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

Voice source dynamics

Fant, G.

journal: STL-QPSR
volume: 21
number: 2-3
year: 1980
pages: 017-037

KTH Computer Science and Communication

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

II. SPEECH PRODUCTION

A. VOICE SOURCE DYNAMICS*

G. Fant

Abstract and Introduction

This is a report on some preliminary studies of voice source parameters in connected speech. This task requires an understanding of the underlying voice production mechanism in its details within a voice fundamental period as well as in the sequence of events within a succession of voice periods and with respect to systematic effects evoked by various modes of laryngeal and supraglottal articulations. Special attention has been devoted to boundary regions between voicing and silence or unvoiced segments. One of the objectives has been to study how speech prosody, e.g., various stress patterns may be correlated to source parameters other than fundamental frequency and timing. In the course of this work it has become apparent that we have much less insight in glottal than in supraglottal articulations and that we still have far to go in understanding some of the observed phenomena. The ultimate practical goal is to develop rules for synthesis that will improve naturalness and enable us to cope with different voice types and speaking styles.

A voice source model

The voice source model adopted by Fant and described in STL-QPSR no. 1 and 3-4/1979 generates a train of volume velocity pulses specified by three shape parameters (Fant, 1979b & c). As illustrated in Fig. II-A-1, these are U_0 = Peak volume velocity flow in cm^3/sec; F_g = glottal frequency defined as $1/(2(T_2-T_1))$, where T_1 is the time coordinate of the pulse onset and T_2 of the pulse maximum. One could conceive of alternative parameters related to pulse duration. K = steepness of "asymmetry factor" related to the ratio of closing and opening speeds.

Initial amplitudes of formant oscillations evoked at the point of main excitation, i.e., at the closing edge of the pulse, are proportional to the rate of change of flow: $U' = dU/dt = U_0 2\pi F_g (2k-1)^{\frac{1}{2}}$, immediately ahead of the discontinuity. Accordingly, all three glottal pulse parameters enter into the expression for formant amplitudes.

Fig. II-A-2 shows the corresponding source spectrum with +6 dB/oct added to include radiation transfer. At constant glottal frequency F_g, the parameter K generates a set of curves which have almost the

---

* Paper given at the ICA X, Sydney, Australia, July 1980.

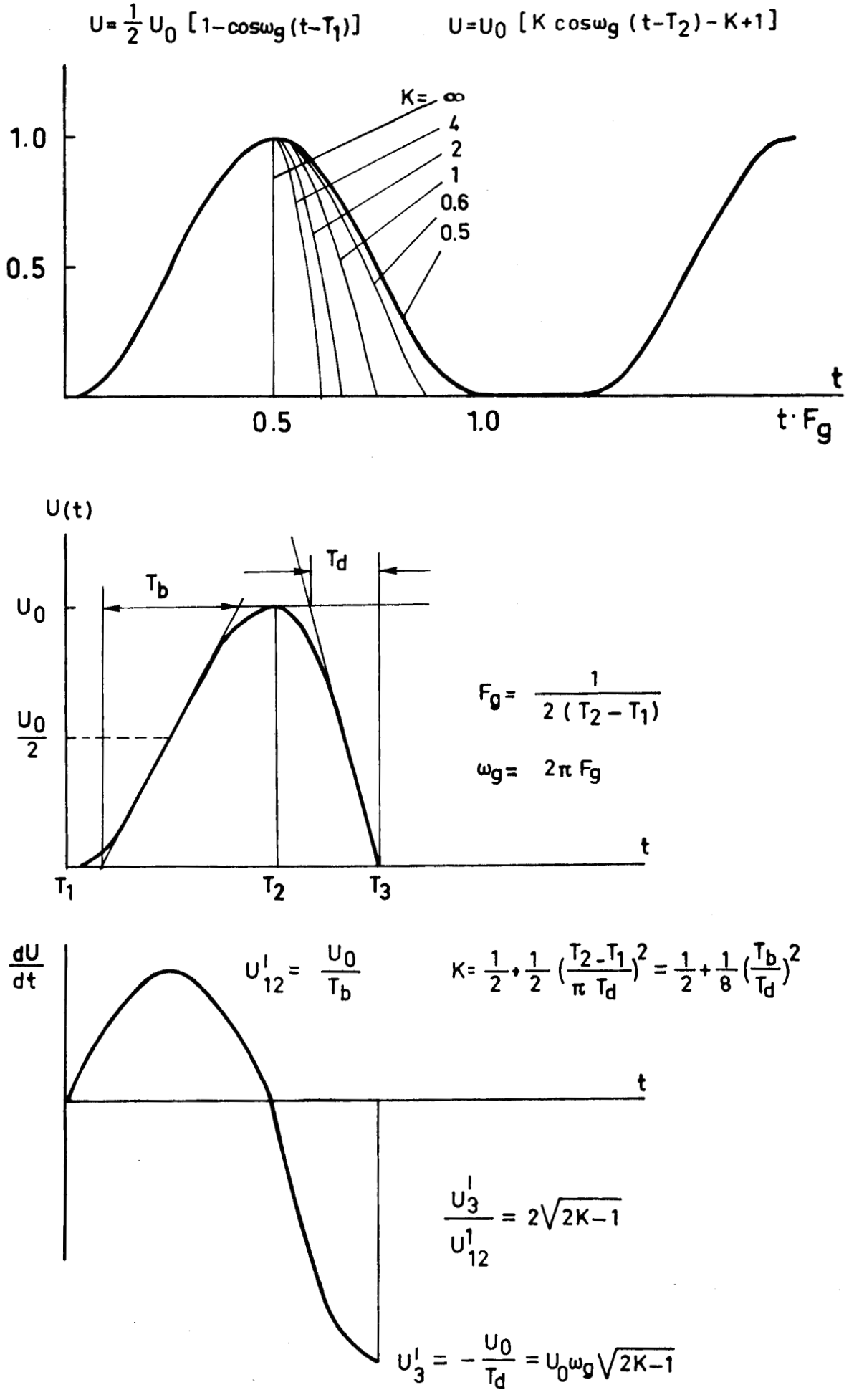


Fig. II-A-1. The voice source model adopted by Fant (1979b).

