Dept. for Speech, Music and Hearing

Quarterly Progress and Status Report

Acoustic analysis of voice source dynamics

Ananthapadmanabha, T. V.

journal:	STL-QPSR
volume:	25
number:	2-3
year:	1984
pages:	001-024

KTH Computer Science and Communication

I. SPEECH PRODUCTION

A. ACOUSTIC ANALYSIS OF VOICE SOURCE DYNAMICS

T.V. Ananthapadmanabha

Abstract

A voice source model intended to capture the variations in the inverse filter output (derivative of glottal pulses) of the acoustic speech data is described. The voice source modelling is proposed in two stages: (a) analysis stage identifying five analysis variables, and (b) synthesis stage for reconstructing source pulses from the analysis variables. With such a distinction, the same analysis variables may be related to the parameters of already existing glottal pulse models or may be used to develop new models with a common set of variables. A specific computational scheme for the measurement of analysis variables from the acoustic data is proposed. Unlike the usual approach, the modelling is done directly on the inverse filter output rather than on its integral, the glottal pulses. Also, an important analysis variable is related to the nonabrupt glottal closure gesture which is highly correlated to the relative abduction state of the vocal folds. The present model thus allows an indirect tracking of the degree of abduction. Male, female and children's speech have accordingly been analysed and synthesised which has rendered high quality natural sounding speech.

1. Introduction

It has long been suspected that the quality of synthetic speech may be improved at the acoustic level by an appropriate model for source during voiced segments of speech. Here at KTH, Fant has addressed this problem on two fronts: (i) by proposing a dynamic three parameter glottal pulse model (Fant, 1979; 1980), (ii) by investigating source-filter (aerodynamic-acoustic) interaction (Fant, 1981; Fant and Liljencrants, 1979; Ananthapadmanabha and Fant, 1982; Fant and Ananthapadmanabha, 1982; Ananthapadmanabha, Nord, and Fant, 1982). In this article we will restrict our attention to the glottal pulse modelling for use in noninteractive source-filter synthesis models.

Fant (1979; 1980) made a detailed period by period measurement of the parameters of his source model on the inverse filter output for one complete utterance and described some general characteristics of voice source dynamics. The author tried to automate this procedure (Ananthapadmanabha, 1982). It was noticed that the three parameter model though economical is too constrained to capture the wide variations in the source pulse shapes. Hence, a five parameter model was proposed. This work which started in 1982 was not formally reported. Only the results were presented in a seminar and cited in an eariler article, (Ananthapadmanabha, 1982). Since then, the procedure has been made more robust and the model has been refined. Also, the analysis and synthesis has been performed on several utterances of male, female and child speakers. These results will now be presented.

2. Voice-source-dynamics

The voice source dynamics has several related aspects: (a) pitch and its variations, (b) relative intensity variations of speech sounds, (c) glottal pulse shape and its variations, and (d) underlying temporal patterns of vocal cord positioning (abduction/adduction). Of these the first two aspects are generally represented and analysis tools are also well developed. The glottal pulse shape has been modelled by several researchers, but, its relation to the physiological state has not been adequately treated. In this article we are concerned with studies of the last two aspects with an emphasis on inverse filtered waveforms and the abduction-adduction dimension.

(A) Measurement Data: The data on phonation can be acquired by direct physiological measurements such as photoglottography or electroglottography. However, such techniques provide us with qualitative information only. Further, for synthesis we require the glottal air flow data which then has to be calculated based on the glottal area and other related physiological variables. A direct method of obtaining the glottal air flow data is by inverse filtering.

The acoustic speech data are recorded with a condenser microphone and digitized into the computer without any hardware preemphasis. We prefer this approach to the recording with a Rothenberg's mask, (Rothenberg, 1973). In the latter approach a low frequency sensitive microphone is placed within a pneumotachograph mask and is supposed to give directly the volume velocity air flow at the lips. But, the following issues have to be noted: (a) The near field pressure within the mask is an approximation to the volume velocity air flow. (b) The radiation characteristics at the lips within the mask has to be properly dealt with. (c) The radiated acoustic pressure from only the mouth, excluding the output from the cavity wall vibrations, represents a pole-zero response. (d) The frequency response of the mask is limited to about 2 kHz. (e) The mask has to be designed carefully to adopt the speaker's physique. In the absence of a condenser microphone and for gross air flow measurements, Rothenberg's mask gives a better signal recording. However, for a careful inverse filtering and glottal pulse modelling, recording the far field pressure by a condeser microphone is preferable. In this work inverse filtering has been performed after formant tracking. Consonant sounds are also inverse filtered with only zeroes, instead of the zero-pole network. An example of inverse filtering of a nasal segment will be illustrated to study the approximation.

(B) Glottal Pulse Modelling: Unlike the acoustic theory of vocal transmission (Fant, 1960), we don't have a solid theoretical formulation to guide glottal pulse modelling. Titze (1984) has presented a sophisticated approach to model glottal pulses considering the effects of source-filter interaction and vocal fold mechanics. But, it presupposes that we have accurate formulations for aerodynamics, acoustics and vocal fold mechanics. Also, signal theory is well developed for describing the spectra by pole-zero transfer functions. But, to describe the finite duration pulse like waveforms there is no established theory. One theoretical approach would be to describe the pulses by orthonormal basis functions such as sine and cosine terms or Laguerre polynomials etc. But, these functions are not piecewise continuous which is required, the convergence may be slow, and mean square error minimization may not be desirable (Gibbs phenomenon). We don't have a significant insight into the perceptual significance of the features of the glottal pulses. Because of these reasons, modelling of glottal pulses has been mainly empirical.

A glottal pulse description given by Rosenberg (1971) is shown in Fig. 1a. Rosenberg (1971) studied several pulse shape models, but, the one referred to has had a significant influence on other researchers. Strube (1974) only reexpressed the same equations differently. Rosenberg (1971) and Holmes (1973) scaled the pulse in duration for pitch variations thus retaining a constant open quotient or pulse shape. Fant (1979) made the pulse shape (Fig.1b) variable by controlling the amplitude of the cosine segment over the closing phase. Some other time domain models are also shown in Fig.1. Also, see Titze (1980).

Ananthapadmanabha (1982) proposed a five-parameter model (Fig.1 d) based directly on the unintegrated inverse filter output instead of glottal pulses, as is usually done. To keep this distinction we refer to the volume velocity of air flow as glottal pulses and the inverse filter output as voice source pulses. This also keeps the physiological function as distinct from the acoustically measured function. Modelling of voice source pulses instead of glottal pulses was based on several considerations: (i) It is well known that the acoustically significant excitation occurs at the epoch near glottal closure which is easily detectable from the inverse filter output, (ii) the integration for obtaining glottal pulses often introduces low frequency drifts and makes measurements difficult, and (iii) optimization of model parameters on glottal pulses is undesirable since the signal has nearly zero level at the acoustically significant instants of the waveform (viz., the glottal onset and glottal closure). Recent modelling efforts by Liljencrants (1984) and Fant (1984) are also based on the derivative of the glottal flow. An important feature of the model (Fig. 1d), proposed in 1982, is the nonabrupt termination of the glottal air flow towards closure. Fant's model (Fig. 1g) represents another method of realizing such a nonabrupt termination. An extension of the earlier model will be discussed in Sec. 3.

(C) Abduction/Adduction is an extremely important activity of the vocal folds during the production of speech (Gauffin, 1972a; 1972b; Fujimura, 1977; Shipp, 1982). In the full extent of abduction, the vocal folds don't collide during phonation, but, rather, leave a gap through out the glottal cycle. The increasing gap area is brought about by abduction and the closing of the gap by adduction of the vocal folds. These actions occur during voice onset (adduction), voicing decay (abduction), production of stops and other consonants (abduction). The photoglottographic record for a final vowel ending (voicing decay) is shown in Fig. 2 to illustrate these remarks. This record from the termination of a vowel in an open syllable derives from an unpublished study by Fant, Gauffin, Kitzing and Löfqvist.

