Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Frequency domain interpretation and derivation of glottal flow parameters

Fant, G. and Lin, Q.

journal: STL-QPSR
volume: 29
number: 2-3
year: 1988
pages: 001-021

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

I. SPEECH PRODUCTION

A. FREQUENCY DOMAIN INTERPRETATION AND DERIVATION OF GLOTTAL FLOW PARAMETERS

Gunnar Fant & Qiguang Lin

Abstract

Glottal flow parameters are generally defined as time-domain entities that specify the shape of glottal pulses or their derivatives. The present study is concerned with the relations of glottal parameters to frequency-domain properties in order to bring out perceptually important aspects. The analysis also aims at techniques to extract frequency- as well as time-domain parameters from frequency-domain representations. This involves frequency-domain inverse filtering, analytical transformations, and analysis-by-synthesis procedures. Frequency-domain processing is recommended as a complement to or a substitute to conventional time-domain analysis. An advantage is the less severe demands on low-frequency recording fidelity. Moreover, already available narrow-band spectral sections may be processed in order to derive major voice source parameters. The frequency-domain matching ensures optimal conditions for Hi-Fi resynthesis. The theoretical analysis also sheds light on time-domain processing techniques suitable to support frequency-domain processing, e.g., selective inverse filtering. The frequency-domain analysis includes studies of covarying formant bandwidths and subglottal coupling effects which become especially apparent in breathy voicing.

Introduction

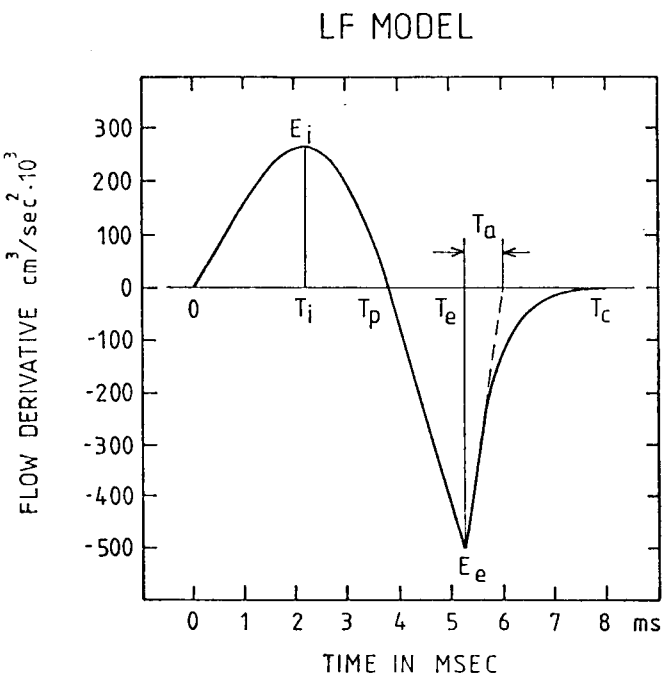
The demands on high-quality synthesis have provoked a renewed interest in voice source analysis and modelling. The model proposed by Fant (1979) and the later LF-model (Fant, Liljencrants, & Lin, 1985) have been exploited by several research groups. At KTH, the LF-model has recently been applied to studies of female speech (Karlsson, 1988) and to studies of the temporal variation of voice source parameters in connected speech Gobl, 1988).

In the course of our work, we have found it necessary to pay a close attention to the complete source-filter analysis and to the final demands of a maximally accurate resynthesis. It is ultimately the combination of source and filter function which is decisive. Since the process of separating source and filter functions is seldom unambiguous, there often remains an uncertainty of how to treat the trading relations; what shall be attributed to the source and what shall be attributed to the filter function (Gobl, 1988). The choice of source parameters thus imposes important constraints on the realization of a corresponding vocal tract filter function and vice versa.

Frequency-domain matching of original and synthesis ensures a proper base for correcting for the difference between the inverse filtering transfer function and a specific synthesizer transfer function. A time-domain derivation of glottal parameters does not necessarily ensure a proper

frequency-domain fit. To pay attention to the frequency-domain consequences of a certain time-domain decision, e.g., of the important return phase parameter T_a, is a recommended technique in our present routines. A final check of the spectral match is also needed. A major motivation for in part incorporating a frequency-domain processing in the voice source analysis is to optimize the choice of T_a. There is also an apparent interest in studying how far one can get with frequency-domain processing alone.

The LF-model



$$E(t) = E_0 e^{\alpha t} \sin \omega_g t \quad (t < T_e)$$

$$E(t) = \frac{-E_e}{\varepsilon T_a} \cdot \left[e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)} \right] \quad (T_e < t < T_c)$$

$$\omega_g = 2\pi F_g \qquad F_g = \frac{1}{2T_p} \qquad T_c = TO = \frac{1}{F0}$$

$$R_g = \frac{F_g}{F0} \qquad R_k = \frac{T_e}{T_p} - 1 \qquad R_a = \frac{T_a}{TO}$$

$$O_q = \frac{T_e + T_a}{TO} \qquad O_q' = \frac{T_e}{TO} \qquad F_a = \frac{1}{2\pi T_a}$$

Fig. 1. The LF-model of differentiated glottal flow.

Any linear source-filter model of speech production is an approximation only. Even though the vocal tract transfer from glottal flow to radiated pressure in all essentials is a linear process, there remains a highly nonlinear transformation from a time-varying glottal area function $A_g(t)$ to a corresponding glottal flow $U_g(t)$. The origin of this nonlinearity is the flow-dependent glottal impedance, the instantaneous value of which is determined by transglottal pressure fluctuations within a pitch period including formant frequency oscilla-tions. Detailed studies of this acoustic interaction are presented in Ananthapadmanabha & Fant (1982), Fant & Ananthapadmanabha (1982), Fant & Lin (1987), and Fant, Lin, & Gobl (1985). The spectral consequences of the interaction, pulse skewing, truncations, and pulse ripple appear as increased formant excitation levels and bandwidths and local spectral distortions such as multiple zeros which cause spectral dips and the tendency of spectral energy being dispersed towards the high-frequency side of a formant peak region. However, perceptual tests performed by Nord, Ananthapadmanabha, & Fant (1984) and, more recent, informal tests indicate that the acoustic interaction whilst adding somewhat to naturalness is not a decisive quality factor.

A linear model should apparently be capable of generating representative overall glottal flow pulse shapes, ignoring ripple effects. The associated filter functions should be tailored towards representative effective bandwidths and formant frequencies that ensure a best overall match to natural speech.

The LF-model (Fant, Liljencrants, & Lin, 1985), see Fig. 1, is defined by the following glottal flow derivative wave shape:

$$E(t)=E_0 e^{\alpha t}\sin\omega_g t \qquad (t<T_e)$$

$$E(t) = -(E_e/\varepsilon T_a)\cdot[e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)}] \qquad (T_e<t<T_c) \qquad (1)$$

It differs from the previous Fant (1979) model in three respects. First, the object of the LF-parameterization is to specify the glottal flow derivative whilst the flow has to be deduced from the integral of the LF time function. Secondly, and this is the main difference, there is included in the LF-model a return phase in the form of an exponential starting at the negative flow derivative peak at time T_e and connected to the onset of the next pulse at time T_o, which is standard practise, or to a fixed time T_c within the "closed phase". The parameter εT_a is a unique function of the other parameters. For small T_a, εT_a is close to 1. The effective duration of the return time T_a is defined by the projection on the time axis of its derivative at time T_e. It can be shown that the essential frequency-domain correspondence of the return phase is a low-pass filter of the first order with cutoff frequency $F_a=1/(2\pi T_a)$, see Fig. 2. This is the main parameter for change of spectral slope. A third feature is that the main body of the LF-source function at $t<T_e$ is represented by a continuous sinusoid with a positive growth factor α, i.e., negative damping. There is accordingly no discontinuity at the flow peak as in the F-model (Fant, 1979) and the spectral slope is more continuous than in the F-model.

The LF-model is inherently a five-parameter model. There ex-ists a large variety of different sets of five parameters that u-niquely define the function. In addition to the direct synthesis parameters E_o, α, ω_g, T_a, and F_o, one can refer to the negative peak E_e and four critical time locations, T_p of max-imal flow, T_e of maximal discontinuity in the flow derivative, the return time constant T_a, and the total period length T_o.

These are the set of parameters that usually evolve from inverse filtering. They directly relate to a set of normalized parameters, as indicated in Fig. 1. Thus, the glottal flow rise time T_p is converted to a "glottal frequency" $F_g=1/2T_p$ which in turn may be related to F_o as the quotient $R_g=F_g/F_o$. R_g is of the order of 1 and usually varies between 0.7 and 1.6.

A steepness factor is defined by how close T_e comes to T_p. This is expressed by $R_k=(T_e/T_p)-1$. A dependent parameter is the open quotient which we may define either as $Q_o=(T_e+T_a)/T_o$ or as $Q_o'=T_e/T_o$. In addition we can normalize T_a by the parameter $R_a=T_a/T_o$.

