3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

VOICE SOURCE PARAMETERS IN CONTINUOUS SPEECH. TRANSFORMATION OF LF-PARAMETERS.

Gunnar Fant, Anita Kruckenberg, Johan Liljencrants and Mats Båvegård

Department of Speech Communication and Music Acoustics, KTH
Box 70014. S-10044, Sweden

ABSTRACT

This is a report on a data reduction scheme for studies of voice source characteristics in connected speech and some results from speech analysis within a segmental and prosodic frame. It is shown that essentials of glottal wave shape are included in U_o/E_e, the ratio of peak glottal volume velocity U_o to the negative peak E_e of glottal flow derivative. Default values of the complete set of LF-source parameters F_0, E_e, R_a, R_k, R_g can be predicted from F_0, E_e and U_o/E_e. Of special importance is the relation of E_e and U_o to F_0. Additional adjustments related to voice type and contextual demands are included. Simplified inverse filtering methods for extraction and direct recording of $U_o(t)$ and $E_e(t)$ in synchrony with $F_0(t)$ are described. The dependency of source parameters on voice fundamental frequency F_0, global phrase structure, loudness, stress and accents, and the influence of segmental type and supraglottal and subglottal interaction effects are briefly discussed.

I. INTRODUCTION

A potentially useful tool for parameterization of the voice source is the LF-model [5], which by now has been adopted by several research groups. However, in spite of several preliminary studies, [6-10] we still lack sufficiently detailed data for a complete control 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.

In practice the data reduction for specifying waveform essentials is combined with a data expansion for deriving the full set of LF-parameters in synthesis. This process involves rules for deviations from the first order default values and is analogous to a principal component analysis, aiming at basic dimensions in the phonatory process. The choice of the peak value U_o of the oscillatory component of glottal flow and the flow derivative E_e at the point of excitation as main descriptive parameters simplifies the inverse filtering process and minimizes the need for supplementary detailed analysis of LF parameters. The significance of U_o is its close relation to the amplitude of the voice fundamental, while E_e is the basic determinant of formant amplitudes. Furthermore, we have found systematic dependencies of E_e and U_o on F_0 which are scaled differently for male and female voices.

II.THE LF-MODEL

2.1 Definitions

The LF-model, [5] see Fig. 1, contains three waveshape parameters, R_k, R_g and R_a and in addition F_0 and the amplitude parameter, E_e. $R_k=(T_e-T_p)/T_p$ is the ratio of the decay time to the rise time of the flow pulse. Tp is the location of the flow peak and T_e is the location of the major discontinuity in the closure phase where the flow derivative attains the peak value $-E_e$, The parameter $R_g=T_0/2T_p$ together with R_k determine 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. The basic measure T_a, defined from the initial slope of the return phase is more directly related to the corresponding cut off frequency, $F_a=1/(2\pi T_a)$, where

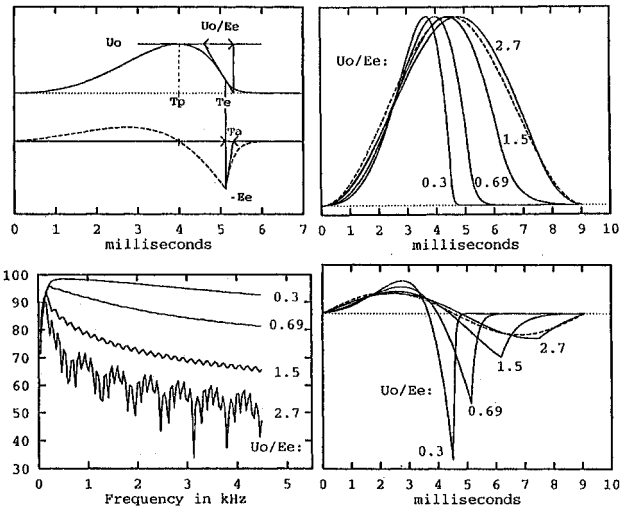

Fig.1 The LF-model with U_o/E_e identified as a declination time, glottal flow, flow derivative, and flow derivative spectrum for four values of U_o/E_e.

the source spectrum attains an extra -6 dB/oct roll-off. F_a can also be expressed as $F_a=F_0/(2\pi R_a)$.