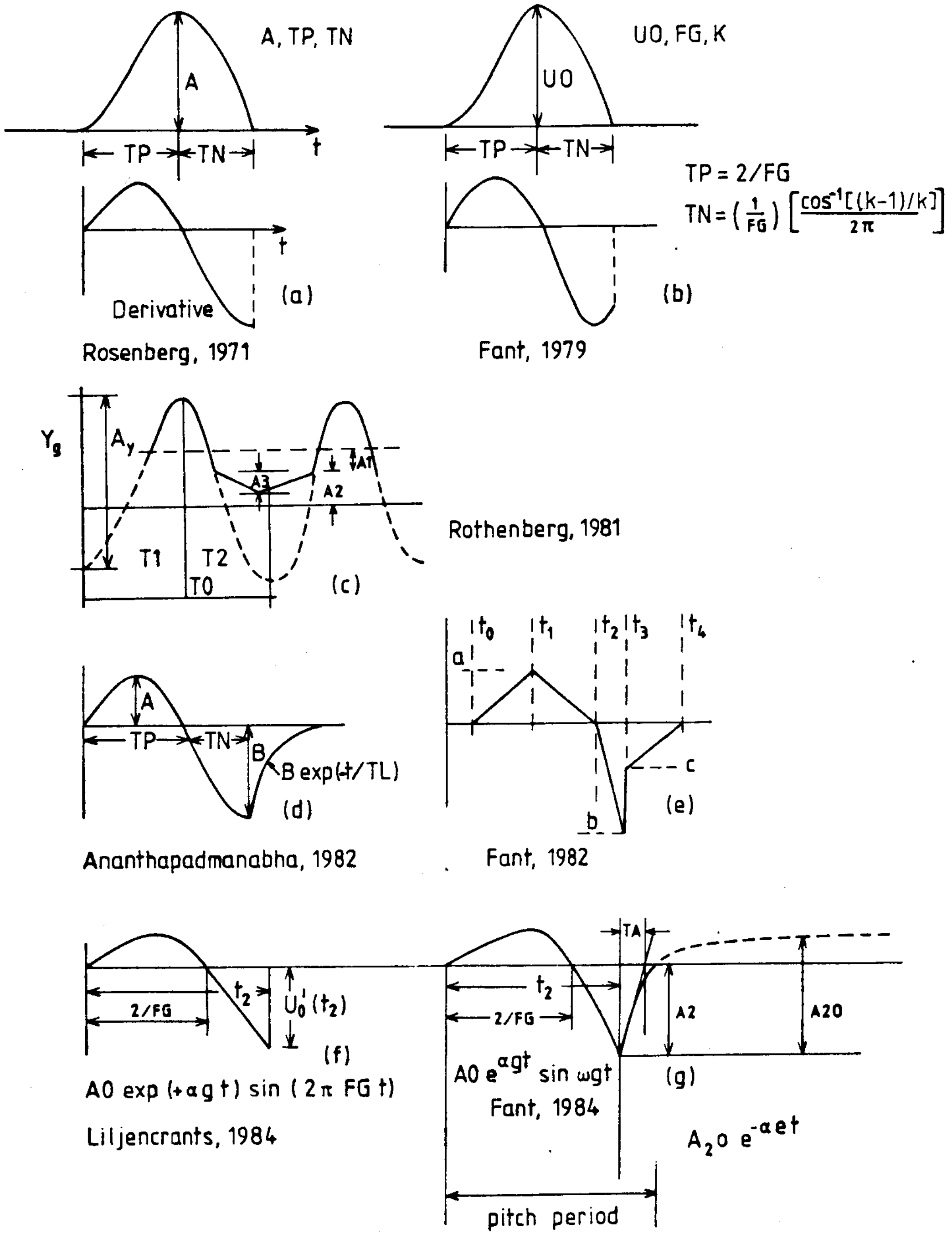


Fig. 1. Some examples of time domain models of glottal pulses.

Abduction is commonly described as a superposition of a low frequency (DC) flow on the regular on-going phonation activity. Since the duration of such an abduction gesture has a duration of about 50 msec, this superposed component will correspond to a frequency of the order of 20 Hz and hence may be deemed to be perceptually unimportant. The abduction component does not tend to be superposition only. The main effect is a rather drastic change of pulse shape. This is illustrated in Fig. 2 by comparing the photoglottographic recording with the inverse filter output. The radiated acoustic pressure of the same vowel sample was inverse filtered (Courtsey: L. Nord). A progressive and significant change occurs in the derivative of the glottal pulses immediately after the negative maximum of the pulse (marked a,b,c in Fig. 2). It may also be observed from the inverse filter output of a voiced /h/-sound (Fant, 1980).

The derivative of the pulses is not strictly zero mean for every pitch period, the bias in the derivative accounting for the DC component. Thus, the abduction activity may be observed in the first place by the relative changes in the decay characteristic of the source pulse after the negative maximum and, secondly, by the variations in the DC bias on a period by period basis. The latter is difficult to monitor and not very reliable. We shall illustrate in Sec. 4 the dynamic variations of the abduction component as measured from the acoustic signal.

3. Proposed voice source model

At present we don´t have any rigorous standards to recommend the use of one model over another. The presence of diverse models makes it difficult to scientifically communicate the results. To avoid this situation, the author proposes to model voice source pulses in two stages: (a) an analysis stage and (b) a synthesis stage. In the analysis stage certain variables are identified from the waveform based on certain features of source pulses and a mathematical theorem for describing finite duration pulses. The author hopes that these variables may be adequate to characterize the source pulses for speech sounds produced by nonpathological phonations. The measurement technique estimates these analysis variables. The synthesis stage consists of a model for constructing the pulse waveform. The parameters of the synthesis model are then related to the analysis variables. Thus, with the same analysis variables, one can construct different pulse waveforms. We will describe one possible model for reconstructing the source pulses.

3.1. Analysis variables for voice source

The terms glottal onset, glottal peak, glottal closure, opening phase, closing phase, open phase, closed phase, glottal abduction etc. have generally been used in describing the glottal pulses. Such terms are related to the physiological mechanism of phonation, the movement of the vocal folds. If we are modelling the glottal pulses obtained from inverse filtering, there may be some deviations between such physiologi-

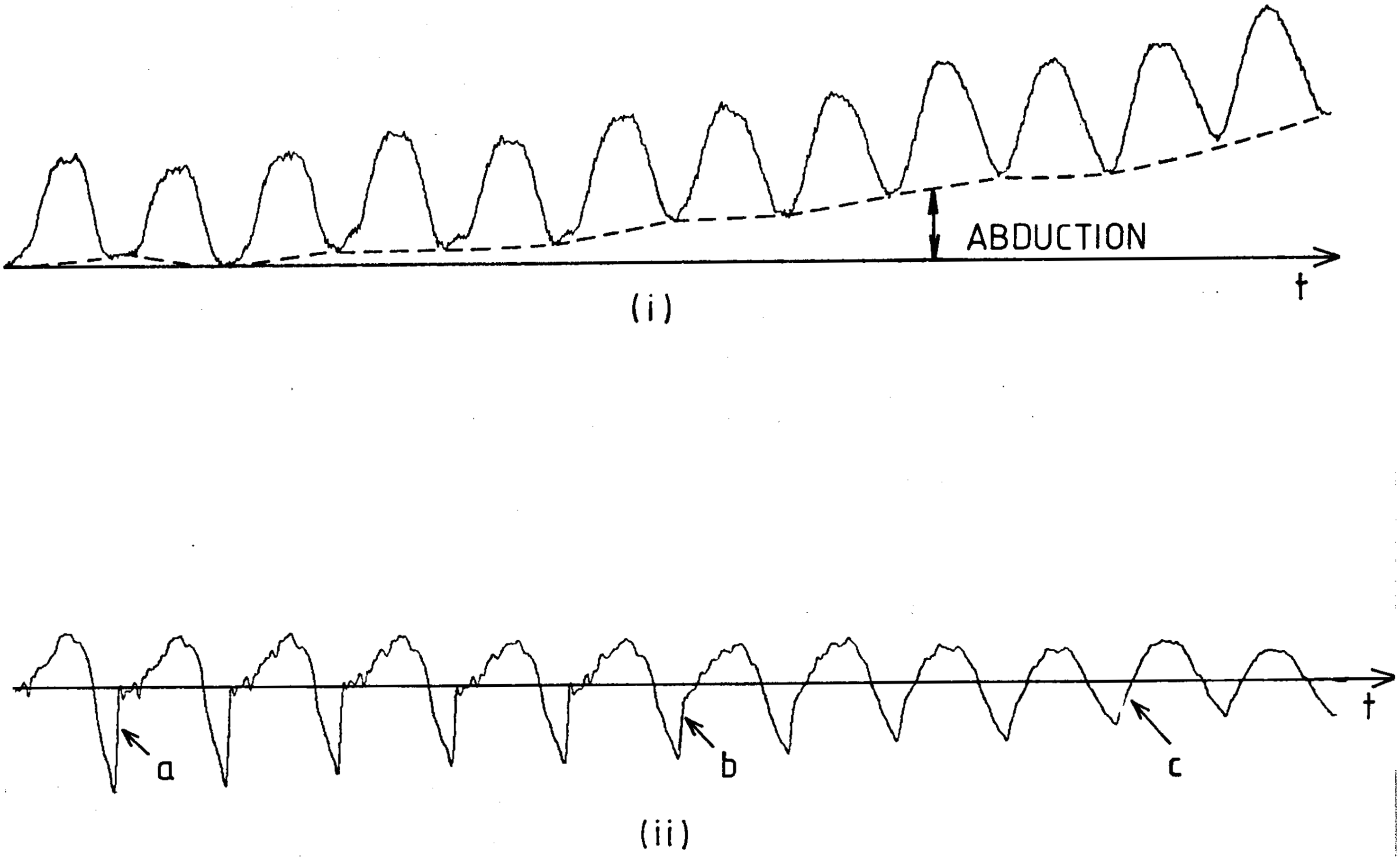


Fig. 2. (i) Glottal area function and (ii) inverse filter output during a segment of vowel termination. Note the effect of glottal abduction on the pulse shape (marked a, b and c).