Much of our data collection (Gobl, 1988) has been concerned with the set of parameters E_e, R_k, R_g, R_a, and F_o. However, for developing source rules we are now looking into an alternative system that is closer related to visual aspects of the glottal flow and which has a closer relation to

general constraints of production and a closer tie to perceptual dimensions. Instead of R_k, we would thus prefer to refer to one of the alternatives E_e/E_i or to U_o/E_e or to its inverse E_e/U_o, where U_o is the peak glottal flow. Instead of R_a, we find $F_a=1/2\pi T_a$ to better suite frequency-domain matching aspects.

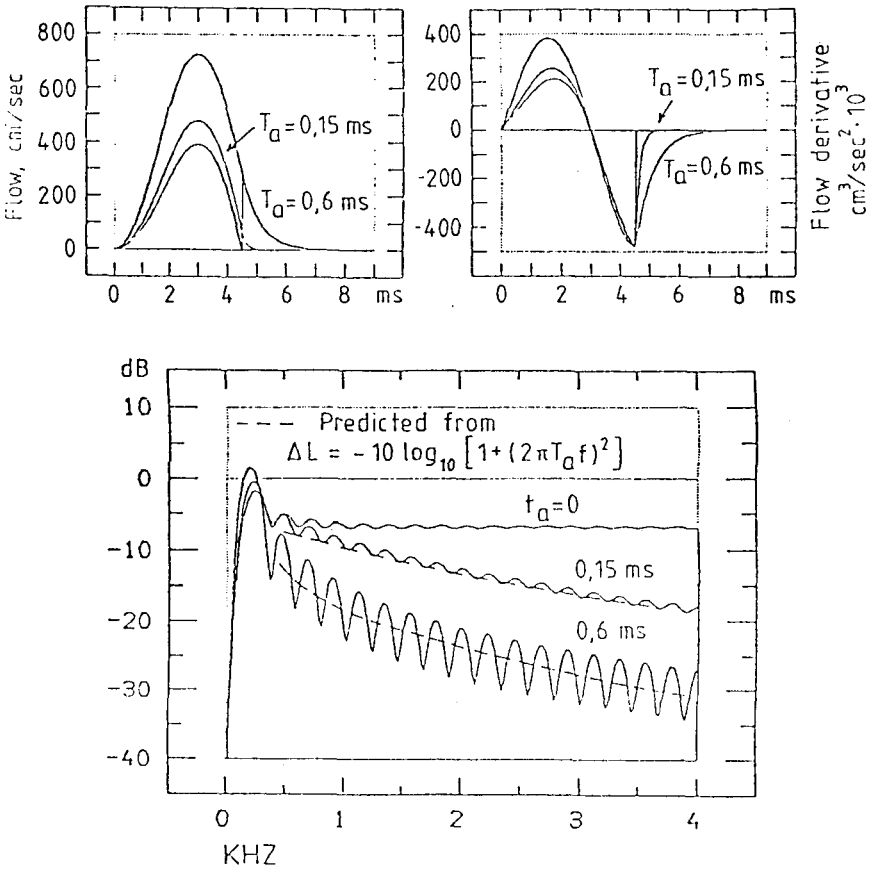


Fig. 2. Wave shapes and spectral changes associated with an increase of the return time parameter T_a.

The spectral consequences of a systematic variation of LF-parameters are brought out in Fig. 3 in which F_g is held constant and in Fig. 4 which illustrates covariation of F_g and R_k whilst both U_o and E_e are constants. In both figures T_a=0.

Fig. 3 illustrates the constancy of the spectral slope, -6 dB/oct, of the differentiated flow spectrum whilst the spectrum level increases with E_e/E_i (or inversely with the underlying R_k). After a second differentiation we are in a better position to discuss the balance between a low frequency level in the vicinity of F_g=125 Hz and the spectrum level at higher frequencies. At constant U_o, the spectrum level above $2{\cdot}F_g$ increases about 10 dB when E_e/E_i varies from 1.5 to 4 whilst the level at F_g is constant.

It follows that at constant E_e and varying glottal flow peak U_o, the spectral level varies over the same 10 dB range in the F_g area and is constant at higher frequencies. Thus, U_o determines the low-frequency level and E_e the high-frequency spectral level. With R_g=1, the level of the source fundamental equals that of upper source harmonics providing E_e/E_i is close to 3 or more precisely, $E_e/E_i=\pi$, as will be derived in Eq. (14). The second harmonic here is a few dB above the fundamental. If we choose a lower F_o than F_g, i.e., a greater R_g, the second harmonic would be even more dominating over the fundamental. This is typical of a low-frequency pressed voice.

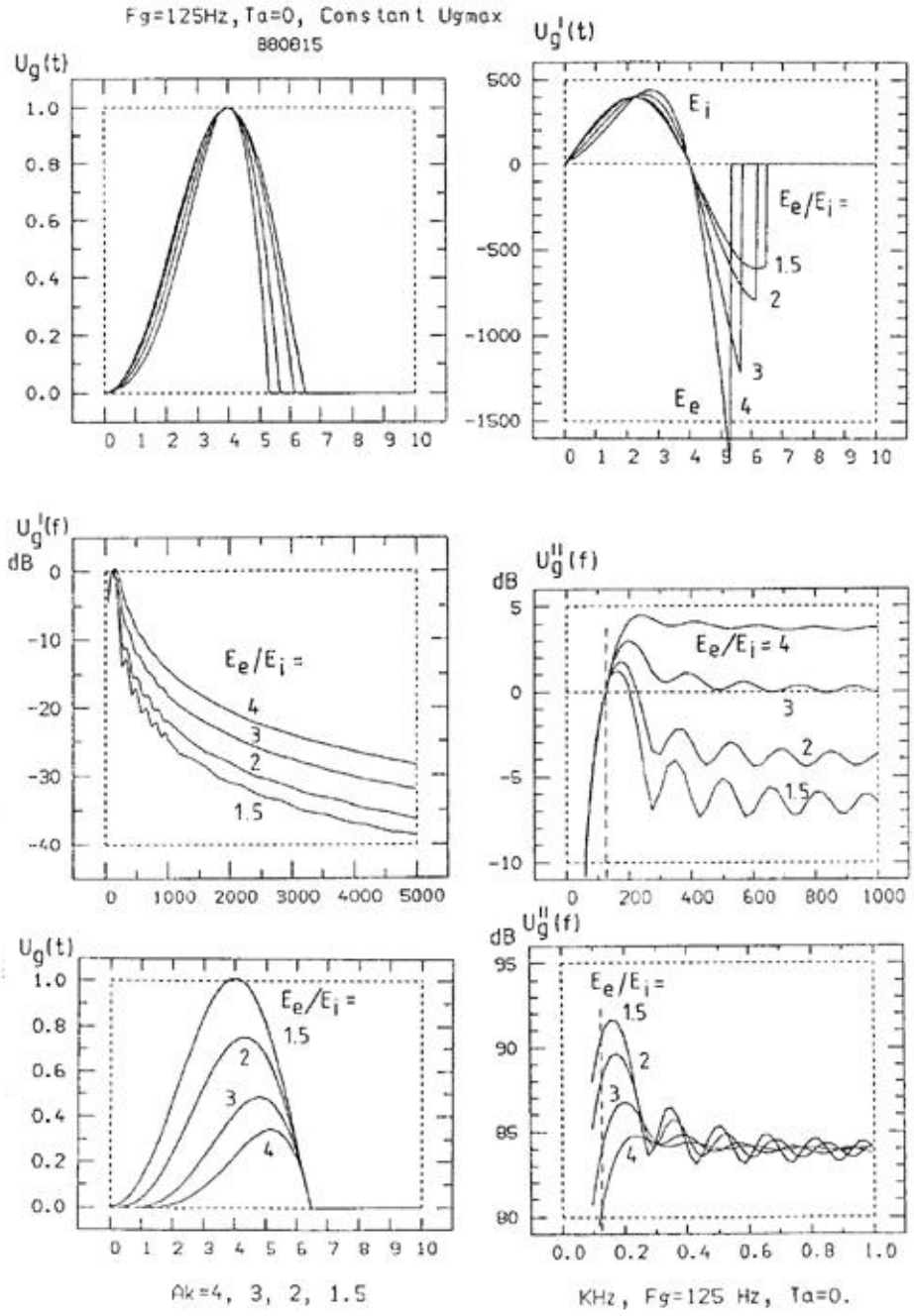


Fig. 3. *In the top two rows: LF-flow, flow derivative, flow derivative spectrum, and the second derivative spectrum at varying E_e/E_i and constant U_o. In the bottom row: LF-flow and its second derivative spectrum when maintaining constant E_e.*

The spectral consequences of these variations can be conceived of as a very selective reinforcement/reduction of low-frequency energy. The perceptual effects are not dramatic.

The specific variations brought out in Fig. 4 are even less apparent. Here, since both U_0 and E_e are fixed, the remaining variation in F_g provides a shift in the frequency range of the second and third harmonic only. In an [a]-vowel context, the effect is barely audible. A much more dramatic effect is the change of T_a as in Fig. 2. In a breathy voice, e.g., in a voiced [h] or in a vowel assimilating glottal abduction, we have measured F_a as low as 100 Hz, which represents a severe low-pass filtering. For average non-breathy female vowels, F_a is in the range of 500-1500 Hz and for males 1000-3000 Hz.

