Voice source properties of the speech code

Fant, G. and Kruckenberg, A.

journal: TMH-QPSR
volume: 37
number: 4
year: 1996
pages: 045-056

http://www.speech.kth.se/qpsr
Voice source properties of the speech code

Gunnar Fant and Anita Kruckenberg

Abstract
This is an outline of the knowledge we need in order to include the voice source in an advanced model of speech production with applications to text-to-speech rules. Recent results from studies of the Swedish language provide information of source properties and source-vocal tract interaction as a function of the segmental and prosodic frame within an utterance and with reference to aerodynamic conditions. Our study discusses source modelling, individual voice qualities, segmental dependencies, influence of stress and accentuation, phrase contours, covariation with F0.

Introduction
A challenge for speech research and technology is to advance our knowledge base for preserving a maximum of efficiency in speech recognition and for improving the quality of text-to-speech synthesis. This is more of a problem in formant or articulatory coded synthesis than in concatenated systems like the PSOLA. However, a knowledge of factors affecting source parameters could also be adopted in concatenating strategies.

The basic problem is not that of the voice source alone. Covarying noise and various source-filter interaction phenomena have also to be considered, e.g. changes of source waveform and amplitude induced by supraglottal constrictions and the joint dependencies of both source and filter functions on the glottal state and overall aerodynamic conditions. Dynamic variations and coarticulatory phenomena within an utterance become more complex to model than speaker specific average data.

There are similarities between the loss of source efficiency, e.g. in amplitude and high frequency content, comparing the effects of supraglottal constriction and glottal abduction before a pause. These relations are best dealt with in the frame of an articulatory synthesis system where all relevant source properties become the automatic consequences of glottal and supraglottal articulations and lung pressure.

The human vocal cords execute a high degree of flexibility but also of instability in their vibratory modes which are sensitive to aerodynamic as well as acoustic interactions. A large number of conditioning factors contribute to seemingly chaotic variations of small details from one voice period to the next which add to the individual character of a voice. These are of secondary importance, but need to be studied.

We are beginning to understand most of the basic phenomena but we lack systematic and sufficiently complete descriptions. A problem is that we have very little experience from perceptual experiments. Much work is needed to reach an insight in the relative perceptual salience of various components of a source rule system.

Voice source parameterization
The basic requirement for developing voice source rules is an efficient parameterization system. We need to describe essentials with a small number of time varying parameters with the option to go into finer details if needed which requires an extended number of parameters. Covarying noise generation and modifications of vocal tract transfer function must also be considered.

How do we define the source? The common concept is the pulsating glottal flow or its derivative disregarding subglottal components and superimposed interaction ripple. The latter is essentially due to the finite presence of formant oscillations in the transglottal pressure drop which determines the glottal flow in a square law dependency. This non-linear transform accounts for the presence of distributed spectral dips, typically a zero around 2F1, and in the time domain an extra positive peak prior to the main peak of the differentiated glottal flow function (Fant, 1986) which complicates the time domain matching of source parameters. It is more apparent in a non-leaky relative strong phonation than in a weak breathy phonation. Another deviation from the ideal LF-model is an extra excitation at the instant of glottal opening which is seen in some voices. To these add the effects of supraglottal coupling in a leaky voice. All these

phenomena add motivation for relying more on frequency domain than time domain derivation of source parameters.

The LF-model

The LF-model (Fant et al., 1985) is a useful approximation that has found a wide use. For maximal correspondence between a natural sample and a synthetic replica it is necessary to adjust the LF-parameters to suit particular constraints such as higher pole conventions in cascaded formant synthesis systems, see (Fant, 1995).

The basic parameters of the LF-model are defined in the time domain but we make a maximal use of frequency domain correspondences (Fant & Lin, 1988) and frequency domain matching (Fant, 1995). A consistent frequency domain system is now used at MIT (Stevens, 1994; Stevens & Hanson, 1994; Hanson, 1995). Transformation between their system and ours have been established.

