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

Vocal source analysis - a progress report

Fant, G.

journal: STL-QPSR
volume: 20
number: 3-4
year: 1979
pages: 031-053

KTH Computer Science and Communication

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

II. SPEECH PRODUCTION

A. VOCAL SOURCE ANALYSIS - A PROGRESS REPORT

G. Fant

Abstract

The general theory of the glottal source and excitation process reported in STL-QPSR 1/1979 has been followed up by model based calculations of source spectra and speech pressure waveforms. The experimental procedure for deriving glottal source parameters from inverse filtering has been tested with a view to evaluate alternative procedures. A special attention has been devoted to transitional regions between fully voiced and aspirated segments in which F1 of vowel [a] is heavily reduced. Other factors such as nasalization also reduce the F1 intensity below the value predicted from the ideal excitation.

The model

The model of glottal volume velocity flow adopted by Fant (1979) is illustrated in Fig. II-A-1. In addition to the voice fundamental frequency $F_o = 1/T_o$ its basic parameters are the peak flow U_o, the glottal frequency $F_g = \omega_g/2\pi$ and the asymmetry factor K. The pulse may be ascribed a starting point $t = T_1$. The rising branch

$$U = \tfrac{1}{2} U_o [1 - \cos \omega_g(t-T_1)] \quad (1)$$

reaches the peak value U_o at $t = T_2$.

$$T_2 - T_1 = \frac{1}{2F_g} \quad (2)$$

The falling branch

$$U = U_o\left[K\cos\ \omega_g(t - T_2) - K + 1\right] \quad (3)$$

hits the zero line after a time

$$T_3 - T_2 = \frac{1}{\omega_g} \arccos\left(\frac{K-1}{K}\right) \quad (4)$$

Providing $k > 0.5$ the termination is abrupt with a slope

$$U_3' = \frac{dU}{dt}\ (t = T_3) = -U_o \cdot \omega_g \sqrt{2K - 1} \quad (5)$$

If $K > 1$ this is the maximum slope during the course of the falling branch. For $0.5 < K < 1$ the maximum slope occurs prior to closure. At $K = 0.5$ the falling branch is symmetrical to the rising branch.

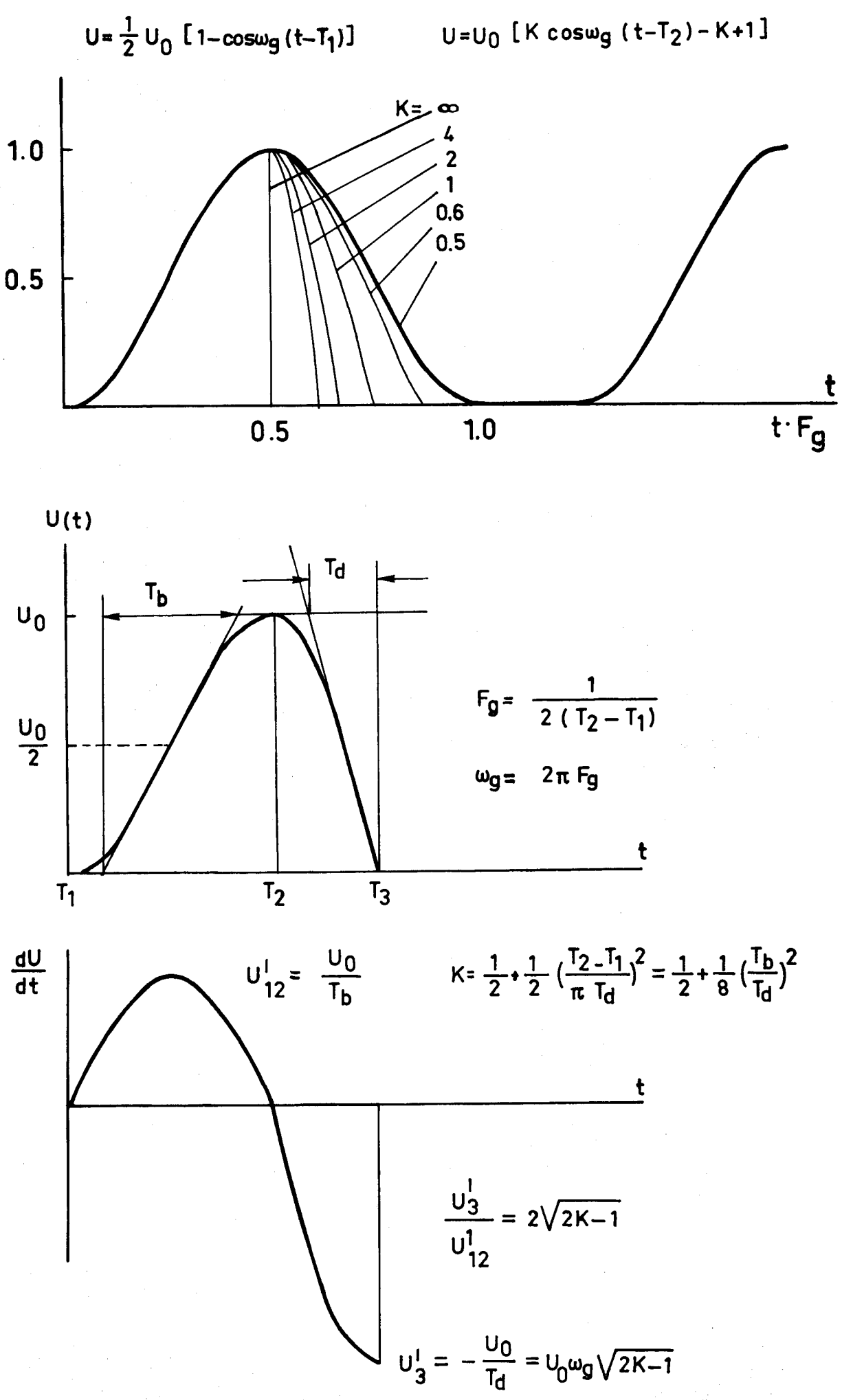


Fig. II-A-1. The voice source model adopted by Fant (1979).

This is the lower bound of k in the present use of the model and represents a lowest degree of excitation strength. Values of $K < 0.5$ would imply a progressive shift of the zero line and a shorter fall than rise.

As will be further discussed in the section on time-domain analysis the slope U_3' is the major source determinant of formant amplitudes.

Frequency domain analysis

The Eq. (56) of Fant (1979) provides the expression for calculating the source spectrum. At non-extreme K-values the source spectrum falls off at -12 dB/octave above the glottal frequency F_g. It may therefore be interesting to normalize the spectrum in order to bring out the deviations from the average trend. For practical purposes we introduce first a correction of +6 dB/octave for radiation and then the approximate +6 dB/octave of our standard preemphasis for speech analysis.

$$H_p(f) = \left[\frac{1 + f^2/200^2}{1 + f^2/5000^2} \right]^{\frac{1}{2}} \qquad (6)$$

which is defined by a zero at 200 Hz and a pole at 5000 Hz. It can also be motivated to include the radiation transfer in excess of the frequency proportionality which Fant (1960) calculated as deriving from the baffle effect of the head and the increase of radiation resistance in excess of frequency square. This correction factor can be approximated by

$$K_T(f) = \frac{1 + f^2/800}{1 + f^2/1200} \qquad (7)$$

It represents an additional emphasis of maximally 6 dB at high frequencies. The validity of this correction has never been fully proved. It is mostly left out in calculations.

The K_T factor is retained in Fig. II-A-2 but not in Fig. II-A-3. Assuming a steepness factor of K = 1 and without K_t this low frequency spectral peak around F_g is about 10 dB above the upper part of the spectrum.