F_a, or the alternative measures T_a or R_a, appears to be the perceptually most significant parameter of the LF-model. Decreasing F_a, i.e. increasing R_a, is usually accompanied by increasing R_k which enhances the relative dominance of the voice fundamental. The parameter R_g is usually close to 1. An increase of R_g and also a decrease of R_k reduces the duration of the flow pulse and thus the open quotient. The spectral consequence is a shift of energy from the fundamental to the second harmonic. An increase of R_g, typical of pressed voice, is sometimes associated with local emphasis in connected speech. For more detailed presentations of the LF-model see also [2], [6], [10].

2.2 Inherent constraints

We have found a systematic first order predictability of LF-parameters from F_0 and the ratio U_o/E_e, which accordingly is a unifying waveshape characteristic. A two-way transformation of parameters is achieved. The complete LF-model is specified by E_e, F_0, R_k, R_a, R_g from which U_o is uniquely determined. The transformed set involves F_0, E_e and U_o/E_e as basic parameters supplemented by orthogonal coefficients k_a and k_g related to R_a and R_g. The transformation back to LF-parameters is supported by statistical data analysis.

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, which is valid for both male and female voices, and can be as high as 3 ms in highly

constricted voiced consonants or in highly abducted prepausal vowel segments.

The flow maximum U_0 is a unique function of R_k, R_g, R_a, E_e and F_0 but is not directly derivable 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 can be expected to increase 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 direct spectral correlate of increasing U_o/E_e is an increase of the ratio of voice fundamental amplitude to formant amplitudes.

III. DATA REDUCTION AND EXPANSION

3.1 Basic dimensions

Eq. 1 expresses an inherent property of the LF-model, but to what extent may each of the LF-parameters R_a, R_k and R_g be predicted from U_o/E_e and F_0 alone? A statistical survey of available experimental data has been carried out. We have relied heavily on the original work of Gobl [6], who has summarized LF-data obtained for vowels in various contexts and voiced consonants. To this data pool we have added some of our vowel data and derived the following regression equations in which R_k and R_a are expressed in percent and U_o/E_e in ms. For R_e smaller than 2.7 we find.

$$R_k = 22.4+11.8R_e \quad (2)$$
(r=0.93)

$$R_a = -1+4.8R_e \quad (3)$$
(r=0.91)

where

$$R_e = (U_o/E_e)(F_0/110) \quad (4)$$

is the new F_0-normalized waveshape parameter.The source cut-off frequency F_a is by definition

$$F_a = F_0/(2\pi R_a) \quad (5)$$

Omitting the constant term in Eq. (3) suggests the approximation

$$F_a = 460(E_e/U_o) \quad (6)$$

which provides a more direct approach than via R_e.

R_g is modeled from general experience as

$$R_g = 3.7\,R_e^2-24R_e+133 \quad (7)$$

The wave shape and spectrum of the LF-model with Re as parameter are included in Fig.1.

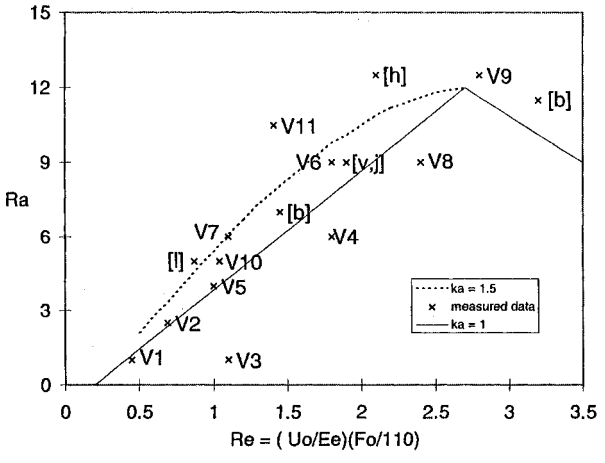

Fig.2 $R_a=T_a/T_0$ versus R_e for vowels, V1-V11, and voiced consonants, [l], [v], [j], [h] and [b]. V10 is the mean of female vowels [a] and [ö], subj F1 of Karlsson [4] and V11 the same for subjects F2-F7. Consonants and V2-V9 are male data from Gobl [6], V3 hard onset, V4 soft onset, V5 after voiced consonant, V6 before unvoiced fricative, V7 and V8 before unvoiced stop.

