Data reduction of LF voice source parameters

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

ABSTRACT

This is a proposal of a datareduction scheme for studies of voice source characteristics in connected speech. The object is to concentrate on a limited number of parameters capable of defining essentials of source waveform, and excitation magnitude characteristics. It is shown that the glottal flow maximum oscillatory amplitude U_o and the flow derivate E_e at the closing discontinuity together meet these requirements and that they can be continuously exctracted and displayed in synchrony with a spectrogram. A statistical analysis allows a fair degree of accuracy in predicting the total set of LF parameters from U_o/E_e and F_0 in a variety of voiced speech sound segments. Furthermore, a substantial part of E_e variations are predictable from F_0.

INTRODUCTION

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. Our effort has now been to reduce the number of descriptive parameters by a data reduction scheme. For this purpose we capitalize on some inherent analytical and functional constraints that govern the covariation of LF parameters.

The LF-model, Fant, Liljencrants and Lin (1985), see Fig. 1, contains three waveshape parameters, R_k, R_g and R_a and in addition the amplitude parameter, E_e. R_k is the ratio of the decay time to the rise time of the flow pulse, $(T_e-T_p)/T_p$ where T_e is the location of the major discontinuity in the closure phase, i.e. of E_e, and T_p is the location of the flow peak with respect to the glottal flow onset. The parameter $R_g=T_0/2T_p$ together with R_k define the open quotient $OQ=(1+R_k)/2R_g$ of the voice fundamental period T_0. The parameter $R_a=T_a/T_0$ is a relative measure of the duration of the return phase. For descriptive purposes one generally refers to T_a, which is less dependent on F_0, or to the corresponding cut off frequency, $F_a=1/2\pi T_a$, where the source spectrum attains an extra -6 dB/oct roll-off. This appears to be the perceptually most significant parameter of the LF-model and shows a significant covaration with the R_k parameter. Decreasing F_a, i.e. increasing R_a, is usually accompanied by increasing R_k which enhances the relative dominance of the voice fundamental versus formant amplitudes, as is apparent in their combined effect of increasing U_o/E_e. The parameter R_g is usually close to 1. An increase of R_g causes a decrease of the flow pulse duration which enhances a spectral region of $F_g=F_0R_g$ in the source spectrum usually boosting the second harmonic. An increase of R_g, typical of pressed voice, is sometimes associated with local emphasis in connected speech. For more detailed descriptions see also Fant (1993), Gobl (1988), Karlsson (1990).

We have found a fair degree of predictablity of LF parameters from the ratio U_o/E_e, which accordingly is a unifying waveshape characteristic. The physical significance of U_o/E_e is a measure of effective decay time, which we may refer to as "declination time", T_d, of the glottal flow pulse. As shown in Fig. 1 it is defined by the projection on the time axis of the tangent to glottal flow at the instant of excitation and up to the level of U_o. For vowels the declination time $T_d=U_o/E_e$ is usually in the range 0.5 to 1 ms and can be as high as 3 ms in supraglottal highly constricted voiced consonants or in highly abducted prepausal vowel segments.

INHERENT CONSTRAINTS

The flow maximum U_o is a unique function of R_k, R_g, R_a, E_e and F_0 but is not immediately accessable in analytic form. An approximation valid within 1.5 dB for U_o/E_e below 3 ms and R_k values below 0.6 and R_a below 0.12 is

$$U_o/E_e = (0.5+1.2R_k)(R_k/4R_g+R_a)(1/F_0) \quad (1)$$

Accordingly U_o/E_e increases with increasing R_a or T_a and with increasing R_k. That this is so follows from the requirement of area balance between the positive and negative parts of the differentiated glottal flow function. The spectral correlate of increasing U_o/E_e is an increase of the ratio of voice fundamental amplitude to formant amplitudes, in case of the R_a increase a progressively higher rate of fall off of high frequency formant amplitudes.