Time frequency-domain relations

We shall now take a more general view of time frequency-domain relations in voice production. The first step will be to quantify a more rigid relation between peak glottal flow U_0 and the amplitude of the voice fundamental A_0 in a harmonic representation and then proceed to the relation of the excitation parameter E_e to various aspects of formant amplitudes and source harmonics. Consider a source-filter radiation representation

$$P(s) = G(s) \cdot H(s) \cdot R(s) \tag{2}$$

with absolute values

$$|P(\omega)| = |G(\omega)| \cdot |H(\omega)| \cdot |R(\omega)| \tag{3}$$

the radiation transfer is

$$|R(\omega)| = (\rho\omega/4\pi a) \cdot k_T(\omega) \tag{4}$$

see Fant (1979). Here ρ is density of air $1{\cdot}14\ 10^{-3}$ g/cm3 and a is the lip-microphone distance in cm. The correction factor $k_T(\omega)$ represents the combined baffle effect of the head and the increase of radiation resistance in excess of ω^2. It is usually neglected in production theory but accounts for a total high-frequency emphasis of about 5 dB from 300 to 4000 Hz. It could be included in a very detailed modelling (Fant, 1979). In the following we shall set $k_T(\omega)=1$.

A basic step is to convert a glottal flow pulse train into a Fourier series. Given the flow peak amplitude U_0 and an open quotient, Q_0, close to 0.5, it can be shown that the flow source fundamental comes close to $U_0/2$. This is exactly so for two conditions. One is to model the glottal flow as a rectified sine wave, omitting the negative parts. The other is to model the glottal flow as a continuous raised sinusoidal wave with the "closed phase" undefined. With an open quotient of 0.4, the half sine wave model produces a 1 dB lower fundamental than $U_0/2$. For an open quotient of 0.3, the correction is -3 dB. On the other side we have a correction of approximately +0.6 dB for open quotients of the order of 0.6-0.8. Denoting the correction factor k and assuming a filter function H(s)=1, i.e., eliminated by inverse filtering or being negligible in the F_0 domain. Eqs. (3) and (4) combined predict a fundamental amplitude of

$$A_0 = U_0 \cdot k \cdot \pi \cdot F_0 \cdot (\rho/4\pi a) \tag{5}$$

in the radiated wave at a distance of a cm.

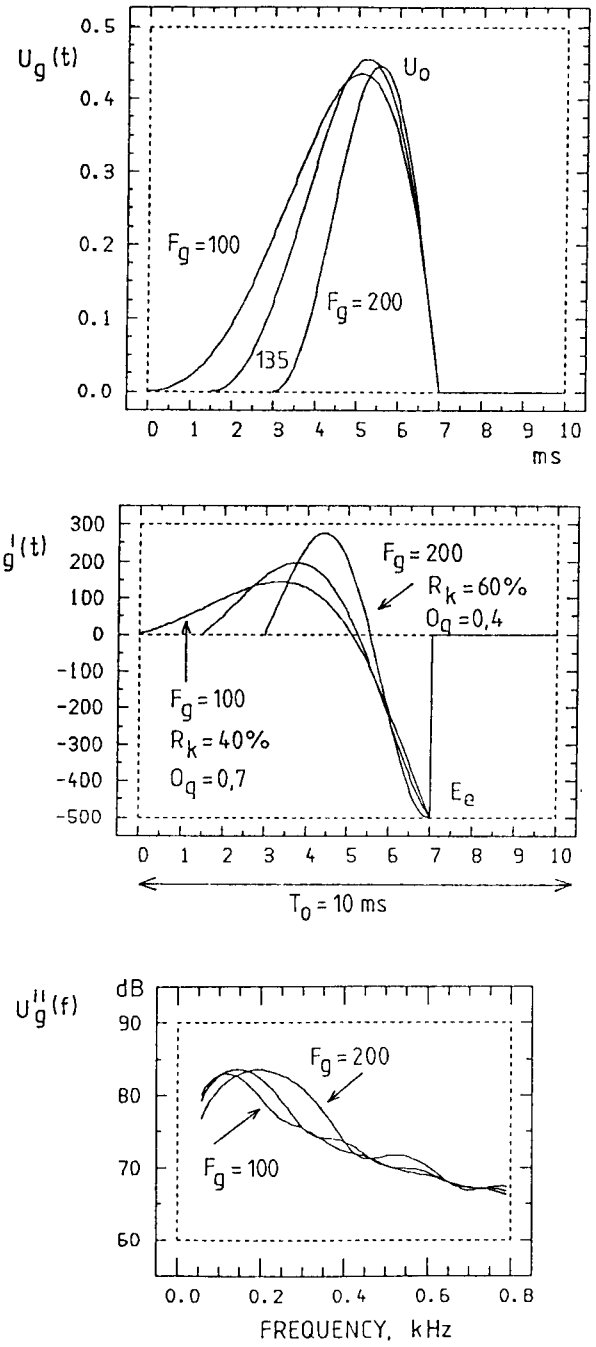


Fig. 4. *Flow, flow derivative, and second derivative spectrum when maintaining both U_o and E_e constant and varying R_g (F_g) and dependently R_k.*

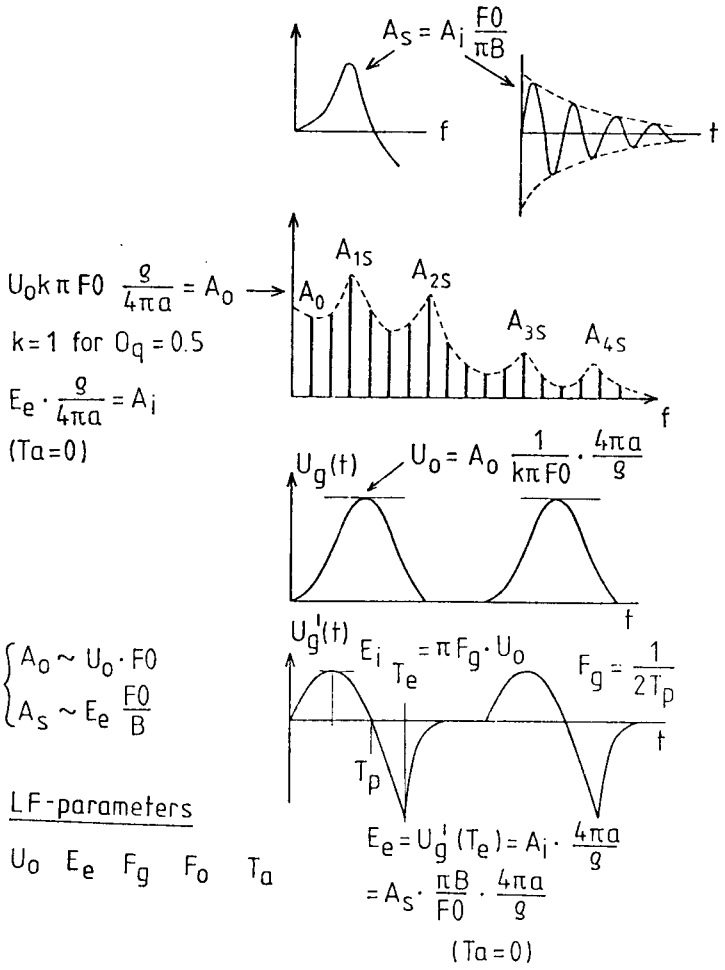


Fig. 5. *Some basic analytical relations between glottal flow wave shape and formant excitation and spectrum.*

For approximate calculations we may usually set k=1. As also indicated in Fig. 5, the voice fundamental is, apart from the influence of the transfer function, proportional to the peak glottal flow and to the voice fundamental frequency.

The derivation of the relation of formant amplitude measures to E_e is more tricky. Assuming T_a=0, i.e., the excitation E_e being a step function in glottal flow derivative or a ramp function in glottal flow, we end up with the following relation between E_e and the initial amplitude A_i of a damped oscillation from a single resonance vocal tract load.

$$A_i = E_e \cdot (\rho/4\pi a) \tag{6}$$

and

$$A_i/A_o = (E_e/U_o) \cdot (1/k\pi F_o) \tag{7}$$

In Eq. (6), the frequency factor of the radiation transfer, Eq. (4), has been traded for the $1/\omega$ factor in the glottal flow derivative spectrum. As illustrated in Fig. 5, A_i is a time-domain entity which needs to be related to a frequency-domain correspondence. As a first step, we choose the peak amplitude of the corresponding resonance curve A_s in a harmonic spectrum. By appropriately handling transform equations for a conjugate complex pair of poles, we arrive at

$$A_s = A_i \cdot (F_o/\pi B_n) \tag{8}$$

Here, πB_n is the distance in the s-plane from the formant frequency $j\omega_n$ to the pole $\sigma_n+j\omega_n$ and F_o enters to ensure a Fourier series property of A_s.