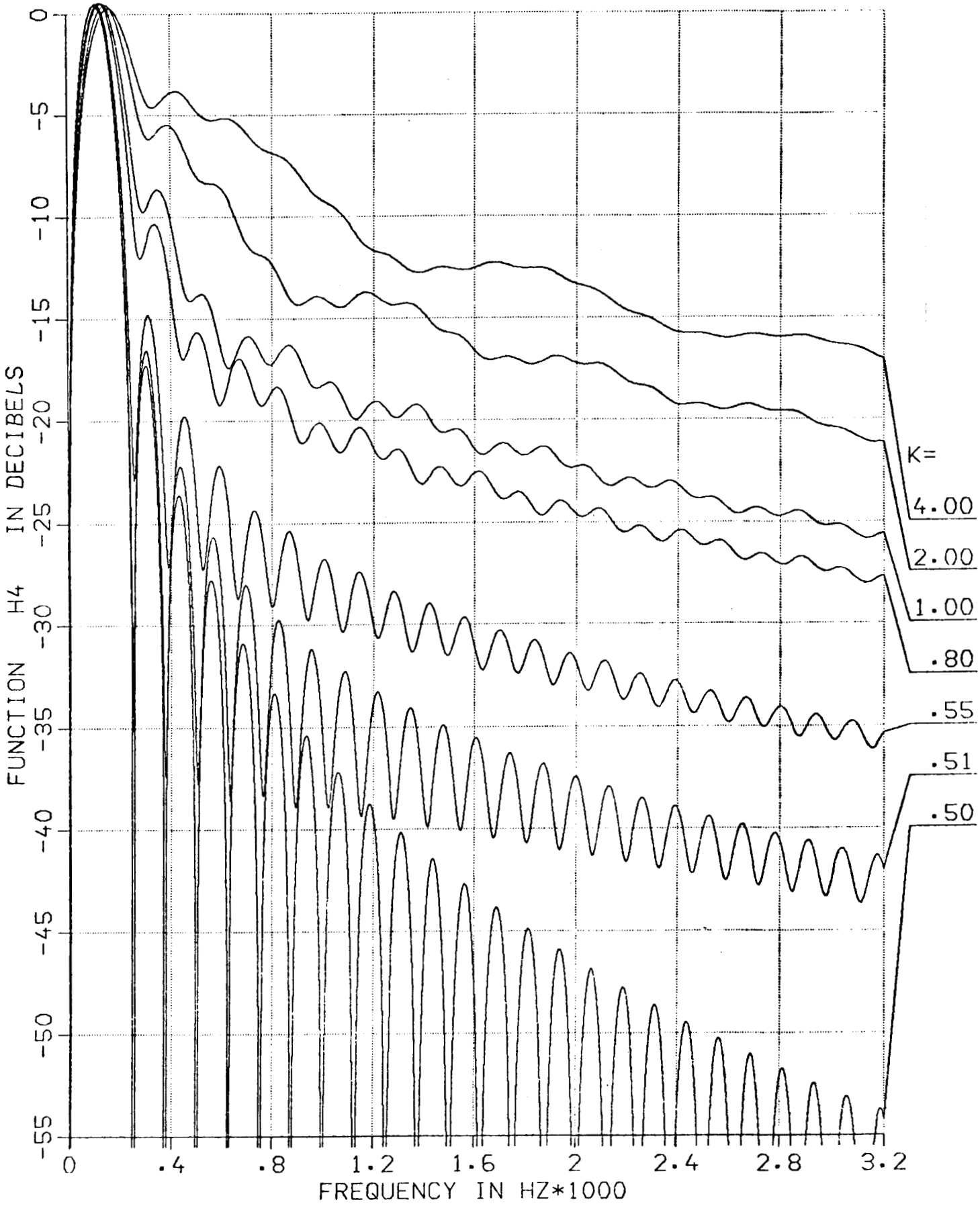


Fig. II-A-2. Voice source spectrum with added radiation correction. +6 dB/oct only.

same rate of fall 6 dB/oct at higher frequencies. The difference induced by the parameter K is the spectrum level in the range well above F_g which rises with K.

At low K-values the low frequency part dominates and accounts for a relative prominence of a base-band area below F_1 in the spectrogram which reinforces the fundamental and, at low fundamental frequencies, also the second and third harmonics.

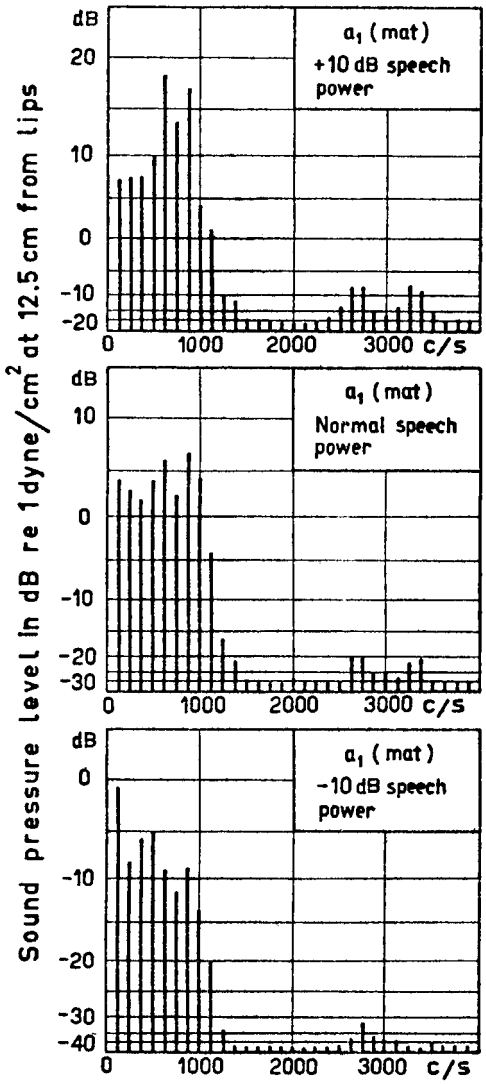
The relative stability of the amplitude of the voice fundamental with increasing voice effort is well known. Fig. II-A-3 from Fant (1959) illustrates how the fundamental in an [a]-spectrum dominates at very low voice effort but is weak compared to first formant amplitude at very high voice efforts. The corresponding overall effects in long-time spectrum of speech are also included in Fig. II-A-3.

These non-uniform spectrum shifts, when viewed in terms of the voice source model, involve changes in both K, F_g, and U_0. The effect of increasing F_g is to rescale the frequency calibration of the source spectrum which implies a horizontal shift of the curves, increasing the spectrum level at higher frequencies. This is one of the stress correlates.

Three different pulse shapes providing the same higher frequency spectrum level but of varying low frequency spectrum shapes are shown schematically in Fig. II-A-4.

They all have the same flow derivative at closure. The low frequency level is basically proportional to the total air volume of the vocal pulse. The low amplitude short vocal pulse would be characteristic of a medial compression of the vocal cords providing a large formant amplitude at low level voice fundamental. The high amplitude and more symmetrical pulse could be typical of a female voice which would conform with a relative dominance of the fundamental.

It should be appreciated that the time domain correspondence in the sound pressure wave of the low frequency source component is the differentiated glottal flow.



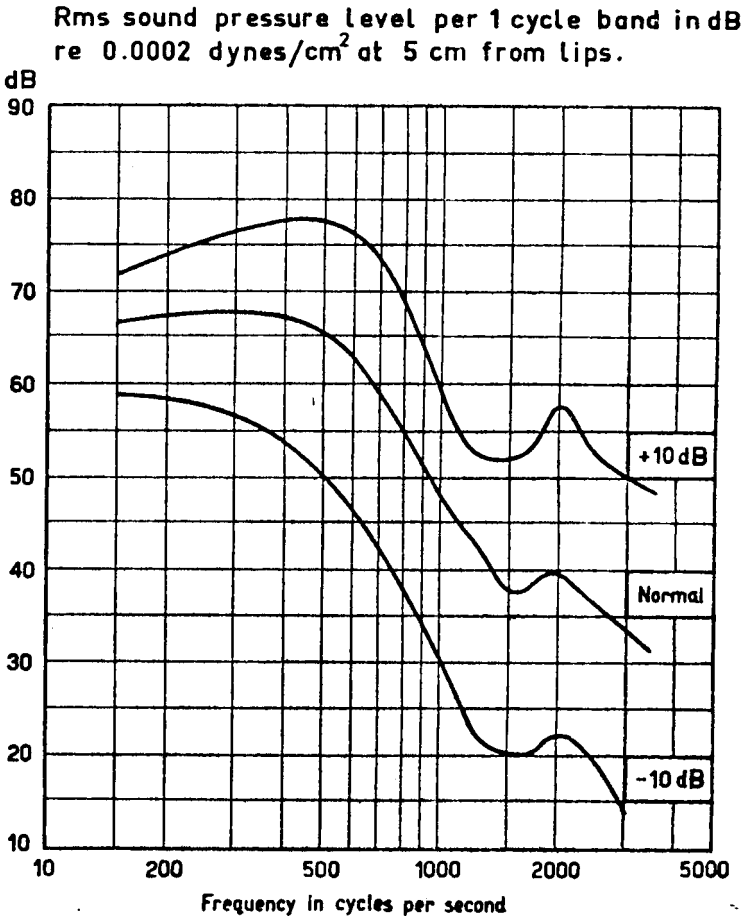
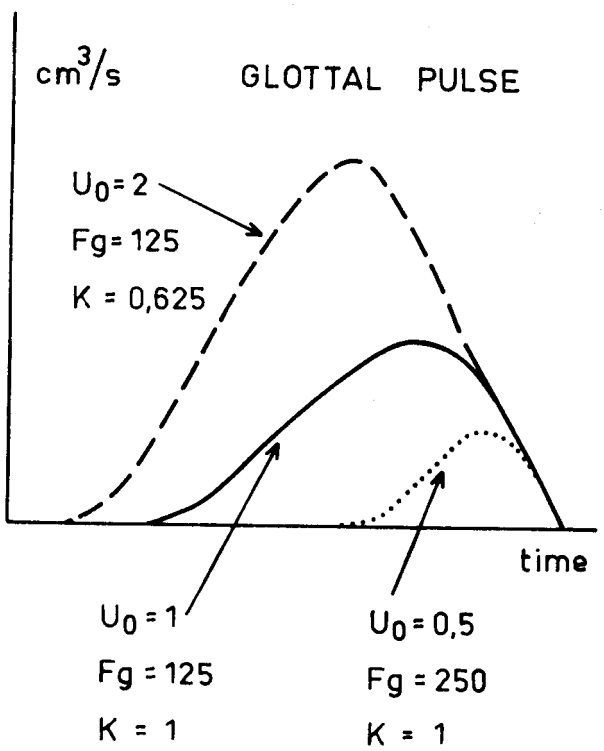


Fig. II-A-3.
Non-uniform spectrum changes with varying voice intensity (from Fant, 1959).



