COMPUTER SCIENCE DEPARTMENT

UNIVERSITY OF CRETE

On the glottal flow derivative waveform and its properties

A time/frequency study

George P. Kafentzis

Bachelor's Dissertation

29/2/2008

Supervisor: Yannis Stylianou

To my parents

Στους γονείς μου

Contents:

1. Introduction

In this work, the glottal flow derivative waveform of speech signal is studied.

The goal of this text is to estimate the glottal flow derivative from speech waveforms, model part of its important features, and review the spectral characteristics of the glottal flow derivative waveform.

The next chapter provides the basic mathematical framework for the linear model of speech production. Then, the basic properties of glottal flow and glottal flow derivative waveforms are illustrated, as well as a model of the glottal flow derivative, called the LF-model. This is followed by the estimation of the glottal flow derivative waveform directly from the speech signal by inverse filtering the speech with a vocal tract estimate obtained during the glottal closed phase. The closed phase is determined through a sliding covariance analysis with a very short time window and a one sample shift. This allows calculation of formant motion within each pitch period predicted by Ananthapadmanabha and Fant to be a result of nonlinear source-filter interaction during the glottal open phase. The timing of the closed phase can be determined by identifying the timing of formant modulation from the formant tracks. Then, the glottal flow derivative is modeled using the LF model to capture the coarse structure. Finally, an analytic formula of the glottal flow derivative is studied and some of its spectral properties are highlighted.

2. All Pole Modeling of Speech Signals

2.1. Time - dependent Processing

It is known that an essential property of speech production is that the vocal tract and the nature of its source vary with time and that this variation can be rapid. However, many analysis techniques assume that these characteristics change relatively slowly, which means that, over a short-time interval of 20-40 ms, the vocal tract and its input are stationary. Stationarity means that the vocal tract shape, and thus its transfer function, remains fixed (or nearly fixed) over this short time interval. In addition, a periodic source is characterized by a steady pitch and glottal airflow function for each glottal cycle within the short-time interval.

In analyzing the speech waveform, we apply a sliding window whose duration is selected to make the short-time stationarity assumption approximately valid. We select a window duration to make a good trade between time resolution and frequency resolution, typically of duration 20-40 ms. Our selected window slides at a frame interval sufficient enough to follow changing speech events, typically 5-10 ms, and thus adjacent sliding windows overlap in time. The shape of the window also contributes to the time and frequency resolution. For example, the rectangular window has a narrower mainlobe than the tapered Hamming window, but higher sidelobe structure.

In performing analysis over each window, we estimate the vocal tract transfer function parameters (vocal tract zeros and poles), as well as parameters that characterize the vocal tract input of our discrete time model. The short-time stationarity condition requires that the parameters of the underlying system are nearly fixed under the analysis window and therefore that their estimation is meaningful.

2.2. Linear Predictive Analysis

At first, we begin by considering a transfer function model from the glottis to the lips output for speech signals with periodic or impulsive source. During voicing, the transfer function consists of glottal flow, vocal tract and radiation load contributions given by the all-pole z-transform:

$$H(z) = AG(z)V(z)R(z) = \frac{A}{1 - \sum_{k=1}^{p} c_k z^{-k}}$$

We have:

$$1 - cz^{-1} = \frac{1}{\sum_{k=0}^{\infty} c^k z^{-k}} = \frac{1}{\prod_{k=0}^{\infty}(1 - b_k z^{-1})}, |z| > |c|$$

which in practice is approximated by a finite set of poles as $c^k \to 0$ with $k \to \infty$. The basic idea is that each speech sample is approximated as a linear combination of past speech samples. We can write:

$$H(z) = \frac{S(z)}{U_g(z)} = \frac{A}{1 - \sum_{k=1}^{p} c_k z^{-k}} \Leftrightarrow S(z)\left[1 - \sum_{k=1}^{p} c_k z^{-k}\right]$$
$$= S(z) - \sum_{k=1}^{p} c_k S(z) z^{-k} = A U_g(z)$$

which in the time domain is written as

$$s[n] = \sum_{k=1}^{p} c_k s[n-k] + Ae[n]$$

where $e[n] = u_g[n]$. The above equation is sometimes referred to as an autoregressive (AR) model. The coefficients c_k are referred to as the *linear prediction coefficients*, and their estimation is termed *linear predictive analysis*. The number of the prediction coefficients p is referred to as the prediction *order*.

In order to estimate the filter $h[n]$ from the speech signal $s[n]$, we set up a least-squares minimization problem where we wish to minimize the error

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k],$$

where $a_k$are calculated estimates of c_k. The total error is given by

$$E = \sum_{R} e^2[n],$$

where the error is to be minimized for the region R. There are many different techniques of linear prediction, based on how $e[n]$ is calculated over the region R. If we assume that the speech signal is zero outside of an interval $0 \le n \le N-1$, then the signal $e[n]$ will be non-zero only during the interval $0 \le n \le N+p-1$, which gives us the region R. This choice will give large errors at the start of the interval, since we are trying to predict non-zero speech samples from zero, as well as at the end, where we are trying to predict zero samples from non-zero data. These assumptions result in the *autocorrelation method* of linear prediction, since the solution to this problem involves an autocorrelation matrix,

$$\boldsymbol{R}\vec{a} = \vec{r},$$

where the $(i,j)^{th}$ term of $\boldsymbol{R}$ is given by $r_{i,j}$, where

$$r_{i,j} = \sum_{n=0}^{N-1-|i-j|} s[n]s[n+|i-j|],$$

where $1 \leq i,j \leq p$. The two vectors are given by

$$\vec{a} = [a_1, a_2, \dots, a_p]^T,$$

$$\vec{r} = [r_{0,1}, r_{0,2}, \dots, r_{0,p}]^T.$$

The primary benefit of the autocorrelation method is that is it guaranteed to produce a stable filter. The autocorrelation technique will calculate the correct filter only if the analysis window is of infinite length, due to the large errors at the beginning and the end of the window. To help reduce the effects of using a finite data window, the data is typically windowed with a non-rectangular window.

If $e[n]$ is calculated over a finite region, with the appropriate speech samples before the window used in the calculation of $e[n]$, the solution to the minimization problem is called the *covariance method* of linear prediction:

$$\boldsymbol{\Phi}\vec{\alpha} = \vec{\psi},$$

where the $(i,j)^{th}$ term of $\boldsymbol{\Phi}$ is given by $\varphi_{i,j}$, where

$$\varphi_{i,j} = \sum_{n=0}^{N-1} s[n-i]s[n-j] : 1 \leq i,j \leq p$$

and the two vectors are given by

$$\vec{a} = [a_1, a_2, \dots, a_p]^T,$$

$$\vec{\psi} = [\varphi_{0,1}, \varphi_{0,2}, \dots, \varphi_{0,p}]^T.$$

This matrix problem can be solved efficiently used Cholesky decomposition because the matrix $\boldsymbol{\Phi}$ has the properties of a covariance matrix.

The benefit of the covariance method is that with its finite error window, a correct solution will be achieved for any window length greater than *p* if no noise is present. Also, since the boundaries are handled correctly, a rectangular window can be used with no ill-effects. For a more detailed discussion of linear prediction, including derivations for the solutions given, see [8].