cal terms and what is seen in the signal waveform. A well known example is the instant of peak of the glottal pulse which does not correpond to the instant of maximum separation of the (projected) vocal folds. In Fig. 3, a typical hand drawn glottal area waveform for a transition from a vowel to a consonant is compared with the inverse filter output. When the vocal folds do not close completely (cycles iii, iv), physiologically, the instant of the glottal closure and the glottal onset coincide. But such instants may not be acoustically significant. Therefore, the important temporal instants which possess significant acoustic influence have to be defined.

Based on these considerations, the author defined the term "epoch" using an asymptotic expansion theorem for finite duration pulses (Papoulis, 1968). This theorem suggests that the important instants (epochs) and variables for a finite duration signal are the instants of discontinuity in the waveform or its higher order derivatives, and the slope and curvature change at these instants. Thus, we can define the instant onset epoch, (TO) the closure epoch, (TE), where these are determined by the signal properties rather than by the physiological considerations. These are shown marked in Fig. 3b. The instant TE can be easily and reliably identified from the voice source pulse as the instant of the maximum negative amplitude within the pitch period. From our knowledge of the glottal area measurements, we know that there has to be a zero crossing (TP) in the derivative between the onset and the closure. This instant TP corresponds to the peak of the glottal pulse. It can be reliably identified from the waveform and is shown marked in the Fig. 3b. The instant of peak in the derivative between TO and TP will be referred to as the point of inflexion (TI), since the second derivative of the glottal pulse changes its sign here.

The above defintions have been given assuming clean inverse filter output. But, the glottal flow has ripple components due to source-filter interaction effects and due to improper inverse filter settings. An example of such an exaggerated case is shown in Fig. 3c. The improperly inverse filtered output shows double maxima and double zero crossings between TO and TE. In such cases TP refers to the zero crossing closest to TE. Also, the instant TO may not be defined clearly. The instants TO and TI are determined by imposing a higher level signal structure which appears reasonable. An inital estimate of TO is made as TO′, Fig.4. It is assumed that the signal between the TO′ and TP is a half sine wave of unknown frequency, FG, and positive amplitude, EI, Fig. 4. The sine wave frequency (FG) and the amplitude (EI) are obtained by an optimization procedure. The instant TO is marked backwards from TP as a half sine wave onset point.

When the glottal closure is not abrupt, there has to be an additional parameter to determine the degree of nonabruptness. Let the signal between TE and the immediately following zero crossing be modelled as a parabolic segment, as shown in Fig. 4. This parabolic segment has a time base equal to TL. The value TL is obtained by minimizing the mean square error between the actual signal and the parabolic segment.

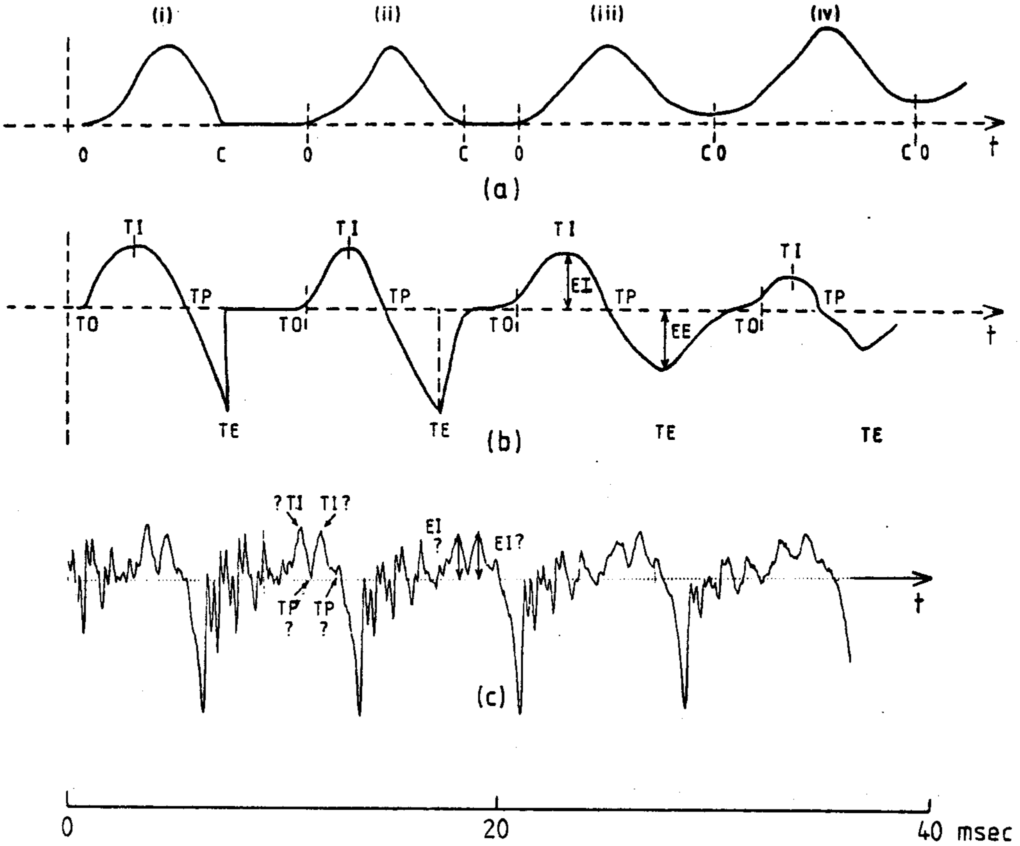


Fig. 3. (a) Typical (hand drawn) glottal area function
(b) Typical (hand drawn) inverse filter output
(c) Noisy inverse filtered data.

o: Physiological onset; c: Physiological closure;
TO: Acoustic excitation epoch near onset;
TE: Acoustic excitation epoch near closure.

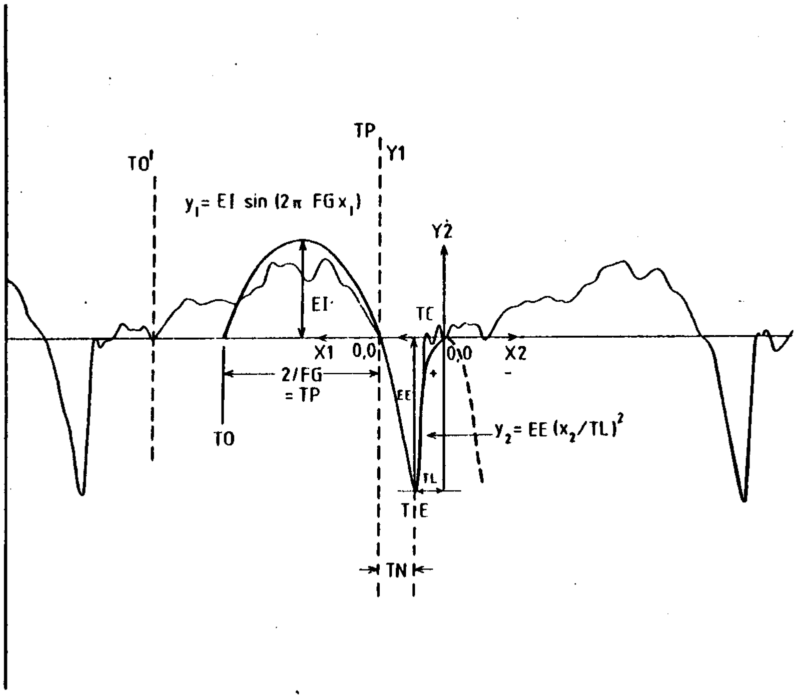


Fig. 4. Voice source analysis variables in relation to the inverse filter output.

The symmetric half sine wave between TO′ and TP and the parabolic segment between TE and TC are assumed only for the purpose of identifying the instant TO and to estimate EI and TL. But, in synthesis any desired pulse shape may be constructed.