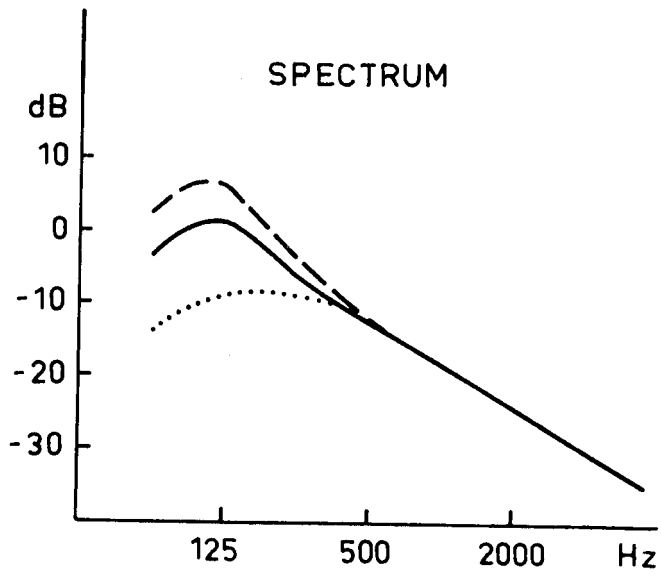


Fig. II-A-4. Glottal flow waveform and spectrum exemplifying the low frequency source spectrum level being proportional to total volume contained in a pulse whilst the spectrum level at frequencies above 2 F_g is proportional to the change in flow derivative at the instant of closure.

A time domain derivation of the sound pressure wave is illustrated in Fig. II-A-5. The sound is decomposed into damped oscillations (one formant illustrated here only) and a component which is the differentiated version of the glottal flow function. The negative spike of this component has ideally the same amplitude as the first formant oscillation. An additional feature, shown in Fig. II-A-5, is the rapid damping out of the formant oscillation in the interval of glottal opening. For K-values less than 1, the maximum of the derivative occurs before closure. It is then a smaller derivative at closure which determines the excitation. For K = 0.5, the derivative is a sine wave without discontinuities.

Inverse filtering experiments

Inverse filtering is a well established technique, Miller (1959), Lindqvist (1965), Holmes (1976), Rothenberg (1973), and Sundberg & Gauffin (1978). The experiments reported here were made by first recording the speech with a high quality condenser microphone in an anechoic chamber on a FM tape recorder. The analysis was later performed at 16 times reduced play-back speed through an analog four-formant inverse filter, especially designed for the scaled-down frequency range. A four-channel Oscillomink was used for the graphic recording. The following functions were selected:

(1) The speech pressure wave,

(2) The output of the inverse filter representing differentiated glottal flow.

(3) The inverse filter output integrated to represent glottal flow. This is what is usually meant by inverse filtering.

(4) The same as (2) but excluding the F1-unit thus simulating the response to vocal tract transfer function composed of F1 alone, as in Fig. II-A-5.

In some instances the order between channels (3) and (4) were reversed.

A typical example is shown in Fig. II-A-6. It conforms well with the time domain derivations in Fig. II-A-5 and the more extensive examples in Fant (1979c). The pulse form departs somewhat from the ideal by a non-flat level between pulses. This is in part a matter of the particular voice mode and in part a manifestation of

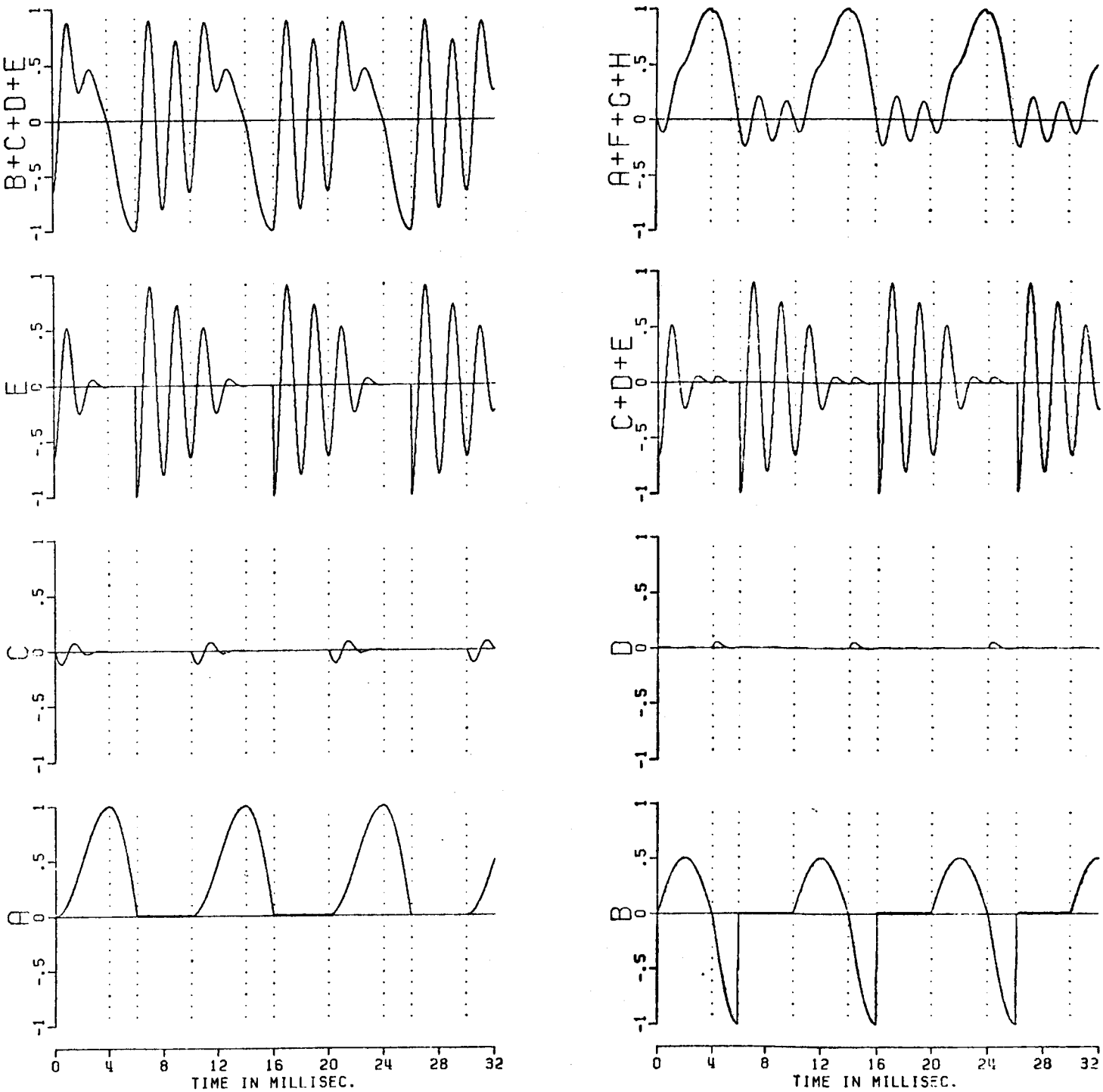


Fig. II-A-5. The same as Fig. II-A-2 with larger glottal bandwidth.

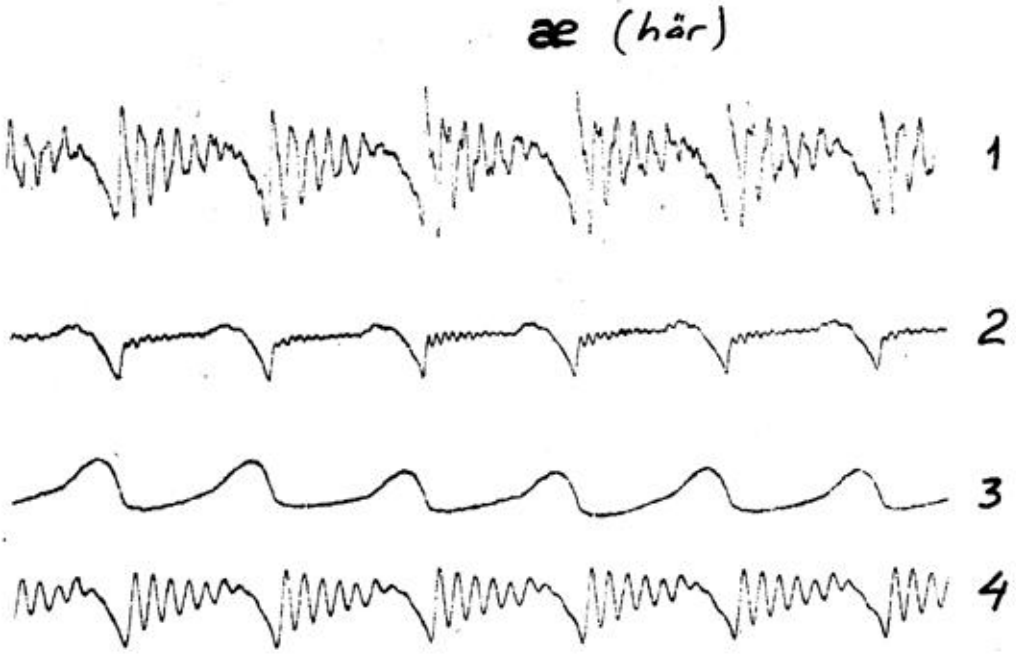


Fig. II-A-6. (1) Oscillogram. (2) Inverse filtered wave without integration. (3) = (2) with integration. (4) = (2) without F1-cancellation.

low frequency noise from the recording affecting the zero-line. The rapid decay of formant oscillations in the glottal open part is apparent as well as the apparent continuity between the negative voicing pulse amplitude (i.e., the flow derivative at closure) and the initial amplitude of the following F1 oscillation.

