discussed further in Section 7.

### 3. Speech Materials

The speech materials used for the analysis was recorded in an anechoic chamber on an FM tape recorder. The speech was picked up by a 1" B&K condenser microphone. This procedure ensured high-quality recording with correct phase response even at the lowest frequencies, thereby minimizing the risk of distorting the glottal pulse shape.

Three adult male Swedish speakers with rather different voice qualities were studied. The utterances analysed were:

- **behslla** from the sentence *vi vill behålla honom*  
  /vivillbehællahɔnɔm/ ("we want to keep him") read with three different stress patterns: focus on vill /'vıl/, focus on behålla /bɛ'holːaː/, and focus on honom /'hɔnɔm/.

- The two sentences *en alldeles utmärkt idé* /ɛnɛldɛlėsˀɪdɛːt, mærkɪtɪdɛː/: ("a perfectly marvellous idea") and *inte i detta århundrade* /intɛɪˈdɛ,tɑːːrɔːndrədɛː/ ("not in this century").

- **ja** /ja:/ ("yes"). **adjö** /a'jɔː:/ ("goodbye")

and only for one of the speakers:

- **Jag heter Johan** /jɑːheːtɛˌjuːan/ ("my name is Johan")

The voice of a ten year old child (male) was also studied. The utterances analysed were:

- **ja**, **adjö** and **Jag heter Johan**

(In the transcriptions, the following symbols were used to denote stress: ' indicates that the following syllable has acute accent, X indicates that the following syllable has grave accent, and , indicates that the following syllable has the secondary stress of a grave accent.)

The speech data was low-pass filtered at 6.3 kHz and sampled at 16 kHz. It was further processed as follows: a high-pass filter at 20 Hz (linear phase) was introduced to remove variations of the zero line,
due to possible superimposed low-frequency pressure fluctuations in the recording-room. Oscillograms of the speech waveform, spectrograms, and spectral sections (for some of the materials) were plotted.

4. Procedure

The source was studied by means of inverse filtering, employing a number of complex-conjugate zeros to cancel the filtering effect of the supraglottal system. Each zero should correspond to a formant of the vocal tract. The complete zero function then represents the inverse supraglottal transfer function at a particular moment in time. With a sampling frequency of 16 kHz and assuming a constant vocal tract length of 17.5 cm for the male subjects, the theoretical number of formants would be eight. No higher pole correction is needed; it is inherent in the digital realization. In practice, nine zeros were consistently used. The reason for the ninth zero and its effect will be discussed later in this section. To get the "true" glottal flow, the radiation at the lips has to be compensated for by a simple real pole at zero frequency (integrator). Otherwise, the differentiated glottal flow is obtained, this being, one might note, the almost exclusive object of study in this paper.

One of the main problems with this method is of course that of finding the correct transfer function of the supraglottal system. Any error in the parameters of the inverse filter will more or less distort the glottal pulse. The pulse shape is very sensitive to erroneous settings of the frequencies and bandwidths of the first formant, especially when F1 is low. Minor errors in the higher formants have little effect on the main pulse shape. One limitation of the technique used here is that, as zero airflow is not indicated, the absolute volume velocity level cannot be calculated. This problem can be solved using the circumferentially vented pneumotachograph mask developed by M. Rothenberg (Rothenberg, 1973). Only the variations in airflow were studied here; for synthesis purposes DC components caused by constant glottal leakage are of minor importance.

To achieve the best possible results, it was thought best not to use automatic formant tracking based on linear prediction techniques. Existing methods are unreliable as they generate frequent errors in formant locations and bandwidth and, often, spurious poles with very large bandwidths. Instead, the time varying filter was manually adapted to the speech waveform utilizing an interactive computer program (INA, by J. Liljencrantz, KTH, cf. Fig. 3) which permits manipulation (with a joy-stick) of frequencies and bandwidths of the zeros (or poles) used for filtering the speech input. Whenever a change is made in the filter settings, the filter-output is updated. The spectrum of the new filter-output is seen directly on the screen. When satisfactory inverse filter
Fig. 3. Program for interactive filtering of speech waveform (INA).

(a) Speech waveform.
(b) Log FFT-spectrum of (a).
(c) Filtered waveform.
(d) Log FFT-spectrum of (c).
(e) Filter configuration. The filter can be manipulated interactively in the s-domain using a joy-stick. The filter output is updated in real-time both in the time domain (the filtered waveform) and in the frequency domain (the FFT-spectrum).
settings have been obtained, these may be stored in a separate data file. Filter functions can be specified for as many frames as are needed to follow the dynamic variations of the vocal tract. The data file for a whole utterance is used for calculating the differentiated glottal output. Filter values in between analysed frames are linearly interpolated and the formant trajectories are smoothed with a low-pass filter. The inverse filter output of a complete utterance is depicted in Fig. 4. Notice the interaction ripple in the glottal open phase of the /ð:/ vowel and the often non-flat "closed" phase (cf. Fant, Lin, & Gobl, 1985b).

Fig. 4. The inverse filter output (differentiated glottal flow) of the utterance (a' jð:). Subject JS. Notice the source-filter interaction ripple in the glottal open phase (particularly in frames 34-47) and the frequently non-flat "closed" phase.
The next step, once the determination of the inverse filter output (differentiated glottal flow) is deemed satisfactory, is to fit the parameter model as closely as possible to the glottal pulses. This was done with a computer program developed at KTH by T.V. Ananthapadmanabha following suggestions by G. Fant, and involved the following procedure: with a joy-stick six time points in the glottal period being analysed are marked out. These are (cf. Fig. 2):

(1) glottal opening ($t_o$);
(2) point of inflexion (maximum positive derivative, $t_i$);
(3) point of maximum flow ($t_p$);
(4) point of excitation ($t_e$);
(5) the projection on the time axis of the derivative at the beginning of the return phase ($t_r$);
(6) next glottal opening ($t_o$).