The LF-model (Fig. 1) operates with three shape parameters, \( R_k \), \( R_g \) and \( R_a \) in addition to the voice fundamental frequency \( F_0 \) and an excitation amplitude \( E_e \). The parameter \( R_k \) determines the degree of pulse skewing; \( R_g \) and to a less degree \( R_k \) influence the duration of the glottal pulse, and \( R_a \) is a relative measure of the duration of the return phase following excitation at closure. \( E_e \) is the negative derivative of glottal flow at the instant of excitation, i.e. of maximal discontinuity.

We usually disregard the \( R_a \) term in (1). However, \( R_a \) is the most important single parameter for spectrum shaping, It determines a cut off frequency

\[
F_a = \frac{F_0}{2\pi R_a} \tag{2}
\]

where the glottal flow derivative spectrum after an initial slope of -6 dB/oct turns into a slope of -12 dB/oct.

The low frequency part of the source spectrum in the region of the fundamental and the second harmonic is determined by the main shape of a glottal pulse. It is influenced by all three LF-parameters, not only by \( R_k \) and \( R_g \) but also by \( R_a \). The upper part of the source spectrum is mainly determined by \( F_a \).

Stevens & Hanson (1994) and Hanson (1995) quantify the low frequency region of the source spectrum by \( H_1^* \), the amplitude of the voice fundamental and \( H_2^* \), the amplitude of the second harmonic where * indicates a property of the source spectrum in distinction to the complete sound spectrum. We have established an empirical relation for deriving \( H_1^* \) and \( H_2^* \) from the open quotient (Fant, 1993).

\[
H_1^* - H_2^* = -6 + 0.27\exp(5.5OQ) \tag{3}
\]

Table 1. Change in each of \( R_k \), \( R_g \) and \( R_a \) needed to increase the level of the fundamental
\( H1 \) by 1 dB and \( H1-H2 \) by 1 dB keeping other parameters constant.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>dR/dH1</th>
<th>dR/d(H1-H2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_a )</td>
<td>1.0</td>
<td>1.25</td>
</tr>
<tr>
<td>( R_k )</td>
<td>3.0</td>
<td>4.5</td>
</tr>
<tr>
<td>( R_g )</td>
<td>-12.0</td>
<td>-10.0</td>
</tr>
</tbody>
</table>

The \( R_d \)-transform

A major extension of the LF model (Fant et al., 1994; Fant, 1995) has been to introduce a data reduction scheme whereby the wave shape parameters \( R_k \), \( R_g \) and \( R_d \) are collapsed into a single parameter \( R_d \).

\[
R_d = (1/0.11)(0.5+1.2 R_k)/(R_k/4 R_g+R_d)
\]

(4)

This is an approximation to the basic physical definition

\[
R_d = (U_0/E_0)(F0/110)
\]

(5)

which is independent of the LF-model. Here, \( U_0 \) is the amplitude of the glottal flow pulse above a possible steady leakage flow. The ratio \( (U_0/E_0)=T_d \) is the effective duration in ms of the falling branch of glottal flow, see Fig. 1. At a normalising \( F0=110 \) Hz there is numerical identity between \( R_d \) and \( T_d \).

An advantage of introducing the \( R_d \)-parameter is that default values of \( R_k \), \( R_g \) and \( R_d \) can be predicted from any particular \( R_d \) value. From statistical analysis (Fant et al., 1994; Fant, 1995) of data mainly from Gobl (1988) we have found

\[
R_{ap} = (-1+4.8R_d)/100
\]

(6)

\[
R_{kp} = (22.4+11.8 R_d)/100
\]

(7)

\( R_{gp} \) is obtained from Eq. 6 and 7 inserted into Eq 4. These default values are summarised below.

Table 2. Default LF-parameters for a range of \( R_d \) values. \( F_{ap}=F_0/(2\pi R_{ap}) \) refers to \( F_0=100 \) Hz.