How is this reflected in actual source data? From the original work of Gobl (1988) who has summarized LF-data obtained from a variety of vowels and voiced consonants we have calculated the following regressions in which R_k and R_a are expressed in percent and U_o/E_e and T_a in ms.:

$$R_k=21+12.7(U_o/E_e)(F_0/110) \quad (2)$$
(r=0.93)

$$R_a=0.6+4(U_o/E_e)(F_0/110) \quad (3)$$
(r=0.91)

alternatively

$$T_a=0.04+0.38(U_o/E_e) \quad (4)$$
(r=0.91)

A general but not always positive korrelation of R_a with R_k was mentioned by Gobl (1998). The limitation is that T_a or R_a is forced to go to zero as the open quotient OQ during a full abduction gesture approaches 1 and the source function degenerates to a sinewave. Accordingly, the return phase is modelled by a straight line from T_e to T_0 if the intended T_a turns out to be greater than (T_0-T_e).

The consequence is that the statistical relation between R_a and R_k is more accurately modelled by a parabolic expression. A suggestion is:

$$R_a= -16.6+0.9R_k-0.0073R_k2 \quad (5)$$

We have tested the general relevance of this regression analysis technique to LF data from an ESPRIT SPEECHMAPS corpus, (personal communication Christer Gobl) containing vowels and consonants [l], [m], [b], [v] from VCV syllables spoken by Italian and French subjects. We obtained results in general agreement with Eq.2-3.

These data illustrate universal constraints. But are the prediction formula also valid for female voices? A test was carried out on vowel data published by Karlsson (1990). U_o/E_e values were first calculated from Eq. 1, given the particular R_k,R_g,R_a and F_0 values. Predicted values of R_k and R_a were then calculated from Eq. 2 and 3. A good fit to measured data is demonstrated in Fig.2. This may seem like a circular proof but involves a data reduction.

An additional noteworthy finding in Fig. 2 is the relatively high U_o/E_e of vowels articulated with labial and/or palatal constriction causing aerodynamic interaction. These differences in vowel inherent source shapes are greater than what we have experienced from analysis of male vowels. An exception is in focally stressed context in connected speech where targets of almost complete supraglottal closure and very high U_o/E_e values are encountered, see Fant and Kruckenberg (1994). It should also be pointed out that the outcome of the prediction may not be as good for other female subjects.,

CONTINUOUS EXTRACTION OF E_e AND U_o

The envelope function $E_e(t)$ can be extracted by continuous inverse filtering, i.e. running the speech wave through a timevariable inverse filter preprogrammed to follow

