Voice source parameters in connected speech. A progress report.

Gunnar Fant and Anita Kruckenberg
Department of Speech Communication and Music Acoustics, KTH
Box 70014. S-10044

ABSTRACT

This is a preliminary report on modelling glottal source characteristics within an integrated segmental and prosodic frame. A method for exctraction and direct recording of the envelope functions of glottal peak amplitude and flow derivative amplitude in synchrony with a spectrogram allows rapid access to essential source parameters in connected speech. The dependency of these parameters on global phrase structure, F_0, loudness, stress and accents, and the influence of segmental type and supraglottal and subglottal interaction effects are discussed with a view of integrating voice source characteristics into a general theory of prosodic realization.

INTRODUCTION

This is a report from an ongoing project on an integrated analysis of prosodic and segmental features. The special issue addressed here is the role of the voice source as a component in models of speech production. A potentially useful tool for parameterization of the voice source is the LF-modell, Fant, Liljencrants, Lin (1985), which by now has been adopted by several research groups. However, we still lack sufficiently detailed data for a consistent use of the model for deriving text-to-speech rules. An inhibiting factor has been the time consuming process of inverse filtering and parameter extraction needed for detailed studies of the temporal characteristics of the voice source within a complete production model and a linguistic frame.

Our approach to overcoming these difficulties has been to concentrate the study of voice source dynamics to two basic parameters: (1) the amplitude E_e of the negative peak of glottal flow derivative which is a proportionality factor for formant amplitudes, and (2) the peak amplitude U_0 of the oscillatory component of glottal flow which is the main determinant of the amplitude of the voice fundamental. The ratio of these two, $T_d=U_0/E_e$, is a unifying measure of glottal waveform from which the pulse steepness factor R_k and the return phase relative time constant $R_a= T_a/T_0$ can be predicted with a fair degree of accuracy as will be described in a separate article, Fant and Liljencrants (1994). The significance of the R_a parameter is its relation to the frequency $F_a=1/2\pi T_a$ above which the spectrum of the glottal source attains an extra -6 dB/oct roll-off. To the benefit of the data reduction thus achieved there adds the relative ease of extraction and display of the two fundamental source parameters, $E_e(t)$ and $U_0(t)$ as envelope functions in connected speech.

EXPERIMENTAL STUDIES AND RESULTS

Our preliminary studies have been concerned with the temporal variation of these two source parameters derived from continuous inverse filtering and their covariation with F_0 in connected speech. We have considered both global aspects such as onsets, focal targets and terminations within a phrase or a sentence and local aspects in terms of segmental dependencies and articulatory interaction, abduction/adduction gestures and voicing boundaries, emphasis/deemphasis, stress and Swedish word accents. We have analyzed passages of read prose and also special "lab sentences" in Swedish constructed for systematic variation of stress patterns, focal emphasis and word accent. Our results are by no means complete but a general view of a production model is now emerging.