<table>
<thead>
<tr>
<th>( R_d )</th>
<th>( R_{ap} )</th>
<th>( F_{ap} )</th>
<th>( R_{gp} )</th>
<th>( R_{kp} )</th>
<th>( OQ )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.44</td>
<td>3600</td>
<td>26</td>
<td>179</td>
<td>35</td>
</tr>
<tr>
<td>0.5</td>
<td>1.40</td>
<td>1590</td>
<td>28.3</td>
<td>137</td>
<td>47</td>
</tr>
<tr>
<td>0.7</td>
<td>2.36</td>
<td>674</td>
<td>30.7</td>
<td>118</td>
<td>55.5</td>
</tr>
<tr>
<td>1.0</td>
<td>3.8</td>
<td>420</td>
<td>34.2</td>
<td>103</td>
<td>65</td>
</tr>
<tr>
<td>1.4</td>
<td>5.7</td>
<td>280</td>
<td>39.0</td>
<td>95</td>
<td>73</td>
</tr>
<tr>
<td>2.0</td>
<td>8.6</td>
<td>185</td>
<td>46.0</td>
<td>93.5</td>
<td>78</td>
</tr>
<tr>
<td>2.7</td>
<td>12.0</td>
<td>133</td>
<td>54.3</td>
<td>98.0</td>
<td>79</td>
</tr>
</tbody>
</table>

The predictability is often remarkably good (Fant et al., 1994; Fant, 1995). An appreciable part of the variance inherent in a complete LF-representation of a number of voice samples, e.g. from different speakers or within an utterance, is reduced by the \( R_d\)-transform Eq. 4, which capitalises on constraints in parameter covariation. In practice, we often need to recover the complete LF-representation which is accomplished by introducing the ratios of true and predicted parameters. These parameter coefficients are

\[
k_a = (R_a/R_{ap})=(F_{ap}/F_a)
\]

(8a)

\[
k_g = (R_g/R_{gp})
\]

(8b)

\[
k_k = (R_k/R_{kp})
\]

(8c)

The value of \( k_k \) is fully predictable from \( [R_g, k_{ap}, k_{gp}] \) which constitutes a wave shape vector. The parameter coefficients Eq. 8 become useful for specifying contextual deviations from default rules or speaker specific glottal wave shapes - a kind of principle component analysis.

A breathy voice is characterised by a high \( k_a \) and a pressed voice a high \( k_k \). Females usually have higher \( R_d \) and \( k_a \) than men which implies lower \( F_{ap} \). Voiced consonants usually have higher \( R_d \) and higher \( k_k \) and lower \( k_g \) than vowels. These data from Fant (1995, 1996) are in general agreement with those reported in (Gobl, 1988; Carlson et al., 1989; Karlsson, 1995).

A sequence of LF source spectra covering the main range of \( R_d \)-parameters is shown in Fig. 2.

The four \( R_d \) values, \( R_d=0.5, 0.7, 1.4 \) and 2.7, are associated with increasing open quotients \( OQ=0.35, 0.55, 0.73 \) and 0.79. They represent base forms of (1) an highly adducted, somewhat pressed male voice, (2) normal male voice, (3) normal female voice, and (4) breathy female phonation or breathy termination of voicing before pause irrespective of voice type. The \( k_a \) factor adds to the breathyness and the \( k_g \) factor to the degree of press.

Voice source rules

The overall strategy for dealing with the influence of individual and contextual variations in connected speech is to submit the voice excitation parameter \( E_c \) and the shape vector \( [R_d, k_{ap}, k_{gp}] \) to a number of operational rules. These are conveniently handled on a decibel scale, which implies a summation of contributions from several factors and levels of analysis. Thus, the value of the source parameters at any instant of time is determined by
an interpolation of data from adjacent target values specified at specific time locations within the utterance. These data are influenced by phonetic segment identities through the degree of supraglottal constriction and the state of the glottal opening, speaker specific constants and habits, degree of emphasis, accentuation and stress, location within the utterance, specially with respect to its boundaries.

The rules take into account systematic covariances of $E_e$ and $R_d$ across segments and within a phrase contour and the extent to which $E_e$ and $R_d$ vary with $F_0$. These $F_0$ dependencies are of basic importance. They enable a prediction of default values of $E_e(F_0)$ valid for non-close vowels uttered at constant subglottal pressure, which determines the basic $E_e(t)$ contour of an utterance prior to the application of contributions from all other factors mentioned above. As a result there is a general tendency of a positive correlation between $E_e$ and $F_0$ which is further enhanced by the simultaneous $F_0$ and $E_e$ dips of voiced obstruents. The envelopes of $E_e(t)$ and $F_0(t)$ show similar contours.