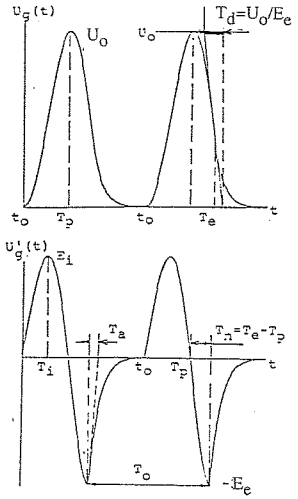

Fig. 1. The LF-model and the Uo/Ee parameter

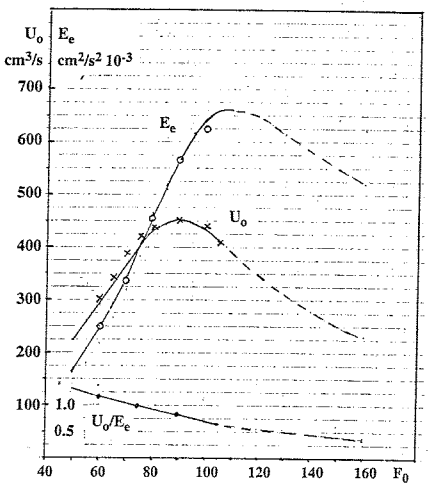

Fig. 3. Ee(F0) and Uo(F0). Prose data, subj. ÅJ

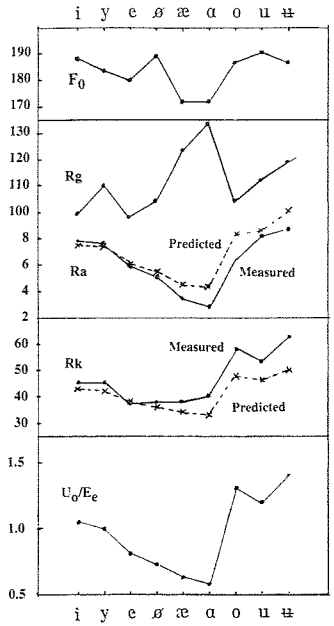

Fig. 2. Data reduction of female LF-parameters

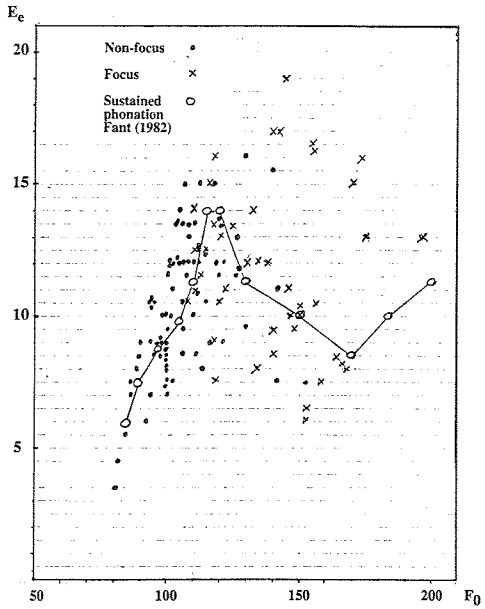

Fig.4. Ee(F0). Accent sentences, subj. GF

formant frequency variations with reasonable accuracy. As previous pointed out, Fant (1993), Fant and Kruckenberg (1993), E_e is rather insensitive to tracking errors as long as these are not highly discontinuous. Fair approximations within a couple of dB can be achieved by a constant setting of the inverse filter appropriate for a neutral vowel. An even simpler approximation to $E_e(t)$ is the sequence of negative peaks of the speech oscillogram, Gobl (1988), Fant (1993). Similarly for extraction of $U_o(t)$ which are the peaks of the integrated inverse filter output one can use neutral vowel setting or simply the integral of the speech wave as an approximation. Systematic errors with the simpler methods are underestimation of E_e and overestimation of U_o at low F_1 locations. One must also look out for possible difficulties in defining the foot of the U_o measures. With these errors in mind one can anyhow produce continuous data on essentials of voice source dynamics in relative large corpora of connected speech. It is recommended to supplement such "poor man's inverse filtering" with spotwise checks with complete determination of LF parameters through a proper inverse filtering, even including zeros.

PREDICTION OF E_e FROM F_0

The excitation function $E_e(t)$ is an important source parameter. Functionally it depends on subglottal pressure, glottal abduction/adduction and tension and supraglottal articulation. As discussed in more detail by Fant and Kruckenberg (1994) the linguistic frame for prediction of E_e includes the entire structure of prosodic and segmental categories, including global phrase contours basically dependent on subglottal pressure and local variations subject to the influence of stress, accents and supraglottal interaction. The latter can be appreciable. Default values of E_e within vowels can be determined from the statistics of E_e as a function of F_0, disregarding interaction effects. Examples of such data are shown in Fig. 3 pertaining to a sentence from our prose corpus, subject ÅJ, and in Fig. 4 pertaining to 15 lab sentences with systematic variations of focal regions and stress. The latter show a high degree of similarity with the data obtained from gliding pitch sustained phonation, Fant (1982). The essential finding is an increase of E_e with F_0 up to about 115 Hz above which E_e decreases with increasing F_0.

U_o displays a similar trend but for a less steep slope up to a maximum which is located at a somewhat lower frequency than the $E_e(F_0)$ maximum. This was also found in the Fant (1982) study. It remains to interpret these functions in terms of underlying subglottal pressure and glottal articulation.

ACKNOWLEDGEMENTS

This research has been supported by a grant from The Bank of Sweden Tercentenary Foundation, the ESPRIT SPEECHMAPS program and by a contribution from Telia Promotor Infovox AB.

REFERENCES

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

Fant, G. 1993."Some problems in voice source analysis". *Speech Communication* 13, 7-22.

Fant, G. and A. Kruckenberg. 1993. "Towards an integrated view of stress correlates". *Lund University. Dept. of Linguistics, Working Papers* 41, 42-45 1993.

Fant, G. & A Kruckenberg. 1994. "Voice source parameters in connected speech. A progress report." 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. 1990. "Voice source dynamics for female speakers". *Proceedings of the 1990 Int. Conf. on Spoken Language Processing,* Kobe, 69-72.