When all points have been set, the program asks whether $t_i$ or $t_p$ is reliable and whether area balance is desired. From these data, the curve-form and the parameter values of the glottal model are calculated (cf. Fig. 5).

Fig. 5. Program for source pulse matching.
(a) Inverse filter output (differentiated glottal waveform).
(b) Log FFT-spectrum of (a).
(c) Calculated curve form of LF model.
(d) Log FFT-spectrum of (c).
(e) Values of analysis parameters of LF model.

The example shows the matching of the second pulse in frame 39 of Fig. 4. Notice that the model is not capable of capturing the source-filter interaction ripple in the glottal open phase.
The position of the maximum positive derivative, $t_i$, is often very much affected by a source-filter interaction ripple. Since the LF model is not aimed at modelling interaction ripples, I have consistently regarded $t_p$ as more reliable than $t_i$. The optional area balance was used throughout the materials.

The curve-form of the model and the inverse filter output are then compared in both the time and the frequency domain. If the divergence between the two is too large, the parameter values can be edited manually and a new calculation and comparison can be made. Alternatively the procedure can be recommenced with a new set of time points. When finally satisfied with the glottal matching, the source parameter values can be stored in the same data file as the filter parameters. This data file was used for plotting the dynamically varying source parameters and (for some of the materials) for creating resynthesized versions of the original utterances. A synthetic voice source corresponding to the inverse filter output in Fig. 4 is depicted in Fig. 6.

Fig. 6. Synthetic voice source corresponding to the inverse filtered waveform in Fig. 4. The time-varying LF model parameters were estimated by means of source pulse matching.
Almost consistently for all of the three adult male subjects studied, the harmonics around 3 kHz (in the region of F4) were found to be enhanced (often by more than 10 dB) in comparison with what would be expected from the theory of an "ideal" acoustic tube with an effective length of 17.5 cm. However, the nomograms in Fant (1960, Fig. 1.4-11) show that F4 and F5 are often relatively close and that F5 is considerably lower than for a uniform tube (4500 Hz). The fact that the levels of formants coming into proximity are increased, is only in part an explanation for the enhancement. There are other possible contributing factors: it could be the consequence of an extra long vocal tract combined with what has been termed the "singer's formant" (Sundberg, 1972). Additionally, or alternatively, it might be that cross-resonances are interfering with plane wave propagation. One of the subjects, JS, is considerably above average height and his vocal tract is therefore very likely to be longer than 17.5 cm. Moreover, he is a trained singer, a fact which is likely to affect his speech. The position of his larynx may be lower compared with speakers without such training, and this would further increase the vocal tract length. Pharyngeal widening is likely to co-occur with larynx lowering (cf. Sundberg, 1972). Together, these factors could result in lower and more salient cross-resonance frequencies. The enhancement is definitely the largest for this subject. However, it is also clearly found for the other subjects, BG and LN, who are not trained singers. Of these two, BG (who is also the taller) exhibits this tendency to a greater degree.

It was because of the enhancement of energy in this region that nine zeros were used in the inverse filter function. Without an anti-resonance for F5 close to F4, there would be a considerable amount of residual ripple in the flow function which could not be explained as acoustic interaction between source and filter. The effects with and without the F5 cancellation (eight and nine zeros respectively) are illustrated in Fig. 7. One might think that it would be sufficient to cancel F1-F5 correctly (as in (b), Fig. 7), and only use three zeros for the region above F5. However, this strategy would also result in ripple components and noise (but of higher frequencies than F5) in the flow function. Due to the missing anti-resonance, levels above F5 would be erroneously emphasized. Disregarding ripple components, the only important effect the ninth zero has on the main pulse shape is a slight increase in the duration of the return phase. For convenience, this filter configuration (nine zeros for the adult males, which theoretically corresponds to a vocal tract length of 19.7 cm) was used for all voiced sounds, and this may have introduced small errors for sounds other than oral (non-nasalized) vowels. But, as pointed out, this mainly affects the return phase when glottal matching is carried out. For the analysis of the child, six zeros were used in the inverse filter function, which corresponds to a vocal tract length of 13.1 cm.
Fig. 7. The filter output (a) without and (b) with F5 cancellation (using eight and nine zeros, respectively) in the inverse filter function for /e/ from the word behálla, subject BG. Notice the residual ripple in the glottal closed phase and the remaining peak in the source spectrum when the F5 anti-resonance is lacking.
5. Results

In this section some of the results and observations from my work with the voice source, mainly in terms of the analysis parameters described in Section 2, is presented. Figs. 9-14 show temporal variations of the glottal parameters $E_g$, $r_g$, $r_k$, and $r_a$. For the sake of comparison, the alternative parameters $Q_g$ and $P_a$ were also calculated and plotted. However, the discussion in the text is focused on the four former ones.

5.1. The Swedish Word Behälla

The word behälla was analysed in three differing stress contexts and for three adult male voices (Figs. 9-10). The differing stress environments arose from repetitions of the Swedish utterance *vi vill behälla honom*, read so that the emphasis was on either vil (referred to as postfocal context), behälla (referred to as focal context) or honom (referred to as prefocal context).