Experimental data underlying Eq. 3 are contained in Fig. 2 which reveals the relative high correlation between measured R_a of vowels and voiced consonants and R_e. However, a major limitation is that T_a and thus R_a are forced to go to zero as the open quotient OQ during an abduction gesture finally approaches 1 and the source function degenerates to a sinewave. This is typical of pre-pause voice termination. As a consequence our model imposes decreasing R_a for R_e greater than 2.7. If the intended T_a turns out to be greater than the duration of the maximally closed phase, (T_0-T_e), the return phase is modeled by a straight line connection from T_e to T_0. For R_e greater than 2.7:

$$R_a = 32.3/R_e = 32.3\,(110/F_0)(E_e/U_o) \quad (8)$$

In this region:

$$OQ = 1-1/(2.17R_e) \quad (9)$$

$$R_g = 596/(7.96-2OQ) \quad (10)$$

$$R_k = 2R_gOQ-100 \quad (11)$$

At the limiting value R_e=2.7 our model prescribes R_a=12, R_k=54 and R_g=95. At higher R_e values R_k continues to increase while R_a decreases linearly with E_e/U_o. The limiting values for very high R_e values, in practice a very small E_e at a less reduced U_o, is R_a=0, R_k=100, R_g =100. It should be pointed out that for R_k greater than 50 the maximum value of the negative part of the flow derivative is the turning point of the main flow sinusoid which occurs prior to T_e and is of greater amplitude. The interval of R_e greater than 2.7 is mainly confined to boundary regions where E_e is small and the spectral shape less crucial.

3.2 Orthogonal dimensions

The procedure above generates default values of the complete set of LF parameters given U_o, E_e and F_0. In practice, to satisfy contextual constraints and to model specific speaker types, we need to take into account two dimensions that operate orthogonally to U_o/E_e or more generally to the R_e parameter. One is a more abducted, i.e. "aspirated" phonotary mode with increased R_a versus an adducted mode with more complete closure and a lower R_a. In order to retain constant R_e this requires a decrease of R_k with increasing R_a. From a differentiation of Eq 1 and a subsequent statistical transform of variables we find

$$dR_k/dR_a = -2.64\,R_e^{0.34} \quad (12)$$

which is of the order of -2.7 at R_e=1.

Similary, an increase of R_g which is often found to accompany stress, requires an increase of R_k if R_e is held constant. This results in a shorter, more symmetrical glottal flow pulse. The model constraints require

$$dR_k/dR_g = 0.126 + 0.0557R_e \quad (13)$$

which is of the order of 0.18 at R_e=1.

In practice these orthogonal variations are handled by coefficients labelled k_a and k_g. Within the domain of R_e smaller than 2.7 the parameter R_a is given an increment

$$dR_a = R_a(k_a-1)(1-R_e^2/2.7^2) \quad (14)$$

which is forced to zero at the turning point R_e=2.7 and remains zero at higher R_e values. The coefficient R_g is handled analoguosly.

A line of k_a=1.5 has been included in Fig 2. The reference, k_a=1, was chosen to favour non-boundary vowels whilst the statistical mean for both vowels and consonants lies between k_a=1 and k_a=1.5.

3.3 A test on female vowels

How do female vowels fit into the model? The mean of the [a] and [ö] vowels of subject F1, Karlsson [9] inserted as point V10 in Fig. 2, is located at k_a=1.1 and the mean of subjects F2-F7 at k_a=2.2.