The next frequency-domain operation is to divide A_s by an estimated $Q=F_n/B_{ne}$. This inverse filtering is performed without an exact knowledge of the underlying bandwidth B_n. Add a differentiation in the spectrum with respect to F_o. The result is an "undressed" and differentiated harmonic source component:

$$A_k' = (F_n/F_o) \cdot (B_{ne}/F_n) \cdot A_s \tag{9}$$

which combined with Eq. (8) provides

$$A_i = \pi \cdot (B_n/B_{ne}) \cdot A_k' \tag{10}$$

In the following we shall drop the correction term B_n/B_{ne} and also assume that the inverse filtering has been made with a correct $F_{ne}=F_n$. Eq. (7) may now be rewritten:

$$E_e/U_o = (A_k'(f)/A_o) \cdot \pi^2 \cdot k \cdot F_o \tag{11}$$

$A_k'(f)/A_o$ is apparently the spectral level at a frequency f versus that of the fundamental in a +12 dB/oct compensated glottal flow spectrum. It is also assumed that the low-pass effects of a finite T_a have been compensated.

The general expressions for a harmonic A_k' in the inverse filtered and f/F_o differentiated sound spectrum originating from a glottal flow ramp termination with the slope E_e is simply

$$A_k' = (2/T_o) \cdot (\omega/\omega_o) \cdot (E_e/\omega^2) \cdot (\rho\omega/4\pi a) = (E_e/\pi) \cdot (\rho/\pi a) \tag{12}$$

Note the similarity to Eq. (6). We thus note a simple time-frequency domain relation

$$A_i = \pi \cdot A_k' \tag{13}$$

Accordingly, Eq. (11) is valid not only for a hypothetical harmonic at a formant frequency but also for any source partial A_k' within the spectrum well above F_0. An estimate of E_e/U_0 from measured A_k'/A_0 can thus be performed from any harmonic in a range, say above $2F_g$, where $F_g=1/2T_p$ as in the F- and LF-models and T_p is the glottal flow rise time. Dealing with harmonics outside formant peaks, we are no longer concerned with the error factor B_n/B_{ne} in Eq. (10). A number of independent estimates of E_e/U_0 may accordingly be made which also adds to certify proper bandwidth and formant frequency estimates.

Instead of U_0 we could refer to the maximum derivative E_i in the rising branch. Because of the quasi-sinusoidal shape of the flow derivative, we may with good accuracy express its integral up to T_p by

$$U_0 = (2/\pi)T_p \cdot E_i = E_i/\pi F_g = E_i/(\pi F_0 \cdot R_g) \tag{14}$$

(The error in U_0 is +2% at $E_e/E_i=2$ and +11% at $E_e/E_i=4$.) Eq. (11) may now be rewritten as

$$E_e/E_i = \pi \cdot (A_k'(f)/A_0) \cdot k/R_g \tag{15}$$

For the specific case of $A_k'=A_0$, which means that the voice fundamental amplitude is the same as that of higher harmonics in the differentiated flow derivative spectrum, we have $E_e/E_i=\pi \cdot k/R_g$ which agrees with the statement made in connection with Fig. 3.

However, as already stated, the derivation above assumes that $A_k'(f)$ values have been corrected for a finite T_a. Let A_k'' denote the uncorrected A_k'.

$$A_k' = A_k'' \cdot (1+f^2/F_a^2)^{1/2} \tag{16}$$

F_a is selected so as to equalize the $A_k'(f)$ contour. From the selective inverse filtering, removing all formants but one, we are used to seeing that the damped oscillation starts at an amplitude equal to that of the negative spike E_e of the flow derivative residue. Here we have a method for time-domain estimates of T_a from the relation of A_i to E_e which will be further discussed in connection with Fig. 13.

An important conclusion of this section is that a frequency-domain inverse filtering has the potential of deriving the main LF-parameters. For this purpose, it is convenient to choose the following set: U_0, E_0/U_0, R_g, F_a, F_0 which except for R_g now have been discussed. R_g is estimated by matching in the domain of the first three harmonics.

It remains to extend this theory to other spectral representations, e.g., broad-band sections or band-pass filtered formant data.

Applications to frequency-time domain conversion

The frequency-domain inverse filtering and differentiation, outlined in the previous section, amount to the following operations on a spectral level basis.

$$L_k'(f) = L(f) + 20\log_{10}(f/F_o)$$
$$-20\log_{10}[(K_{rr} \cdot (f) \cdot \prod_{n=1}^{r} \cdot |H_n(f)|)] \quad (17)$$
$$|H_n(f)| = [(1-x_n^2)2 + x_n^2/Q_n^2]^{-1/2}$$
$$x_n = f/F_n$$

where L(f) is the input spectrum level in decibels with possible preemphasis removed and $K_{rr}(f)$ the correction for poles higher than no. r. For r=5, we have

$$k_{rr} = 0.433\, x_1^2 + 0.000712\, x_1^4 \quad (18)$$
$$x_1 = f/f_{ref} = c/4\,\ell$$

The higher pole correction should be scaled to the particular vocal tract length, ℓ_t, Fant (1959; 1960), as derived, e.g., from F4. $\ell_e=(7/4)\cdot(c/F_4)$.

We are now in a position to process any harmonic spectrum. We shall start with an attempt to reconstruct in absolute physical scales glottal flow parameters from a study of Swedish vowels undertaken at the Ericsson Telephone company in 1946-1947 reported by Fant (1948) and published by Fant (1959). This may seem a rather "archaeological" undertaking but is motivated by the fact that great care was taken in preserving absolute sound pressure values of spectrum components. However, this is one of the very few studies in the history of speech analysis in which both frequency and amplitude of formants and individual harmonics have been reported and it is the only one I know of with absolute calibration.

Seven males and seven females served as subjects phonating steady-state vowels in an anechoic chamber with a distance of 12.5 cm to a Brüel & Kjaer condenser microphone. Harmonic spectra were recorded on line by a sweep-frequency method. The subjects had to sustain the vowels for about 5 sec. Most of them were amateur singers.

Fig. 6 shows the variation of the voice fundamental amplitude L_o within a sequence of vowels ordered in terms of increasing F_1 in back vowels followed by decreasing F_1 of front vowels and ending with an increasing F_1 series of rounded front and mid vowels. Subscript 1 stands for phonemically long vowels and subscript 2 for short vowels and 3 for long pre-r variants. The u_1 is a maximally rounded front vowel [ʉ:] and u_2 is a mid vowel [ɵ] (Fant, 1983).

Females tend to have higher L_o than men, especially in low F_1 vowels, which may be explained by the relative proximity of F_o to F_1 enhancing L_o of females. As seen in Fig. 7, the situation is reversed after undressing the transfer function, the males showing about 1 dB higher L_o than females. A typical male-female difference is the higher first formant amplitude L_1 versus L_o, as seen in the bottom part of Fig. 6.

There are systematic trends in Fig. 7 that deserve some comments. The amplitude of the corrected source fundamental L_{oc} is somewhat lower in vowels with low F_1 compared to vowels of higher F_1. This is typically so for o_1=[u:] and i_1=[i:] which may be anticipated from production theory. Both the loss of transglottal driving pressure and the pulse skewing caused by a rather extreme vocal tract constriction would reduce the magnitude of glottal flow. This would also be expected for the vowel $å_2$=[ɔ] which is produced with a narrow pharyngeal constriction.

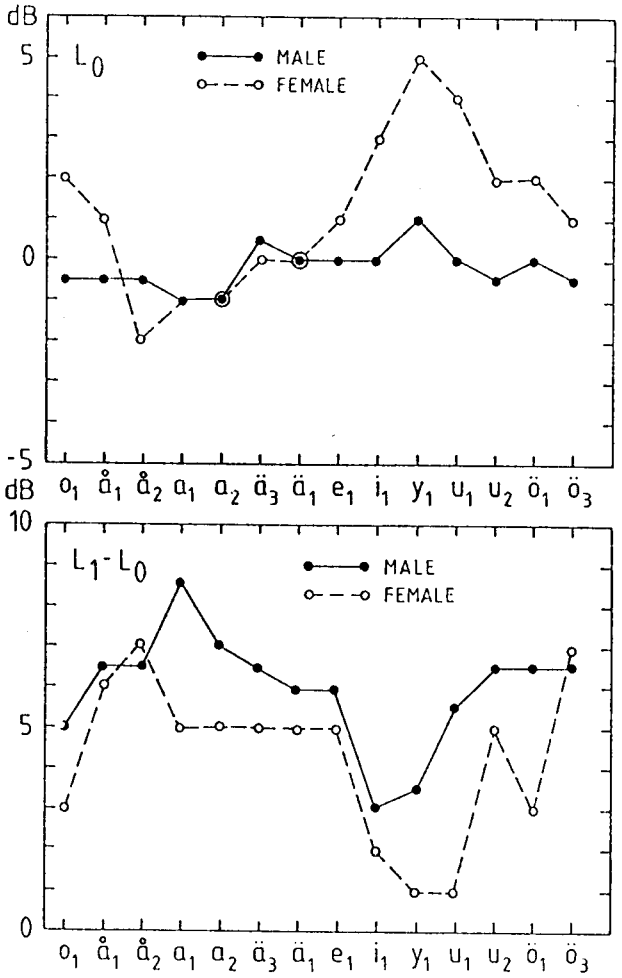


Fig. 6. Original data from Fant (1959) on the spectrum level of the fundamental L_0 and its relation to the first formant spectrum level L_1-L_0 within a sequence of vowels.