5.1.1. Variation Due to Different Stress Patterns

The main influence of the stress patterns is on the excitation parameter, $E_g$. Tables Ia, Ib and IIa, IIb, derived from Figs. 9-10, summarize some of the important aspects of this parameter. Not surprisingly, $E_g$ tends to be stronger for vowels and weaker for consonants, regardless of stress environment. However, when behälla occurs in focal context, the excitation-spike exhibits a larger dynamic range: as would be expected, vowels tend to be relatively stronger, but there is also a clear tendency for voiced consonants to be relatively weaker, so that the distinction between the two types of sounds is enhanced (cf. Fant, Nord, & Kruckenberg, 1987).

In order to highlight the dynamic range of $E_g$, Table Ia gives the maximum value for each vowel and the minimum value for each consonant in behälla across stress environments and speakers ($E_g$ is measured in arbitrary units). Table Ib shows for each sound in behälla, the relative difference (in dB) between the different stress contexts. Table IIa shows the relative difference between adjacent segments and between the vowels /ɔ/ and /ɛ/. Finally, Table IIb shows the relative difference in contrast between the segments in Table IIa, for the different stress contexts.

* One could question whether this is an adequate way of representing the excitation-strength of the different sounds. It can sometimes be misleading: compare, for example, the values in Table Ia with the corresponding Fig. 9a (postfocal context, subject JS). Note that the vowel /ɔ/ actually exhibits, for most of its duration, a weaker excitation than the preceding voiced consonant /h/. Other approaches could be, for instance, to take the average of the $E_g$ values within a segment, or simply to take the value in the middle of each segment. However, problems in finding the "true" segment boundaries would influence the result for these methods. Furthermore, the full dynamic range of the excitation would not be captured. As the dynamics of the voice source is the centre of attention in this study, the method used was found to be the most appropriate.
Table Ia. Values of $E_e$ (excitation-strength) for the segments in the word behälla in the three stress contexts.

<table>
<thead>
<tr>
<th>SUBJECT: JS</th>
<th>$E_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>postfocal</td>
<td>450, 1240, 1840, 1900, 940, 950</td>
</tr>
<tr>
<td>focal</td>
<td>300, 2130, 70, 1930, 770, 1550</td>
</tr>
<tr>
<td>prefocal</td>
<td>340, 2040, 1140, 2210, 1250, 1790</td>
</tr>
</tbody>
</table>

Table Ib. The relative difference in $E_e$ between the three stress contexts for the segments in the word behälla.

<table>
<thead>
<tr>
<th>SUBJECT: JS</th>
<th>(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>postfocal</td>
<td>-3.5, 1.4, -26, 1.6, -1.3, 4.3</td>
</tr>
<tr>
<td>focal</td>
<td>-1.1, 0.3, -28, -1.2, 8.7, -1.1</td>
</tr>
<tr>
<td>prefocal</td>
<td>2.4, -1.1, 1.4, -2.9, -2.9, 5.5</td>
</tr>
</tbody>
</table>

Table IIIa. The relative difference in $E_e$ between adjacent segments and between the vowels /s/ and /e/.

<table>
<thead>
<tr>
<th>SUBJECT: JS</th>
<th>(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>postfocal</td>
<td>-2.8, 1.2, -19, 0.1, -2.9, 0.5</td>
</tr>
<tr>
<td>focal</td>
<td>-0.6, 3.1, -20, 1.5, -3.8, -0.6</td>
</tr>
<tr>
<td>prefocal</td>
<td>2.2, 2.0, -0.6, 1.4, -1.0, -1.4</td>
</tr>
</tbody>
</table>

Table IIIb. The relative difference between the three stress contexts, in the degree of contrast between the segments in Table IIIa.
Fig. 8. Temporal variations of the analysis parameters for the Swedish word behålla from the sentence vi vill behålla honom. Subject JS(a) Postfocal context (emphasis on vill); (b) Focal context (emphasis on behålla); (c) Prefocal context (emphasis on honom).
A tendency was mentioned above of a stronger excitation for the vowels in focal context. This, it must be pointed out, is not absolute. One exception is the final /a/ where one finds that $E_e$ is slightly stronger in prefocal than in focal context for all three speakers. This is perhaps not too surprising. It is the second syllable in behálla which carries the word stress. Thus, when the word occurs in prefocal context the unstressed /a/ immediately precedes the emphatically stressed first syllable of homon, and is clearly somewhat affected by it. Note however, that /E/ in behálla is also unstressed; yet, when it is immediately preceded by the emphatically stressed vill (postfocal context), it is not similarly affected. A very tentative hypothesis at this point might be that an unstressed vowel is more affected by a stressed element to its right than to its left.*

The other gross corollary of stress mentioned above was a weakening of $E_e$ for voiced consonants. Again, this statement needs to be refined somewhat. The effect is quite different for the three consonants:

1. For all three subjects, /h/ exhibits the largest weakening. For JS and LN it is very extensive; $E_e$ is about 20-25 dB weaker compared to the non-focal contexts.

2. In the case of /l:/, the effect is not always consistent, and the weakening is of considerably lesser degree (the weakening is here never more than ca. 4 dB, as can be seen in Table Ib).

3. For /b/ there is no clear evidence of a weakening (the reason for this could be that /b/ does not pertain to a potentially stressed syllable in any case). Differences found are small and inconsistent.

At this point, it might be worth considering the possible production factors that would give rise to the observed variations of $E_e$. The stronger $E_e$ associated particularly with the vowels of the word in focal context could be explained by either (or a combination of) two factors:

1. Increased respiratory effort and subsequently higher subglottal pressure. Increased activity of the internal intercostals was demonstrated to immediately precede the stressed syllable (Ladefoged, Draper, & Whitteridge, 1958).