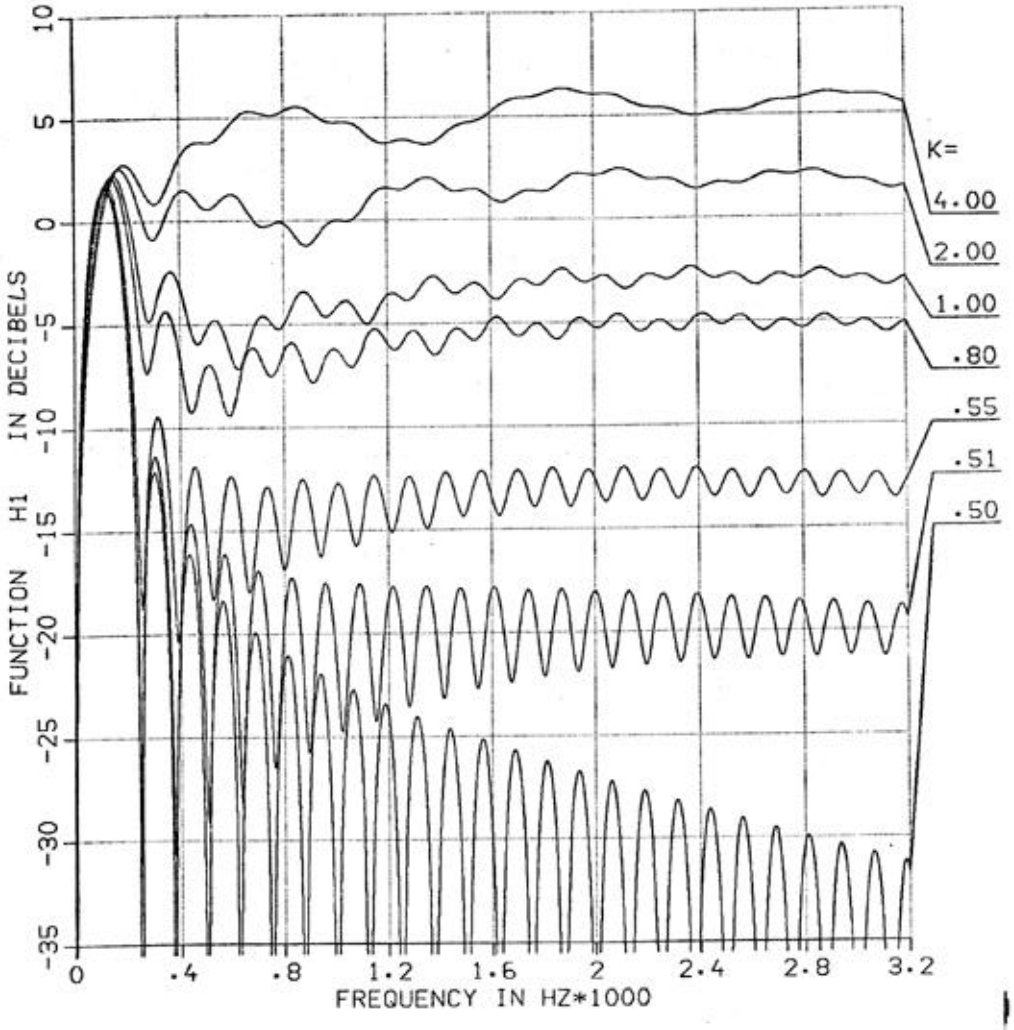


Fig. II-A-2. The voice source spectrum with three added corrections:
(1) +6 dB/oct for radiation transfer,
(2) theoretical baffle effect, Fant (1960), with the head as a sphere of radium 9 cm,
(3) our standard preemphasis for speech analysis, approximately +6 dB/oct.

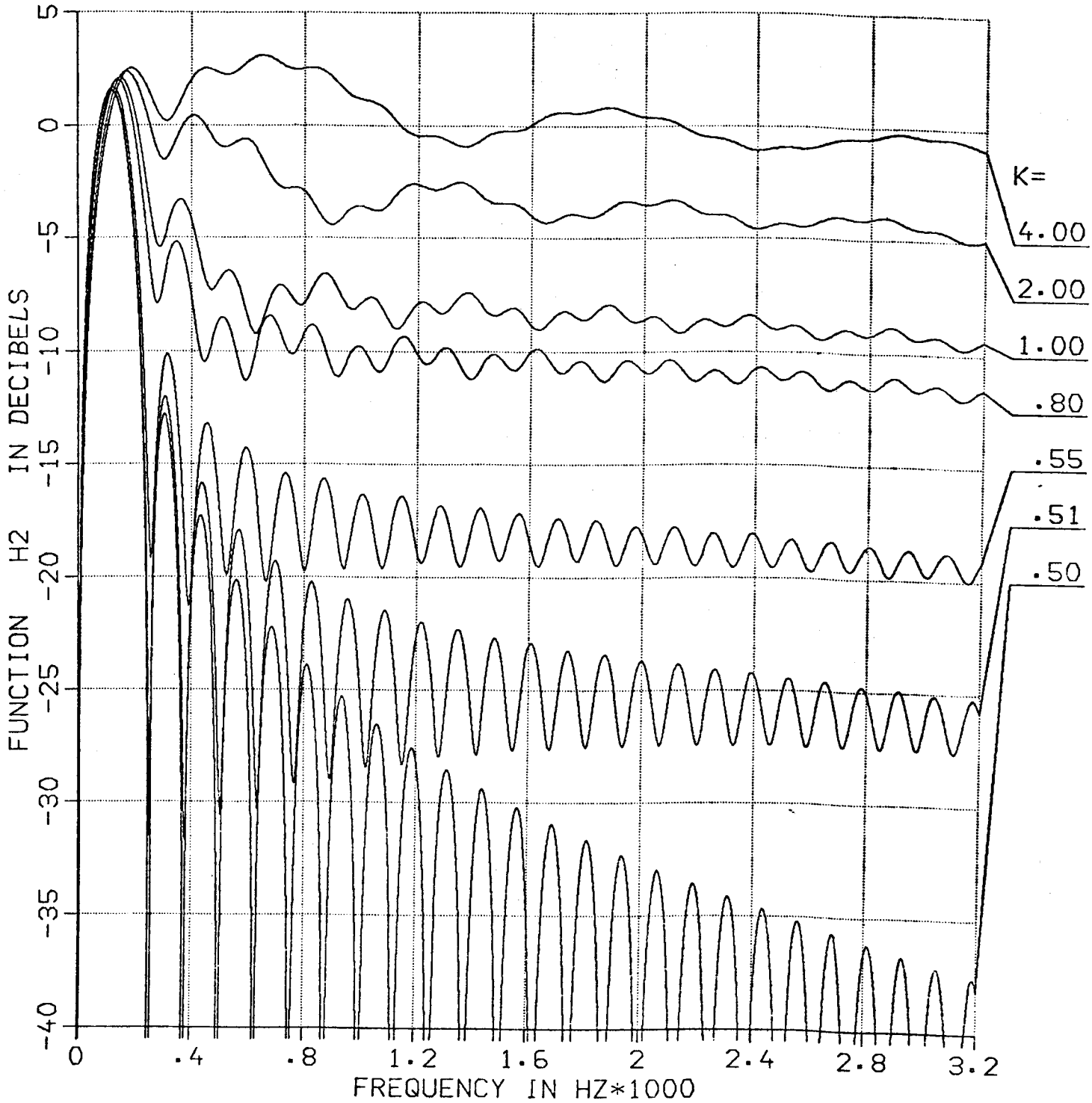


Fig. II-A-3. Same as Fig. II-A-2 without head baffle correction.

It should be noted that the glottal frequency F_g is by no means constant. It is partially locked to F_o and about 1.7 F_o. A rise in F_g causes not only the cutoff frequency of the spectrum to shift up proportionally but it also shifts the spectrum level in proportion to F_g, as seen from the expression for the glottal pulse slope, Eq. (3). Fig. II-A-4 is the source spectrum corrected with a +6 dB/octave slope and K_T but without the standard preemphasis. This is what a sound spectrum would look like if the vocal transfer function was constant at all frequencies.

Finally, in Fig. II-A-5 we add (without K_T) the transfer function for a neutral tube resonator with resonance frequencies at 500, 1500, 2500 etc. Hz and bandwidths of B_1 = 70 Hz and $B_n = (2n - 1)^{\frac{1}{2}}$ 70 Hz. Here we see clearly how the glottal spectrum peak gains prominence at K = 1 or smaller but is still noticeable at K = 2 and K = 4. The K_T factor, Eq. (2), would have added about 3 dB to the level of the first formant.

A limitation in the derivation of a vowel spectrum as the product of a source spectrum and a filter function is that the transfer function is not time invariant. The damping and even the frequency of the first formant may substantially differ in the closed and open phase. The major effect is the sometimes excessive damping in the glottal open period. Therefore the spectrum level at a peak submitted to a high degree of glottal damping will be somewhat reduced in the averaged spectrum.