All vowels are sustained at one and the same subject's preferred F_0 which averaged 125 Hz for males and 217 Hz for females. Inherent F_0-variations were thus eliminated and cannot explain the L_{oc} variations.

The L_{oc} of two reference subjects are shown in Fig. 8. Subject G was a well known Swedish phonetician Olof Gjerdman and subject M a well known Swedish soprano singer and voice specialist, Marianne Mörner. Here, we have evidence for a special aspect of acoustic-aerodynamic interaction modelled by Fant & al. (1985) and initially observed by Rothenberg (1985).

When F_1 is very close to F_0, as in the subject M's vowel o_1 with F_0=256 Hz and F_1=270 Hz or in u_1 with F_0=256 Hz and F_1=300 Hz, there exists economy not only by optimal filtering but also a minimum of glottal flow since an F_1 oscillation component opposes the transglottal pressure drop and minimizes the air consumption whilst a relative high E_e excitation is retained. The large step in subject M's L_{oc} from o_1 to $å_1$ of 4 dB and from u_1 to u_2 of 5 dB is explained by the about 1.5 F_1/F_0 ratio in $å_1$ and u_2, which according to Fant & al. (1985) conditions the opposite effect, i.e., an increased air consumption. Since this effect is much more pronounced in the female voices than in the male voices, it seems to be a plausible explanation in addition to variations of supraglottal constriction.

What about glottal flow parameters and absolute scale values? We selected the vowels a_1=[ɑ:], a_2=[a], $ä_3$=[æ:], and $ä_1$=[ɛ:] and processed the tabulated F_0,L_0 F_1,L_1 F_2,L_2 F_3,L_3, and F_4,L_4 data from Fant (1959) in accordance with the inverse filtering equations (17) and (18). Glottal peak flow U_o was then calculated from Eq. (5) with k=1 and the appropriate F_0. Next a representative

value of F_a was estimated from Eq. (16) to account for the spectral drop off in regenerated source levels L_k' from F_1 over F_2 to F_3. After this spectral correction, a value of A_k'/A_o could be estimated to be inserted in Eq. (11) to provide the E_e/U_o estimate.

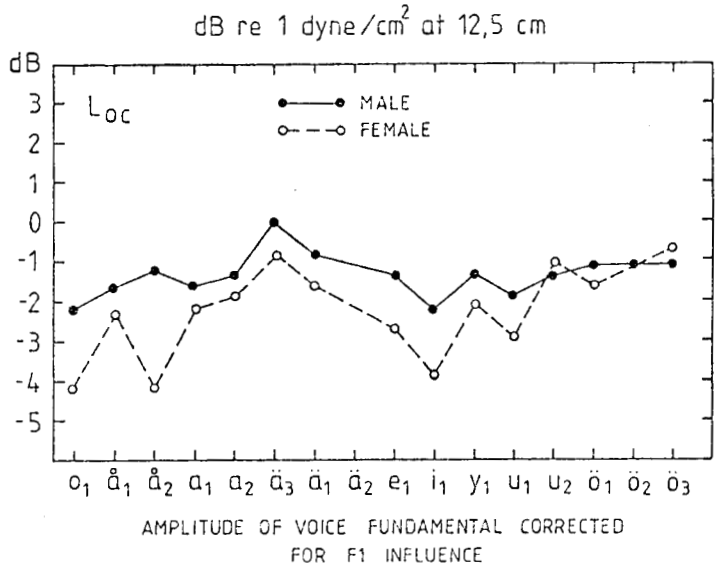


Fig. 7. Male and female average data on the voice fundamental amplitude corrected for influence of F_1 and higher formants.

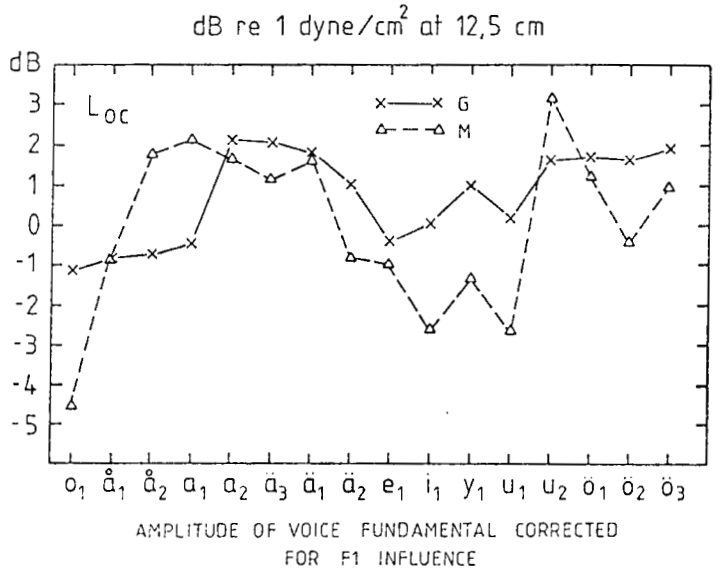


Fig. 8. The same as Fig. 7 for a male subject G and a female subject M.

This procedure has one snag. The formant amplitude levels initially measured by Fant (1959) were attained by root mean square summations of partials within the formant domain. In order to translate these A_e measures to most likely spectrum envelope peak values A_S, we performed a correction by a factor

$$A_S/A_e = (1\text{-}e\text{-}y) / (1\text{-}e\text{-}2y)1/2 \qquad (19)$$

where $y=\pi B_n/F_0$, derived by Fant (1959). We found $A_S/A_e=0.83$ for male average and 0.72 for female average.

Average glottal parameter data for the four vowels have been documented in Table I together with data from other surveys of interest. Because of lack of data on the second harmonic, we could not estimate R_g.

There are remarkable similarities across studies. None of the studies, except possibly Holmberg, Hillman, & Perkell (1988), reveals a significant female/male difference in U_0/E_e and the same is true of the related parameter R_k. The U_0/E_e is of the order of 1.1 ms for the Fant (1959) and the Holmberg & al. (1988) data and of the order of 0.7 ms in the Fant, Gobl, & Karlsson (1987) data. The higher U_0/E_e in the first two studies implies according to Eq. (11) a greater dominance of the voice fundamental amplitude. This is plausible considering the sustained almost singing mode in the Fant (1959) phonations which would enhance the fundamental (Sundberg & Gauffin, 1979). In the Holmberg & al. study (1988), the relative high U_0/E_e might be explained as an abduction assimilation from the unvoiced [p] to the following vowel [æ] in their test words or else originating from a low-pass smoothing of glottal flow termination inherent in the mask technique.

Why do we in spite of the relative constant U_0/E_e observe a typically higher level of the voice fundamental amplitude versus the level of the first formant in females compared to men, see Fig. 10, pertaining to a vowel [ɑ:]? For subject JS, we may note $L_1\text{-}L_0=14$ dB and for subject MS, $L_1\text{-}L_0=7$ dB. The answer is that the F_0 factor in Eq. (11) contributes with 5 dB and that the difference in the L_0 reinforcement from F_1 and higher formants of the vowel [ɑ:] is 2 dB in favor of the female voice. Instead of referring to Eq. (11) we may refer to Eq. (15) in which E_i/E_e takes the place of U_0/E_e and the R_g factor enters. However, it seems to be more direct to refer to the 5 dB difference in F_0 entering Eq. (11) than to refer to a 3 dB difference in E_i/E_e and a 2 dB difference in R_g in Eq. (14).

The absolute magnitude of the peak flow derived from the spectrum data of Fant (1959) is reasonable. The 0.42 liters/sec noted for males and 0.21 liters/sec for females may be compared to the Holmberg & al.(1988) 0.23 liters/sec for males and 0.14 for females. However, the experimental conditions were quite different and one may note that the latter study reported an average of 0.1 liters/sec additional steady air leakage. Our values range well within typical data reported in the literature, e.g., Rothenberg (1973); Sundberg & Gauffin (1979).

Both U_0 and E_e are typically 5 dB higher for males than for females. This constant E_e/U_0 ratio does not imply that the shape of the female glottal flow pulse is a proportional down scaling of the male pulse. Proportionality would have implied linear scaling in both amplitude and time, and that the amplitude reduction would be the same as the reduction of the time base in terms of $T_p=1/2F_g$. Under such conditions, the female E_e would equal that of the male E_e, and the E_e/U_0 would have increased by the inverse of the time scale proportionality factor. In reality, the 3 dB higher F_g of

females is compensated by a 3 dB lowering of E_e at constant U_o associated with the wave form change induced by a relatively longer return phase.