We now have the following variables:

(a) TP: The interval between the estimated TO and TP as a per cent of pitch period (1/F0)
(b) TN: The interval between TP and TE as a per cent of (1/F0)
(c) TL: The time base of the parabola as a per cent of (1/F0)
(d) EI: The amplitude of a modelled half sine wave between TO and TP
(e) EE: The absolute amplitude of the voice source pulse at TE.

(Note that both the instant and the interval are referred to as TP as we assume TO to be the time origin equal to zero, KTH Glottal waveform Nomenclature convention, 1984.)

The variables have been chosen so that an error in the estimation of one variable does not significantly affect the other variables. The choice of the variables has been dictated by the goal of obtaining a general characterization with a minimum number of variables, with a possibility to estimate these variables reliably, and to be able to edit and interpolate the variables by examination only. To capture the variabilities due to different types of phonation (including pathologies), due to different sounds and different speakers, more number of variables may be required. However, we have obtained reasonable results with the above choice. Especially for the purpose of synthesis the above choice appears adequate.

3.2 Synthesis of glottal pulses

The same analysis variables as defined above may be used for synthesizing glottal pulses in various ways. Also, it is possible to obtain three or four parameter variations from the above analysis variables. The simplest case is to approximate the pulse shape by piecewise linear segments (Fant, 1983, Fig. 1e). The three-parameter Fant source (1979) (see Fig. 1b) imposes the following constraints:

$$U0 = (EI)/(\pi\, FG)\ ;\ K = (1/2 + (EE/EI)^2/8);\ FG = 1/(2TP)$$

Recently, Liljencrants proposed another three-parameter model (Fig. 1g). The parameters of his model can also be obtained from the analysis variables. The parameters of Fant's model (1984), Fig. 1g, can also be obtained from the same analysis variable. We call this procedure as mapping of the analysis variables to other source models in contrast to the direct matching of the model signals. In our experiments we have found both the matching and mapping to give very similar results. Thus, the same set of analysis variables can be used as a standard set of reference parameters.

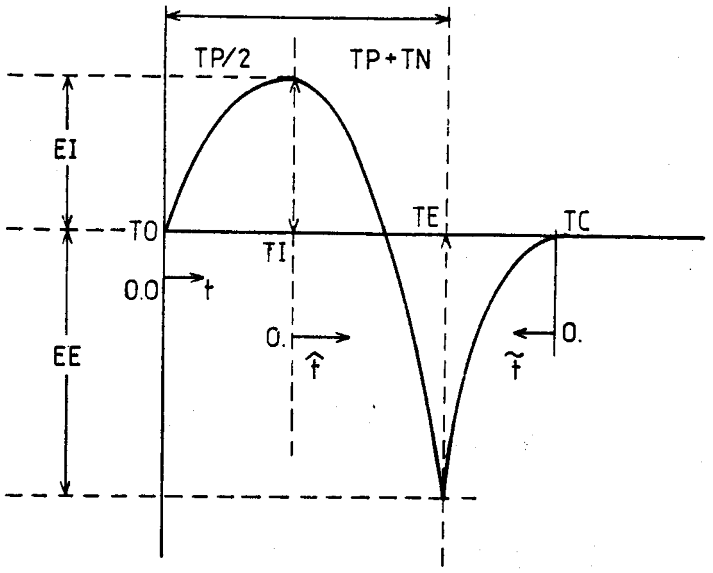
We shall now describe in some detail one particular five-parameter model for synthesis. This model is shown in Fig. 5. Although TP was estimated for the signal between TO and TP, the signal between TO and TI is considered as one segment. The signal between TI and TE is considered as a single continuous segment. This is judged from the results of source-filter interaction studies. Ananthapadmanabha and Fant (1982) calculated the true glottal flow starting from an assumed glottal area function. The derivative of the no load glottal flow and the true glottal flow for two vowel sounds are shown in Fig 6. Although the no load glottal flow (Fig. 6a) has a discontinuity in the second derivative at the zero crossing, the flow derivative with load (Fig. 6b) between TI and TE appears as a single segment. Also, the analytical analysis by Ananthapadmanabha and Fant (1982) has shown that the hypothetical inductance of the nonlinear time varying glottal impedence has a postive value during the opening phase and negative value during the closing phase. This will render the effective sinusoidal frequencies for the opening and closing phase to be different.

The average value of the derivative of the glottal pulse is usually assumed to be zero over one pitch period. But this constraint is not imposed in the present model for the following reasons. If the matching on a period by period basis is good with a zero mean error, then the synthesized pulses are assumed to be acceptable. The mean value may not be zero for every pitch period especially in the state of dynamic variation of glottal abduction. For speech data prepared with high pass filtering to 5 Hz or 10 Hz, we can expect the mean value to be zero only over time scales of the order of about 200 or 100 msec. However, if one desires, the input signal from the pitch period can be adjusted to be zero-mean before matching.

3.3 Spectral features-of glottal pulses

We have discussed glottal pulse modelling in the time domain. Frequency domain modelling based on the spectral information is not so desirable due to the following reasons: (a) the spectrum computation is time consuming, (b) the spectrum computation without artefacts is difficult, (c) the glottal pulse for synthesis cannot be reconstructed based on spectral information only, (d) the relevant spectral features are not well identified, and (e) the spectral features are indirectly related to the physiological parameters. Nevertheless, it is instructive to study the spectra of source pulses since the auditory impression is directly related to spectral features. We shall now present some spectrum calculations.

Several researchers have given spectral description of the glottal pulses, e.g., Flanagan (1958), Miller (1959), Mathews et al. (1961), Monsen and Engebretson (1977), Fant (1979), Sundberg and Gauffin (1979). A canonical high frequency spectral roll-off as multiples of 6 dB/octave is usually assumed. But, for a piecewise continuous function this rule may not be obeyed, as shown by several researchers. Another spectral feature is related to the amplitude of the spectrum at fundamental



SEGMENT	EQUATION	
TO to TI	$EI \sin\left(\frac{\pi}{2}\frac{t}{TP}\right)$	$0 \le t \le TP/2$
TI to TE	$(EI+EE)\cos\left[\frac{\pi}{4}\frac{\hat{t}}{TN}\right] - EE$	$\hat{t} = t - TP/2$
TE to TC	$-EE\left[\frac{\tilde{t}}{TL}\right]^2$	$\tilde{t} = [TP + TN + TL - t]$

Fig. 5. Synthesis of voice source pulse.

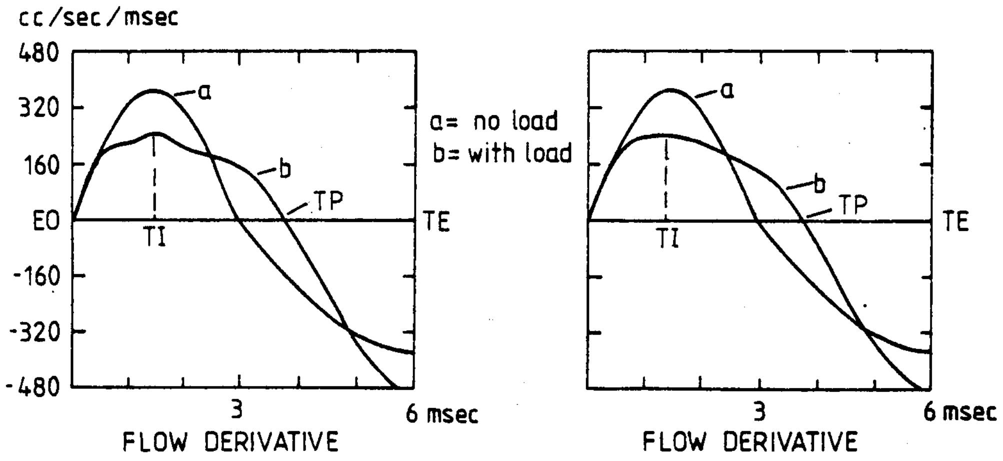


Fig. 6. Model calculation of the derivative of glottal air flow with and without source-filter interaction.

(Sundberg and Gauffin, 1979). Fant (1979) emphasizes the significance of the low frequency spctral maximum of the derivative of the source pulse. Another spectral feature of the source pulses is related to the distribution of the nulls (zeroes) in the spectrum, their locations and bandwidths. Mathews et al. (1961) and Matausek and Batalev (1980) model the glottal spectrum both by a large number of zeroes and by right half plane poles with large bandwidths.