This phenomenon which inherently preserves the proportionality between excitation and formant amplitude has been experimentally verified by Gauffin and Sundberg in their paper (in this issue of STL-QPSR) on stationary voice qualities. The interesting feature about voice source dynamics is that this proportionality does not hold in boundary regions affected by vocal cord abduction, e.g., devoicing and aspiration. Nasalization causes similar effects of reducing first formant amplitude, whilst the peak amplitude of the differentiated glottal flow remains relatively constant. Fig. II-A-7 illustrates an intervocalic voiced [h]. In the following [ae] there is a perfect synchrony between negative source spike and the pulse edge at closure. Moving back to the [h]-sound, the spikes are seen to occur more in the down hill of the falling branch and the differentiated flow as well as the speech wave attains the shape of a full wave rectified sine wave, which lacks or shows very weak traces of F1 and has a fine structure of random noise. The F1-oscillation amplitude grows in the opposite direction looking back towards the following vowel [ae]. The bottom part of Fig. II-A-7 is the result of an analysis performed by Martin Rothenberg with a subject speaking into a flow mask. This provides a different picture, probably because of a relatively greater degree of devoicing.

In Fig. II-A-8, the F1-reduction in the first period after the [f] in the word "Fant" is apparent and also the gradual F1 amplitude reduction when the nasal boundary is approached. The F1-reduction is also quite apparent in the vowel [a] preceding the [k]-occlusion in the words "Axel" and "Aksell" of Fig. II-A-9. This can be conceived of as a matter of F1-cut back in the pre-closure aspiration and may, in part, be explained by increased formant damping associated with glottal abduction. A change in the excitation function is probably also involved.

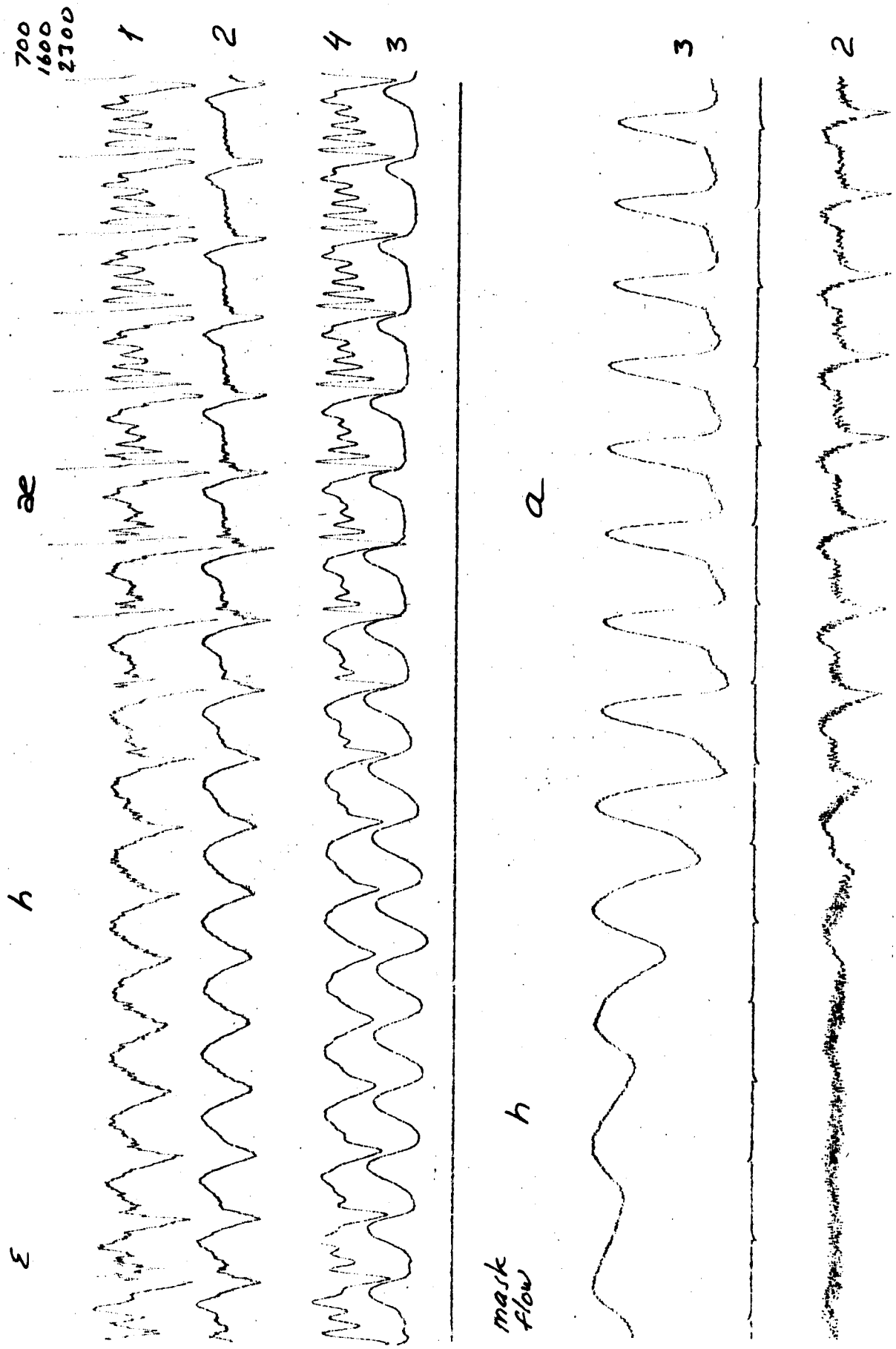


Fig. II-A-7 a) The same functions as in Fig. II-A-6 with (3) and (4) exchanged illustrating source dynamics in intervocalic [h] and a following vowel. Inverse filtering setting was constant. b) Rothenberg mask output and inverse filtered function of intervocalic [h] with a following vowel [ɑ].