2. Increased medial compression of the vocal folds. By bringing about a more abrupt closing phase of the glottal cycle, this mechanism could conceivably yield a higher $E_e$ without any necessary contribution from respiratory factors. Data from the word "adjö" in Fant

* For subject JS /b/ is also stronger in prefocal than in focal context. Note, however, that the focal case is nevertheless characterized by a greater contrast between the vowel and adjacent consonant segments (Table IIb).
Fig. 9. (a)(b)(c) Subject BG. Otherwise as in Fig. 8.
(1979a) showed evidence of increased medial compression in the stressed syllable (for a discussion, see also Fant 1981; 1982a).

Of the two possible explanations, the present data suggest the former to be the more likely. If the latter were implicated, one should probably find (along with increased $E_e$) a reduced $r_3$ value (less dynamic leakage) and a smaller open quotient, and these were not attested here.

As regards the very great weakening of $E_e$ for /h/ in the focal context, it would appear to reflect the fact that /h/ tends to be substantially voiceless in this environment (in Swedish /h/ is normally assumed to be voiced in intervocalic position), whereas it is fully voiced in the non-focal contexts.* From this we can infer that there is likely to be a greater degree of vocal fold abduction associated with /h/ in the focal context. See also Fant (1987) for a discussion of this point. A greater degree of glottal abduction for voiceless consonants in stressed syllables was also reported by Ni Chasaide (1985; 1987) - a conclusion based on photo-electroglottographic data and more recently supported by EMG data (personal communication).

In the literature there is evidence suggesting that stress may involve increased muscular activity at every level of production: respiratory, laryngeal and articulatory (see, for instance, Ni Chasaide 1987, and references therein). In the case of /1:/, respiratory and articulatory correlates of stress may somewhat counteract each other in determining the $E_e$ level. Increased respiratory levels should yield a higher $E_e$ as for the vowels. On the other hand, if the lateral is articulated with greater force, one might expect it to be executed more rapidly, to involve (perhaps) greater oral occlusion, and to have a longer duration. A greater occlusion and increased duration should reduce the transglottal pressure drop and thus $E_e$. That /l:/ is longer in focal context can be ascertained from Figs. 8-10. Note also in these figures that in the stressed cases there is a sharper, more abrupt drop in $E_e$ from /3/ to the following /l:/: This suggests that the articulatory gesture is indeed more rapidly executed. Furthermore (regardless of environment), the minimum tends to occur at the end of the segment, an indication that the aerodynamic consequences of oral occlusion do seem to cause a depression of $E_e$.

The combined effect of stronger $E_e$ for vowels and weaker $E_e$ for voiced consonants in focal context means that the contrast between these types of sounds is enhanced (see Table IIb). Put another way, it means that the nucleus of the stressed syllable is much more prominent relative to the surrounding consonants. Past researchers have tended to

*For speaker BG, /h/ is still fully voiced in focal context and the weakening is consequently much less, only 2-4 dB, as can be seen in Fig. 9 and Table Ib. The tendency towards a weakening of consonants in focal position appears to be generally less salient and also less consistent for this speaker.
Fig. 10. (a)(b)(c) Subject LN. Otherwise as in Fig. 8.
regard amplitude variation (between stressed and unstressed syllables) as one of the less important correlates of stress, being apparently less important perceptually than other cues such as FO or duration differences. Such conclusions were based on experiments where the absolute amplitudes of syllables were manipulated in a rather gross fashion (see Lehiste, 1970, and references therein). The data reported here suggest that stress may be less reliably manifested by absolute increases in intensity than by the increased contrast of syllable nucleus and margins. For example, looking in detail at $E_o$ levels of the vowel /ɑ/ for subject LN across stress environments (Fig. 10), note that there is very little difference in absolute values. However, when stressed, /ɑ/ is very prominent relative to the surrounding consonants, something which does not hold for the unstressed environments. Formulated in terms of a syllable nucleus/margin contrast (rather than in absolute peak levels of the syllable) it might well be that amplitude variation would emerge as a more important production/perception correlate of stress than it is currently thought to be. This suggestion is one I hope to investigate in the near future.

One final comment: note that peak $E_o$ levels for /ɛ/ tend to be slightly higher than for /ɑ/, irrespective of whether the word occurs in focal/non-focal context (Tables Ia and IIa). This might appear surprising: /ɛ/ is always unstressed, whereas /ɑ/ is the potentially stressed syllable of the word. However, it may not be wise to attach too much importance to this finding at this stage, as two different vowel qualities are involved in the comparison. The differences found here could simply be due to different intrinsic voice source characteristics of the vowels concerned.* More controlled data will be required in order to be able to distinguish between stress and vowel quality effects.