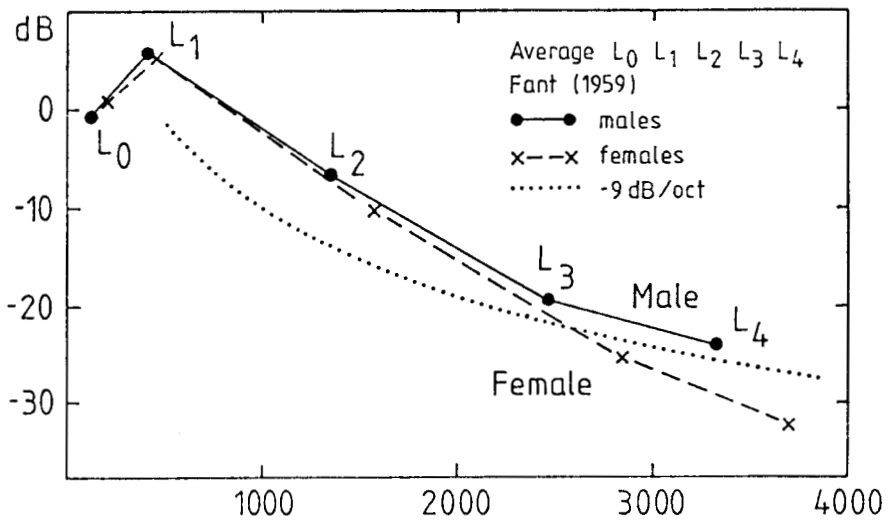


Fig. 9. Fant (1959) vowel data averaged over all vowels.

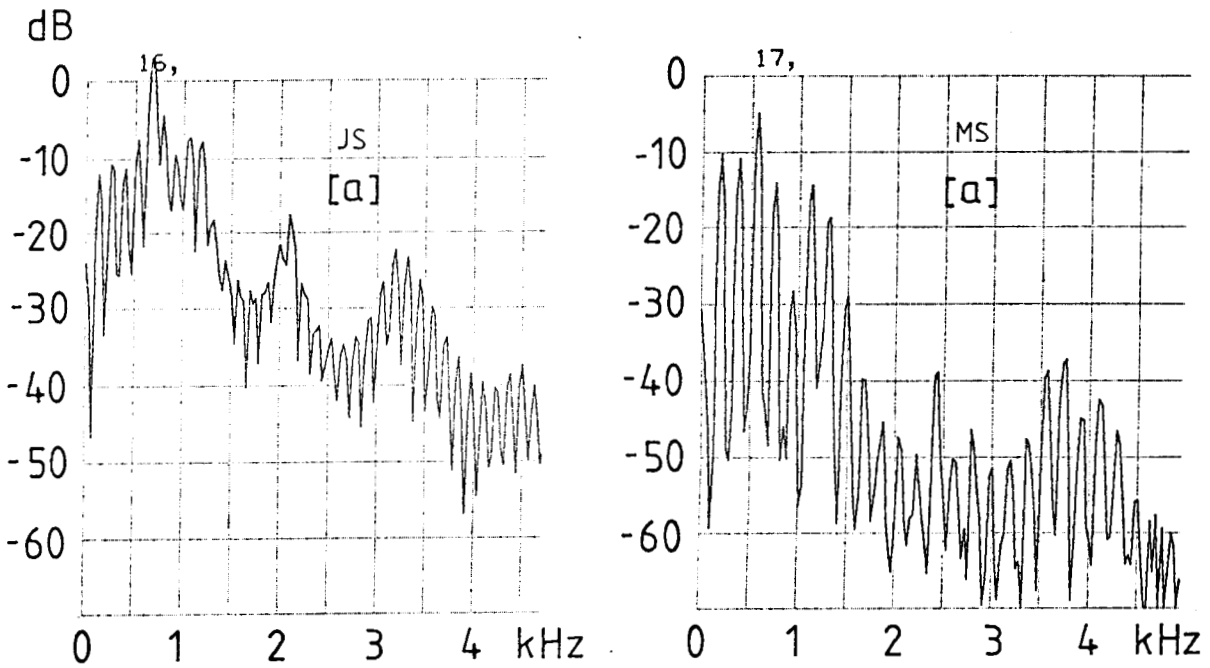


Fig. 10. Spectra of the vowel [ɑ:] from the word "ja" produced by a male JS and a female MS.

	U_o cm^3/sec	E_e 10^3 cm^3/sec^2	U_o/E_e ms	E_e/E_i	R_g %	R_k Hz	F_a
1. Fant (1959), Present study							
Males, F_0=125	420	360	1.15				2100
Females, F_0=217	210	190	1.10				1200
2. Holmberg & al. (1988)							
Males, F_0=116	230	240	0.95				
Females, F_0=213	140	120	1.15				
3. Fant & al. (1987)							
Males, F_0=125			0.65	3.2	1.1	30	1000
Females, F_0=217 (from 1).			0.70	2.2	0.9	30	430
4. F-domain analysis, Present study							
JS (male), F_0=133			0.65	3	1.2	28	3000
MS (female), F_0=188			0.70	2.4	1	30	2000

Table 1 . Glottal flow data

The lower E_e of the glottal pulses of females compared to men does not imply that the formant amplitudes are proportionally lower (Fant, 1982). The conversion from E_e to spectrum envelope peak includes the factor F_0/B_n, see Eq. (8). The F_0 part is a general consequence of the conversion from a Fourier integral to a Fourier series and B_n is obvious from the Q-concept of a resonance.

The 5 dB lower E_e of females compared to males is thus fully compensated by the 5 dB higher F_0 and there remains due to smaller bandwidths about 1.5 dB higher formant levels in the male spectra in addition to the difference in the spectrum slope associated with F_a. An averaging of voice fundamental and formant frequency and amplitude data from the Fant (1959) study is shown in Fig. 9. The rather small differences between the male and female data are consistent with the reasoning above and the discussion in connection with Eq. (19). This is also consistent with the general observations of Monsen & Engebretson (1977).

We shall now exemplify the technique of frequency-domain inverse filtering and parameter extraction from complete harmonic representations of the male JS's vowel and the female MS's vowel, Fig. 10. We have already commented on the relative dominance of L_0 in the female spectrum. After frequency-domain inverse filtering and differentiation, Eq. (17), we have thus generated a +12 dB/oct boosted glottal flow spectrum, i.e., a +6 dB/oct emphasized glottal flow derivative spectrum shown in Figs. 11 and 12. In addition to the separate harmonics, we have included data points for each formant envelope peak. A fair continuity of the source spectrum levels at formants and neighboring harmonics ensures that no major errors in formant frequencies and bandwidths have remained. The bandwidths thus derived will be optimal for resynthesis. Estimated F_a-filtering curves, Eq. (15), are included in Figs. 11 and 12.

This procedure of determining formant frequency and bandwidth values is similar to the initial stage of our present time-domain inverse filtering in which a single voice period is extracted for spectral analysis (Gobl, 1988). However, truncation errors and artifacts occasionally cause spectral distortions that may be larger than the fluctuations in narrow-band sectioning over a window length of 30 ms. A closer analysis of these problems is needed.

The source spectra in Figs. 11 and 12 display an irregular detail structure which could be expected from acoustic interaction (Fant & Lin, 1987). The zero between F_1 and F_2 is especially apparent for the female subjects. However, it is also frequent in male voices, see e.g., Fig. 22 in Fant (1959) and Carré (1987). It is probably caused by acoustic interaction.

A more conspicuous aspect of MS's source spectrum is the relatively large dip of 10-15 dB in the F_3 region. This causes an ambiguity in the choice of F_a. Ignoring F_3, the source level at F_4 F_5 would prescribe an F_a of at least 2000 Hz whilst if we put some weight on L_3, we could select something like F_a=1000 Hz. A time-domain analysis according to our routines gave F_a=850 Hz. The advantage of the frequency-domain parameterization is that it provides a free choice influencing resynthesis.

As a spinoff of these studies, we have proposed a novel, alternative method of F_a-determination from the time domain that can be directed to provide a match in any formant region. This is the ratio of the glottal flow derivative spike E_e to the initial amplitude A_i of a single formant remaining after selective inverse filtering of all other formants.

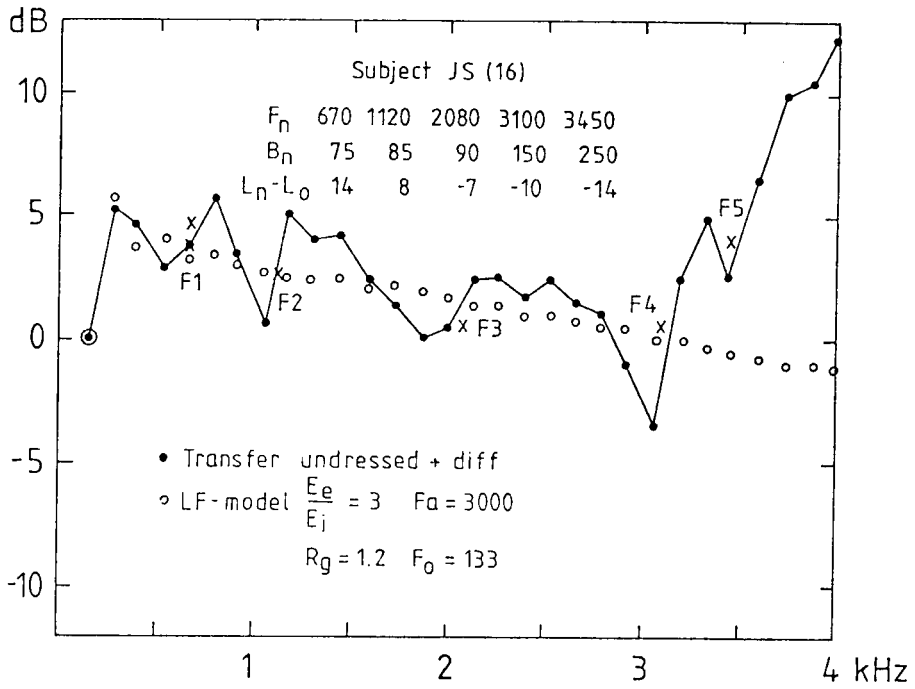


Fig. 11. Frequency domain inverse filtering and differentiation of the male [ɑ] vowel from Fig. 10. This technique is especially useful for determining F_a.