From a spectral standpoint, linear prediction attempts to match the power spectrum of the signal $s[n]$ to the predicted filter given by the a_i's. In particular, the error function $e[n]$ is given in the frequency domain by:

$$E(\omega) = \frac{P(\omega)}{\hat{P}(\omega)},$$

where $P(\omega)$ is the power spectrum of the signal $s[n]$, and the $\hat{P}(\omega)$ is the power spectrum of the estimated filter. If the excitation function has a non-uniform spectrum, the a_i's calculated will be influenced to result in a spectrum $\hat{H}(z)$ that matches $H(z)E(z)$.

2.3 Inverse Filtering

We can estimate the excitation signal $e[n]$ from the speech signal $s[n]$ and the estimated vocal tract response given by the a_i's:

$$\hat{e}[n] = s[n] - \sum_{i=1}^{p} a_i s[n-i],$$

or in the frequency domain,

$$\hat{E}(\omega) = S(z)\frac{1}{\hat{H}(z)} = E(z)H(z)\frac{1}{\hat{H}(z)}.$$

These equations describe a process called *inverse filtering*, in which the estimated vocal tract response is removed from the speech to yield an estimate $\hat{e}[n]$ of the source function.

2.4 Pre-emphasis

Speech signals are commonly *pre-emphasized* before linear prediction analysis is performed. Pre-emphasis is the process of filtering the speech signal with a single zero high pass filter:

$$s_p[n] = s[n] - \beta_p s[n-1],$$

where β_p is the pre-emphasis coefficient. The value used for β_p is typically around 0.9 to 0.95.

While it is difficult to find reasoning for using pre-emphasis in the literature, we give two reasons here. As discussed above, the filter estimated by linear prediction will match the power spectrum of the combined excitation and vocal tract. The excitation has a spectral shape which has more energy at low frequencies than high

frequencies, as will be seen below. In order to approximately remove the large-scale spectral contribution of the source, the speech signal is pre-emphasized. The resulting spectrum is a closer representation of the vocal tract response, and thus the filter calculated through linear prediction is a better match for the vocal tract response.

The other reasoning for pre-emphasis is an argument based on the spectral properties of the error function minimized. As was seen earlier, the error is the ratio of the two power spectrum, which results in uniform spectral matching in a squared sense regardless of the energy at any particular frequency. Speech spectra are typically viewed on a log or dB plot, however, which will show better matching for high energy regions of the spectrum than for low energy regions. Since speech tends to have a decrease in energy at high frequencies, the high-pass filter effect of pre-emphasis will help achieve more uniform spectral matching in a log sense across the entire spectrum.

3. Glottal Flow and Glottal Flow Derivative Waveform

3.1. The Glottal Flow Waveform

According to the anatomy and physiology of speech production, the glottal flow is the airflow velocity waveform that comes out of the glottis and enters the vocal tract. If we were to measure the flow velocity at the glottis as a function of time, we would obtain a waveform approximately similar to that illustrated below:

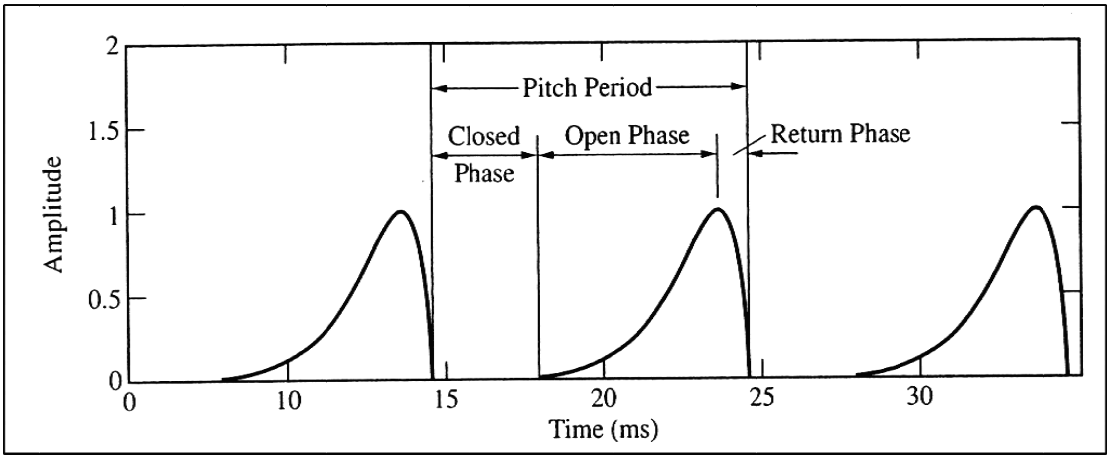

Figure 1: Glottal Airflow Model

Typically, with the folds in a closed position, the flow begins slowly, builds up to a maximum, and then quickly decreases to zero when the vocal folds abruptly shut. The time interval during which the vocal folds are closed, and no flow occurs, is referred to as the glottal *closed phase*; the time interval over which there is nonzero flow and up to the maximum of the airflow velocity is referred to as the glottal *open phase*, and the time interval from the airflow maximum to the time of glottal closure is referred to as the *return phase.* The specific flow shape can change with the speaker, the speaking style, and the specific speech sound. In some cases, the folds do not even close completely, so that a closed phase does not exist.

The time duration of the one glottal cycle is referred to as the *pitch period* and the reciprocal of the pitch period is the corresponding *pitch*, also referred to as the *fundamental frequency*. In conversational speech, during vowel sounds, we might see

typically one to four pitch periods over the duration of the sound, although the number of pitch periods changes with numerous factors such as stress and speaking rate. The rate at which the vocal folds oscillate through a closed, open, and return cycle is influenced by many factors. These include vocal folds muscle tension (as the tension increases, so does the pitch), the vocal fold mass (as the mass increases, the pitch decreases because the folds are more sluggish), and the air pressure behind the glottis in the lungs and trachea, which might increase in a stressed sound or in a more excited state of speaking (as the pressure below the glottis increases, so does the pitch). The pitch range is about 60 Hz to 400 Hz and typically the males have lower pitch than females because their vocal folds are longer and more massive.

3.1. The Glottal Flow Derivative Waveform

The glottal flow derivative waveform and its relation to the glottal flow are illustrated below:

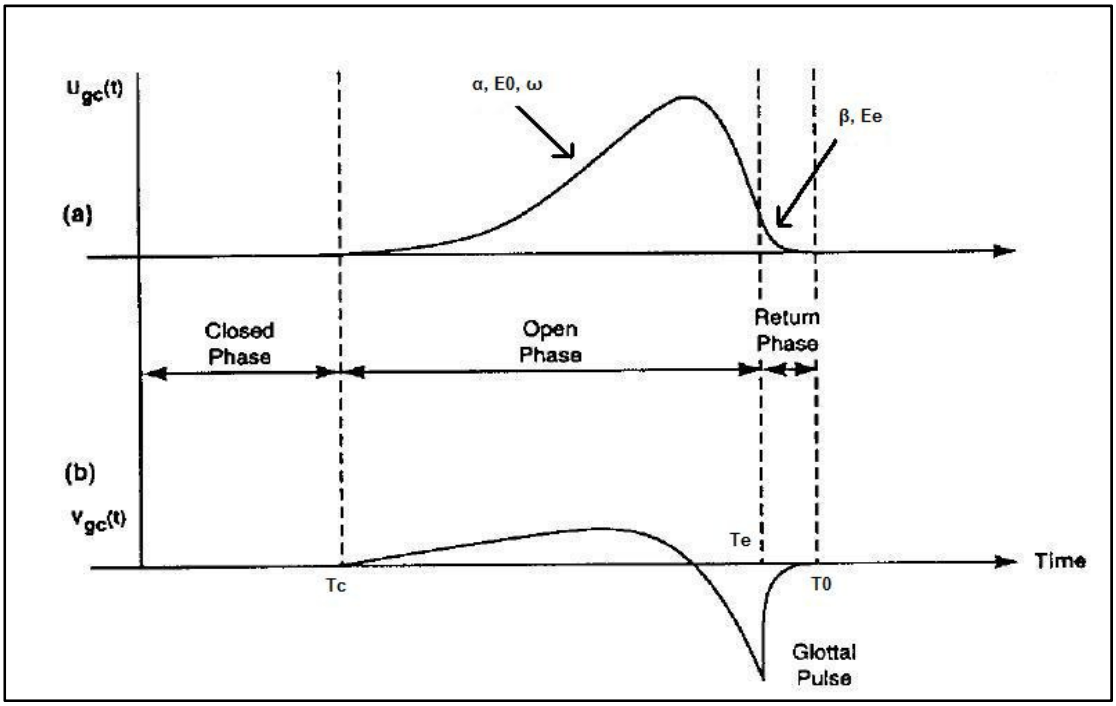

Figure 2: Glottal Flow and Glottal Flow Derivative

In order to simplify the problem of representing the glottal flow derivative, we can separate it into two main parts, the coarse and the fine structure of the flow. The coarse structure includes the large-scale portions of the flow, primarily the general shape. The fine structure includes the ripple and aspiration. We will consider only the coarse structure for this text.

Vowel production can be viewed as a simple linear filtering problem, where the system is time invariant over short time periods. Under these assumptions, the glottal

flow, acts as the source, while the vocal tract acts as a filter. The glottis opens and closes pseudo-periodically at a rate between approximately 50 and 300 times per second. As we have already mentioned, the period of time during which the glottis is open is referred to as the open phase, and the period of time in which it is closed is referred to as the closed phase. The *open quotient* is the ratio of the duration of the open phase to the pitch period, and is generally between 30 and 70 percent. The closing of the glottis is particularly important, as this determines the amount of high frequency energy present in both the source and the speech, this period of time is called the return phase.

Under steady-state non-interactive conditions, the glottal flow would be proportional to the glottal area. The time-varying area of the glottis, and source-filter interaction modify the flow in several ways. The first change is the skewing of the glottal flow to the right with respect to the glottal area function. The air flowing through the glottis increases the pressure in the vocal tract, which causes loading of the glottal flow. This loading results in pulse skew to the right, as the loading slows down the acceleration of air through the glottis. Since closing the glottis eliminates loading, the glottal flow tends to end suddenly.

If we apply the radiation effect to the source rather than the output speech, the rapid closure caused by pulse skew results in a large negative impulse-like response at glottal closure, called the glottal pulse, which was illustrated above. The glottal pulse is the primary excitation for speech, and has wide bandwidth due to its impulse-like nature. From the glottal flow derivative, we can see the reasoning for the term return phase. After the peak of the glottal pulse, it takes some time for the waveform to return to zero. Fant has shown that for one model of the return phase, the effect is to filter the source with a first order lowpass filter. The more rapidly the glottis opens, the shorter the return phase. If a glottal chink or other DC glottal flow is present, the return phase will be lengthened.

As we mentioned, we consider the glottal flow derivative as currently described to be the coarse structure of the source. The features of the source tend to have a smooth spectral content, and are of fixed positioning in relation to the glottal pulse. The extent of the features determines their timing in relation to the glottal pulse. For example, a glottis that closes slowly will result in a longer return phase, but it is not possible for the return phase to occur before the pulse.

3.2. The Liljencrant-Fant Model (LF Model)

The Liljencrants-Fant model provides a parameterized version of the coarse structure of the glottal flow derivative. The coarse structure is dominated by the motion and size of the glottis and pulse skew due to loading of the source by the vocal tract. The features we want to capture through the coarse structure include the

open quotient, the speed of opening and closing, and the relationship between the glottal pulse and the peak glottal flow. The open quotient is known to vary from speaker to speaker, and has been shown empirically to adjust the relative amplitudes of the first few harmonics. Breathy voices tend to have large open quotients, while pressed voices have smaller open quotients.

The relationship between the peak glottal flow and the amplitude of the glottal pulse indicate the efficiency of the speaker. As mentioned previously, the glottal pulse is the primary excitation for voiced speech. Thus it is the slope of the glottal flow at closure, rather than the peak glottal flow that primarily determines the loudness of the speaker. Ripple can also play a role in efficiency, if the ripple is timed such that the supra-glottal pressure is at a maximum at the same time as the glottal flow. In this case, the ripple will tend to lessen the glottal flow, but not impact the rate of closure.

The model we use is described by the following equations:

$$V_{LF}(t) = \begin{cases} E_0 e^{at} sin\omega_g t, & 0 \le t \le T_e \\ \dfrac{-E_0}{\beta T_a}[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}], & T_e \le t \le T_c \\ 0, & elsewhere \end{cases}$$

where T_e, T_0, T_c are illustrated on the figure above. The model is considered a four parameter model. Three of the parameters describe the open phase; they are $E_0, \omega_g, a,$ with one parameter describing the return phase, T_a. In order to ensure continuity between the open and return phases at the point T_e, β is dependent on T_a. While the relationship between β and T_a cannot be expressed in closed form, $\beta \approx T_a$ for small values of T_e. Generally, it is assumed that T_0 coincides with T_c from the previous pitch period, requiring only that the timing of T_e in relation to T_0 to be known. This assumption results in no period for which the glottis is completely closed; however, a small T_a will result in flow derivative values essentially equal to zero, due to the exponential decay during the return phase. The parameter T_a is probably the most important parameter in terms of human perception, as it controls the amount of spectral tilt present in the source. The return phase of the LF model is equivalent to a first order low-pass filter [6] with a corner frequency of

$$F_a = 1/(2\pi T_a).$$

This equation illustrates the manner in which the parameter T_a controls the spectral tilt of the source, and thus the speech output. The parameter a determines how rounded the open phase is, while the parameter ω_g determines how rounded the left side of the pulse is. These parameters primarily influence the relationships between the first few harmonics of the source spectrum.

In order to express the model in a closed form, an assumption can be made that $\beta = 1/T_a$, for small values of T_a, while, generally, $\beta T_a = 1 - e^{-\beta(T_0-T_e)}$. Also, the

time variable is normalized during the open phase by the time difference between T_0 and T_e, which at time T_e gives the equation

$$E(t) = E_0 e^a sin\omega_g.$$

4. Calculation of the Glottal Flow Derivative Waveform Estimate

The theory for the production of voiced speech suggests that an accurate vocal tract estimate can be calculated during the glottal closed phase, when there is no source/vocal tract interaction. This estimate can then be used to inverse filter the speech signal during both the closed and the open phases. Any source/vocal tract interaction is thus lumped into the glottal flow (or its derivative), the source for voiced speech, since the vocal tract us considered fixed.