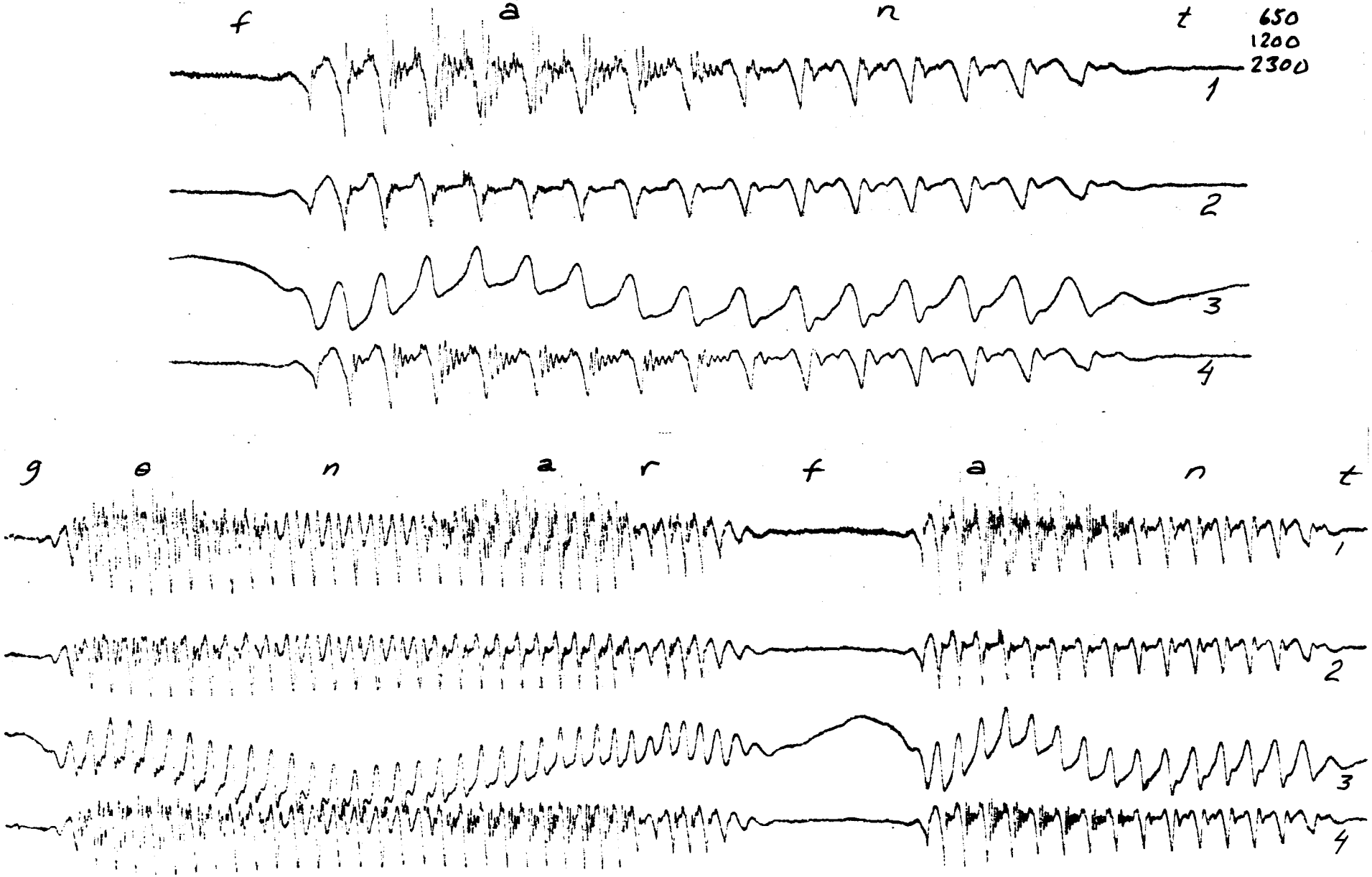


Fig. II-A-8. Function the same as in Fig. II-A-6. Text above "Fant", below "Gunnar Fant". Note the reduction of F1-ripple in vowel boundary regions, after a fricative consonant and before and after a nasal consonant. Compare with spectrogram of Fig. II-A-10. The F1-reduction causes the source spectrum max.F_g to gain rel.prominence.

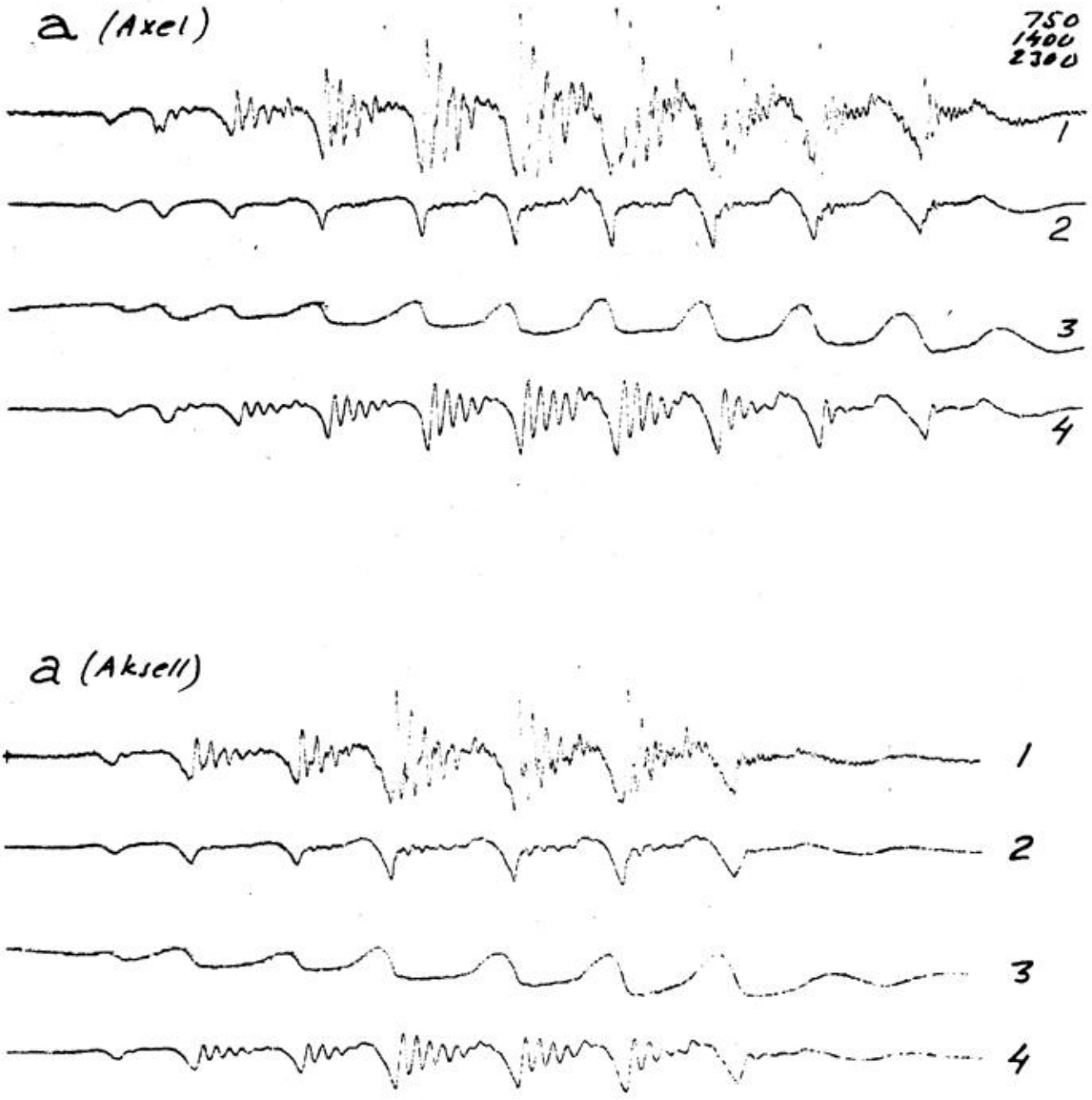


Fig. II-A-9. The same functions as in Fig. II-A-6. The [a] vowels in the stress-contrastive words: "Ax$\overset{4}{\underset{0}{e}}$l" and "Aks$\overset{0}{\underset{4}{e}}$ll". Note the pre-occlusion aspiration. Source function is more prominent in [$\overset{4}{a}$] than in [$\overset{0}{e}$].

The ascending part of the negative excitation pulse rises more gradually which reflects a continuation of vocal cord movements towards closure. This might be the effect of asynchrony between the mucous layer and the inner parts of the cords. It is also conceivable that there is some interior "inverse filtering" from the coupling to the subglottal tract which should have an anti-resonance near the frequency of F1 of [a].

The nasal system apparently imposes some internal inverse filtering on F1. As F1 is reduced, the F_g base-band component of the source attains prominence and is seen in spectrograms as a "base-band formant" which may occur alone or together with a sub-F1 nasal formant. Thus, what has been believed to be a base-band nasal formant in the nasalized vowel is often a source feature.

Prosody correlates

The source correlates of lexical stress patterns are not very distinct. The Swedish words $\overset{4\ 0}{\text{Axel}}$ and $\overset{0\ 4}{\text{Aksell}}$ differ in terms of conventional stress assignment as 4-0 and 0-4, respectively. The greater stress of the [a] in Axel is usually manifested by increased duration, as well as the higher F_0 and F_g and, to some extent, larger K and thus excitation strength. However, the main difference between the two contrasting words above appears to be the duration of the second vowel, Fig. II-A-10. In the examples of Figs.II-A-11 and II-A-12, it may be seen that the vocal pulse amplitude and shape is almost identical for the [e] in Axel and Aksell. There are also examples of the [a]-vowel having the same duration. One aspect of the stress in the [a] of Axel is the increased duration of the [ks]-cluster. This appears to be rather consistent with speakers. The same rules apply to the lengthening of the [l] after a stressed vowel.

Voiced consonants except nasals appear to have reduced glottal pulse height and flow derivative which account for a lower F1-amplitude than adjacent vowels. If the consonantal constriction is appreciable, the effect is maximal and it appears that the interval between flow peaks lacks a flat "closed" part.