* This should not be confused with the well-known fact that different vowels have different intrinsic intensities (see review in Lehiste, 1970). There is not necessarily a correlation between intrinsic vowel intensity and voice source characteristics (i.e., $E_o$ in this particular case). The intensity of a vowel is determined by both its formant pattern and its source characteristics. Of these two factors, the formant pattern probably contributes more to the intrinsic vowel intensity. However, it is still quite likely that there are appreciable intrinsic voice source differences between different vowels or vowel categories. It has been shown that the vocal tract load, under certain conditions, does affect the glottal source quite substantially. Interaction effects such as superposition of formant energy from one fundamental period to the following (Fant & Ananthapadmanabha, 1982; Lin, 1987) and inertive loading of the vocal tract (Rothenberg, 1981; 1983), may be particularly important to the $E_o$ parameter. Model simulations have shown differences in glottal pulse shape for different vowels (Ananthapadmanabha & Fant, 1982). However, estimations of intrinsic voice source characteristics in natural speech have not been carried out satisfactorily. Mechanical interaction between articulators and larynx muscles, effects of different degrees of vowel height, and other phenomena not taken into account in the present simulations, may cause a discrepancy between source pulses of natural speech and the simulated ones. I hope to be able to clarify some aspects of this topic in future research.
Beside the variation in excitation-strength, $E_e$, it is difficult to find reliable source properties related to stress. One might have expected a shorter return phase in focal context because of greater medial compression of the vocal folds. However, as was discussed above, this expectation was not really borne out in these results. Subject LN shows higher $r_a$ values for the vowels in focal context (Fig. 10); for the other two subjects the differences sometimes go in the expected direction but are small. Further analyses of more speakers and contexts are needed.

5.1.2. FURTHER OBSERVATIONS MADE FROM THE WORD BEHÅLLA

There are further observations to be made from the temporal parameter variations of this word beside those pertaining to the stress pattern. For instance, one might note that, regardless of speaker or stress environment, $E_e$ drops during the /b/ reaching a minimum just before release: this is naturally a consequence of the oral closure, and of the consequently increasing oral pressure and decreasing transglottal pressure drop. The diminishing transglottal pressure makes the vocal fold vibration more and more difficult to sustain and is directly reflected in the excitation-strength. The decrease of $E_e$ is generally accompanied by an increase of $r_a$ and $r_k$. Specifically for $r_a$: the initial value is relatively high (6-8%) compared with the typical value for the adjacent vowel (2-3%), and then gradually increases until release (10-13%) where it drops rapidly. The correlation between $r_a$ and $r_k$ is considerable here, as well as in many other positions.

Also, notice the higher dynamic leakage for /h/ as compared to the surrounding vowels. Here the duration of the return phase is often three to four times as long. There is also generally less skewing of the pulses (larger $r_k$).

Occasionally $r_a$ increases during /l:/ and then decreases rapidly at the transition to the vowel. The increase is analogous to (but much less pronounced than) that of the voiced stop /b/; the peak value for /l:/ is here typically 4-6%, and never higher than 8%. In focal context this is found for all three speakers. In postfocal context it is found for speakers BG and JS but not for LN. In prefocal context, where the distinctions between the lateral and the adjacent vowels are much more reduced, it is not clearly found for any of the speakers (sometimes there is a slight increase, but without the sharp drop at the transition to the vowel). The increase of $r_a$ is typically concomitant with the decrease of $E_e$ that was mentioned in Section 5.1.1.

What about the "rising-time" parameter, $r_g$? This parameter varies irregularly and within a narrow range, and is sometimes almost constant, as in focal context, subject JS (Fig. 8b). Excepting the rise in /l:/ and the initial low part in /b/, the $r_g$ value is very stable just around 100% for this speaker. As a possible explanation for the generally
narrow range of $r_g$ variation, I would suggest that the opening time in normal phonation may be more or less conditioned by the time constant of the mass and compliance of the vocal folds, whereas the closing time is affected by the Bernoulli forces. However, this would not be a complete explanation; the reasoning above concerns the glottal area and not the flow. The inertive loading of the vocal tract (Rothenberg, 1981) and the glottal inductance, delay the peak glottal airflow as compared to the peak glottal area (skewing to the right). Furthermore, Cranen and Boves (1985) have found the vertical phase lag between the upper and the lower parts of the vocal folds to contribute to the skewing of the flow pulse.

For these speakers, $r_g$ values are typically ca. 100-120%, which means that the glottal frequency, $F_g = 1/2t_p$, is normally slightly higher than the fundamental frequency. The only subject that occasionally shows a systematic variation for this parameter is LN. For instance, in focal context there is some covariation between $r_g$ and $E_e$ (Fig. 10b). Except for the final /a/, $r_g$ is higher for the vowels than for the consonants. But the variations are small and probably not perceptually important.

5.2. TWO COMPLETE SENTENCES

The two following Swedish sentences were studied: en alldeles utmärkt idé, with sentence stress on the word utmärkt, and inte i detta århundrade, with sentence stress on the word detta (Figs. 11 and 12). The subjects were the same as in Section 5.1. Here only the most general observations will be mentioned.

Throughout these sentences and for all three speakers there is a strong correlation between excitation-strength, $E_e$, and the negative amplitude of the speech waveform.

The parameter values of $E_e$ decrease and the values of $r_a$ and $r_k$ increase before the voiceless consonants /s/, /t/, /k/ and at the termination of the utterance. There is reason to believe that these changes, causing a smoother and more sinusoidal glottal pulse shape, occur whenever voicing is terminated; whether it is before a voiceless consonant, a pause, or the end of an utterance (disregarding special cases such as glottal stops and vocal fry). One could argue that this is a direct consequence of glottal abduction occurring prior to devoicing. In a study combining airflow recordings and inverse filtering (Ni Chasaide & Gobl, 1987), there is strong evidence of a correlation between glottal abduction and a change towards such voice source characteristics for devoicing prior to voiceless consonants. Furthermore, earlier airflow recordings of female voices (Ananthapadmanabha, 1984) show evidence of similar source characteristics for prepausal devoicing.

Usually $r_a$ and $r_k$ values are higher for the voice onset as well; both initially and within an utterance after voiceless sounds. However,
Fig. 11. Temporal variations of the analysis parameters of the Swedish sentence en alldeles utmärkt idé. Subject LN.