A test of data reduction applied to female vowels is demonstrated in Fig. 3 which pertains to the 9 Swedish vowels of Karlsson [10], subj. W1. Given the full set of LF-parameters and F_0 the U_o/E_e values were first computed from Eq.1. The subsequent prediction of R_k and R_a from U_o/E_e and F_0 combined into R_e, Eq.2-3, came out fairly well as seen in Fig. 3. A value of k_a=1 was assumed but k_a=0.9 would have provided a somewhat better match. An accurate prediction of R_g would have required vowel specific k_g values.

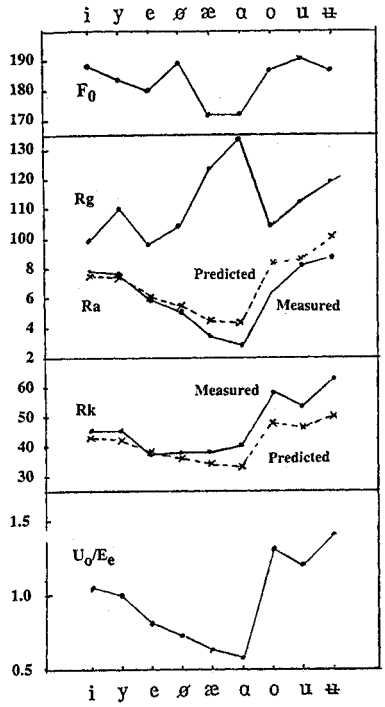

Fig.3 Data reduction of female vowels. Measured LF-data from Karlsson (1990) and R_a and R_k values predicted from F_0 and calculated U_o/E_e.

An additional noteworthy finding in Fig. 3 is the relatively high U_o/E_e of vowels articulated with labial and/or palatal constriction causing aerodynamic interaction, i.e. [i:], [y:], [u:], [u:]. 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, [4].

Further evidence for the universality of the data reduction scheme was found when processing data from analysis of French and Italian VCV sequencies made available by A. Ní Chasaide, C. Gobl and P. Monahan within the ESPRIT SPEECHMAPS project. Quite similar regression coefficients were obtained. Ongoing work on Swedish continuous speech also supports the validity of the reduction scheme.

IV. PARAMETER TRACKING

$U_o(t)$ and $E_e(t)$ within a phrase can be continuously traced by inverse filtering operating on a matrix of time localised frames containing predetermined formant frequencies and bandwidths, if neccessary extended to a complete pole-zero specification. However, E_e and even more so U_o are rather insensitive to errors in inverse filter tuning. It was already found by Fant [2], that a neutral vowel setting provided almost the same $E_e(t)$ contour of a phrase as a proper inverse filtering. Moreover, the negative part of the speech oscillogram also provided a fair substitute. These findings have now been extended to $U_o(t)$. A simple integration of the speech wave provided very much the same outline as the proper inverse filtering. Systematic errors are underestimation of E_e and overestimation of U_o at low F_1. This also holds for neutral vowel constant settings but to a less degree. On the other hand, a discontinuity in the decision of what is F_1 in the formant tracking of complete inverse filtering may introduce a local discontinuity which is more disturbing than the systematic but more continuous errors in neutral vowel inverse filtering.

$U_o(t)$ and $E_e(t)$ can be displayed in synchrony with a spectrogram. We have encountered some problems in the stability of the zeroline of $U_o(t)$. With an expanded time scale this problem is reduced. Period by period manual analysis of U_o and E_e is exemplified in Fig. 4. and continuous recording of E_e together with spectrogram and F_0-curve in Fig. 6 to be discussed below.

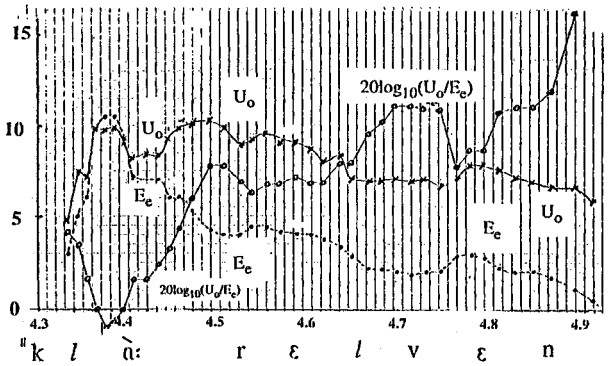

Fig.4 Manually extracted U_o, E_e and $20log_{10}(U_o/E_e)$ from inverse filtering of the sentence final word "Klarälven". Observe the minimum U_o/E_e in the stressed vowel [a:] and the maxima for [l],[r] and [v] superimposed on the gradual increase of U_o/E_e towards the end of the utterance.