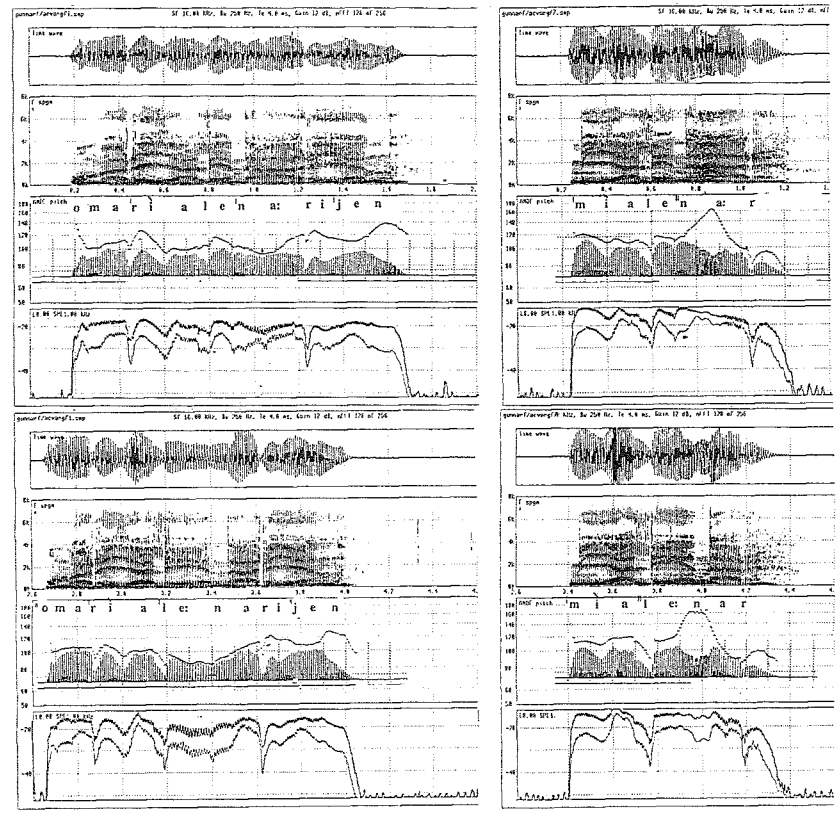

Figure 1. *Oscillogram, spectrogram, F_0 on log scale, E_e linear scale, log intensity (SPL), log intensity with high frequency emphasis. Four sentences, top left: "Å Maria Lenár igen.", bottom left:" Å Maria Lénar igen," top right:"Mia* ***Lenár****," bottom right: Mia* ***Lénar***

A basic finding, not unexpected, is that the $E_e(t)$ contour within a sentence shows some tendencies of following the intonation contour, $F_0(t)$. However, a negative covariance is often seen as illustrated by Fig. 1, which pertains to a male subject GF reading four "accent sentences" contrasting in stress patterns and place of focal accent. A statistical analysis reveals a rise of E_e with F_0 up to a frequency of the order of about 115 Hz and then a fall. The general tendency was the same for another male subject reading a sentence from our prose corpus. A rise-fall relation was also found for $U_o(F_0)$ in the prose sentence but here the rising branch was less steep and the maximum occurred at a slightly lower frequency than in the $E_e(F_0)$ curve, see Fant and Liljencrants (1994). A close match in $E_e(F_0)$ was also found comparing our accent sentence data and a study of sustained phonation at a gliding pitch, Fant (1982).

The general conformity between $F_0(t)$ and $E_e(t)$ is marked by local dips in F_0 and in E_e at voiced occlusives such as [r] and [l]. A similar instance of articulatory interaction is found in the maximally closed phase of stressed [u:], [ʉ:], [i:], [y:] which are pronounced with a diphtong towards a homorganic voiced fricative, e.g. the [j] element of Maria, [mari:ja]. Here E_e is reduced more than may be predicted from F_0. Other departures from average $E_e(F_0)$ are high values in sentence initial positions and an overall declination within a breathgroup which seems to follow subglottal pressure.

What about the glottal waveshape parameter, U_o/E_e? It has a general tendency of decreasing with increasing fundamental frequency but this trend is counteracted by a simultaneous increase of R_k and R_a, Fant and Liljencrants (1994). For both male and female vowels Uo/Ee is accordingly of the order of 0.7 ms.