Time domain analysis

The source filter analysis of voice production introduced by Fant (1979) is more general than that of earlier presentations, since it enables a realistic derivation of the complete speech wave and accounts for the time variable processes within the glottal open period. The procedure followed involves the traditional derivation of the direct and then the inverse Laplace transform of the process which generates a complex signal of the general form.

$$\sum_{i=1}^{3} \sum_{n=0}^{r} A_{ni} \cdot e^{-\alpha_n (t-t_{ni})} \cos \left[\omega_n (t-t_{ni}) + \varphi_{ni} \right] \qquad (8)$$

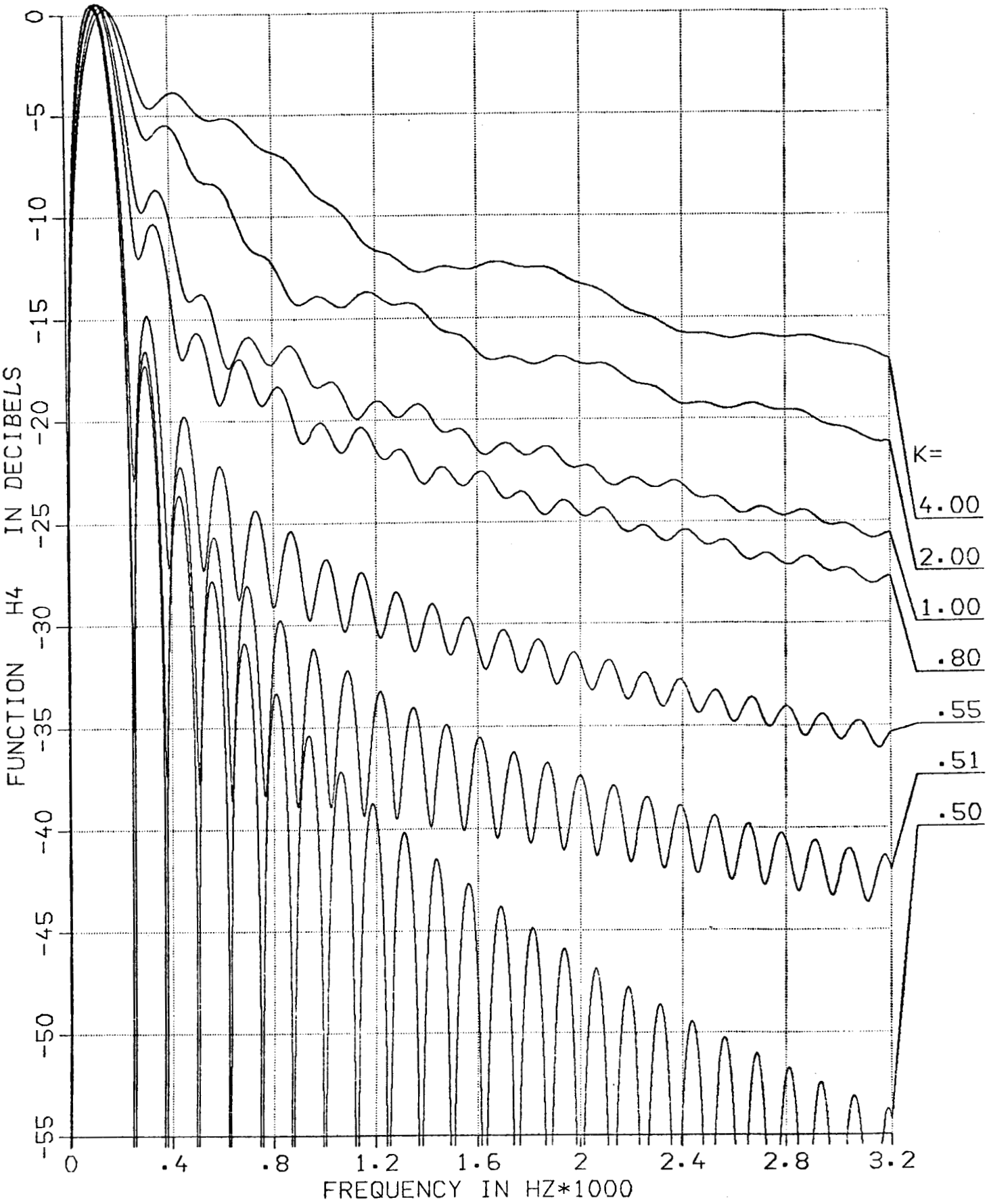
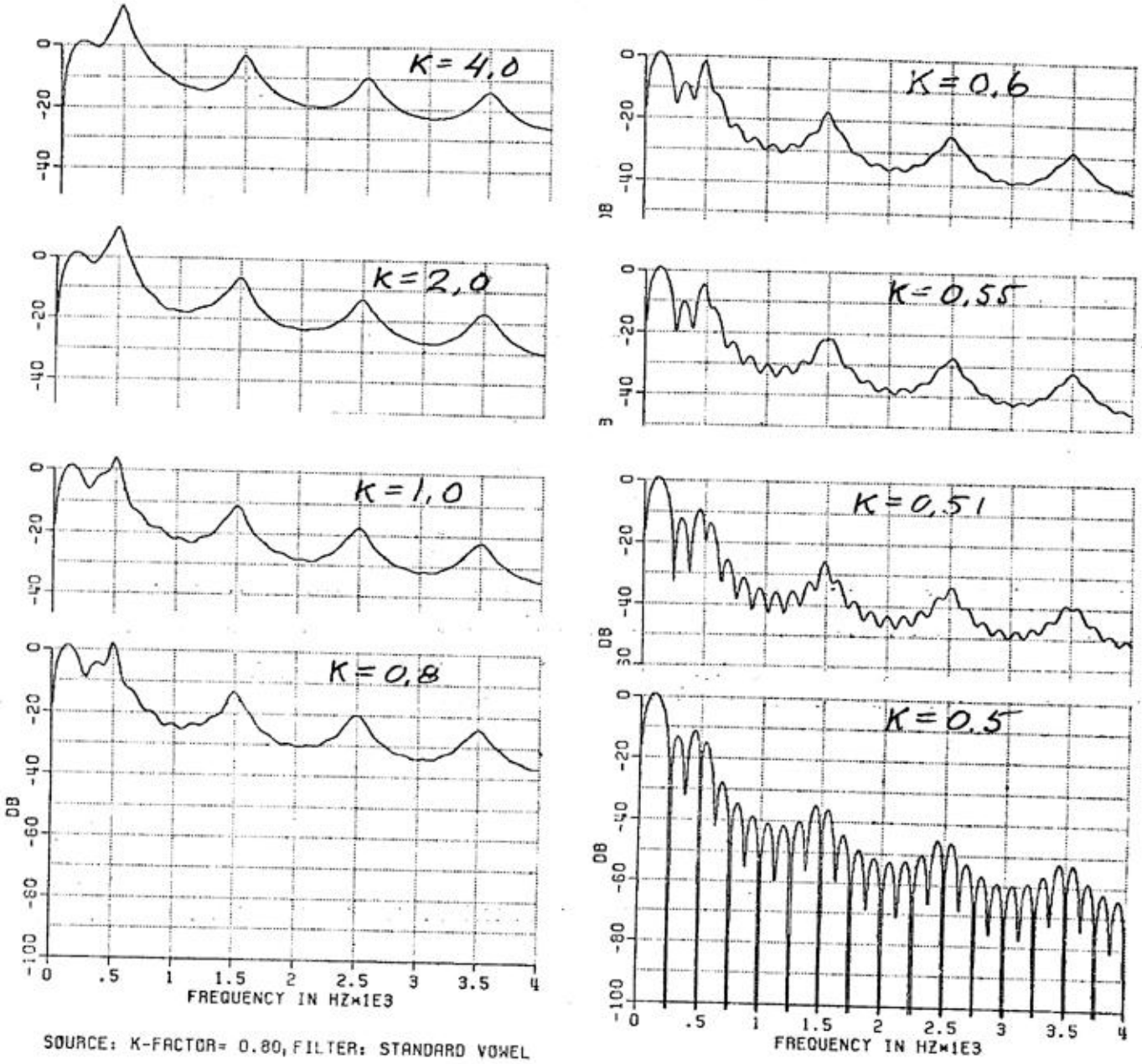


Fig. II-A-4. Voice source spectrum with added radiation correction. +6 dB/oct only.



F_g = 125 Hz

Fig. II-A-5. Combination of vocal tract single tube transfer function and source with F_g=125 Hz and various K-values.

This is the superimposed response of wave functions generated at glottal opening i=1, the glottal peak i=2, and at the glottal closure i=3. The novelty is that the voice source component in the speech wave is included and is represented by an undamped (α_2=0) sinusoid of frequency $\omega_n = \omega_g = 2\pi F_g$, which has a function equivalent to any formant

$$\begin{cases} F_n = \omega_n/2\pi \\ B_n = \alpha_n/\pi \end{cases} \qquad (9)$$

and follows the same rules with respect to frequency location patterns and formant amplitudes.

In the waveform calculations I have for simplicity assumed time invariant formant frequencies, F_n, but time variable bandwidths

$$B_n(t) = B_{on} + \frac{1}{t}\int_{T_1}^{T_1+t} B_{max}\frac{Ag(t)}{Agmax}dt \qquad (10)$$

where B_{on} represents supraglottal sources of bandwidths and B_{max} the additional damping through the glottis at the peak of glottal area Ag(t).