4.1. Determination of the Closed Phase

The first and most difficult task in an analysis based on inverse filtering from a vocal tract estimate calculated during the closed phase is identification of the closed phase. A rough approximation of the beginning of the closed phase can be determined through inverse filtering the speech waveform. Since linear prediction matches the spectrum of the signal analyzed, inverse filtering a signal $S(z)$ with a filter $\hat{S}(z)$ determined by linear prediction by linear prediction will result in an approximately white signal:

$$\left| S(z) \frac{1}{\hat{S}(z)} \right| \approx 1.$$

For periodic speech signals, inverse filtering will result in impulses that occur at the point of primary excitation, the glottal pulse. The exact timing of these pitch pulses can be identified by finding the largest sample approximately every T_0 samples, where T_0 is the pitch period. This procedure is known as *peak picking.* The return phase shows that complete glottal closure does not occur until a short time after the glottal pulse, so additional processing is needed to find the onset of the closed phase.

Determination of the glottal opening is much more difficult, since the glottal flow develops slowly, and glottal opening does not cause a significant excitation of the vocal tract. As discussed earlier, formant modulation will occur when the glottis is

open. By tracking the formants during a pitch period, the time at which the formants begin to move can be identified. This will be when the glottis begins to open.

To identify the closed phase, a two step procedure is therefore used:

I. Identify glottal pulses through peak picking of an initial whitening of the speech. This provides a frame for each pitch period in which to identify the closed phase.
II. Determine the closed phase as the period during which formant modulation does not occur. This formant modulation occurs due to source-filter interaction whenever the glottal opening is changing.

4.1.1. Initial Glottal Closure Estimate

In order to ease the analysis, pitch estimates and voicing probabilities are required as input to the system, along with the speech. The pitch estimates and voicing probabilities are generated with one estimate every 10 ms and an analysis windows of length 30 ms. Most any pitch estimator could be used. This pitch information is used to perform a pitch synchronous linear prediction. The covariance method of linear prediction is used, because it will generate a more accurate spectral match. The goal of this initial linear prediction is not an accurate model of the vocal tract, rather, the goal is an inverse filtered waveform amenable to peak picking.

The size of the rectangular analysis window is two pitch periods, and the window shift is one pitch period. The location of the glottal pulse within this window is not controlled. This initial analysis is used to inverse filter the waveform. The resulting source estimate tends to be very impulse-like, easing the identification of the glottal pulse. The figure below shows an example:

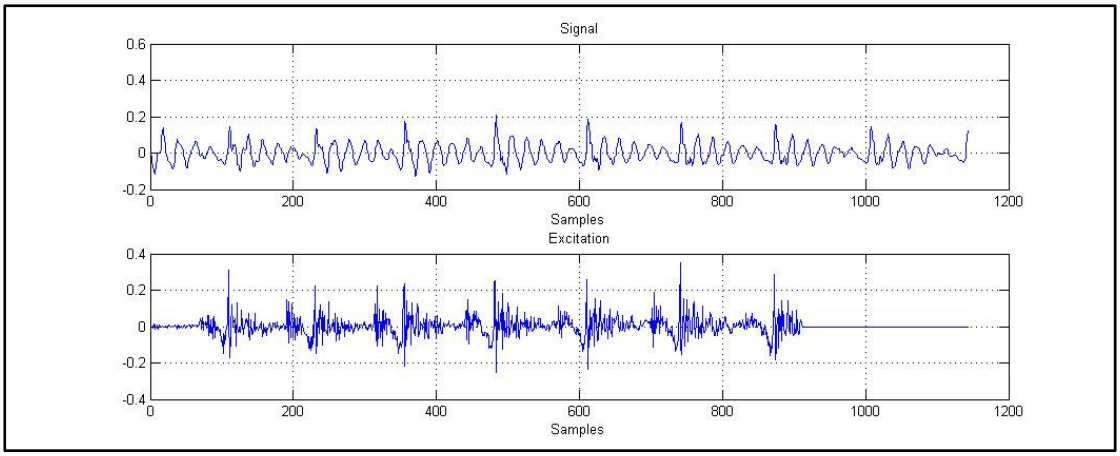

Figure 3: Speech signal and signal excitation

The peaks of the inverse filtered waveform are identified as follows: The voicing probabilities taken as input to the system are used to identify voiced regions in the speech. Each voiced region will consist of one or more voiced phonemes, such as the entire word “man”. In order to identify all the glottal pulses, we will first identify one pulse which we expect to identify with a good deal of accuracy. The remaining glottal pulses will be identified in small regions around where the pitch estimates predict they should occur.

For each voiced region, the largest peak is found; this is considered to be a glottal pulse. The pitch information provided as input to the system is used to give an estimate of the location of the glottal pulse. A small window around this estimated location is searched for the largest peak, whose location is considered to be the timing of the next glottal pulse. This is continued until the end of the voiced region, and then repeated for the voiced region before the initially identified voiced region.

4.1.2. Sliding Covariance Analysis

The glottal closure estimates provide a frame for each pitch period, since each closed phase must be entirely contained between two consecutive glottal closures. This frame enables identification of the closed phase based on changes which happen each period. The formant frequencies and bandwidths are expected to remain constant during the closed phase but will shift during the open phase. For voiced in which the glottis never completely closes, such as breathy voices, a similar formant modulation will occur. During the nominally closed phase, the glottal opening should remain approximately constant, resulting in an effect on the formants of stable magnitude. Due to the nonlinear nature of the source-filter interaction, the formants will vary even with a constant glottal area as present during the closed phase of a breathy speaker. When the glottis begins to open, the formants will move from the relatively stable values they had during closed phase.

To measure the formant frequencies and bandwidths during each pitch period, a sliding covariance based linear prediction analysis with a one sample shift is used. Each formant is a free resonance of the vocal tract system, thus the corresponding time signal can be written as a sum of complex resonances, as follows:

$$r[n] = \sum_{l=1}^{p/2} \rho_l^n (A_l e^{i\theta_l n} + \hat{A}_l e^{-i\theta_l n}).$$

where F_s is the sampling frequency, l is the index of a particular formant,$\theta_l, -\pi \leq \theta \leq \pi$, is the normalized formant frequency, $\rho_l, 0 \leq \rho < 1$, determines formant damping, and A_l is the complex formant amplitude. The above equation holds because $r[n]$ is real-valued and, therefore, the formant resonances occur in complex-conjugate

pairs. The z-transform of the time signal assuming a half-infinite sequence starting at$n = 0$, is given by:

$$R(z) = \sum_{k=1}^{p} \frac{A_k}{1 - \rho_k e^{i\theta_k} z^{-k}} = \frac{b_0 + b_1 z^{-1} + \cdots + b_{p-1} z^{-(p-1)}}{a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}}.$$

Note that due to the arbitrary formant amplitudes A_k, $R(z)$ is not necessarily the z-transform of an all-pole transfer function. However, $R(z)$ can be regarded as the z-transform of the impulse response of an infinite impulse response filter. The formant frequencies F_k and bandwidths B_k can be derived from the roots $\rho_k e^{i\theta_k}$ of the prediction polynomial

$$A(z) = a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}.$$

The formant frequencies F_k and bandwidths B_k in Hz are given [16] by:

$$F_k = F_s \theta_k / 2\pi$$

$$B_k = -F_s \ln(\rho_k)/\pi.$$

The size of the rectangular analysis window is constrained to be slightly larger than the prediction order, while still being several times smaller than the pitch period. In particular, the length of the analysis window is chosen for each frame to be

$$N_w = N_p/4,$$

with upper and lower bounds of

$$p + 3 \leq N_w \leq 2p,$$

where N_w is the size of the sliding covariance analysis window, N_p is the length of the pitch period as calculated by the time between the glottal pulses identified above, and p is the order of the linear prediction analysis, 14 for this study. Window lengths less than $p + 3$ cause occasional failure of the Cholesky decomposition, while using more than $2p$ points will not make the estimate significantly more accurate but will decrease the time resolution. The first analysis window begins immediately after the previous glottal pulse, while the last analysis window ends the sample before the next glottal pulse. There are thus a total of $N - N_w$ windows for each pitch period. This sliding covariance analysis gives one vocal tract estimate per sample in the pitch period. Formant tracking is performed in each pitch period on the formants calculated from the vocal tract estimates. This provides estimates of each formant during both the closed and open phases, enabling identification of the time of glottal opening based on formant modulation.