In a first approximation $R_d$ is independent of the basic $F_0$-contour. The covariation of $R_d$ and $E_e$ follows variational rules that supplement the segment specific data, specially in transitions between extreme values. A common rule is that 1 dB increase in $E_e$ is associated with 0.5 dB decrease of $R_d$. According to Eq. 5, at constant $F_0$, this implies that $U_0$ varies with the square root of $E_e$. At a hard voiced onset $U_0$ and $E_e$ vary proportionally and $R_d$ is constant. In the abduced termination of voicing before a pause or before an unvoiced consonant we find a tendency of constant $U_0$, i.e. a 1 dB decrease of $E_e$ is accompanied by 1 dB increase of $R_d$. On a variational basis, we may thus define a relation

$$\Delta \log(1/R_d) = K_s \Delta \log E_e$$

where the constant $K_s$ has a normal range of variation from 0 to 1. In a transition towards a more efficient abduced phonation, an $E_e$ increase may be attained at a lower $U_0$ and the constant $K_s$ attains negative values.

**Rule sequence**

1. Define the domain and boundaries of a phrase or a breathgroup. Derive accentuations, relative stress levels and segment specific data.

Fig. 2. LF source spectra at $F_0=100$ Hz and four representative $R_d$ values.
2. Select speaker specific data. These include a reference midfrequency F0r, of the order of 100-150 Hz for a male voice and 150-300 Hz for a female voice, and in addition a specification of the modulation range by an upper and a lower limiting F0. Select a source shape vector [Rd, ka, kg] and also formant frequency scale factors (Fant, 1975) appropriate for the sex of the speaker and the length of the vocal tract.

3. From data in (1) and (2) construct a first approximation to the basic F0 contour within the utterance. Remap the default normalised Ec(F0) with respect to the speakers particular F0r and calculate the corresponding basic Ecb(t)-contour.

4. Add onset and offset Ec contours and deviations from the overall Ec declination not predictable from F0 alone, i.e. due to the decline of subglottal pressure, see Fig. 7.

5. Introduce tabulated segment specific reference values of data on Ec and [Rd, ka, kg]. Make note of the time locations within each phonetic segment where these reference data apply.

6. Add modifications with respect to the overall voice intensity level and to local variations of emphasis, accentuation and stress.

7. Interpolate Ec and the source shape vector [Rd, ka, kg] within the utterance, frame by frame or pitch synchronously. Take care to preserve coarticulations within and across segments. Observe the particular locations of the targets which may not coincide with segment boundaries or segment centers. The abduction gesture of an aspirated unvoiced stop may start already at the onset of the preceding stressed vowel.

8. Add aspiration noise and formant band-width broadening as a function of Rd and kg and the particular phonetic segment. Add covariation of the vocal tract transfer function if needed.

Supporting data

The voice source rules above constitute a framework that suggests a principal procedure. Some of the detailed contents will now be discussed.

Speaker specific data

A typical Rd vector for male vowels (Fant, 1995) is [Rd=0.7, ka=1, kg=1] which implies Ra=2.4%, Fd=680 Hz at F0=100 Hz, Rk=31% and Rg=118%. According to Eq. 8, H1*-H2*/= 0.2 dB. The open quotient OQ without Rd is 0.555 and with Rd included 0.58. The corresponding spectrum is shown to the left in Fig. 3.

The right hand figure pertains to a somewhat more sonorous voice of the same Rd=0.7 but with Fd=2700 Hz, Rg =177% and Rk=47% which implies ka=0.25, kg=1.5 and Rk=1.5. The greater relative dominance of the second harmonic, H1-H2=-3.4 dB is predictable from the smaller OQ==O.415 inserted into Eq. 7. Perceptual tests indicate that the main quality difference lies in the Fd-domain, i.e. in the less spectral tilt in the upper diagram. Here the overall intensity is also 2 dB higher than in the lower diagram. This is an example of a greater efficiency at the same Ec, U0 and F0, often found in a stressed vowel situated late in a phrase.