Fig. 12. Temporal variations of the analysis parameters of the Swedish sentence inte i detta århundrade. Subject BG.
the onset differences are not as consistent as for the devoicing and are
normally not of the same magnitude and duration.

Typical values before a voiceless fricative are 8-10% for \( r_a \) and
40-50% for \( r_k \). One might expect slightly lower values of these two
parameters before voiceless stops: as the stops involve complete oral
closure, this might conceivably lessen the extent to which glottal
abduction is needed to effect voicelessness. This is sometimes seen,
sometimes not. The \( r_a \) value before voiceless stops vary as much as 3-
18%, but a reasonable rule would be 3 times the value of the preceding
vowel, say 6-9%.

Maximum values of \( r_a \) are normally found at the termination of the
sentence. The increase extends over a longer time interval than in-
creases occurring within the utterance. This more gradual termination of
voicing may be related to the fact that final segments lengthen in any
case. Devoicing often starts in the middle of the final vowel, which is
here ca. 200-300ms long. Within an utterance the time interval over
which \( r_a \) increases is about 20-50ms.

The above comments on the voice source do not hold for cases where
an utterance ends in vocal fry. The glottal flow for creaky phonation
differs considerably and was not dealt with here.

5.3. COMPARISON OF ADULT MALE AND CHILD VOICES.

The Swedish utterance jag heter Johan was analysed for an adult
male (JS) and a child (AN) (Figs. 13 and 14). The subjects were told to
read the utterance in a natural way. Otherwise, no further instructions
were given.

The most striking difference is perhaps in reading styles: for JS
sentence stress falls on the last

word, the name Johan, whereas for AN
all three words are equally (fairly strongly) stressed. For the child
this means longer phonemes and less reduced pronunciation, especially
for the two first words.

Apart from distinctions pertaining to reading styles, there are
obviously differences in the glottal pulse shape as well. Again we can
see that the return phase, \( r_a \), plays an important role. Except before
the stop consonant, the adult has \( r_a \) values ranging between 1% and 4%
for the vowels, but the child’s values are much higher; 5-12%.

Disregarding the overall level differences, the gross pattern of
dynamic variation is roughly similar. At first glance the terminations
can look rather dissimilar. However, this is a consequence of the way
the model works and of the fact that the source function becomes approx-
imately sinusoidal, particularly for the child. The glottal pulses
towards the termination become less and less skewed, which means that \( r_k \)
values become larger. For small values of \( r_k \), there is a positive
correlation between it and \( r_a \). As \( r_k \) increases beyond a certain thresh-
old (normally around 50%, when \( r_g \) is around 100%) the correlation is
Fig. 13. Temporal variations of the analysis parameters of the Swedish utterance *jag heter Johan*. Subject JS (adult male).

Fig. 14. Subject AN (child). Otherwise as in Fig. 13.
reversed and \( r_a \) drops. For \( r_k \) values of more than 50\%, the main excitation, defined as the maximum discontinuity, does not coincide with the maximum negative flow derivative. The excitation occurs later and weakens as \( r_k \) increases (and \( r_a \) decreases). Here, reduced values of \( r_a \) are accompanied by a reduction in excitation-strength and therefore will not lead to stronger higher harmonics. When \( r_k \) and \( r_g \) are equal to 100\%, there is no excitation at all and there cannot, per definition, be a return phase (\( r_a = 0 \)) (cf. Section 2). For source pulses which are almost sinusoidal it is hard to locate the main excitation, and it is also very unlikely that there is only one dominating excitation. In these cases the source matching was done to give the best spectral resemblance and a reasonable continuity of parameter values. In other words; one will not always find an increase in \( r_a \) where one might have expected it (e.g. when \( r_k \) is increasing). If \( r_a \) is already high (as typically for the child's voice) and if the glottal pulse is becoming more sinusoidal, further change in the same direction may yield little increase or even a decrease in \( r_a \).

The onset for subject JS shows an exception to the rule of covariation between \( r_a \) and \( r_k \) that is not due to any property of the model. Here we have a hard (glottalized) onset which is reflected by the relatively low value of \( r_a \) (1\%) initially. Even though \( r_a \) is small and rising rather than falling towards /a/, \( r_k \) is falling (though not very much). In the child's voice, where the onset is much softer, \( r_a \) falls from 20\% to 5\% and \( r_k \) from 57\% to 38\%. Generally, \( r_k \) is slightly higher for the child than for the adult. Typical values for vowels are 25-35\% for the three adult males and 30-40\% for the child.

6. Summary and Conclusions

The distribution of the values of the analysis parameters (except \( E_e \)) of the LF model is shown in Fig. 15. These results are from 12 seconds of adult male speech and 2.5 seconds of a child's voice. Two alternative parameters, \( F_a \) and \( O_q \), were also calculated. \( F_a \) is inversely proportional to the absolute time of the return phase, thus independent of the fundamental period. Excepting cases where there are very large variations in \( F_0 \), it can be assumed that \( F_a \) increases when \( r_a \) decreases. The "open quotient", \( O_q \), is here defined as the relation between the time from glottal opening to main excitation and the fundamental period, \( t_e/T_0 = (1 + r_k)/2r_g \). This means that small values of \( r_g \) and large values of \( r_k \) correlate with large values of \( O_q \).

The histograms in Fig. 15 illustrate the range of values for the parameters in normal speech. Parameter values are mostly confined to a narrow part of the distribution. Glottal pulse shapes that differ considerably from the "typical" shape still constitute a small percentage of the total number. It is most likely, however, that they contri-
Fig. 15. Distribution of \( r_g \), \( r_k \), \( r_a \), \( F_a \), and \( O_q \) values, as a percentage of the total duration.