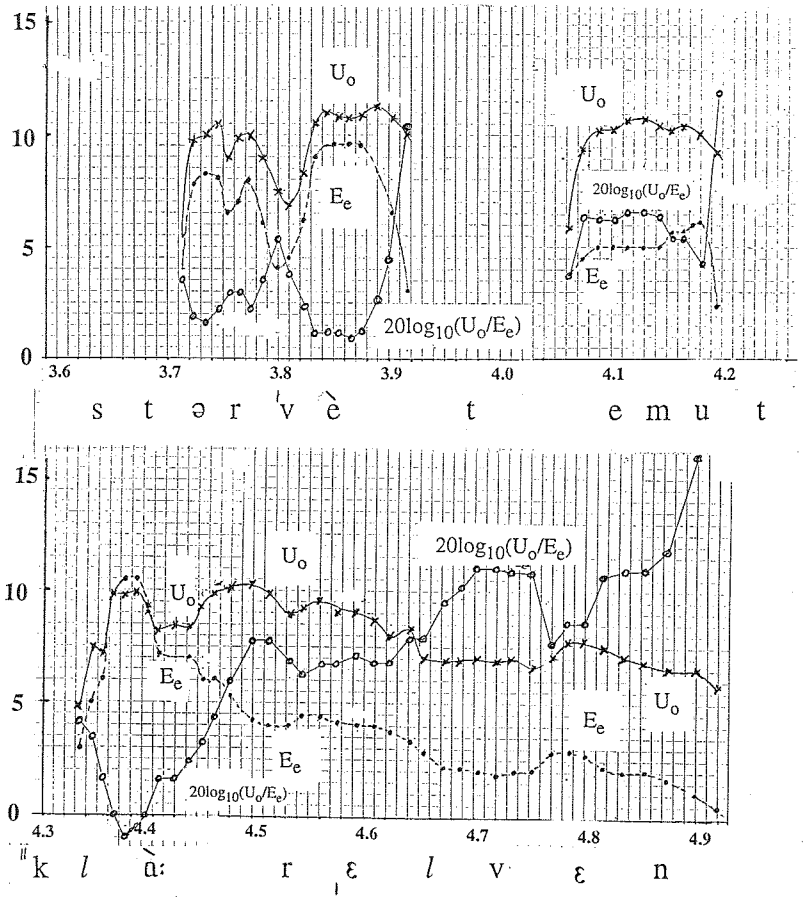

Figure 2. *E_e and U_o on linear scale and U_o/E_e in dB traced through a phrase:" (vars fön)ster vette mot Klarälven."*

A study of U_o/E_e within a sentence of the prose corpus, a part of which is shown in Fig.2, and data from Gobl (1988) have provided tentative values for various classes of voiced consonants ranging from 0 to 15 dB above that of normal vowels, the higher values for voiced occlusions in stops and the lower values for nasals. These are in general agreement with data of Karlsson and Neovius (1994). The covariation of U_o and E_e is on the average such that a shift in E_e by 2 dB is accompanied by about 1 dB in U_o. The relative stability of U_o is especially apparent in the termination phase of a breathgroup, see Fig. 2.

Time constants of $E_e(t)$ and $U_o(t)$ transitions at voiced/voiceless boundaries vary between 15 and 50 ms, the larger values associated with glottal abduction for an aspirated unvoiced stop anticipated already in the beginning of the preceding vowel.

On the average, stressed syllables have about 2 dB higher intensity than unstressed syllables. However, an increase of E_e is not a necessary component of increasing stress in an accented syllable, not even if it is in a focal position. The underlying cause is that F_0 comes above the point of maximum E_e at about 115 Hz. Thus, E_e no longer follows F_0 as seen in the sentences of Fig. 1. On the other hand, half an octave rise in F_0 is associated with about 2 dB gain in intensity which adds to the stress contrast. A local increase of subglottal pressure appears to be necessary to cause an E_e increase above the F_0-predicted value. The pressure component should be included in future investigations.

We frequently observe a relative gain in the upper part of the voice source in stressed vowels, (higher F_a). In addition there are typical segmental boundary effects associated with stress, i.e. a greater contrast in intensity between voiced consonants and vowels as already mentioned by Gobl (1988). Thus, the sizes of the F_0, E_e and intensity dips of an [r]-flap or an [l] segment increase with overall stress and emphasis in the following vowel. In our experience the most consistent physical correlate of stress is duration, while the primary correlate of focal versus nonfocal stress lies in F_0.

There does not appear to be any specific source changes other than those associated with F_0 that could add to the accent 1 versus accent 2 distinction in Swedish.

To sum up our experience, the voice source rules in connected speech can to a first and quite powerful approximation be organized as follows:

(1) Choose a global rise-declination contour in subglottal pressure, P_s, for a breathgroup. Include one or more focal regions and derive segmental durations and the F0 contour.

(2) Establish E_e and U_o/E_e as default functions of F_0 and P_s.

(3) Determine the local decrease/increase of E_e and increase/decrease of U_o/E_e associated with segment type, degree of supraglottal narrowing and glottal abduction.

(4) Translate U_o/Ee values into the full set of LF-parameters. Include rules for deviations of R_g from the default value R_g=1.

(5) Work out speaker specific rules.

This is the general strategy we are following in our work.

ACKNOWLEDGEMENTS

This research has been supported by a grant from The Bank of Sweden Tercentenary Foundation and by contributions from Telia Promotor Infovox AB and the Ericsson Radio Systems AB.

REFERENCES

Fant, G. 1982. "Preliminaries to the analysis of the human voice source". *STL-QPSR* 4/1982, 1-27.

Fant,G. & J. Liljencrants. 1994. "Data reduction of LF voice source parameters". In this volume.

Fant, G., J. Liljencrants & Q. Lin. 1985. "A four-parameter model of glottal flow". *STL-QPSR* 4/1985, 1-13.

Gobl. C. 1988. "Voice source dynamics in connected speech". *STL-QPSR* 1/1988, 123-159.

Karlsson, I.& L. Neovius. 1994. "VCV-sequencies in a preliminary text-to-speech system for female speech". *STL-QPSR* 1/1994