Examples with F_o = 100 Hz, F_g = 125 Hz, K = 1, F_1 = 500 Hz, B_{10} = 25 Hz, and B_{max} = 180 Hz respectively B_{10} = 35 Hz and B_{max} = 690 Hz were treated by Fant (1979). Here follows in Figs. II-A-6 - II-A-10 computer calculated waveforms including the examples above and some results from additional parameter variations. The upper left curve in each figure is the total speech wave and the upper right is the integrated version representing the flow output from the mouth (zero-order approximation to inverse filtering. Below follows in the left column the damped oscillation evoked at glottal closure, at glottal opening, and at the bottom the glottal flow. In the right hand column, second from top is the sum of all damped oscillations in the pressure wave. The third oscillogram is the damped oscillation derived from the peak of the flow. The bottom right hand graph shows the glottal component in the speech pressure wave which is proportional to the derivative of the input glottal flow. The pressure wave is essentially the linear sum of this voice source component and formant oscillations evoked at glottal closure. The Figs. II-A-6 - II-A-10 include one-formant system functions only.

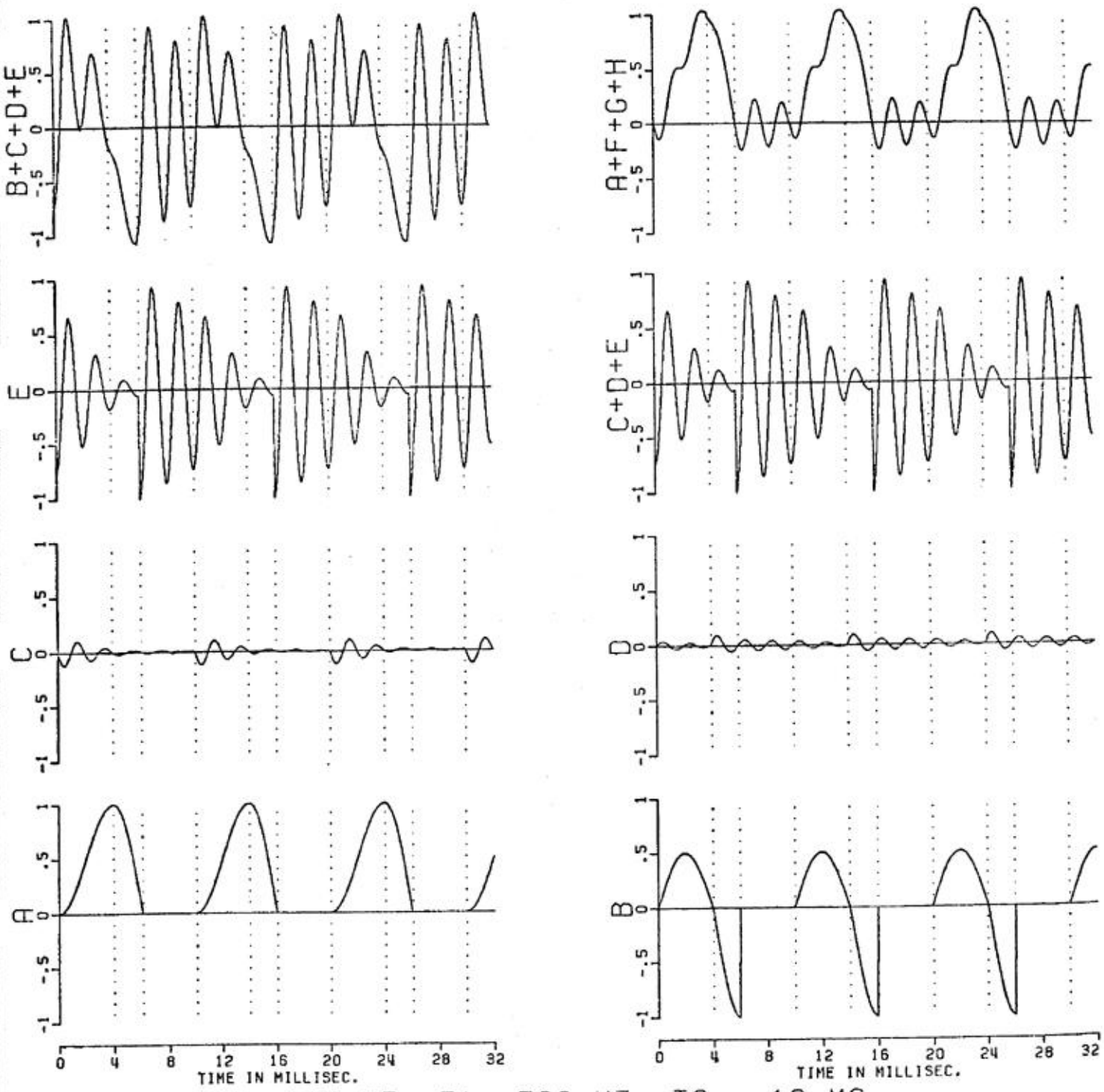


Fig. II-A-6. Example of model generated first formant and source components. A = source volume velocity, B = derivative of A = source component in the speech pressure wave. C, D, and E are F1-components of the pressure wave of glottal opening, peak and closing, respectively. B+C+D+E = total speech wave, A+F+G+H = integrated speech wave. BM = maximum instantaneous glottal bandwidth in Hz.