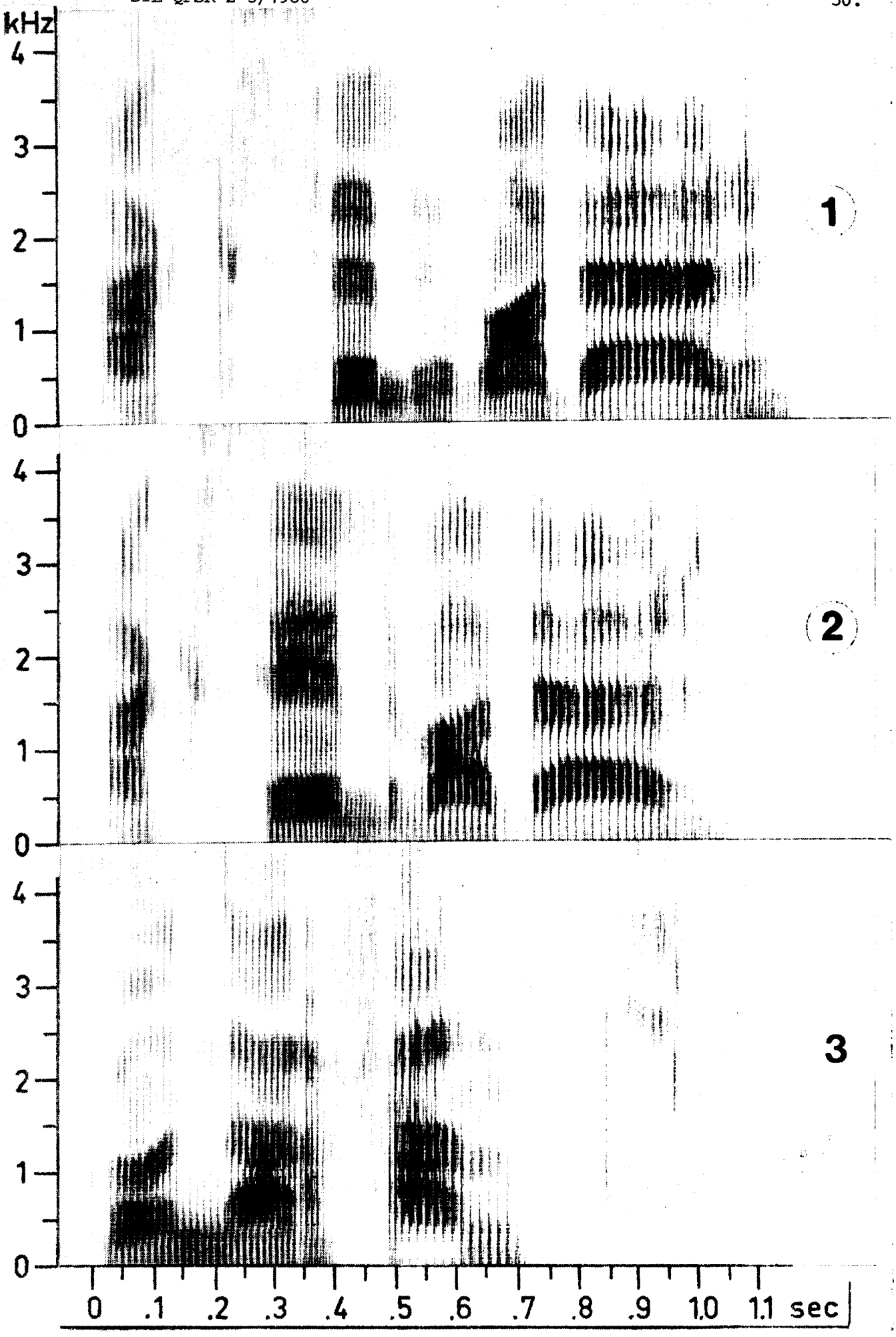


Fig. II-A-10. Spectrogram of the same utterances as in Figs. II-A-8 and II-A-9. (1) "Axel var där", (2) "Aksell var där", and (3) "Gunnar Fant".

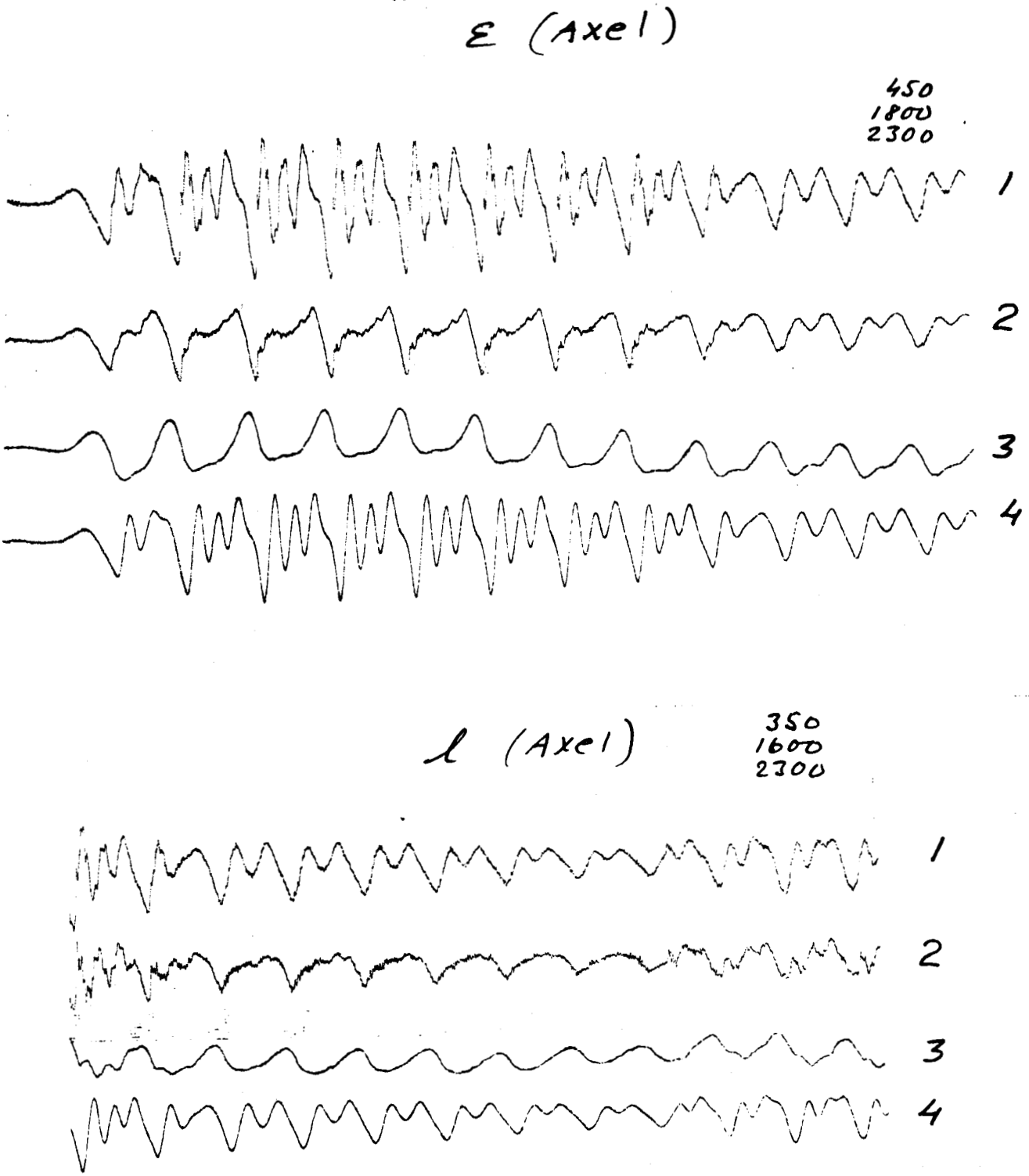


Fig. II-A-11. Inverse filtering with one setting for the second vowel of "Axel" and one for the final [l]. Functions the same as in Fig. II-A-6.