![Fig. 3. Synthesised spectra of vowel [a:]. both with Rd=0.7, F0=100 Hz. Left: default values, k_a=1, k_g=1, OQ=55.5%. Right: k_a =0.25, k_g=1.5, OQ=41.5%.](image-url)
RD values of the order of 0.5-1.5 are found for male voices, which overlaps the distribution for female voices with RD in the range of 0.8-2.5. Increasing RD implies increase of RK and RA, F0 decreasing and RG on the whole decreasing.

Female voices usually have larger ka and thus lower FA than men (Karlsson, 1992; Karlsson & Neovius, 1994; Karlsson, 1995; Fant, 1995). This is especially true of breathy, soft female voices, which also show a substantial glottal leakage and aspiration noise (Fant & Lin, 1988; Klatt & Klatt, 1990; Fant et al., 1991; Karlsson, 1992; Karlsson & Neovius, 1994; Stevens & Hanson, 1994; Fant, 1995; Hansson, 1995; Karlsson, 1995).

A voice attribute outside the domain of the LF-model found in some voices is the appearance of subglottal formants introduced by pole-zero patterns (Fant & Lin, 1988; Karlsson & Neovius, 1994). An other is the occurrence of an extra excitation from a discontinuity at the instant of glottal opening which adds somewhat to the spectral shape and to irregularities of waveforms. It has been shown (Båvegård & Fant, 1994) that perturbations induced by a complete articulatory synthesis add to the perceived quality.

**Segmental data and emphasis/deemphasis**

The left hand part of Fig. 4 from Fant (1995) shows log EE in dB, referred to as EE, versus log(1/Rd) in dB, referred to as EUF, for a corpus of Swedish vowel and consonants originating from a Swedish subject. The overall span from open vowels to stop consonant is of the order of 10 dB in 1/Rd and 20 dB in EE. This is an example of a more general rule of covariation of 1 dB in 1/Rd with 2 dB in EE, i.e. KS=0.5 which we have found to be typical of dynamic variations within and across segments and as a consequence of varying voice effort. The right hand part of Fig. 4 contains a limited set of vowels from a French speaker (Karlsson, 1995). The trend is similar but for a linear relation of EE and 1/Rd.

In a vowel, increased voice effort or emphasis will cause increased EE and lowered RD and KG, and often a higher F0 which contribute to an increase of FA and thus a high frequency emphasis (Fant, 1959; Fant & Kruckenberg, 1994; Fant, 1995; Sluijter, 1995), often combined with increased KG and thus decreased open quotient (Fant, 1995; 1996). A common trend is that RD and EE covary with a KS=0.5. However, a high narrow vowel might reach a more extreme target which may counteract this trend. Lexical stress without an F0 mediated accentuation will have a rather small effect on the source parameters.

Increased emphasis adds to segmental contrasts not only in terms of formant patterns but also in terms of properties of the voice source (Fant, 1993). This is the consequence of more extreme targets, open vowels become more open whereas consonants become more constricted. Voiced consonants attain a higher RD. The narrowing of the supraglottal constriction in a voiced consonant will cause a reduction of EE as well as 1/RD which counteracts an increase due
to raised subglottal pressure. In addition, there enters an intensity loss due to formant frequency shifts.

The underlying mechanism of the source tract interaction is a loss of transglottal pressure due to the increase of supraglottal flow resistance (Bickley & Stevens, 1986; Strik & Boves, 1992; Fant, 1995). Similar reductions of 1/R_d and E_c occur at boundaries towards termination of voicing before a pause or before an unvoiced consonant. A common denominator is a lowered subglottal pressure and abducted vocal folds (Strik & Boves, 1992). The local drop in P_{sub} in a stressed vowel preceding an aspirated stop was demonstrated in Fig. 2 of Fant et al. (1996).

As seen in Fig. 4, voiced consonants have higher R_d and lower E_c values than vowels. This is also true of narrowly articulated vowels compared to open vowels. In Swedish, emphasis of a close long vowel will cause a more extreme target, e.g. a [j] element within an /i:/ segment, see Fig. 8.