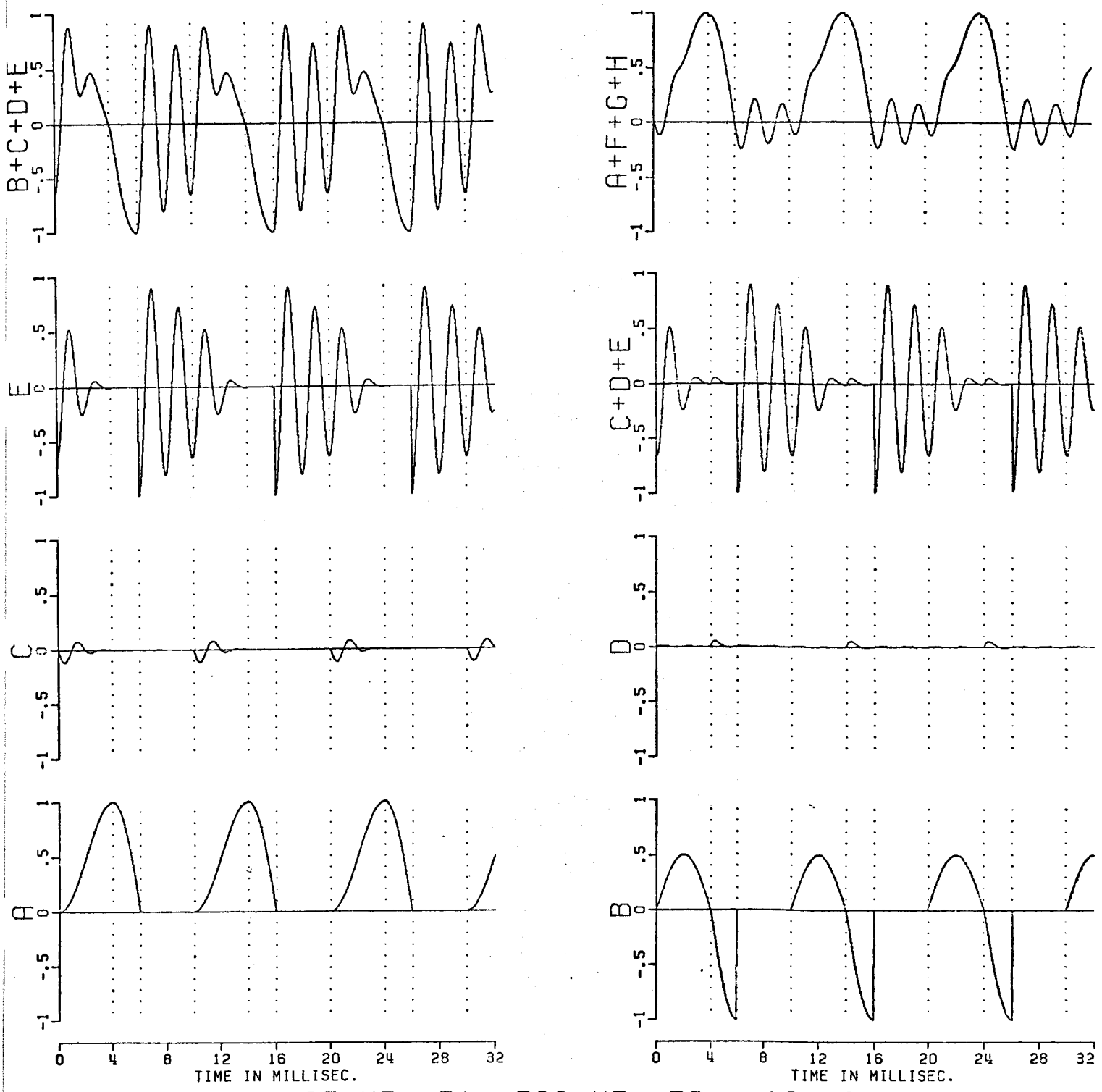


Fig. II-A-7. Same as Fig. II-A-6 with larger glottal bandwidth

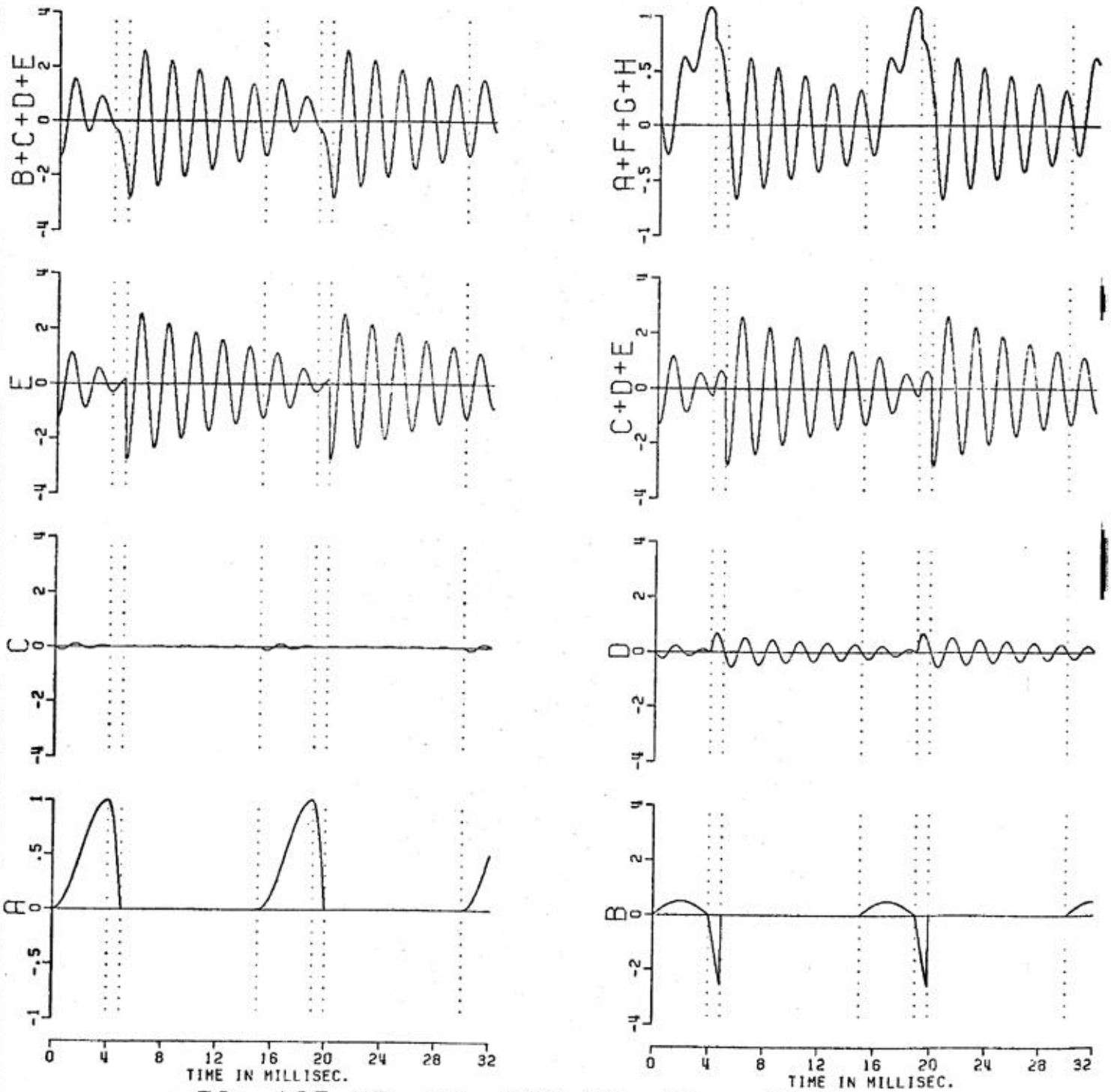


Fig. II-A-8. See Fig. II-A-6. Low F_o, high K-value, long duration of period of glottal closure.

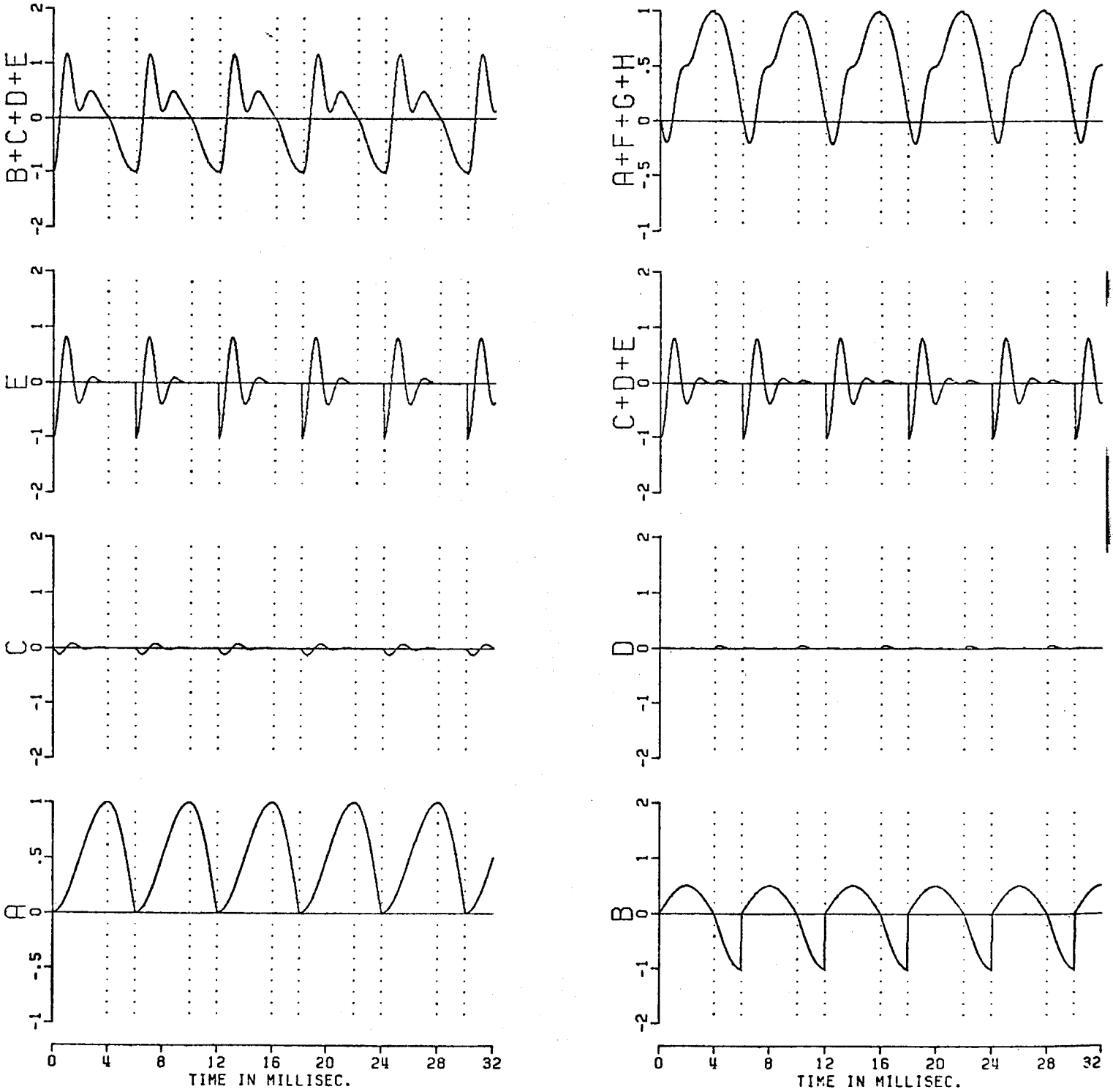


Fig. II-A-9. Extreme combination of zero duration glottal closure and large glottal damping, K = 1.