V. EXPERIMENTAL STUDIES AND RESULTS

Our preliminary studies have been concerned with the temporal variation of U_o and E_e 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 preliminary only but a general view of a production model is now emerging.

A basic finding, adding the predictability, is that the $E_e(t)$ contour within a sentence shows clear tendencies of following the intonation contour, $F_0(t)$. However, a negative covariance is often seen at high F_0 levels. A statistical analysis confined to non-obstructed vowels, see Fig. 5, shows a rise of E_e up to F_0=115 Hz followed by a fall. U_o as a function of F_0 shows a similar trend but for a less steep ascending part and a maximum at 90 Hz. These $E_e(F_0)$ and $U_o(F_0)$ functions are similar to earlier findings from a study of sustained phonation at gliding pitch, [1]. Data from a second male subject have provided similar results. For a female subject we have found the same type of $E_e(F_0)$ but for a frequency displacement with a maximum at F_0=215 Hz and a more peaked appearence than for the male voice.

An approximate proportionality between $E_e(t)$ and $F_0(t)$ in continuous speech is demonstrated in Fig. 6, which pertains to a male reading two short sentences. Here the $E_e(t)$ recording is displayed just below the F_0-contour. The conformity is marked by the synchrony of E_e dips with the F_0-dips of [r] and [l] and the center part of the vowel [i:] which has a [j] target. In these instances, because of supraglottal interaction, the E_e dips are greater than what can be predicted from F_0 alone.

The two-word sentence in the lower part of Fig. 6, "Mia Lé:nar" was uttered with a focal accent realized by an F_0 rise in the [é:n] region. The associated E_e-minimum may at the first instance seem paradoxal but is a natural consequence of the E_e versus F_0 relation, with a maximum at about 125 Hz for this speaker. The same phenomenon was even more apparent for the female subject reading the same sentence. This is an interesting property of the phonatory mechanism, that deserves to be studied in greater detail.

Other departures from average $E_e(F_0)$ are high values in sentence initial position and an overall declination within a breathgroup which seems to follow subglottal pressure.

What about the glottal waveshape parameter, U_o/E_e? For a single speaker it has a general tendency of decreasing with increasing F_0 but across speakers this trend is counteracted by a simultaneous increase of R_k and R_a, [3]. For both male and female vowels U_o/E_e is accordingly of the order of 0.7 ms.

A study of U_o/E_e within a sentence of the prose corpus, a part of which was shown in Fig.4, together with data from [6-7] have

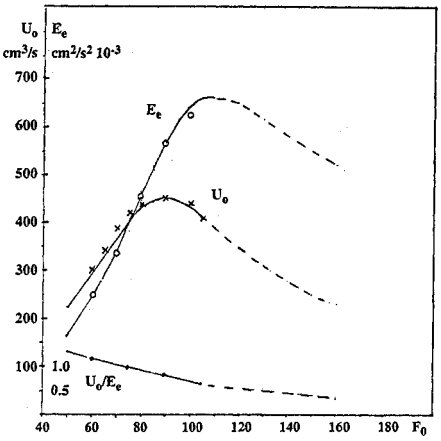

Fig.5 E_e and U_o as a function of F_0 in nonobstructed vowels. Average data for a sentence, male subject ÅJ.