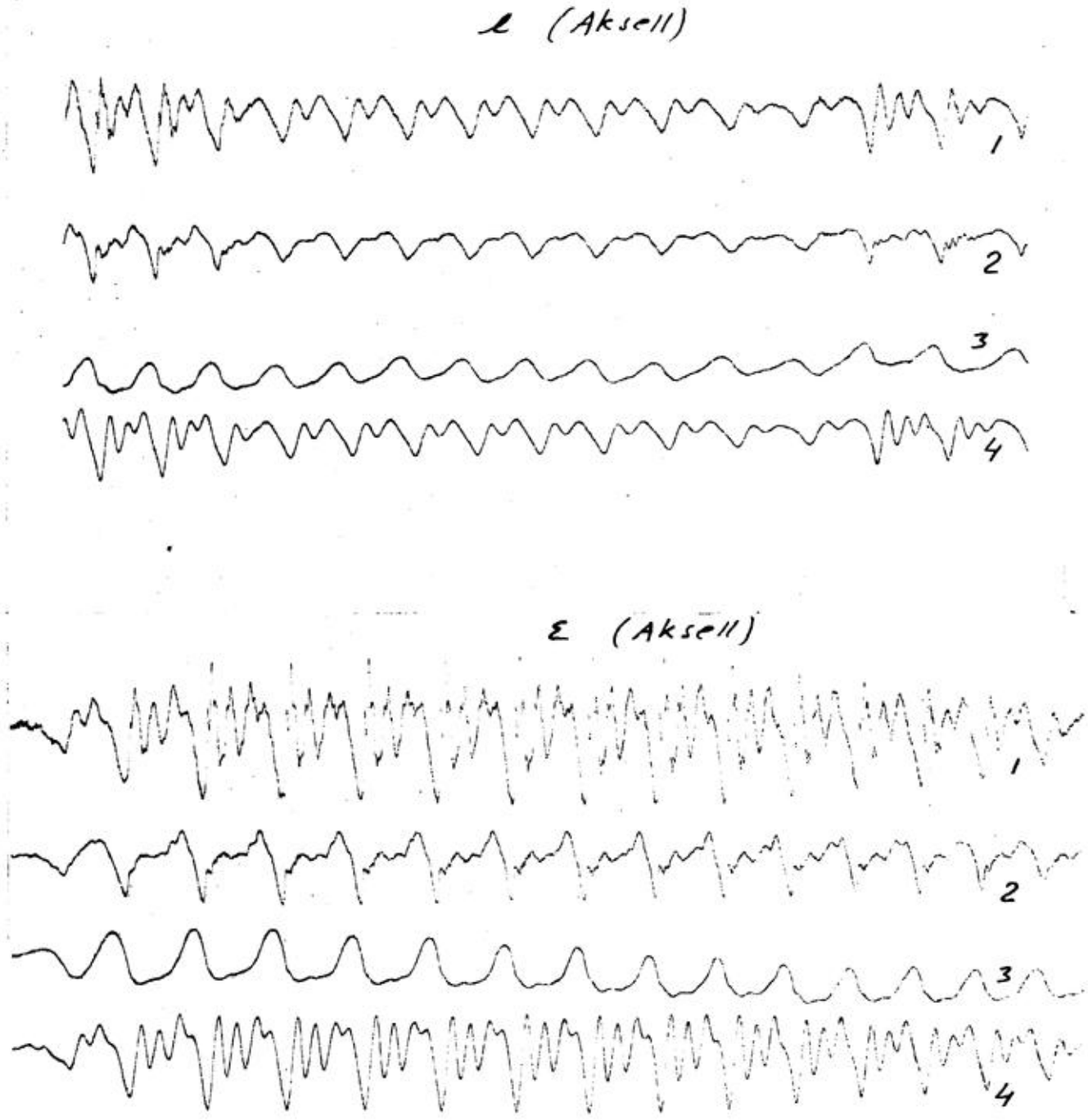


Fig. II-A-12. Inverse filtering with one setting for the [l] and one for the second vowel of "Aksell". The [e] of Fig. II-A-11 and the [e] of Fig. II-A-12 have almost identical glottal flow but differ in duration and formant frequencies.

There was no significant difference in the source excitation parameters for [l] of the two words.

Fig. II-A-13 shows a synchronous recording with inverse filtering and optical glottography. The latter instrumentation of type BFJ, Denmark, would hopefully portray glottal area function. The inverse filter, on the other hand, shows the product of area and velocity. Apparently, the area function pulses are more invariant than the flow pulses. There is a reduction of the flow during the voiced consonant [v] and a tilting of the [v]-pulse towards the right. The area function, on the other hand, tilts backwards, the opening time being shorter than the closing time. As pointed out by Rothenberg (1980), the asymmetry varies with different vowels. The total low frequency inductance of the vocal tract within and above glottis causes an integration of the flow and thus a tilt forward which is more apparent for [a] than for [ɛ].

Fig. II-A-14 illustrates qualitatively how the vocal source characteristics vary within a simple one-word fully voiced sentence "adjö" = [ajø:]. Two apparent features are the relative low glottal peak flow amplitude during the interval of the main stress in the second vowel and that formant amplitudes decrease whilst vocal pulse amplitude may increase towards the termination of the utterance. The same sentence spoken by a different subject was analyzed by Fant (1979). It shows that the stress in the [ø:] is a matter of increase in both F_g and K, see Fig. II-A-15.

Final remarks

This work represents a limited exploratory study only. It may be seen as a follow-up of the work of Rothenberg et al (1974). Several of the mechanisms of voice production and source-filter interaction, e.g., excitation during aspiration, are not sufficiently well understood at present. Data from synchronous glottal flow and glottal area function, subglottal pressure, and electroglottogram would be useful.

The voice source model accordingly has to be refined when we have gained more experience.

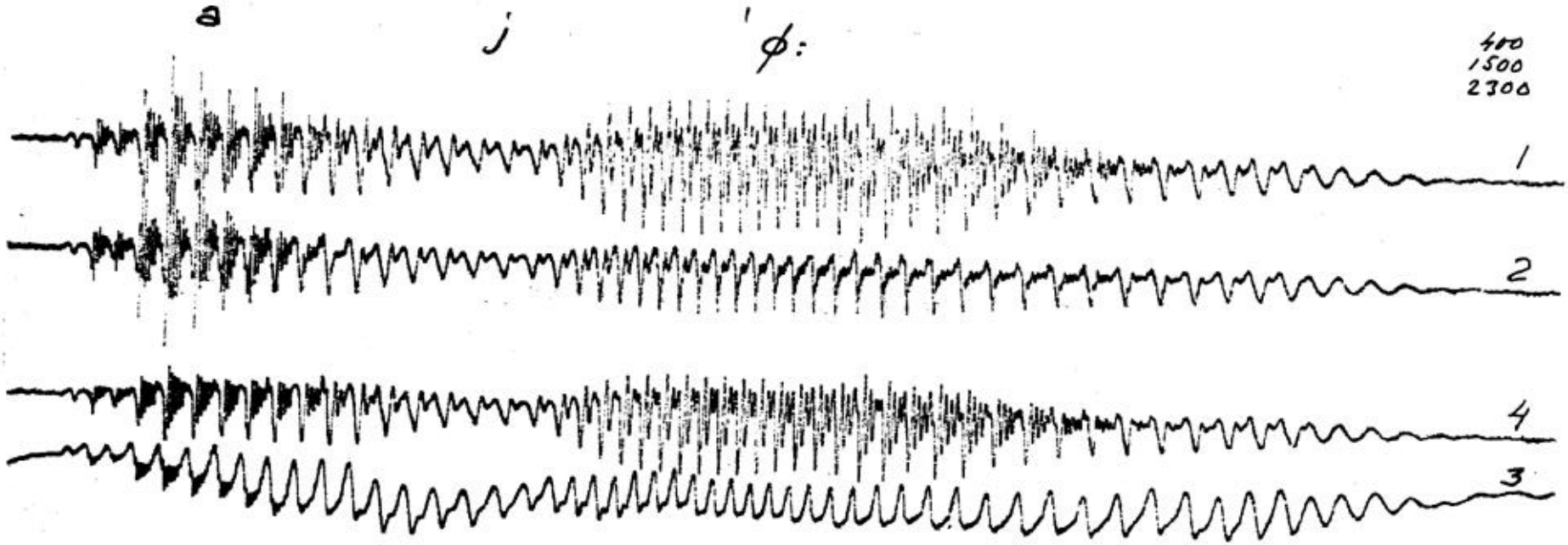


Fig. II-A-13. Single utterance "adjö" with constant inverse filtering setting optimal for [ø:]. (1) oscillogram. (2) = inverse filtering without integration, (3) = (2) without F1-cancellation (4) = (2) with integration. Observe the relative stability of the rate of air consumed (4) in contrast to voice excitation strength (negative pulses of (2)).