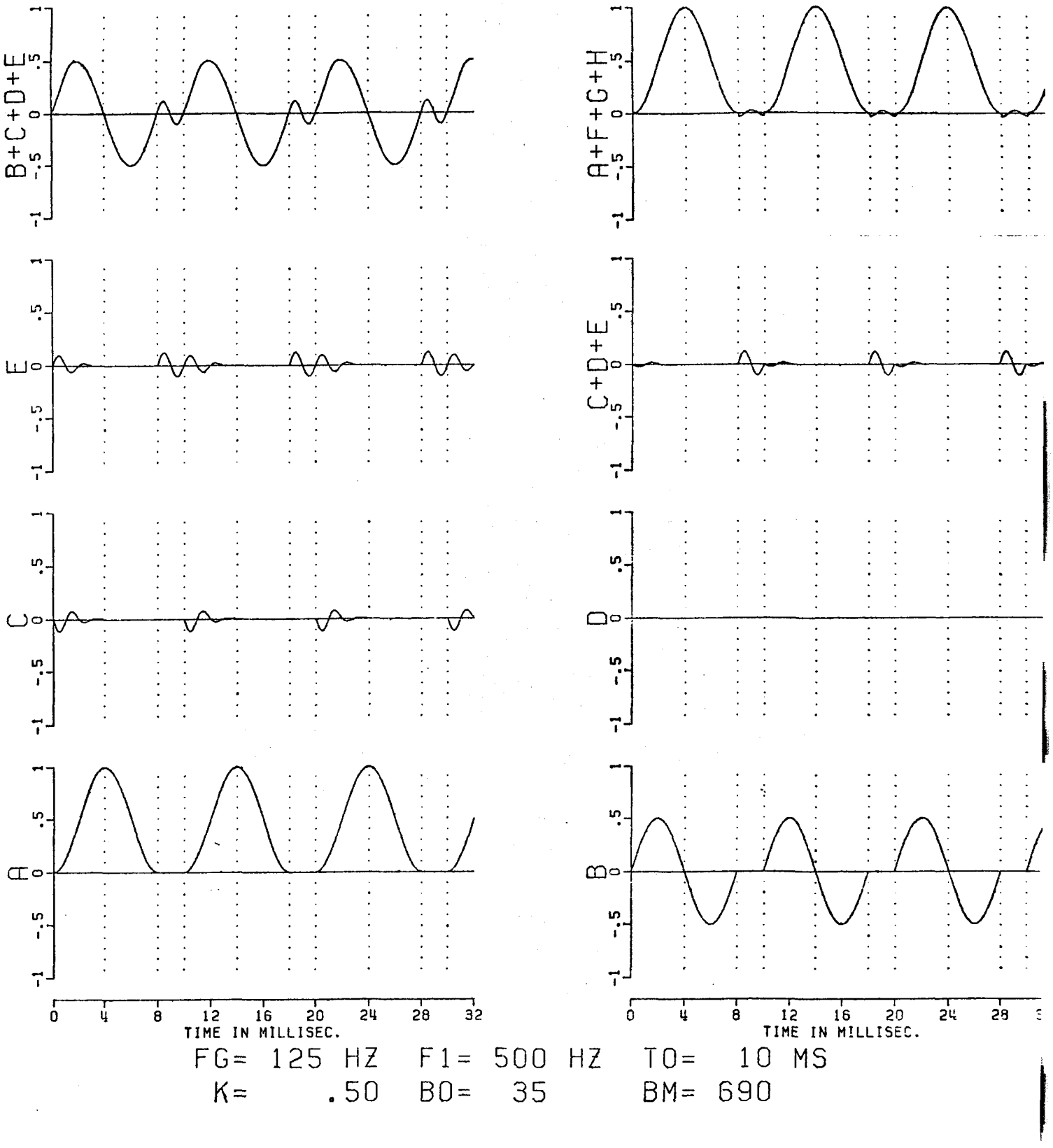


Fig. II-A-10. K = 0.5 and finite glottal closure phase.

Under identical conditions of a one-formant vocal tract system function the proportionality between formant amplitudes and glottal pulse slope at closure is illustrated by the continuity in the pressure wave of the initial amplitude of the F1-oscillation and the negative peak of the glottal component.

Two representative values of B_{max} in the glottal opening part have been adopted. The higher value, 690 Hz, could represent the F1-damping of the vowel [a] which is discussed by Wakita and Fant (1978) and could be thought of as an impedance match between the glottal resistance and the characteristic impedance of the pharynx as a terminating acoustic transmission line. This explains the truncated shape of the [a] vowel F1-oscillation envelope. However, even in case of a moderate glottal damping the carry over of energy from one period to the next is rather small and usually below -15 dB.

Fig. II-A-8 represents a case with a relatively small relative duration of the glottal open phase and a high steepness factor K of the falling branch. Here, the glottal component in the speech wave is less apparent. The integrated speech wave, on the other hand, retains the essential outline of the glottal pulse with added F1-ripple during glottal closure starting with a minus sine phase. In Fig. II-A-9 the closed glottal interval is zero and the pressure wave is clearly dominated by the glottal component. The integrated speech wave is similar to the glottal flow but for the superimposed F_1-component.

Fig. II-A-10, finally, illustrates a symmetrical glottal pulse shape with K = 0.5 combined with high glottal damping. The differentiated glottal flow pulse has the shape of a complete period of a sinewave which dominates the pressure wave. Because of the continuity there is no F1-excitation at the glottal flow peak and the components at closure and opening are of the same starting amplitude. The resulting F1-oscillation is small and confined to the glottal closure interval.

Analytical expressions. Generalized F-pattern concept

The analytical expressions, Fant (1979), underlying the graphical presentation above can be summarized as follows.

Any component in the pressure wave is a function of three major properties: source, radiation, and vocal transfer. These enter into the expressions for the formant oscillations as well as the glottal pulse residue in the speech wave. Extending the analysis to any number of conjugate complex poles (formants) we find for the main excitation, i=3 at glottal closure

$$v_n(t) = \sum_{n=1}^{n=\infty} A_n \cdot e^{-\alpha_{nt}} \cos(\omega_{nt} + \varphi_n) \qquad (11)$$

$$A_n = A_{ns} \cdot H_R \cdot H_{Tn} = (\text{source})\,(\text{radiation})\,(\text{transfer}) \qquad (12)$$

$$A_{ns} = -U_o \omega_g \sqrt{2K-1} \cdot \sqrt{1+d^2/b^2} = U_3' \cdot \sqrt{1+d^2/b^2} \qquad (13)$$

where $d = \omega_g(K-1)/\omega_n$ and $b = (2K-1)^{\frac{1}{2}}$; attain importance at $K < 0.7$ and ω_n close to ω_g.

$$H_R = \frac{\varrho}{4\pi a} \cdot K_T(\omega) \qquad (14)$$

contains the distance a from the mouth of the speaker and the K_T factor of Eq. (7).

For the F1-oscillation and a four formant system function

$$H_{T1} = \left[N_{g1}\right]^{-1} \left[1-F_1^2/F_2^2\right]^{-1} \left[1-F_1^2/F_3^2\right]^{-1} \left[1-F_1^2/F_4^2\right]^{-1} k_{r4}(F_1) \qquad (15)$$

$$H_{T2} = \left[N_{g2}\right]^{-1} \left[1-F_2^2/F_1^2\right]^{-1} \left[1-F_2^2/F_3^2\right]^{-1} \left[1-F_2^2/F_3^2\right]^{-1} k_{r4}(F_2) \qquad (16)$$

$$H_{T3} = \left[N_{g3}\right]^{-1} \left[1-F_3^2/F_1^2\right]^{-1} \left[1-F_3^2/F_2^1\right]^{-1} \left[1-F_3^2/F_4^1\right]^{-1} k_{r4}(F_3) \qquad (17)$$

$$H_{T4} = \left[N_{g4}\right]^{-1} \left[1-F_4^2/F_1^2\right]^{-1} \left[1-F_4^2/F_2^2\right]^{-1} \left[1-F_4^2/F_3^2\right]^{-1} k_{r4}(F_4) \qquad (18)$$

$$N_{gn} = \left[(1-F_g^2/F_n^2)^2 + (F_g/F_n Q_n)^2 \right]^{\frac{1}{2}} \qquad (19)$$

$k_{r4}(F_n)$ is the higher pole correction at the frequency F_n. The glottal component is the combination of functions at all three points of excitation resulting in the two successive parts: the rising and falling branches of the glottal pulse.

$$v_t = \frac{A_g}{2} \sin\left[\omega_g(t - T_1) + \psi_1\right] \quad (20)$$

$(T_1 < t < T_2)$

$$v_g(t) = -A_g\left[\sin\ \omega_g(t - T_2) + \psi_1\right] \quad (21)$$

$(T_2 < t < T_3)$

$$v_g(t) = 0 \quad (22)$$

$(t > T_3)$

$$A_g = A_{gs} \cdot H_R \cdot H_{Tg} \quad (23)$$

$$A_{gs} = U_o \cdot \omega_g \quad (24)$$

$$H_{Tg} = \left[N_{g1}\right]^{-1}\left[N_{g2}\right]^{-1}\left[N_{g3}\right]^{-1}\left[N_{g4}\right]^{-1} \cdot k_{r4}(F_g) \quad (25)$$

$$N_{g2} = \left[1 - F_g^2/F_2^2\right] \text{ etc.} \quad (26)$$

N_{g3} and higher factors are negligible. N_{g2} would be significant with very low F_2 only.