Usually the spectrum of the glottal (volume velocity) pulse is illustrated for discussion. But, we prefer to discuss the spectral features of the voice source pulse (inverse filter output pulses). This is motivated both by computational considerations (lower dynamic range, absence of bias of integration etc.) and by auditory consideration (actual effective spctrum of the source in speech). A typical hand drawn glottal spectrum is shown in Fig. 7. We identify the following spectral features shown marked in the figure.

(a) Source Spectrum Peak: FL: This is the frequency at which the voice source pulse attains the maximum value. This may be different from FG of Fig. 4 since FG by definition is determined only by the opening phase, whereas FL is determined by the entire pulse shape (see Figs. 8-10).

(b) Source Peak Prominence: DL: This is measured as the dynamic range in dB between the spectrum at FL and the null immediately following.

(c) First Zero: ZL: This is the frequency of the first null after the source peak. This is an indirect measure of the effective open duration.

(d) Bandwidth of Source Peak: BL: Since the spectrum falls off steeply after the first maximum, we have chosen the 12 dB down bandwidth.

(e) High Frequency Prominence: DH: The dynamic range in dB between low frequency source peak maximum and a reference high frequency location give a relative measure of the high frequency dominance. This is also related to the spectrum roll off in dB/octave.

(f) Ripple Content: DR: The dynamic range of the second spectrum peak at low frequency is a measure of the ripple content of the glottal pulse. Also this determines the relative excitation strengths at onset and closure epochs. Larger the DR, deeper are the nulls, thus determining the bandwidths of the nulls of the spectrum.

We shall now present spectra of source pulses for selected variations in the analysis variables. It is very important to remember that the spectra are specific to the model chosen here. The variable EE primarily determines the energy of the source pulse and, hence, has been chosen as the reference amplitude. The log spectra of a single source pulse corresponds to the Fourier transform of the signal. The variable (TP+TN) is kept fixed since changes in this value only implies scaling in the time domain or stretching of the spectrum in the frequency domain.

The effect of varying the ratio EI/EE is shown in Fig. 8a with the analysis variables EE, TP and TN fixed, and TL=0. Although TP and TN

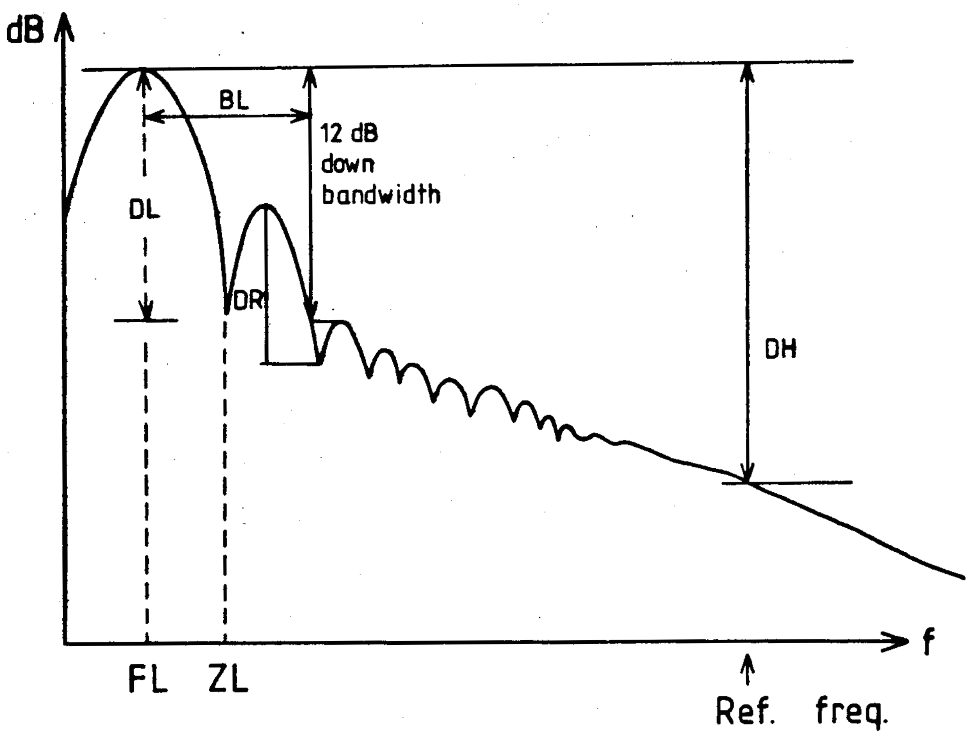


Fig. 7. Log spectrum of a typical voice source pulse and spectral features.

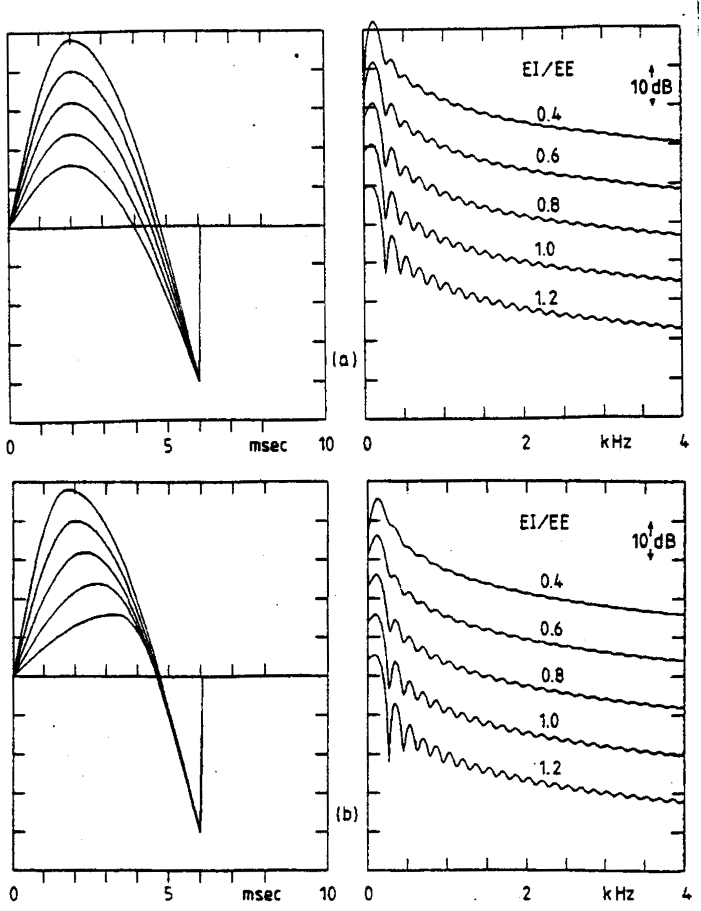


Fig. 8. Effect of varying EI/EE on the log spectrum of the voice source pulse.
(a) With the analysis variables TP and TN fixed.
(b) With the synthesis variables TP and TN fixed.

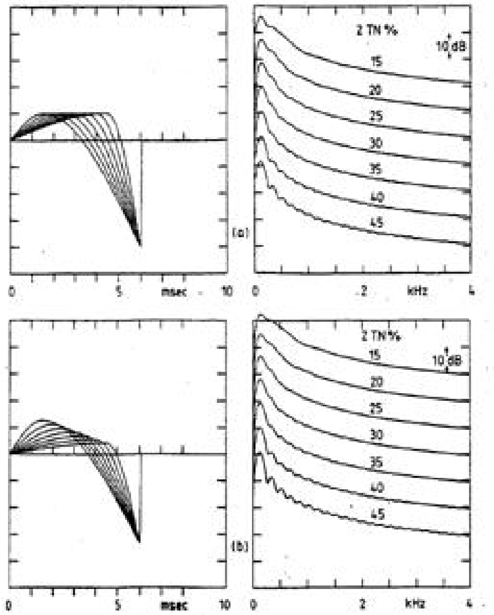


Fig. 9. Effect of varying TN on the log spectrum of the voice source pulse.

(a) without zero mean level, and
(b) with zero mean level.

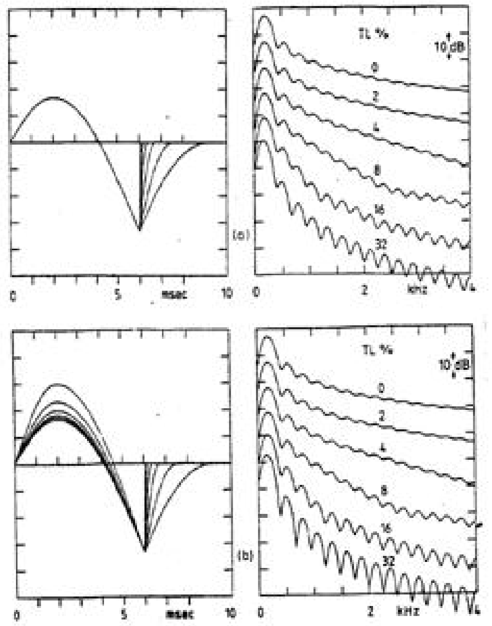


Fig. 10. Effect of varying TL on the log spectrum of the voice source pulse.