While a mathematical framework for calculating the expected modulation of the formant frequencies and bandwidths was developed in [10], we have found a large variety in the frequency and bandwidth changes that occur in the open phase. Also,

due to different fixed glottal openings from speaker to speaker, the amount of formant modulation that occurs during the closed phase will vary from speaker to speaker. This varying amount of formant modulation during the closed phase makes it difficult to set a threshold for an amount of formant modulation that indicates glottal opening. Because of these two problems, we have chosen to take a statistical approach to identifying the glottal opening. The approach taken is also a more practical approach, in that we want to estimate the vocal tract when the formant values are constant. The basic idea is to find a region during which the formant values vary minimally, while outside this region the formant values change considerably.

A small region of sequential formant samples is determined in which the formant modulation is minimal as defined by the sum of the absolute difference between successive formant estimates:

$$minD = \sum_{i=n_0}^{n_0+4} |F(i) - F(i-1)| : 1 \leq n_0 < N - N_w - 5,$$

where D is the sum of absolute differences to be minimized, n_0 is the first sample if this small region, which is varied to minimize D, F are the formant values calculated for each sample in the pitch period, and N is the number of samples in the pitch period. The size of the initial stable region is five formant samples, which ensures meaningful statistics are available to extend the region.

Once an initial stable region is identified, the mean and standard deviation of the formants within this small region are calculated, and the region is grown based on the following criteria: if the next sample is less than two standard deviations from the mean, it is included in the stable region and the mean and standard deviation are recalculated before continuing on to test the next point. A slightly different algorithm is used to extend the window to the left. The final mean and standard deviation from extending the stable region to the right are kept constant, and the region is grown to the left until a sample is more than two of these standard deviations from the mean. The closed phase is considered to include every speech sample which was used to calculate the stable formant values. Since each formant value is calculated from N_w speech samples, the total length of the closed phase will be $n_2 - n_1 + N_w$ samples, where n_1 is the time of the first formant in the stable region and n_2 is the time of the last formant in the stable region.

There are two primary reasons for the different techniques used to identify the glottal opening and closure. First, after the region has been extended to the right to identify the glottal opening, the statistics have been estimated from sufficient data and extending the window to the left will not improve those estimates. More importantly, we have found that the glottal opening tends to result in sudden formant shifts, while gradual formant shifts are found when extending the region to the left towards glottal closure. This may be because the sub- and supra-glottal pressures are approximately

equal during the return phase, which combined with the minimal flow results in little influence on the vocal tract estimate. If we attempted to update the statistics during a gradual change in the formant estimate, the statistics would likely incorporate this change, and glottal closure would not be identified.

Identifying a small initial stable region allows the algorithm to adapt to the variability of the formants for each frame. If there is more aspiration or ripple during the closed phase, the initial standard deviation calculated from this window will reflect the greater variability that will occur in the formant estimates due to the nonlinear source-filter interaction. When the glottis begins opening from its maximally closed position the interaction will increase, and the standard deviation limits will be exceeded, indicating the glottis has begun to open.

In the above discussion, the specific parameter used for the formant estimates was not stated. According to the theory presented in [10], all of the formants will undergo modulation of both their frequencies and bandwidths. The first formant shows these modulations clearer than other formants, in part because the energy of the first formant is greater and estimates of it tend to be less effected by noise. In general, both the formant frequencies of it tend to increase during the open phase, while they remain relatively constant during the closed phase.

Experiments have shown that the best measure to use in determining formant modulation is the frequency of the first formant. The first formant is more stable than higher formants during the closed phase and exhibits a more observable change at the start of the open phase. Also, the sliding covariance and formant tracker tend to make more errors for higher formants; the figure below illustrates the above discussion, for a phoneme /a/ taken from speech out of the CMU Database:

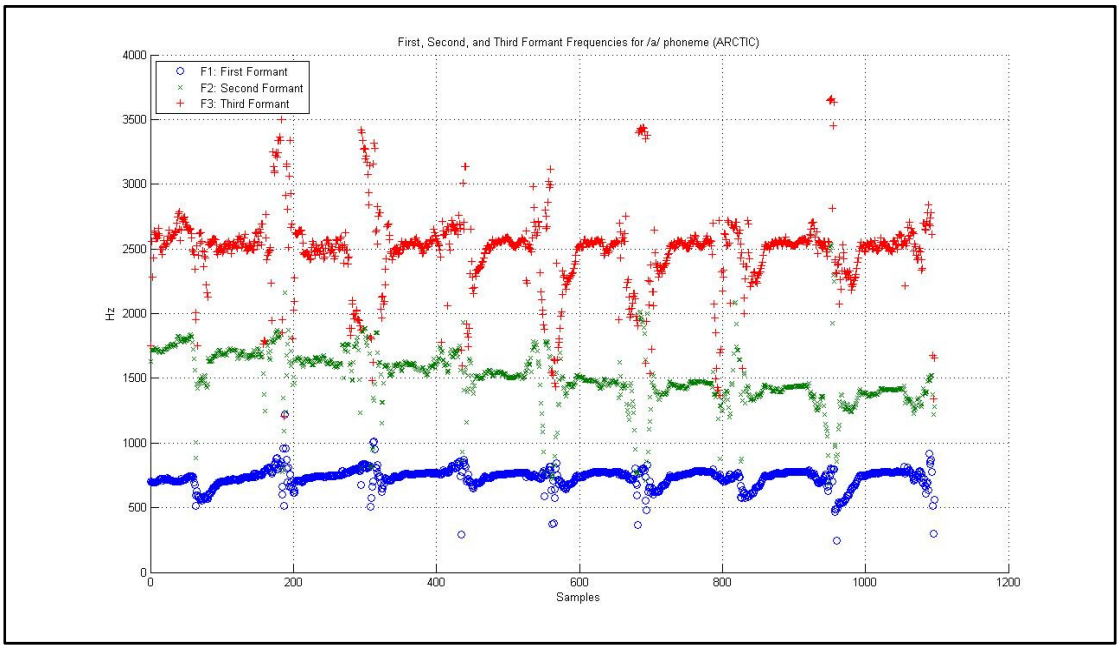

Figure 4: Formant tracking of the first three formants

4.1.3. Examples

Here, we show some examples from voiced speech, where formant tracks, closed phase formant samples and closed phase speech samples are illustrated.