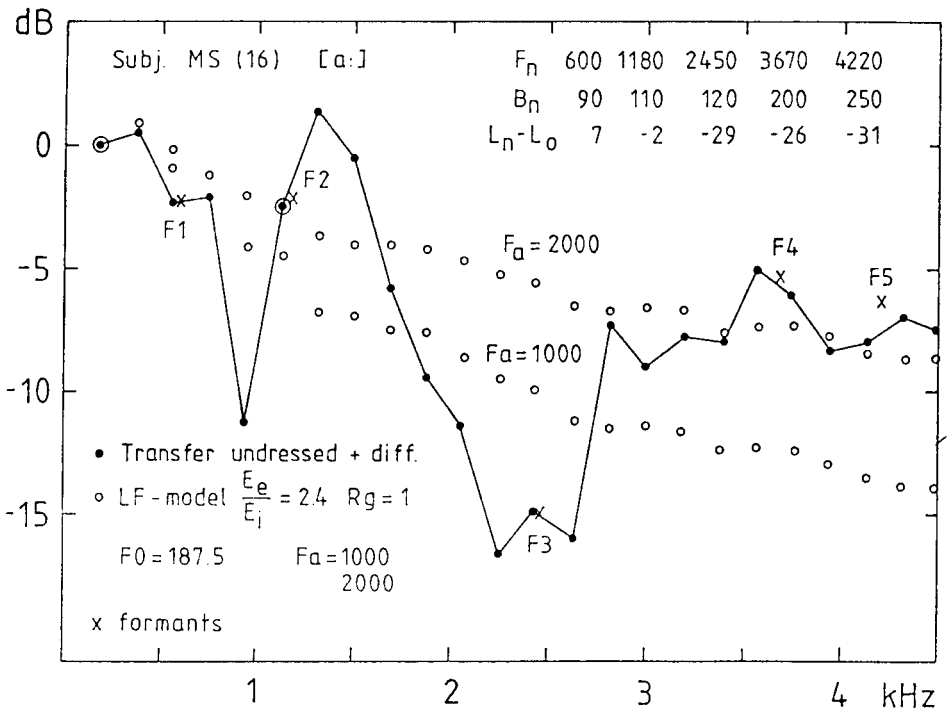


Fig. 12. The same for the female subject MS. Two different values of F_a are suggested.

This ratio is approximately the gain of the F_a filter, Eq. (16), at the frequency of the formant. An example is brought out in Fig. 13 which refers to an [h]-vowel sequence (Fant, 1980). We may here follow the temporal variation of F_a from 90 Hz in the [h] to 1200 Hz in the following vowel. The associated bandwidth as determined from envelope fitting varies in this context from something larger than 300 Hz to 52 Hz in the vowel. Such bandwidth changes are important to preserve in the resynthesis. A quantitative analysis of glottal damping (Fant, 1960; Badin & Fant, 1984) can be extended by specific rules for specific vowel categories. Because of the glottal inductance and reactive components of the rest of the vocal tract, the glottally induced bandwidth increase is minimized for frequencies above 1000-1500 Hz. With this in mind, the data of Fant (1960, p. 136), here labelled B_{ref}, may be generalized for any particular glottal flow by the relation

$$\Delta B(t) = B_{ref} \cdot U_g(t)/55 \qquad (20)$$

where $U_g(t)$ is the glottal flow in cm^3/sec, see also Fant (1979).

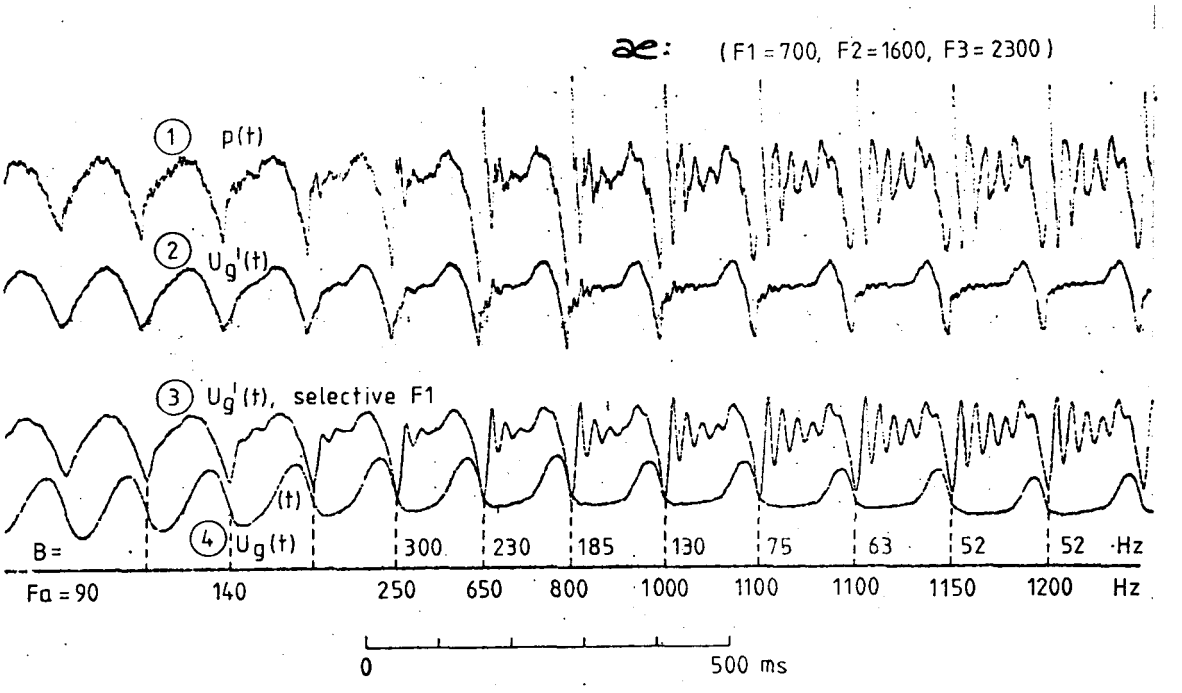


Fig. 13. Selective F_1 inverse filtering; curve 2 from the bottom, provides means of determining both B_1 and the spectrum parameter F_a.

One aspect of increasing glottal abduction or leakage is that the subglottal system no longer can be neglected in vowel production. A system function distortion enters. It can be specified by the appearance of extra poles and zeros and some finite detuning of poles of the original uncoupled stage. The effect is exemplified in Fig. 14 which pertains to a simulation of the Russian vowels [e] and [a] with the subglottal system of Badin & Fant (1984). The transfer function was obtained with a pressure source introduced in the glottis. The coupled state corresponds to a glottal opening of 0.2 cm^2 and a lung pressure of 2 cm H_20, while the uncoupled state was simulated with a very high glottal impedance. The [e] vowel displays an extra peak at 1500 Hz,

which can be identified with the 1400 Hz extra formant in the female [ɛ] spectrogram at the right in Fig. 14. More extra peaks appear in the [a] transfer function. These effects are very much the same as measured by Fujimura & Lindquist-Gauffin (1971) on a live subject with the sweep-frequency method, see also Fant, Ishizaka, Lindqvist-Gauffin, & Sundberg, 1971). It is important to note that overall shifts in spectrum levels occur. The relative attenuation of F_2 of [e] could shed light on the relative weak F_3 of subject MS in Fig. 12.