(a) without zero mean level, and
(b) with zero mean level.

are fixed, the reconstructed pulse shape shows variation of TP and TN for variations in EI due to the particular synthesis procedure (Fig. 5). But, in Fig. 8b we have kept TP and TN in the synthesized pulse nearly constant. As EI increases, the ripples (DR) in the spectrum become stronger due to the relatively greater importance of excitation at the onset. The lowering of FL may also be noted. For very low EI/EE, the spectrum approaches that of an exponential pulse.

The effect of varying TN for fixed (TP+TN), EI and EE is shown in Fig. 9a. The main effect on the spectrum is to change the 12 dB down bandwidth BL which decreases as TN increases. Also, the prominence of the spectrum peak, DL, increases. For very low TN, the spectrum approaches that of an exponential pulse. The frequency FL decreases slightly as TN increases. This shows that FG and FL could be different. In Fant's model (1979), the same sinusoidal frequency is used over the entire cycle which makes FG to be determined by the opening phase only. As TN is varied, the signal pulse is not zero mean, but has been rendered nearly zero mean by changing EI, as in Fig. 9b. However, the variations in the spectral shape are still mainly due to the variations in TN.

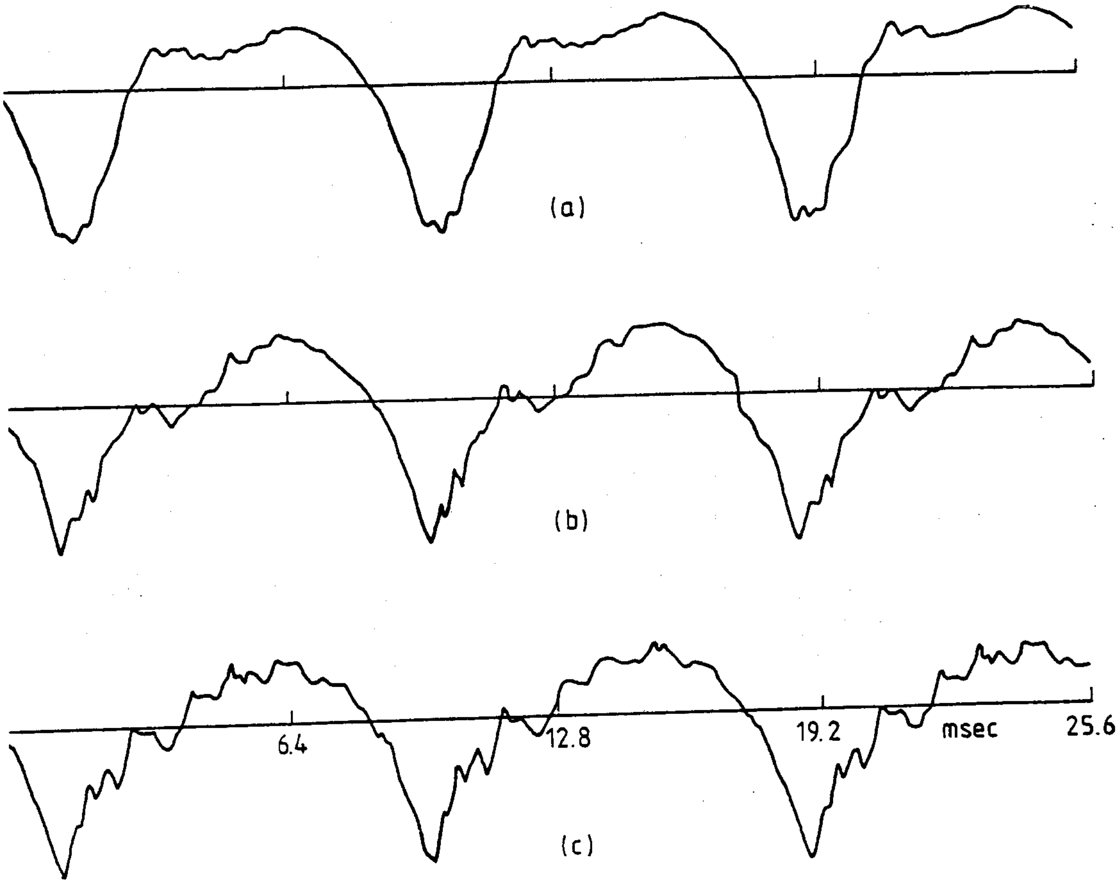
The effect of varying TL keeping all other analysis variables fixed is shown in Fig. 10a. The main effect is to reduce the high frequency prominence (DH). The relation of the decrease in DH to TL seem to obey a logarithmic law as equal changes in DH are obtained for powers of two of TL. As TL increases, the location of FL decreases for the same FG and the prominence of the spectrum peak increases. Also, the ripple components become stronger, as shown by increased values of DR. For large TL the effective excitation strength at closure epoch decreases, thus becoming comparable to the onset excitation resulting in spectrum ripples. The mean level has once again been nearly compensated by changing EI, Fig. 10b. But, the effect of TL variation dominates.

It appears that for small DC level changes, the spectrum of the pulse is not altered significantly. The spectrum shape is mainly determined by the competing excitation strengths at the onset and closure epochs, Ananthapadmanabha and Yegnanarayana (1977).

4. Results

We shall now present some results for real speech data.

The inverse filtering should be strictly done with a zero-pole network for nasals and laterals. But, we have used only zeroes in the inverse filter. The consequent effect on the inverse filter output will now be illustrated. Using the INA program of Liljencrants, Nord has inverse filtered a nasal segment in the context of the utterance /i:na/. The result is shown in Fig. 11b. The inverse filtering obtained using only zeroes is shown in Fig. 11c. A comparison shows that inverse filtering with only zeroes is adequate for the purpose of model matching. This result is not unexpected since the uncancelled nasal pole-zero pairs correspond to an almost flat spectral trend with nearly zero phase.



	F Hz	B Hz	Q
CZ	214	142	1.5
CP	314	354	.9
CZ	385	329	1.2
CZ	866	156	5.6
CP	1055	318	3.3
CZ	1509	99	15.2
CZ	2108	310	6.8
CZ	2941	129	22.8
CZ	3737	415	9.0
CZ	4572	288	15.9

(d)

	F Hz	B Hz	Q
CZ	258	150	1.7
CZ	1614	183	8.8
CZ	2906	235	12.4
CZ	3477	463	7.5
CZ	4128	1685	2.4

(e)

·Fig. 11. Inverse filtering of a nasal segment.

(a) Speech waveform and the output of inverse filter
(b) for zeroes and poles,
(c) for zeroes only.

Our aim here is to determine if the five analysis variables and the synthesis model proposed are adequate to capture a variety of source pulse shapes. All the results to be presented have been obtained by using the computer software package VISHRANTH developed by the author for speech analysis and synthesis. Some editing in an user interactive mode has been made since our aim here is to test the model and not the algorithms used for matching.

The vowel sound /a/ in the Swedish utterance "Jag heter..." spoken by a male, a female and a child is now considered for illustration. The matching is performed only for one selected pitch period, but the pulses are synthesized assuming a periodic signal with the given analysis variables. This is to check that the pulses are appropriately connected in a dynamic situation. The inverse filter output and the synthesized pulses are shown in Figs. 12-14. The selected period for matching has dotted line markers indicating the initial estimates of TO, TI, TP, TE and TE of the next pitch period. The error signal for this period alone has to be considered. The log spectra of one pitch period of the signal are shown. The inset shows the pitch in Hz, and the analysis variables.

The matching in time domain appears reasonable for all the three speaker categories. The error signal is approximately zero mean and appears to contain the ripple components only. The source pulses tend to be sinusoidal for female and child voices. This can be seen both by the large value of the variable TL and the low ratio of EI to EE. Near the frequency FL, the log spectrum shows a narrow peak for male voice, and this peak is considerably broadened for female and child voices.

There are small discrepancies with regard to spectral matching. These arise since the inverse filtering has been performed with inappropriate bandwidth settings derived from automatic analysis and in case of nonabrupt closure, the effect of subglottal system might influence the results. An improper setting of the bandwidth is equivalent to an excitation of the formant, either as a pole or as a zero depending on whether the estimated bandwidth is too narrow or too broad. The cumulative effect of several formants can introduce a gross spectral roll-off. The optimal spectral matching may be expected only for models with source-filter interaction. Since the matching is acceptable in the time domain, we thus consider the small mismatch in the spectral domain to be of lesser importance.