The loss of vowel/consonant contrast associated with deemphasis occurs more frequently in connected speech than would be implied by the phonological structure of an utterance. This is also an individual speaker attribute. Incomplete oral closure causes nasal consonants to be realised as nasalised vowels and stops as approximants. These phenomena are conveniently handled in articulatory synthesis but are more difficult to structure in formant synthesis.

Continuities and coarticulation

Glottal parameters are not constant within a phonetic segment. They follow general gesture patterns of continuity and coarticulation (Gobl, 1988; Gobl & Ní Chasaide, 1988; Ní Chasaide et al., 1994). A typical example is that the onset of voicing usually has a smaller time constant than the offset towards a following unvoiced segment. This is especially the case of pre-occlusive aspiration (in the transition from a stressed vowel to an unvoiced aspirated stop) which not only produces noise at the offset of the vowel and at the initiation of the occlusion, but also imposes the typical abduction correlates of increased R_d and k_d, increased bandwidths mainly affecting the first formant, and traces of subglottal formants (Fant & Lin, 1988; Fant et al., 1991). These features are fully developed at the boundary, but may be detectable already at the onset of the vowel (Gobl & Ní Chasaide, 1988).

Calculations (Fant, 1995, 1996) of the effects of glottal leakage, see Fig. 5, show that a glottal area as small as a few mm² causes a substantial increase of the formant bandwidths, mainly of B1. The following empirical formula for bandwidth increments have been suggested (Fant, 1995).

\[ \Delta B_1 = 250(F_1/500)^2 \frac{R_d}{12} \]  
\[ \Delta B_2 = (\Delta B_1)(F_1/F_2)/2 \]

The weakening of F1 from a substantial bandwidth increase is a consistent spectral attribute of abduction which explains the “F1-cut back” in the transitional phase from the release of an unvoiced consonant to a following open vowel and it also occurs regularly in voice termination before a pause.

Subglottal pressure, E_c and F0

Preliminary results from a study of subglottal pressure, P_S, in speech, measured directly from a tracheal probe, was reported in (Fant, 1996; Fant et al., 1996). More recently we have gained some further insights in the covariation of subglottal pressure and speech wave parameters including inverse filtering. Data were obtained from glissando phonations, single vowel utterances and connected speech.

Covariation of E_c and P_S in glissando phonations are shown in Fig. 6. They support the earlier reported tendency (Fant, 1982; Fant & Kruckenberg, 1994; Fant et al., 1994) of E_c rising with the second power of F0 up to the speakers midfrequency, F0=130 Hz for this particular male subject, at F0>F0, followed by a decrease or increase depending on P_S. An
Fig. 6. Covariation of $E_e$ and subglottal pressure $P_s$ with $F_0$ in glissando phonations of a vowel [ae]. Data from a continuation of Fant et al. (1996).

Fig. 7. Stylised outline of the $P_s$ dependent component of the $E_e$ contour within a phrase (breathgroup).

Fig. 8. Recordings of subglottal and supraglottal pressures added to our standard analysis display. The text is “Ingrid fick brev från Arne”. “Ingrid” and “brev” have accent 1. “Arne” has accent 2. At the bottom SPL with LP 1000 Hz and SPLH, high frequency preemphasis.
analysis of the covarying subglottal pressure revealed a rise of $P_s$ proportional to $F_0^{0.7}$ below $F_0$, and above $F_0$ a considerable variation. $P_s$ both increasing and decreasing.

The glottal maximum area as studied by means of fiberscope filming showed a maximum just below $F_0$, with an increase in proportion to $P_s$ at $F_0<F_0^r$ and a decrease in proportion to $F_0^r$ at $F_0>F_0^r$. The total increase of $E_e$ from $F_0=75$ Hz to $F_0=130$ Hz was of the order of 10 dB. In this region we found $E_e$ to be proportional to $F_0^{1.35}$ and to $P_s^{1.2}$. These relations also provided a good prediction for single vowel utterances and could be used as a guide for analysis of connected speech.

Other main findings from these aerodynamic studies are that the voice onset after a pause requires a minimum $P_s$ of about 3.5 cm H2O while vocal cord vibrations can continue down to very low transglottal pressures at an abducted voice offset. Voiced consonants as well as semiclosure targets in Swedish long stressed vowels show up as a significant local rise in supraglottal pressure which reduces the transglottal pressure.