(a) 12 seconds of male voices. Three subjects.
(b) 2.5 seconds of a child's voice.

Note that in several of the histograms the x-axis is not consistently linear.
bute significantly to the perceived naturalness, an assumption corroborated by the resynthesis that was carried out. I do not claim that these limited samples of speech are necessarily representative for all of normal speech. There might indeed be significant differences between this type of speech materials and more casual or spontaneous speech. Nevertheless, they provide a good initial estimate of the necessary control range of these parameters.

The limited data in this study obviously preclude statistical analysis of the differences between adult male voices and children's voices in general. Bearing this in mind, the differences between the values of the parameters $r_a$ and $F_a$ are, however, striking. It is very likely that a larger dynamic leakage is a characteristic feature of children's voices compared to adult male voices. Other tendencies found in these data are higher values of $r_k$ and $O_q$ for the child's voice. Especially the very high values of $r_k$ ($>60\%$) are more frequent for the child. One might also note, for both the adult males and the child, that $O_q$ values of less than 50% are very infrequent, i.e. the closed portion (or the flat portion if there is a constant leakage through the glottis) of the glottal cycle is almost always shorter than half the fundamental period. The values of $r_d$ in these data are slightly lower for the child than for the adult males.

It transpires that a lot of information about the glottal pulse shape can be inferred directly from the negative amplitude of the speech waveform. There is a strong correlation between it and the excitation-strength, $E_e$ (as one might expect intuitively). $E_e$ is the main determinant of the initial amplitude of $F_1$, which in turn is the main determinant of the negative amplitude of the speech waveform. (Phenomena such as superposition of formant oscillations from previous periods mean that the relationship will not be perfectly linear).

The return phase $r_a$, for its part, varies inversely with the excitation, which can be understood in the following way: if, at the main excitation, the vocal folds do not make contact completely along their entire length, this will cause a residual flow. When the excitation is strong, the vocal folds are closing rapidly, and therefore the time from excitation to complete closure (or maximum closure, if there is constant leakage) will be short. Assuming a similar closing gesture, a weak excitation implies a slower rate of vocal fold closure, and therefore the time from excitation to complete closure will be longer.

As stated earlier, there is a covariation between $r_a$ and $r_k$ which can be explained if $r_q$ is fairly constant. When $r_a$ is small (and, from the reasoning above, $E_e$ large), the closing-time from maximum flow to the excitation will be short and cause a skewed pulse and thus a low $r_k$ value. Assuming similar opening and closing gestures and about the same peak flow, $U_p$, a large value of $r_a$ and a small value of $E_e$ implies a longer falling-time and therefore a high value of $r_k$. For one of the
<table>
<thead>
<tr>
<th>SOUNDS</th>
<th>$r_a$ (%)</th>
<th>$r_k$ (%)</th>
<th>$r_g$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>adult</td>
<td>child</td>
<td>adult</td>
</tr>
<tr>
<td>Vowels</td>
<td>2-3</td>
<td>5-10</td>
<td>25-35</td>
</tr>
<tr>
<td>Nasals and liquids</td>
<td>normally like vowels, see comment (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voiced stops: at oral closure</td>
<td>6-8</td>
<td>10-13</td>
<td>25-45</td>
</tr>
<tr>
<td>Voiced stops: at oral release</td>
<td>10-13</td>
<td>13-16</td>
<td>50-70</td>
</tr>
<tr>
<td>Voiced fricatives (except [h]); at maximum oral constriction</td>
<td>8-10</td>
<td>10-15</td>
<td>40-50</td>
</tr>
<tr>
<td>Voiced glottal fricative, [h]</td>
<td>10-15</td>
<td>10-20</td>
<td>40-60</td>
</tr>
<tr>
<td>Devoicing: before voiceless stops</td>
<td>6-9</td>
<td>10-15</td>
<td>35-60</td>
</tr>
<tr>
<td>Devoicing: before voiceless fricatives</td>
<td>8-10</td>
<td>10-15</td>
<td>40-60</td>
</tr>
<tr>
<td>Devoicing: prepausal</td>
<td>10-15</td>
<td>15-20</td>
<td>40-60</td>
</tr>
<tr>
<td>Voice onset: after voiceless stops and fricatives</td>
<td>2-6</td>
<td>8-12</td>
<td>25-45</td>
</tr>
<tr>
<td>Voice onset (soft); postpausal</td>
<td>4-8</td>
<td>10-20</td>
<td>40-50</td>
</tr>
<tr>
<td>Voice onset (hard); postpausal</td>
<td>0-2</td>
<td>3-6</td>
<td>40-50</td>
</tr>
</tbody>
</table>

Table III. Typical values for the parameters $r_a$, $r_k$, and $r_g$, for different sounds in different contexts, generalized from the inverse filtered data. See text for comments.
subjects (LN, cf. Fig. 10b) \( r_g \) covaried somewhat with \( E_e \), which means that the opening-time gets shorter as the excitation gets stronger. This will of course counteract the correlation between \( r_k \) and \( r_a \) and make the skewing more independent of the excitation-strength. However, in these materials the covariation between \( r_k \) and \( r_a \) is the dominant trend.