The terminating negative value at $t = T_3$ is $2\sqrt{2k-1}$ larger than the maximum at $t = \frac{1}{2}(T_1 + T_2)$.

With high Q-resonances the initial phases of the wave functions are small. They have been neglected in the present numerical calculations. When F_1 approaches F_g the effects may be more apparent, see expressions in Fant (1979). The glottal component and the first formant will execute a mutual reinforcement similar to the case of two formants approaching in frequency. A close analysis has not been made yet. One could anticipate a relative greater importance of this effect in female voices where F_g has a greater tendency to come into the F_1 range.

Another conclusion is that the voice source components in the pressure wave may differ from those derived by a simple differentiation of the actual glottal flow or from a complete inverse filtering. With the inverse filter tuned to the F_1 and B_1 of the glottal closed period there will result an incomplete cancellation of the remaining F_1 ripple in the glottal open period. This effect is generally rather small.

The source model is defined to be independent of the vocal tract transfer function and will thus by definition exclude superimposed formant ripple. The areodynamic interaction between glottal mechanics and the vocal tract is assumed to be taken into account before the source parameters are determined.

Some experiments

FM taperecordings were made of sentences spoken through a Brüel & Kjaer condenser microphone and played back at 16 times reduced speed through an inverse filter with two modules, one for F1 and one for F2, designed for the reduced speed and tuned to F_1 = 710 Hz and F_2 = 1370 Hz effective frequency.

Oscillographic recordings were made on a Mingograph run at a speed of 10 cm/sec, i.e. effective 160 cm/sec. The four functions selected for analysis were the inverse filter input and output, the integrated input, and the output with F2-cancellation only. The integration was performed with a 10 Hz RC low pass filter.

The two [a]-samples in Fig. II-A-11 were the stressed and unstressed initial vowels of the sentences "Axel var där" and "Aksell var där", respectively, see spectrograms in Fig. II-A-13. Typical of [a]-vowels the oscillogram reveals the presence of the differentiated glottal flow component, the latter part of which executes a clear negative spike as expected and the tendency of F1 being damped out effectively already in the opening phase. The last periods before the occlusion of the following consonant [K] exhibit a lack of F1-oscillation, which may seem remarkable in view of the presence of clear glottal spikes with K = 1. The most probable explanation is that the vocal cords executed incomplete closure in these transition periods towards an abducted stage, which greatly increases the first formant bandwidth and causes immediate cancellation.

The first few periods display a gradual transition from closure to opening which reduces the slope of the rise and, accordingly, increases the calculated K-values.

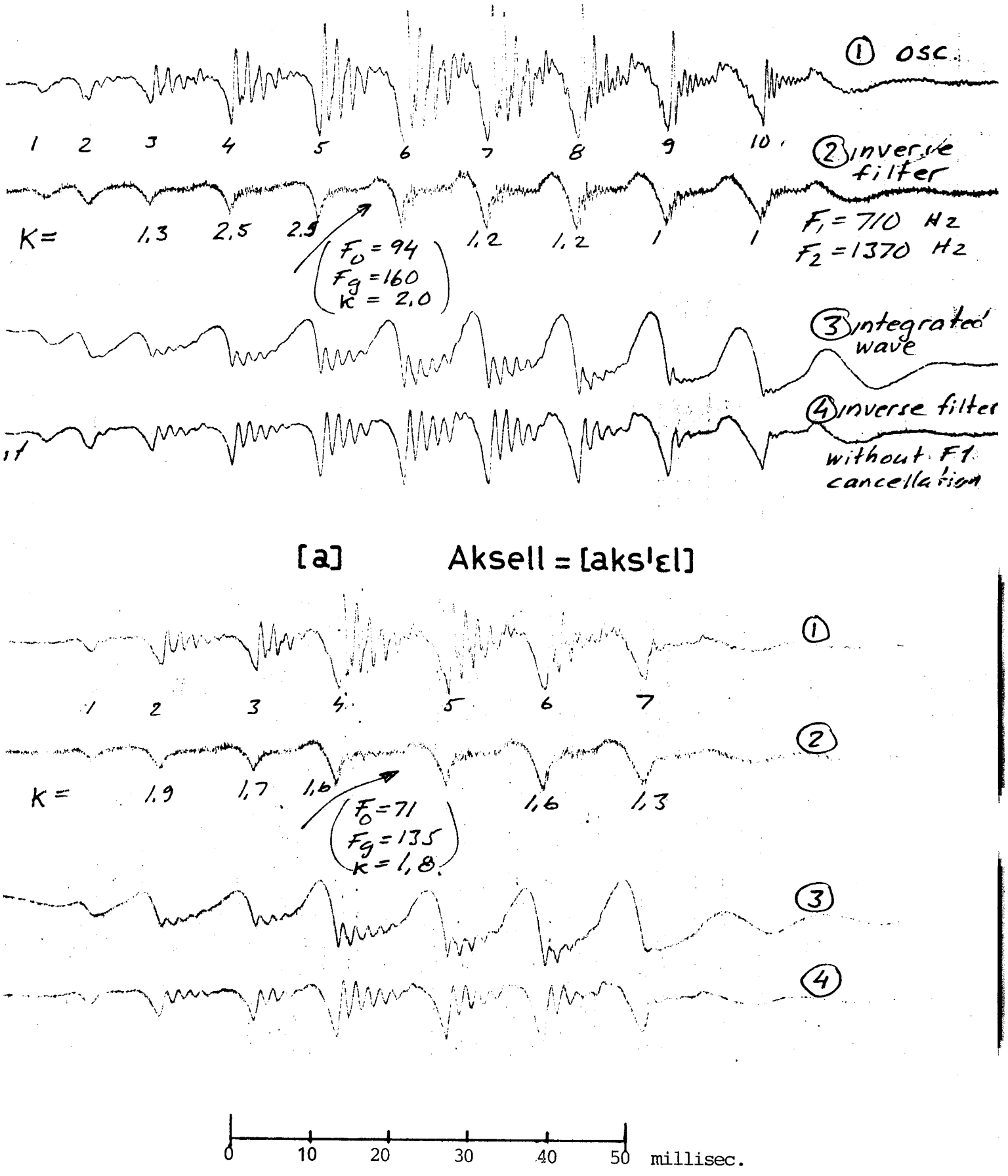


Fig. II-A-11. Oscillographic recordings of stressed vowel [a] top and unstressed vowel [a] below. (1) Speech pressure wave. (2) Inverse filtering with fixed F_1 and F_2 and differentiation to bring out glottal component in the speech pressure wave. (3) Integration of speech pressure wave. (4) Inverse filtering of F2 only to bring out F1.

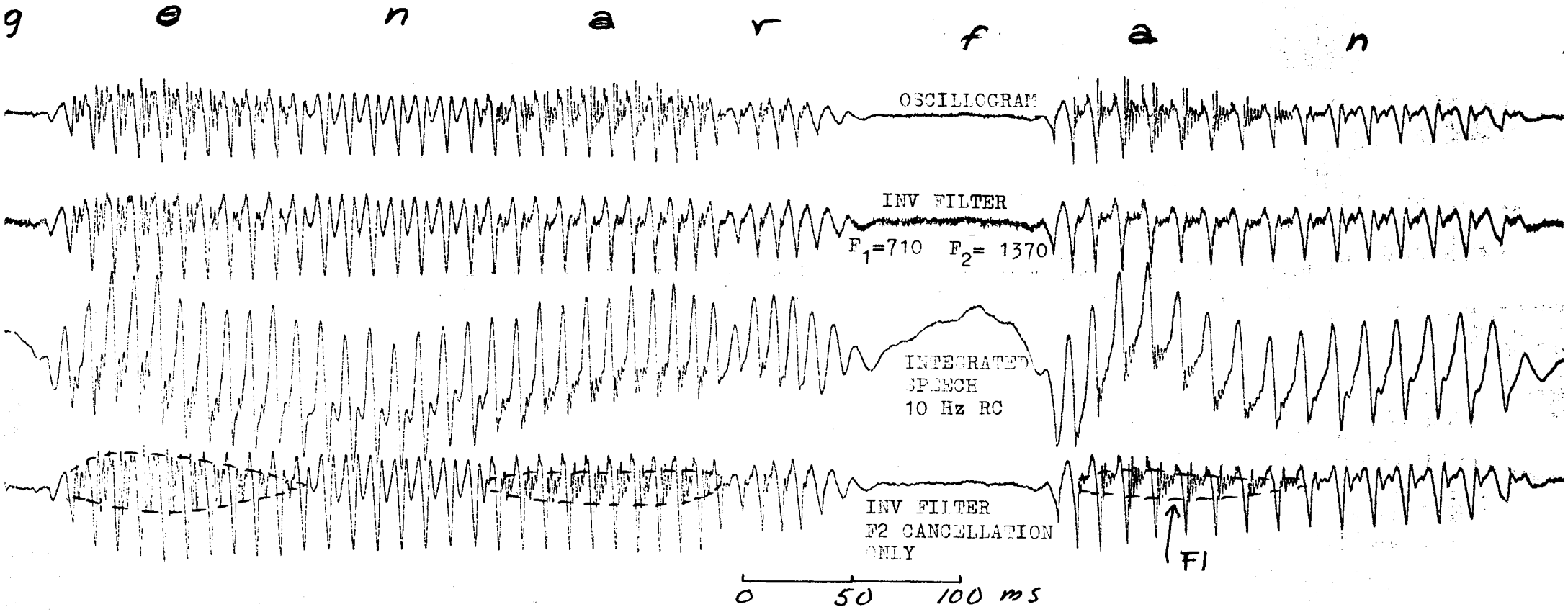


Fig. II-A-12. Same analysis as in Fig. II-A-11 applied to the text "Gunnar Fant". Notice F1 amplitude reduction in vowels adjacent to nasal consonants and in the first voice period after the consonant [f] .