The results for the sound /h/ for the above male voice in the same Swedish utterance is illustrated in Fig. 15. The TL variable assumes a large value in this case. The value of TL is comparable to that for vowel /a/ of a child voice. But, the spectrum of the /h/ sound of the male voice still shows a narrow spectral peak due to the high ratio of EE/EI. The spectral ripple at high frequency in the input signal may be due to the aspiration noise of the /h/ sound.

In keeping with the title, we shall now present an example of the dynamic variations of the analysis variables for connected speech. The inverse filtered waveform over a part of a sentence is illustrated in

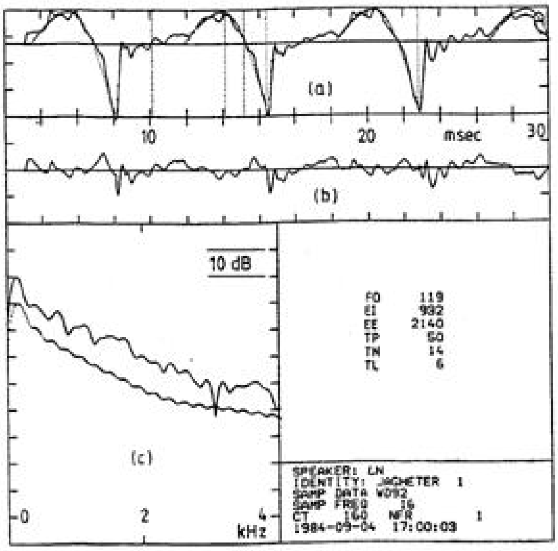


Fig. 12. Model matching results for vowel /a/, adult male speaker.

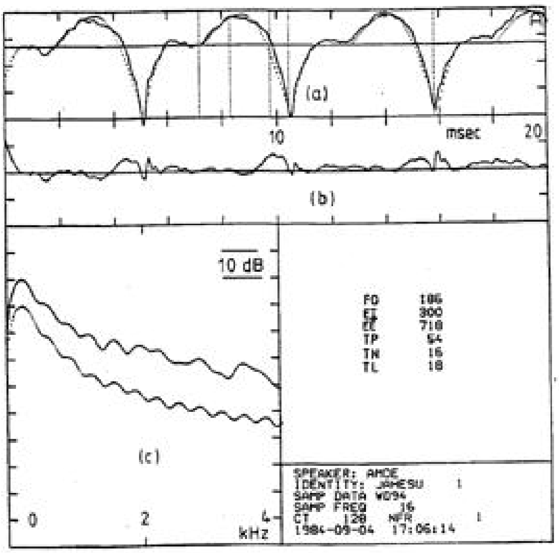
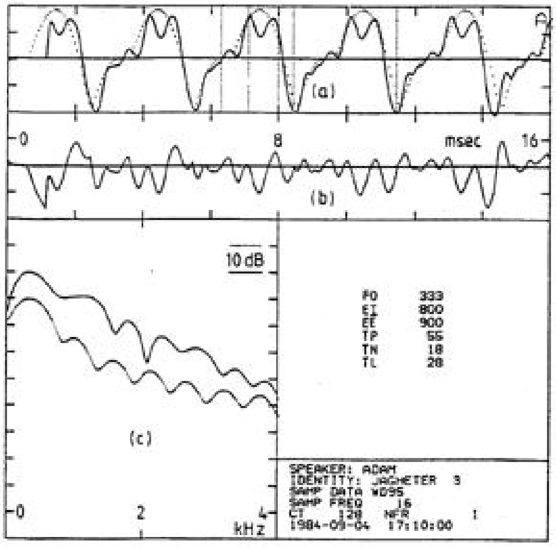


Fig. 13. Model matching results for vowel /a/, adult female speaker.

(a) Inverse filtered signal (full) and model signal (dotted).
(b) Error signal.
(c) Kig spectra of inverse filtered signal (full) and model signal (dotted).



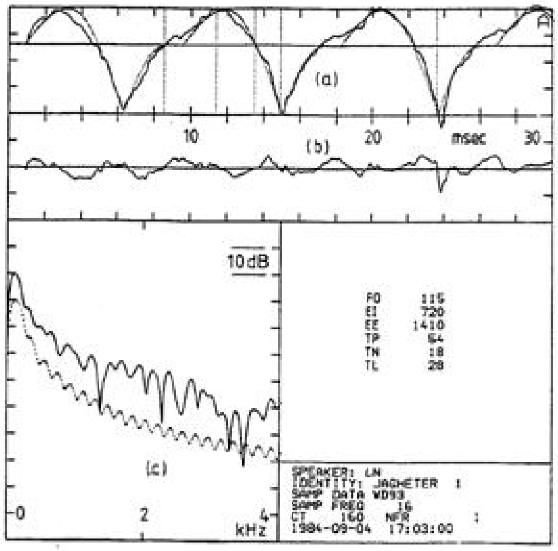


Fig. 14. Model matching results for vowel /a/, child speaker.

Fig. 15. Model matching results for vowel /a/, adult male speaker.

(a) Inverse filtered signal (full) and model signal (dotted).
(b) Error signal.
(c) Log spectra of inverse filtered signal (full) and model signal (dotted).

Fig. 16a. Because of the automatic analysis procedure and erroneous bandwidths, uncancelled formant ripples are seen in the inverse filtered waveform. However, the program to estimate the voice source analysis variables is robust against such errors. The souce pulses are reconstruced after a running analysis to estimate the souce variables, Fig. 16b. It may be seen that the dynamic variations in the pulse shapes are well reproduced in the reconstructed source pulses. Also, the nonabrupt glottal closure for consonant sounds may be noted.

The utterance in Swedish "Jag heter Lennart" spoken by an adult male speaker is then considered for analysis. The source variables and the formant frequencies estimated by a running analysis procedure are shown in Fig. 17. Several interesting points may be noted. The TL parameter shows large values over the voiced consonants, /j/, /h/ and /n/ contrasting with the low values over vowel segments. Also TL is large at the vowel onset after /t/ and in the final ending vowel. This highly correlates with the glottal abduction activity. The variable EE appears to follow the same trend as the probable course of transglottal pressure variation (not shown in the figure) over the utterance. This variable EE has a low value when there is an oral constriction as for /j/, and /l/. Also this variable shows a sudden increase over a decreasing general trend for the vowel where the primary stress of the sentence occurs (at the position marked by arrow for label EE in Fig.17). The variable EI is dominating over EE for the final ending vowel where TL is decreasing with an increase in TP. This indicates that the vowel onset (adduction) and the vowel decay (abduction) may have different characteristics.

APPLICATIONS: One important application of estimating the source dynamics is in the synthesis. We have analysed and synthesized utterances spoken by male, female and child voices. Compared to the use of either exponential pulses or fixed shape glottal pulses, or pulses with abrupt glottal closure, the synthesis with a dynamic voice source model gives a more natural sounding synthetic speech which retains the identity of the speaker. This procedure can be used in the development of a Text-to-Speech system, in speech coding (voice response) and in speech communication (vocoders).

More elaborate experiments on the relative perceptual importance of the source variables, pulse shapes etc. are needed. Systematic acoustic-phonetic studies for deriving dynamic source parameters are planned. Simultaneous recordings of the radiated acoustic pressure, oral pressure, subglottal pressure, photoglottograhs are available (Fant, Gauffin, Kitzing and Löfquist). This data will be analyzed to correlate the physiological variables, such as glottal abduction, transglottal pressure variations etc. with the voice source variables measured acoustically.

The variables EE and TL together with pitch variations can be used for segmentation of continuous speech. This could be useful in speech recognition and measurement of the phoneme durations.