The overall declination of subglottal pressure within a phrase is of the order of 2 dB/sec. With $E_e$ approximately proportional to $P_s$ at constant $F_0$ we would therefore expect the same rate of declination of $E_e$ within the phrase, see Fig. 7. It includes a 50 ms risetime and a faster than average declination in the last 0.5 seconds. This is the correction to the general $E_e(F_0)$ to be added to the rule (4) for calculating the initial $E_e(t)$ contour of the phrase. A typical example from a subglottal pressure recording is shown in Fig. 8.

**Accentuation and stress**

A general observation of interest is how $P_s$ varies with stresses and accentuations within an utterance. We found a build up of $P_s$ early in the onset of a focally accented syllable, i.e. in the

Fig. 9. Illustration of focal accent 1, female subject, AK [15,18]. Observe how the intensity curves, SPL and SPLH reflect the shape of an $E_e(F_0)$ with maximum at $F_0=F_0^r=215$ Hz, and an intensity minimum at the peak $F_0=320$ Hz.
region of an anticipated syllable P-center, followed by a decrease at a higher rate than the overall $P_g$ declination. This tendency was also found in non-focal accentuation. It explains the tendency reported in (Fant & Kruckenberg, 1994, 1995) that $E_c$ does not follow the F0-contour in the focal F0 peak area but stays constant or shows a minimum at the maximal F0, and maxima whenever F0 passes through the F0$_r$ value.

We may now compare our earlier findings (Fant & Kruckenberg, 1995) in Fig. 9, with data from the recent aerodynamic study (Fant et al., 1996) in Fig. 10. The sentence is the same. Fig. 9 pertains to a female subject, AK, where we estimated an F0$_r$ of 215 Hz. For the male subject in Fig. 10, we estimated an F0$_r$ of 130 Hz. Both showed a pronounced intensity minimum in the F0-peak area of the [α] in

![Waveform and pressure plots](image)

*Fig. 10. The same sentence as in Fig. 9 produced by the male subject, SH (Fant et al., 1996). Observe the dip in subglottal pressure synchronized with the peak F0, and similarities with Fig. 9.*
“Lenar”, which for the male voice was initiated by a local drop in subglottal pressure followed by a pressure restoration in the final word “igen”, marking a prosodic boundary.

Another noteworthy observation comparing Fig. 9 and 10 is that the semitone F0-scale preserves an almost identical shape and size of the main F0 peak of the female and male utterances.

The main acoustic correlate of focal accentuation in Swedish is the local dominance of the F0 peak usually combined with a significant increase of intensity, high frequency emphasis and increased duration of the stressed word. However, individual variations are large. The difference between focal and nonfocal stress may be signaled by F0 alone On the whole, duration and F0 appear to be the main stress correlates. Excluding focal accentuation the average increase of F0 at and F1 are of the order of 2 dB only (Fant & Kruckenberg, 1994, 1995). In emphatic focal accentuation we encounter 3-5 times larger values. Lexical unaccented stress may be realised by duration alone.

**Conclusions**

In order to develop a more profound insight in the theory of speech production as a guide for future developments of speech technology we need to learn more about the voice source and its role of an integrator of segmental and prosodic structure. We have suggested a framework for the further developments of voice source rules and exemplified their contents This modelling is best understood with reference to aerodynamic constraints and the continuous gestures of pulmonary activity and glottal and supraglottal articulations.

The rules have not yet been implemented in a text-to-speech system. Also, it remains to test the relative perceptual salience of the various components of our system. Some details may not be very significant or perceptually masked by deficiencies in other parts of the rule system. We need a deeper insight in the dependencies of individual voice qualities on source functions and segmental and prosodic structures, specially with respect to temporal dynamics. A mere change in average values of glottal parameters is not sufficient.

**Acknowledgements**

This work has been financed by grants from TFR. The Swedish Research Council for Engineering Sciences, the Bank of Sweden Tercentenary Foundation, the Carl Trygger Foundation and support from the Telia Promotor AB. The authors are indebted to Johan Liljencrans for valuable support in data processing.

**References**