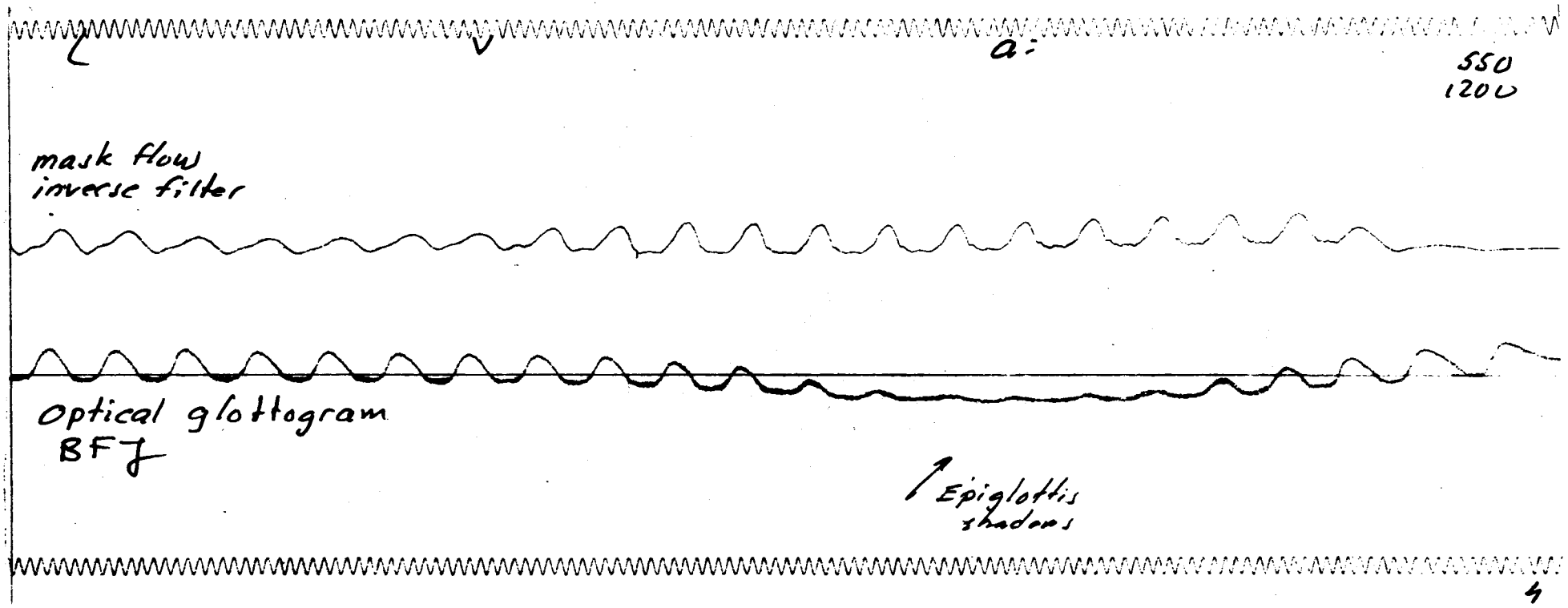


Fig. II-A-14. BFJ optical glottogram together with inverse filtering from Rothenberg type flow mask. Observe the relative stability and left hand asymmetry of the otpically traced glottal pulses.

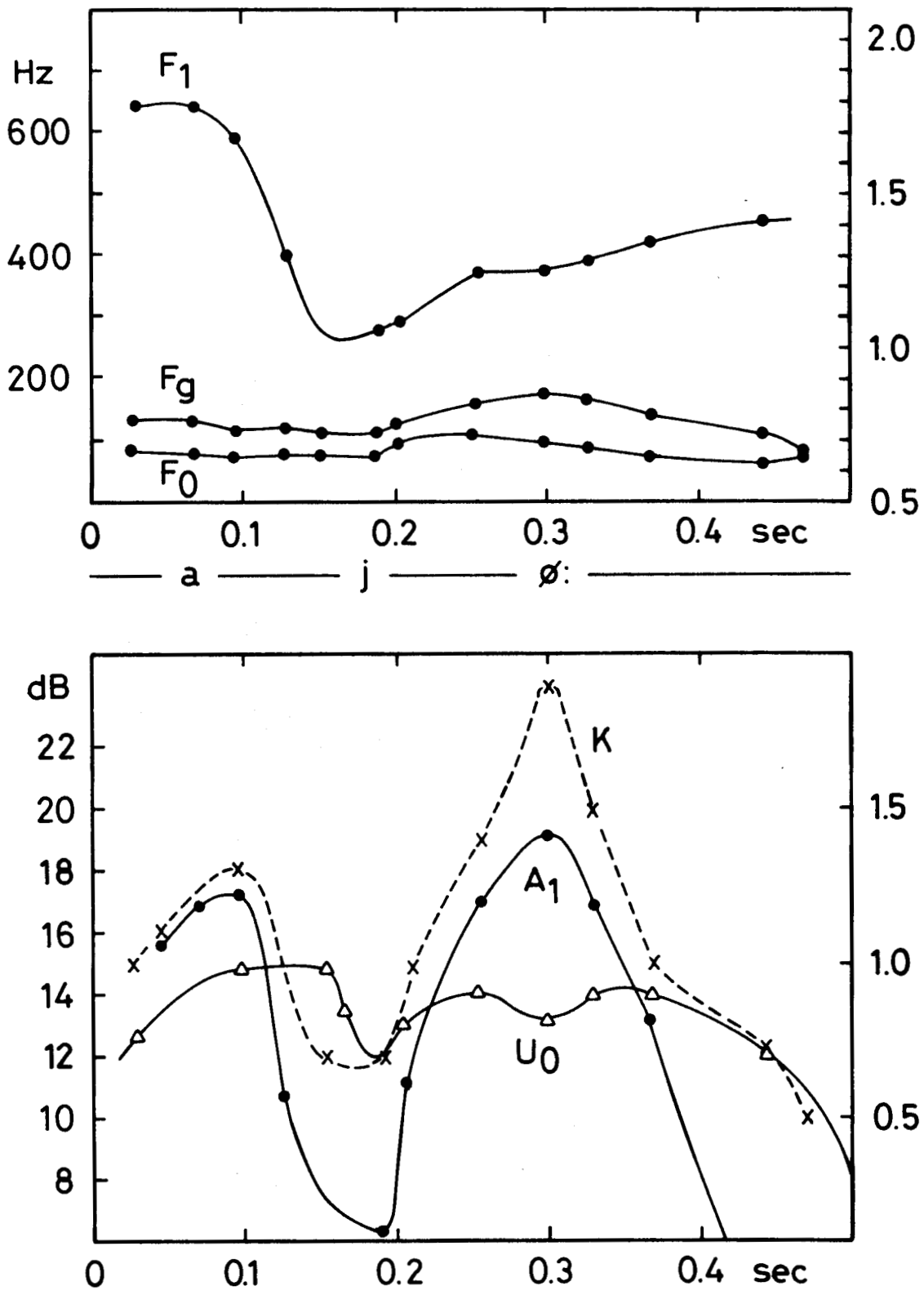


Fig. II-A-15. Variation of first formant frequency F_1 and initial A_1 and the source parameters, F_0, F_g, U_o, and K. (From Fant, 1979b).

Acknowledgments

Stimulating discussions with Martin Rothenberg, Johan Sundberg, and Jan Gauffin have contributed to clarify some of the problems.

References

FANT, G. (1959): "Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Technics, No. 1.

FANT, G. (1979a): "Temporal fine structure of formant damping and excitation", paper presented to the 50th Meeting of the Acoustical Society of America, Cambridge, MA, USA, June.

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

FANT, G. (1979c): "Vocal source analysis - A progress report", STL-QPSR 3-4/1979, pp. 31-53.

GAUFFIN, J. & SUNDBERG, J. (1980): "Data on the glottal voice source behavior in vowel production", in this issue of the STL-QPSR.

HOLMES, J.N. (1976): "Formant excitation before and after glottal closure", IEEE Conf. on Acoustics, Speech and Signal Processing, Philadelphia, PA, USA, April.

LINDQVIST, J. (1965): "Studies of the voice source by means of inverse filtering", STL-QPSR 2/1965, pp. 8-13.

MILLER, R.L. (1959): "Nature of the vocal cord wave", J.Acoust. Soc.Am. 31, pp. 667-677.

ROTHENBERG, M. (1973): "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing", J.Acoust. Soc.Am. 53, pp. 1632-1645.

ROTHENBERG, M. (1980): "Acoustic interaction between the glottal source and the vocal tract", to be publ. in the Proc. of the Conf. on Vocal Fold Physiology, Kurume, Japan, Jan.

ROTHENBERG, M., CARLSON, R., GRANSTRÖM, B. & LINDQVIST-GAUFFIN, J. (1974): "A three-parameter voice source for speech synthesis", pp. 235-243 in Speech Communication, Vol. 2 (ed. G.Fant) (Proc. SCS-74, Stockholm), Almqvist & Wiksell Int., Stockholm 1975.

SUNDBERG, J. & GAUFFIN, J. (1978): "Waveforms and spectrum of the glottal voice source", STL-QPSR 2-3/1978, pp. 35-50.