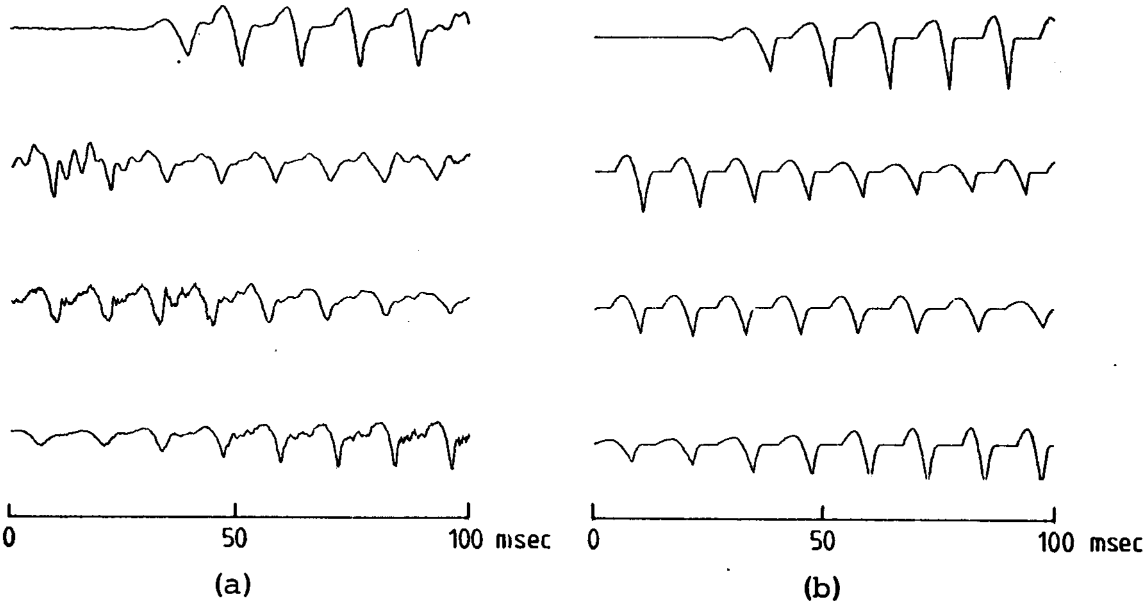


Fig. 16. (a) Inverse filter output and (b) reconstructed model pulses for connected speech: /elva/ in the utterance "aks<u>ell</u> <u>va</u>r där". (Automatic procedure.)

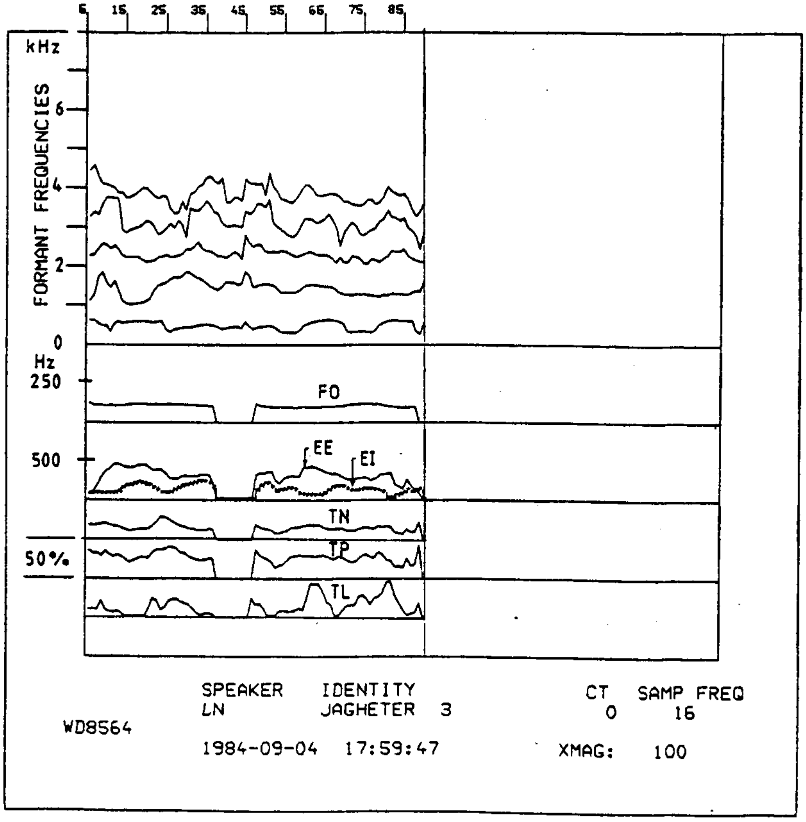


Fig. 17. Source variables and formant frequencies for one complete utterance.

4. Conclusions

A dynamic voice source model is proposed. Two stages are considered, an analysis stage and a synthesis stage. The author hopes that the analysis variables proposed may receive a general consent amongst speech researchers so that we can communicate our results more easily. An extension of the analysis variables to different types of phonations, including voice pathologies, needs to be made. We have proposed one possible scheme for reconstructing source pulses. The error signal between the inverse filter output and the reconstructed pulses could be used to estimate a high frequency noise component for a mixed noise and voice source excitation. An important outcome of the present study is the acoustic measurement of the glottal abduction parameter and the probable correlation of the EE variable with the transglottal pressure variations.

Acknowledgment

The author is grateful to Prof. Fant for his critical comments and encouragement. These comments have contributed significantly to the author's thinking. The author would also like to thank Lennart Nord who bore much of the burden of the data preparation, and participated in the synthesis of male, female and child voices, and also gave many useful feedbacks during the development of this work.

References

Ananthapadmanabha, T.V. (1982): "Intelligibility carried by speech-source functions: Implication for theory of speech perception", STL-QPSR 4/1982, pp. 49-64.

Ananthapadmanabha, T.V. and Fant, G. (1982): "Calculation of true glottal flow and its components", Speech Communication 1, nos. 3-4, Dec., pp. 167-184.

Ananthapadmanabha, T.V., Nord, L., and Fant. G. (1982): "Perceptual discriminability of nonexponential/exponential damping of the first formant of vowel sounds", pp. 217-222 in (R.Carlson & B.Granström, eds.) The Representation of Speech in the Peripheral Auditory System, Elsevier Biomedical Press, Amsterdam.

Ananthapadmanabha, T.V., and Yegnanarayana, B., (1977): "Zero-phase inverse filtering for extraction of source characteristics", in Conf. Rec., IEEE Int. Conf. on Acoust., Speech and Signal Processing, pp. 336-339.

Fant, G. (1960): Acoustic-Theory-of-Speech-Production, Mouton, The Hague.

Fant, G. (1979): "Vocal source analysis - a progress report", STL-QPSR 3-4/1979, pp. 31-54.

Fant, G. (1980): "Voice source dynamics", STL-QPSR 2-3/1980, pp. 17-37.

Fant, G. (1981): "The source-filter concept in voice production", STL-QPSR 1/1981, pp. 21-37.

Fant, G. (1982): "The voice source - acoustic modelling", STL-QPSR 4/1982, pp. 28-48.

Fant, G. and Ananthapadmamabha, T.V. (1982): "Truncation and superposition", STL-QPSR 2-3/1982, pp. 1-17.

Fant, G. and Liljencrants, J. (1979): "Perception of vowels with truncated intraperiod decay envelopes", STL-QPSR 1/1979, pp. 79-84.

Fant, G. (1984): Personal communication.

Flanagan, J. (1958): "Some properties of glottal sound source", J. Speech Hear. Res. 1, pp. 99-116.

Fujimura, O. (1977): "Physiological functions of the larynx in phonetic control", Int.Congr.Phon.Sciences, Miami, Dec. 1977.

Holmes, J.N. (1973): "The influence of glottal waveform in the naturalness of speech from a parallel formant synthesizer", IEEE Trans. Audio and Electro Acoust. AU-21, pp. 298-305.

Liljencrants, J. (1984): "Glottal waveform modelled with unstable filter", (unpublished report), Personal communication.

Lindqvist-Gauffin, J. (1972a): "A descriptive model of laryngeal articulation in speech", STL-QPSR 2-3/1972, pp. 1-9.

Lindqvist-Gauffin, J. (1972b): "Laryngeal articulation studied on Swedish subjects", STL-QPSR 2-3/1972, pp. 10-27.

Matausek, M.R. and Batalev, V.S. (1980): "A new approach to the determination of glottal waveform", IEEE Trans. ASSP-28, pp. 616-622.

Mathews, M.V., Miller, J.E., and David, E.E. (1961): "Pitch synchronous analysis of voiced sounds", J.Acoust.Soc.Am. 33, pp. 179-186.

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

Monsen, R.B. and Engebretsson, A.M. (1977): "Study of variation in the male and female glottal wave", J.Acoust.Soc.Am. 62, pp. 981-993.

Papoulis, A. (1968): Systems and Transforms with Applications in Optics, McGraw-Hill, New York, ch. 7.

Rosenberg, A.E. (1971): "Effect of glottal pulse shape on the quality of natural vowels," J.Acoust.Soc.Am. 49, pp. 583-590.

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. (1981): "An interactive model for the voice source", STL-QPSR 4/1981, pp. 1-17.

Shipp. T. (1982): "Aspects of voice production and motor control", in (S. Grillner, B. Lindblom, J. Lubker, and A. Persson, eds.) Speech Motor Control, Wenner-Gren Symposium series-Vol. 36, Pergamon Press, Oxford.

Strube, H.W. (1974): "Determination of the instant of glottal closure from the speech wave", J.Acoust,Soc.Am. 56, pp. 1625-1629.