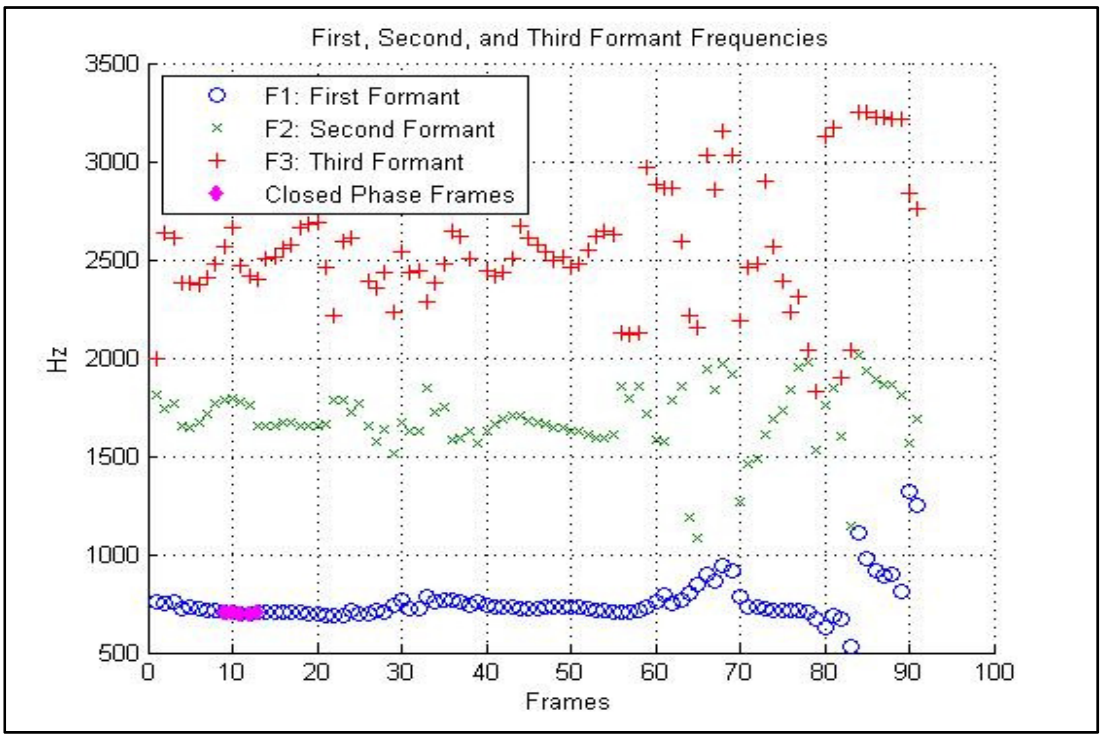

Figure 5: Formant tracking and formant stable region

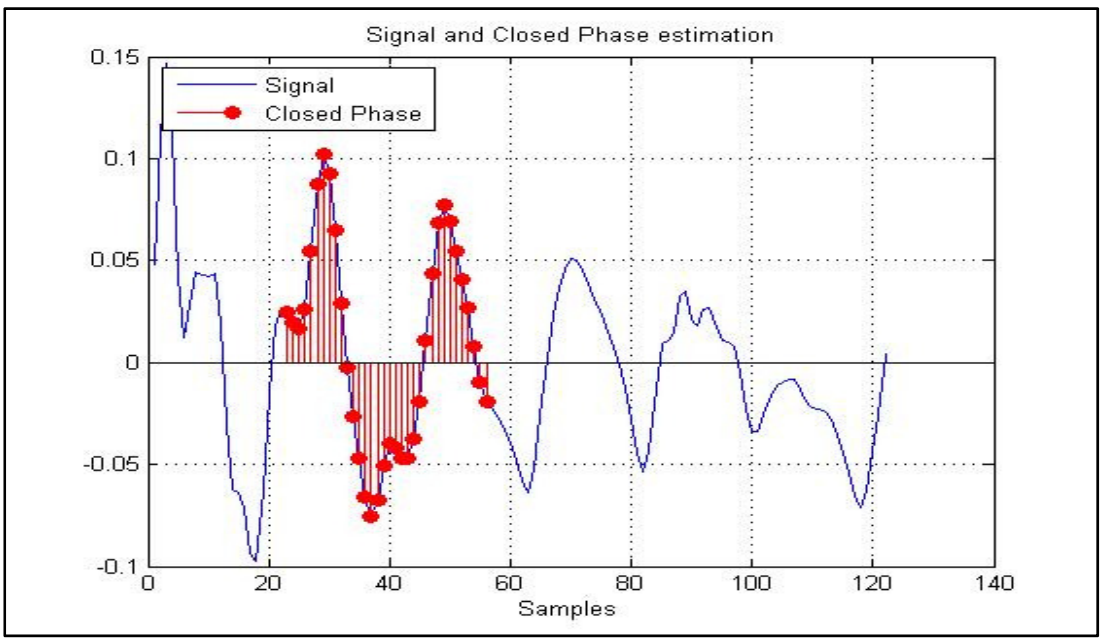

Figure 6: Closed Phase speech samples

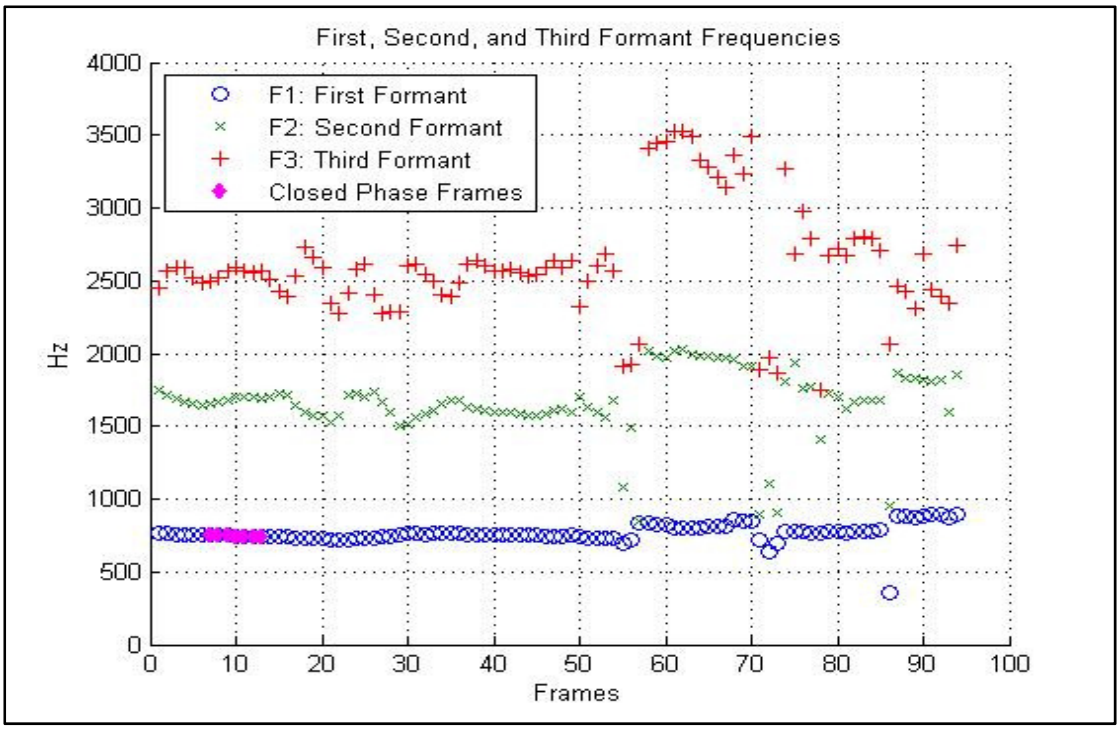

Figure 7: Formant tracks and formant stable region

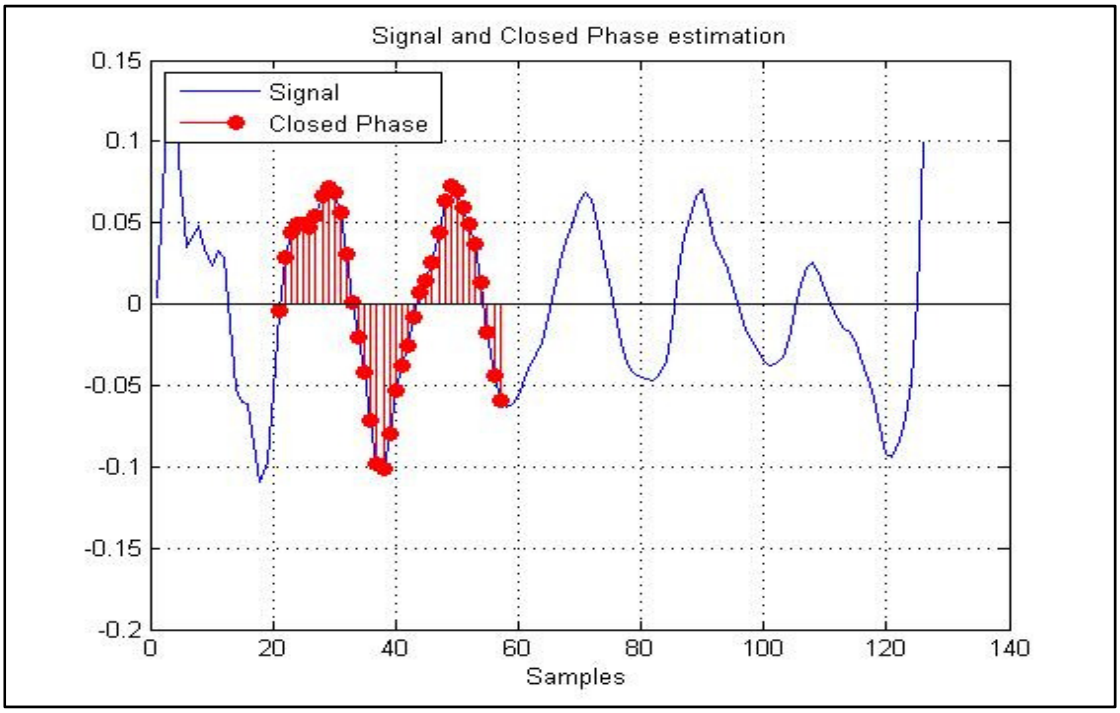

Figure 8: Closed phase speech samples

4.2. From Closed Phase to Glottal Flow Derivative

Once the closed phase is determined, the vocal tract response s calculated, and then used to inverse filter the speech signal to generate the glottal flow derivative waveform.

4.2.1. Vocal Tract Response

The vocal tract response is calculated from a rectangularly windowed region of the speech signal bounded on the left by the glottal closure and on the right by the glottal opening, as determined in the preceding section. The vocal tract is estimated using a covariance based linear predictor, with an adaptive pre-emphasis. To determine the pre-emphasis coefficient, a first-order autocorrelation linear prediction is performed on the analysis window, including the preceding samples required to initialize the covariance analysis. This filter is then used to pre-emphasize the data. It is found this adaptive pre-emphasis to work better than a fixed pre-emphasis filter.

4.2.2. Inverse Filtering

There is some uncertainty as to what region to inverse filter with a particular vocal tract response. This problem arises due to the fact that the vocal tract is estimated during the closed phase but must be used to inverse filter both the open and the closed phase. This can create a problem, since the difference equation implementing the inverse vocal tract filter is changed at the start of the analysis window, where there is significant energy in the speech signal, and thus significant energy in the inverse filter. This sudden change of filter artificially excites the formants, and sometimes results in a large output shift.