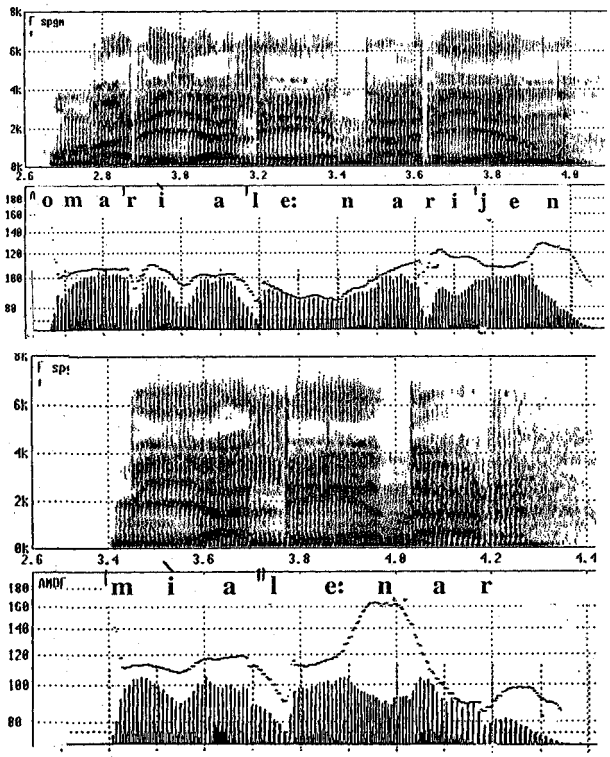

Fig.6 Spectrogram and $E_e(t)$ extracted from inverse filtering superimposed on the pitch contour $F_0(t)$. Above, "Å Maria Le:nar igen " neutral prosody, below "Mia Lé:nar" focal F_0 rise on [é:n] with a minimum in the $E_e(t)$.

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 [8]. 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 specially apparent in the termination phase of a breathgroup, see Fig. 4.

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 as already exemplified. 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 pronounced increase of E_e and intensity is mainly encountered in higher degrees of prominence and requires a subglottal pressure increase.

We frequently observe a relative gain in the upper part of the voice source spectrum of 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 in [6]. Thus, the size of the F_0, E_e and intensity dips of an [r]-flap or [l] increase with overall emphasis or with stress in the following vowel. In our experience the most consistent physical correlate of stress in Swedish 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, [4].

VI. VOICE SOURCE RULES

To sum up our experience, the voice source rules in connected speech can 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 F_0-contour.
(2) Establish E_e and U_o/E_e as default functions of F_0 and P_s.
(3) Determine local increments in E_e and U_o/E_e associated with segment type, stress level, degree of supraglottal narrowing and glottal abduction. Add context specific values of k_a and k_g if motivated.
(4) Determine proper time constants for segment and phrase boundary regions
(5) Add aspiration noise and VT transfer function modifications caused by glottal abduction.
(6) Add speaker specific modifications
(7) Translate the F_0, U_o/E_e, k_a and k_g values into the full set of LF-parameters.

This is the general strategy we are now following in the search for more specific rules.

ACKNOWLEDGEMENTS

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

REFERENCES

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

[2] G. Fant, "Some problems in voice source analysis," Speech Communication 13, pp. 7-22, 1993.

[3] G. Fant and J. Liljencrants, "Data reduction of LF voice source parameters," Working Papers 43, Lund University, Dept. of Linguistics, pp. 62-65, 1994.

[4] G. Fant and A. Kruckenberg," Voice source parameters in connected speech. A progress report," Working Papers 43, Lund University, Dept. of Linguistics, pp. 58-61, 1994.

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

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

[7] R. Carlson, G. Fant, C. Gobl, B. Granström. I. Karlsson and Q. Lin, "Voice source rules for text-to-speech synthesis," ICASSP 1989, Vol. 1, pp.223-226.

[8] I. Karlsson and L. Neovius, "Speech synthesis experiments with the GLOVE synthesizer," Proc. Eurospeech 93, pp. 925-928.

[9] I. Karlsson, "Glottal waveform parameters for different speaker types." Proc. Speech 88, 7t FASE Symposium, Edinburgh, pp. 69-72, 1988

[10] I. Karlsson, "Voice source dynamics for female speakers,"Proc. I.C.S.L.P. Kobe, pp. 69-72, 1990.