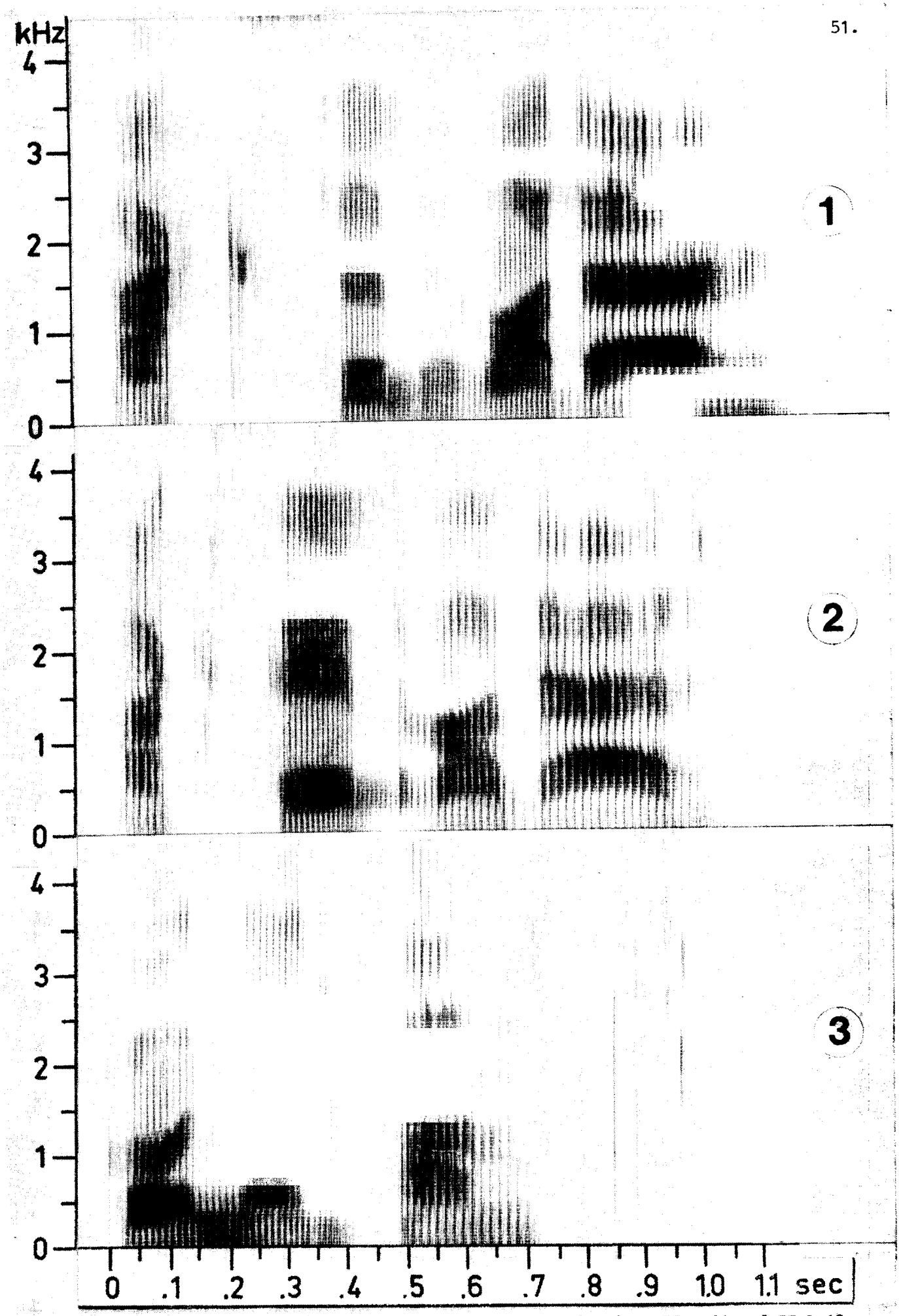


Fig. II-A-13. Spectrogram of the same utterances as in Figs. II-A-11 and II-A-12. (1) "Axel var där", (2) "Aksell var där", and (3) "Gunnar Fant".

We do not yet have any statistics of the variation of glottal pulse parameters with lexical stress patterns. In the example of Fig. II-A-11, the stressed vowel [a] differs from the unstressed vowel [a] by about 30% greater values of both duration, first formant amplitude, glottal slope U_3', and F_o, and about 20% higher F_g and K.

Excessive F1-damping does not only occur in voiced, aspirated segments. It is a basic characteristics of nasalization. The sentence "Gunnar Fant", see Figs. II-A-12 and II-A-13, shows the progressive F1-damping in the nasalization of vowels before and also after [n]. The basic conclusion is that what is seen as a baseline formant below F_1 in the spectrogram, is not a nasal formant but merely a glottal spectrum component which remains the same as in the unnasalized parts of the vowels but gains relative prominence when F1 is attenuated. Because of the low frequency deemphasis in the spectrograph, this "glottal formant" is not always seen.

A few words should be said about the procedure of measuring the glottal parameters. The spread in determining F_g and K is of the order of 10% for a single measurement and often greater, especially in the starting and terminating periods of voicing. There exist two methods of determining K from the glottal flow, and one from the differentiated flow. In the first instance, we draw a tangent to the glottal pulse at the point of closure and adopt Eq. (5).

$$K = \frac{1}{2} + \frac{1}{2}\left(\frac{U_3'}{U_o\,\omega g}\right)^2 = \frac{1}{2} + \frac{1}{2}\left(\frac{T_2 - T_1}{\pi \cdot T_d}\right)^2 \qquad (27)$$

where $T_d = U_o/U_3'$ is the effective closing time.

A second approach is to measure the ratio of the slope at closure and the maximal slope of the rising branch, either from the glottal flow curve or from the differentiated function. This ratio is:

$$a = U_3'/U_{12}' = T_b/T_d = 2(2K - 1)^{\frac{1}{2}} \qquad (28)$$

$$K = \frac{1}{2} + \frac{a^2}{8} = \frac{1}{2} + \left(\frac{T_b}{T_d}\right)^2 \cdot \frac{1}{8} \qquad (29)$$

Identity of Eg. (29) and (27) requires

$$T_b = \frac{2}{\pi} (T_2 - T_1) \qquad (30)$$

as an inherent property of the model which may match the actual flow with various degree of accuracy. The maximum error in T_b is apparently $(2/\pi)$ when the rising branch is a straight line instead of a double sinusoid. A symmetrical triangular flow pulse would have $K = 0.55$ according to Eq. (27) and $K = 0.625$ from Eq. (29).

Because of the uncertainty in determining K and F_g one could choose the parameters U_o F_g and U_3' instead of U_o F_g and K in which case K if needed is calculated from Eq. (27). Other alternative parameter sets exist, e.g. U_o $(T_2 - T_1)$, and T_c or U_o $(T_3 - T_1)$, and T_c. These are uniquely convertible to the basic model parameters U_o K and F_g and might have a more general significance. We need much more experience from analysis of different types of voices and contexts in order to test the validity of the models and the significance of the parameters.

An instrumental source of error lies in an imcomplete or otherwise incorrect representation of the correction for higher poles. In our experiments with the provisional inverse filter we could derive an error of about 10-20% in K-values from a low pass filter of 1000 Hz inherent in the passive network. One should also see to that the cutoff frequency of an integrator should be below 10 Hz.

References:

Fant, G. (1960); Acoustic Theory of Speech Production, Mouton, The Haague; 2nd edition 1970.

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

Acknowledgments

The author is indebted to Thomas Murray for the computer calculations of spectra and waveforms.