The decay of a linear filter with zero input contains components at pole locations. For speech, we have

$$s[n] = e[n] + \sum_{i=1}^{p} a_i s[n-i].$$

Considering $e[n]$ to be zero (superposition allows us to add in the response to $e[n]$ later), we have

$$s[n] - \sum_{i=1}^{p} a_i s[n-i] = 0$$

Difference equations are easily solved through the z-transform, giving

$$S(z) - \sum_{i=1}^{p} a_i \left(S(z)z^{-i} + \sum_{k=1}^{i} s[-k]z^{1-k}\right) = 0,$$

where the inner sum is due to the initial conditions. Rearranging in the form required for partial fraction expansion, we have

$$\frac{S(z)}{z} = \frac{\sum_{i=1}^{p} a_i s[-i]z^{p-i}}{\sum_{i=0}^{p} a_i z^{p-i}} = \frac{\sum_{i=1}^{p} a_i s[-i]z^{p-i}}{\prod_{i=1}^{p}(z - z_i)},$$

where $a_0 = 1$, and z_i are the complex pole locations. The partial fraction expansion of the above equation will generally be of the form

$$S(z) = \sum_{i=1}^{p} \frac{C_i z}{z - z_i},$$

where the C_i's are due to the initial conditions. A slightly different form of the above equation will result under the unusual condition of repeated poles. The inverse Fourier transform of the above equation is of the form

$$s[n] = \sum_{i=1}^{p} C_i z_i^n u[n],$$

where $u[n]$ is the unit step function. Under the normal condition of complex pole locations z_i, poles will appear in complex conjugate pairs, with their responses combining to form a decaying sine wave. The above equation shows that the only possible output is a combination of decaying sine waves at the pole frequencies. Since the only possible outputs are at the pole frequencies, if the filter is suddenly changed, the energy in the filter must be redistributed to the new frequencies. Experiments have confirmed that this redistribution can cause excitation of some of the formants.

4.3. Examples

Here, we show some examples of glottal flow derivatives taken from an /e/ phoneme of an utterance of the ARCTIC CMU Database.