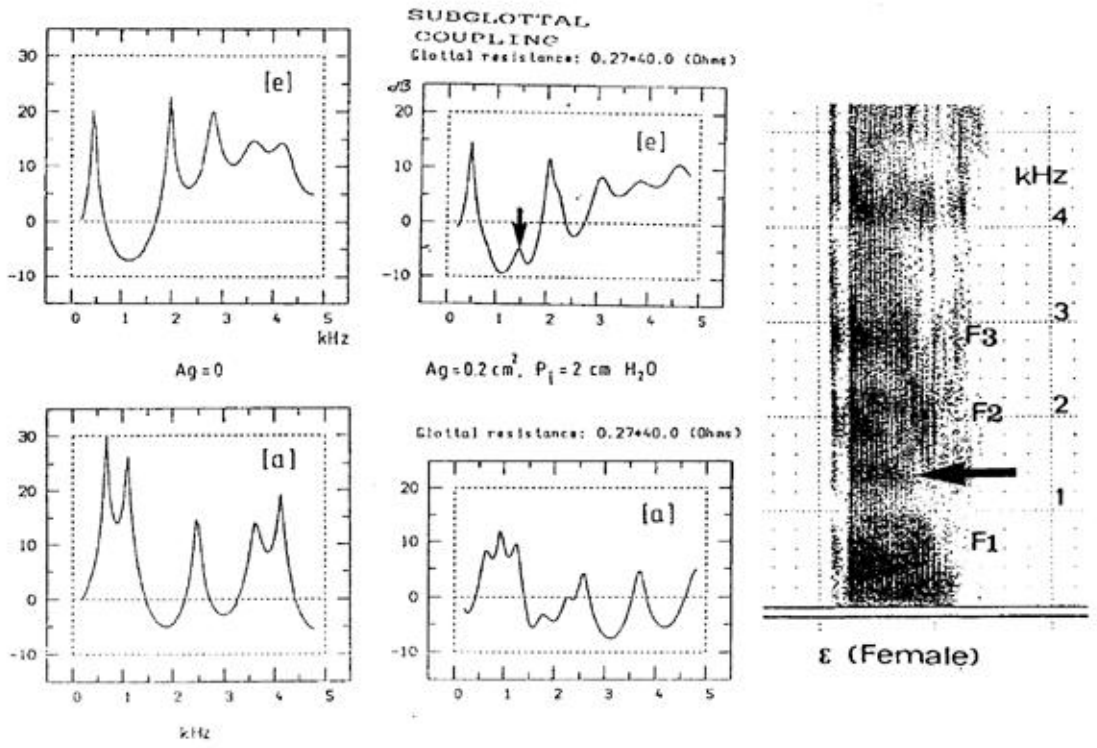


Fig. 14. Finite and large subglottal coupling creates spectral distortion of vowels. Additional formants and changes in overall spectrum are apparent. Note for comparison the extra formant in the spectrogram of the female post [d] and pre [t] breathy vowel [ɛ]

The underlying cause appears to be a zero associated with the third subglottal resonance which appears in this region. We have definite examples of relative prominent extra subglottal formants occasionally seen in spectrograms of female voices. The weak spectral peak at 2800 Hz is the MS [a] spectrum, Fig. 10, could derive from subglottal coupling. However, extra peaks of a few dB are also a common effect of acoustic interaction (Fant & Lin, 1987). The main indication of a subglottal coupling is thus the relatively low F_3 compared to the rest of the spectrum.

We shall end this survey by reference to the frequency-domain inverse filtering performed by Mártony (1965). Fig. 15 shows the mean and extremes of his subjects' source spectra. Before we have more evidence from analysis and modelling, we should execute some caution when attempting to explain the details. The main point is that a typical voice source spectrum may deviate systematically from a uniformly sloping line.

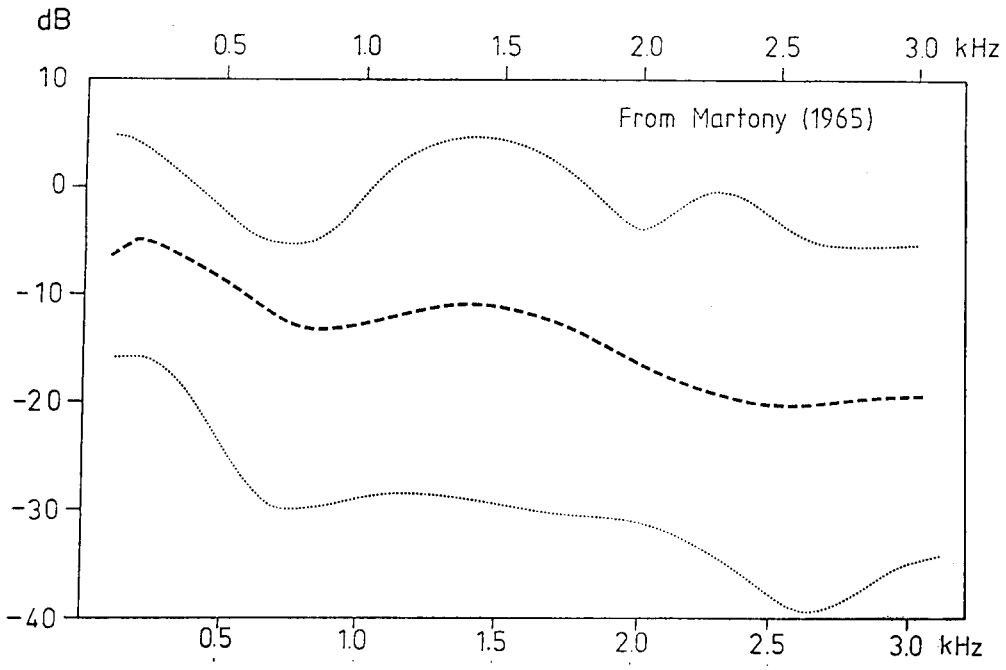


Fig. 15. Extreme and mean voice source spectra, from Mártony (1965).

Conclusions

We have here touched upon some of the potentialities and difficulties in inverse filtering and voice source parameterization. A frequency domain approach as we have outlined avoids the need of high fidelity low-frequency phase correct recordings and allows a direct processing of harmonic spectra. The technique could be extended to broad-band spectral analysis. The basic advantage of the frequency-domain processing is a closer tie with resynthesis needs. However, it remains to perform a more systematic analysis of the unavoidable differences between true human speech with all interaction effects and linear synthesis.

A non-interactive source-filter system based on best spectral match with the LF-model will probably stand up to rather high quality standards.

References

Ananthapadmanabha, T.V. & Fant, G. (1982): "Calculation of True Glottal Flow and its Components", Speech Communication **1**, pp. 167-184.

Badin, P. & Fant, G. (1984): "Notes on Vocal Tract Computation", STL-QPSR 2-3/1984, pp. 53-108.

Carré, R. (1987): "Review of French Work on Vocal Source - Vocal Tract Interactions, pp. 371-375 in *Proc. XI ICPhS, Vol. 3*, Academy of Sciences of the Estonian SSR, Tallinn.

Fant, G. (1948): "Analys av de svenska vokalljuden", L M Ericsson protokoll H/P 1035.

Fant, G. (1959): "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", Ericsson Technics No. 1.

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

Fant, G. (1979): "Glottal Source and Excitation Analysis", STL-QPSR 1/1979, pp. 85-107.

Fant, G. (1980): "Voice Source Dynamics", STL-QPSR 2-3/1980, pp. 17-37.

Fant, G. (1982): "Preliminaries to Analysis of the Human Voice Source", STL-QPSR 4/1982, pp. 1-27.

Fant, G. (1983): "Feature Analysis of Swedish Vowels - A Revisit", STL-QPSR 2-3/1983, pp. 1-19.

Fant, G. & Ananthapadmanabha, T.V. (1982): "Truncation and Superposition", STL-QPSR 2-3/1982, pp. 1-17.

Fant, G. & Lin, Q. (1987): Glottal Source - Vocal Tract Acoustic Interaction", STL-QPSR 1/1987, pp. 13-27.

Fant, G., Gobl, C., & Karlsson, I. (1987): "The Female Voice - Experiments and Overviews", J.Acoust.Soc.Am. **82**, p. S92(A).

Fant, G., Ishizaka, K., Lindqvist-Gauffin, J., & Sundberg, J. (1972): "Subglottal Formants", STL-QPSR 1/1972, pp. 1-12.

Fant, G., Liljencrants, J., & Lin, Q. (1985): "A Four-parameter Model of Glottal Flow", STL-QPSR 4/1985, pp. 1-13.

Fant, G., Lin, Q. & Gobl, C. (1985): "Notes on Glottal Flow Interaction", STL-QPSR 2-3/1985, pp. 21-45.

Fujimura, O. & Lindqvist-Gauffin, J. (1971): "Sweep-tone Measurements of Vocal-tract Characteristics", J.Acoust.Soc.Am. **49**, pp. 541-558.

Gobl, C. (1988): "Voice Source Dynamics in Connected Speech", STL-QPSR 1/1988, pp. 123-159.

Holmberg, E.B., Hillman, R.E., & Perkell, J.S. (1988): "Glottal Air Flow and Transglottal Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice", J.Acoust.Soc.Am. **84**, pp. 511-529.

Karlsson, I. (1988): "Glottal Wave Form Parameters for Different Speaker Types", pp. 225-231 in *Proc. of SPEECH 88*, 7th FASE Symp., Edinburgh.

Mártony, J. (1965): "Studies of the Voice Source", STL-QPSR 1/1965, pp. 4-9.

Monsen, R.B. & Engebretson, A.M. (1977): "Study of Variations in the Male and Female Glottal Wave", J.Acoust.Soc.Am. **62**, pp. 981-993.

Nord, L., Ananthapadmanabha, T.V., & Fant, G. (1984): "Signal Analysis and Perceptual Tests of Vowel Responses with an Interactive Source Filter Model", STL-QPSR 2-3/1984, pp. 25-52.

Rothenberg, M. (1973): "A New Inverse Filtering Technique for Deriving the Glottal Air Flow Wave Form During Voicing", J.Acoust.Soc.Am. **53**, pp. 1632-1645.

Rothenberg, M. (1985): "Cosi Fan Tutte and What It Means", draft for discussion, Fourth Int. Vocal Fold Physiology Conf., New Haven, CT.

Sundberg, J. & Gauffin, J. (1979): "Wave Form and Spectrum of the Glottal Voice Source", pp. 301-322 in (B. Lindblom & S. Öhman, eds.) *Frontiers of Speech Communication*, Academic Press, London.