The above discussion on the interdependencies between the different glottal parameters is somewhat simplified and is mainly intended to give the reader an intuitively graspable explanation of the covariations found in these data. A linear relationship between the glottal area and the airflow through the glottis is assumed, i.e. influences of the vocal tract load and the glottal inductance are not taken into account. Furthermore, the peak flow, \( U_p \), is assumed to be proportional to the glottal opening time (and consequentially also to \( t_p \)). For some of the analysed materials, parameters did not covary exactly in accordance with the general trends just described, showing that the simplifying assumptions made here will not always hold. Further, for other voice types (and for non-modal voice qualities) one might well find other correlations between the model parameters.

In Table III typical values for the parameters of the LF model in different contexts are presented. They may serve as a starting-point for a useful dynamic voice source governed by rules. A first tentative approach to control the LF model in a text-to-speech system could simply be to let \( E_e \) be controlled by the amplitude parameter of the conventional exponential pulse source. The other parameters should then be determined from the value of \( E_e \), with different proportionality coefficients to give the best fit with the typical parameter values in Table III. As the main purpose of the table is to lay down the broad outlines for developing a source rule system, I have also included parameter values for sounds that were not analysed. These are of course only an approximate first guess based on the values of similar sounds, and will eventually need to be specified more precisely.

Comments on Table III:

(1) If devoicing is effected by a glottal stop, \( r_a \) should be reduced by a factor of 2-4, \( r_k \) should be increased by a factor of 1.5-2, and \( r_g \) should be increased by a factor of 1.3-2.

(2) The /l/ sound may exhibit a pattern similar to the voiced stops. In those cases, the \( r_a \) value at oral release should be 1-3 times the initial \( r_a \) value. This pattern is more likely to be found in stressed syllables, and is normally accompanied by a decrease in the excitation, \( E_e \). The \( r_a \) value for nasals may be increased by a factor of 1-2.

(3) \( r_g \) may covary moderately with the excitation, \( E_e \).
7. Discussion

The materials analysed here mainly shed light on the general behaviour of the voice source in a normal dynamic context. For future studies it would be interesting to investigate whether there are any significant inherent differences in glottal pulse shape as for different voiced sounds. Obviously vowels differ considerably from most voiced consonants, but there could conceivably also be physiologically related dependencies that might lead to differences in voice source characteristics from one vowel to another, or at least to differences between groups of vowels, say, back and front vowels (cf. Footnote on p. 144). (For effects of various degrees of vocal tract constriction on the voice source, see Bickley & Stevens, 1986.) One might also ask as to whether there might be inherent voice source differences between nasalized and non-nasalized vowels, as well as source differences which correlate with voice loudness (see, for instance, Holmberg, Hillman, & Perkell, 1987), vocal fold tension, FO, etc. (For FO dependencies, see Fant 1982a; Pant & Ananthapadmanabha, 1982; Monsen & Engebretson, 1977.) To be able to answer some of these questions, controlled speech materials will be needed where variables that affect the glottal airflow, other than the one under investigation, are held as constant as possible. This procedure will ensure that differences found are really related to the feature studied, and not caused by other covarying properties.

Alternative analysis parameters have been mentioned in Sections 2 and 6. The question remains as to which quantities are phonetically the most relevant. Unfortunately it is not a simple decision. To give an example: originally we defined the open quotient as \( \frac{t_e + t_a}{T_o} = O_q + r_a \) according to the traditional definition. However, with \( r_a \) included, it resulted in open quotient values depending very much on the return phase. Because of the importance of the return phase for the source spectrum, it seems reasonable to describe it with one independent parameter \( (r_a, F_a, \text{or} t_a) \). To use another parameter including the return phase would give redundant information and obscure other dynamic source properties. Therefore the open quotient is defined as \( \frac{t_e}{T_o} \). Furthermore, the values of \( O_q \) seem closer to earlier data on the ratio between glottal open period and fundamental period, where the return phase was normally disregarded.

The excitation-strength is of paramount importance and should naturally be described by a single parameter, here \( E_e \). The return phase is also very important, but it is not obvious which of the representations \( r_a, F_a, \text{and} t_a \) is preferable. \( r_a \) has the advantage of providing intuitively graspable information on the glottal pulse shape. \( F_a \) which is the inverse of \( t_a \) multiplied by a constant, is independent of FO but instead directly correlated to the source spectrum. For the two remaining parameters there exist several candidates. Three have already been
mentioned above \( (r_k, r_g, \text{and } 0_q) \), but any of the following might have been used: \( U_p, A_i=E_i/E_e, t_p/t_n=1/r_k, \) closed quotient \( C_q=1-O_q, R_g, t_n, \) etc. The two parameters, \( r_k \) and \( r_g \), mainly affect the low frequency energy of the source spectrum. Simplistically one might say they represent level and frequency of the "glottal formant", even though the properties "level" and "frequency" cannot be described by independent time domain parameters.

Some of the utterances were resynthesized, using formant values taken from the inverse filtered data and source parameter values from the matched source data. The child's voice is most strikingly improved with the more sophisticated model, but also for the adult male voices synthesis quality is considerably enhanced.

Most of the work here was concerned with adult male voices. It would be useful in the future to focus on the voices of women and children if we are to get a more complete picture of both voice source properties and the versatilities of the LF model.

Acknowledgments

The presented work was supported by the Swedish Telecom Administration (Televerket). The research was carried out at the Dept. of Speech Communication and Music Acoustics, KTH.

I wish to thank Prof. Gunnar Fant for his invaluable help and guidance. I am also indebted to Johan Liljencrants and T.V. Ananthapadmanabha for providing the computer programs, to Lennart Nord who did the recordings, and to Gudrun Tannergård who drew most of the figures. I am grateful to Ailbhe Ni Chasaide for discussing many aspects of this work and for critical comments on the manuscript.

References