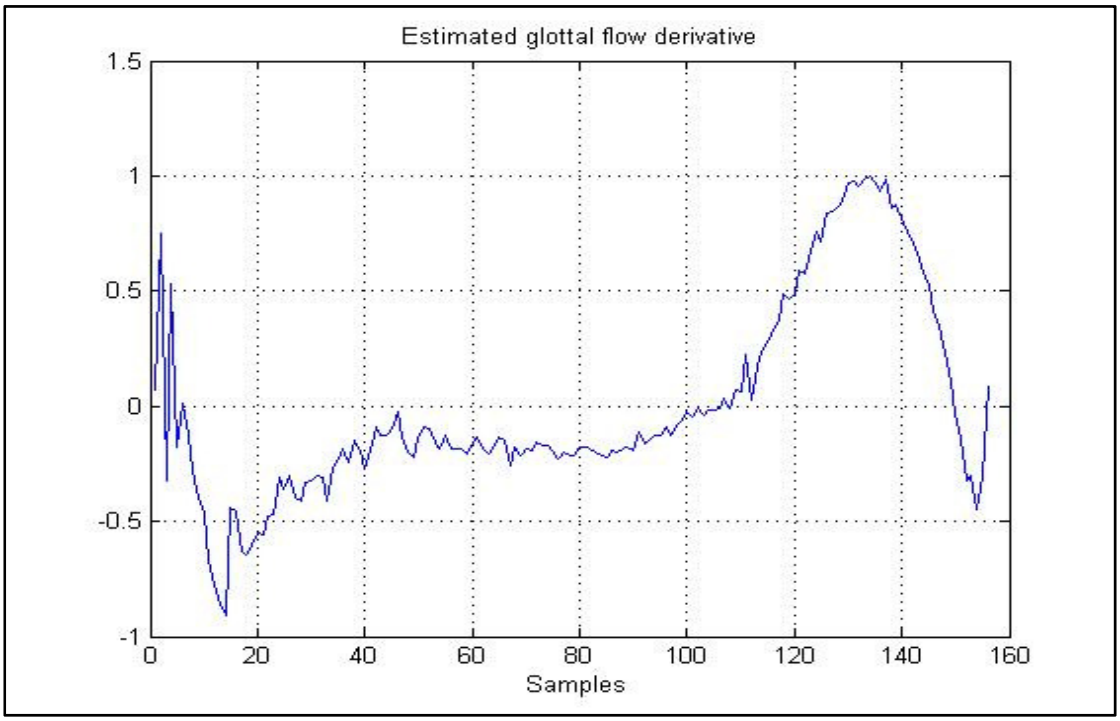

Figure 9: Glottal Flow Derivative Estimate

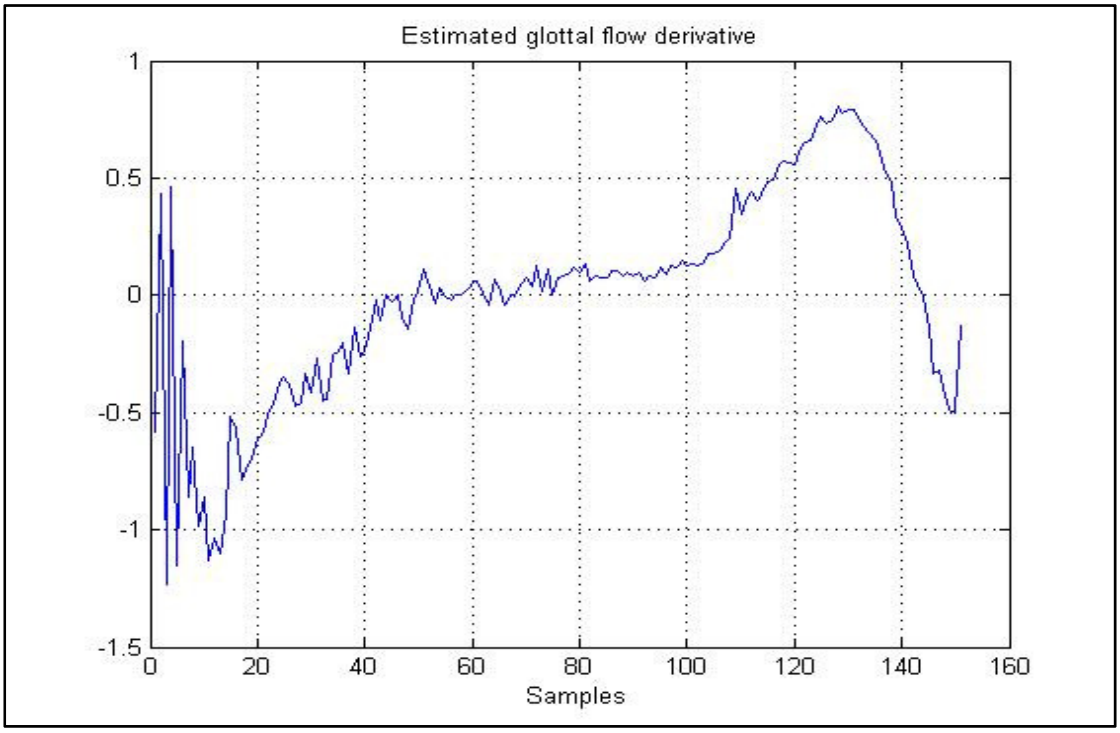

Figure 10: Glottal Flow Derivative Estimate

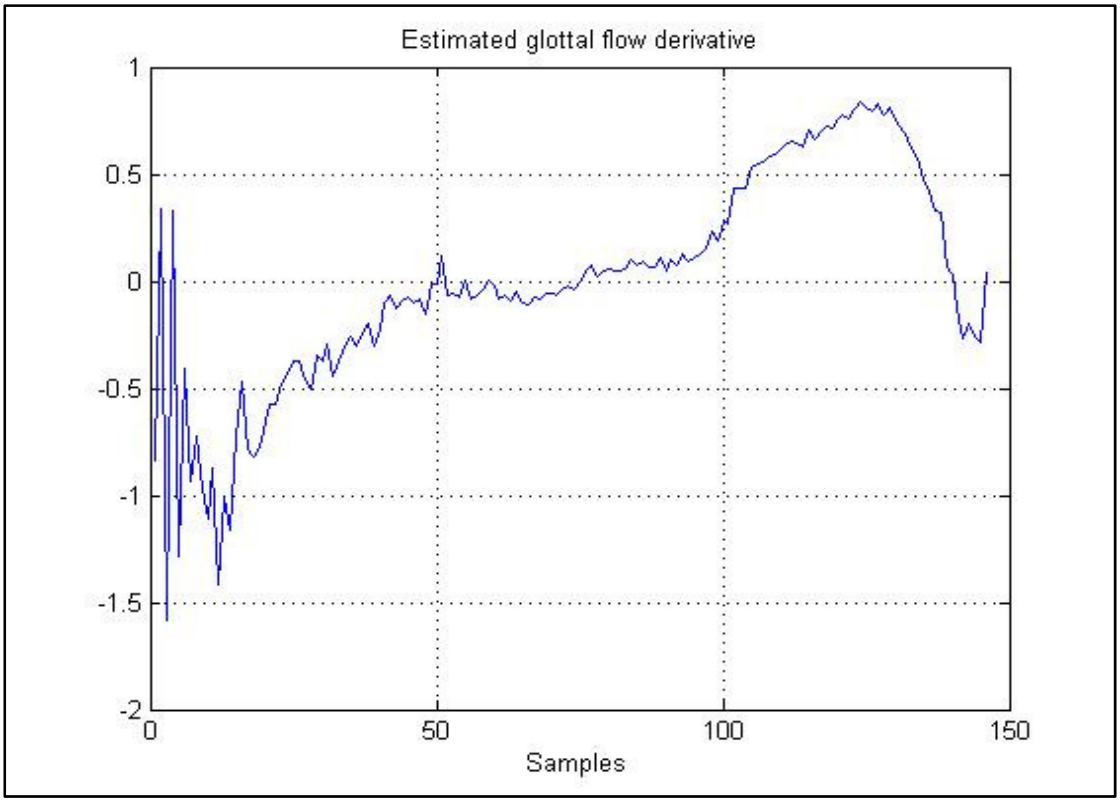

Figure 11: Glottal Flow Derivative Estimate

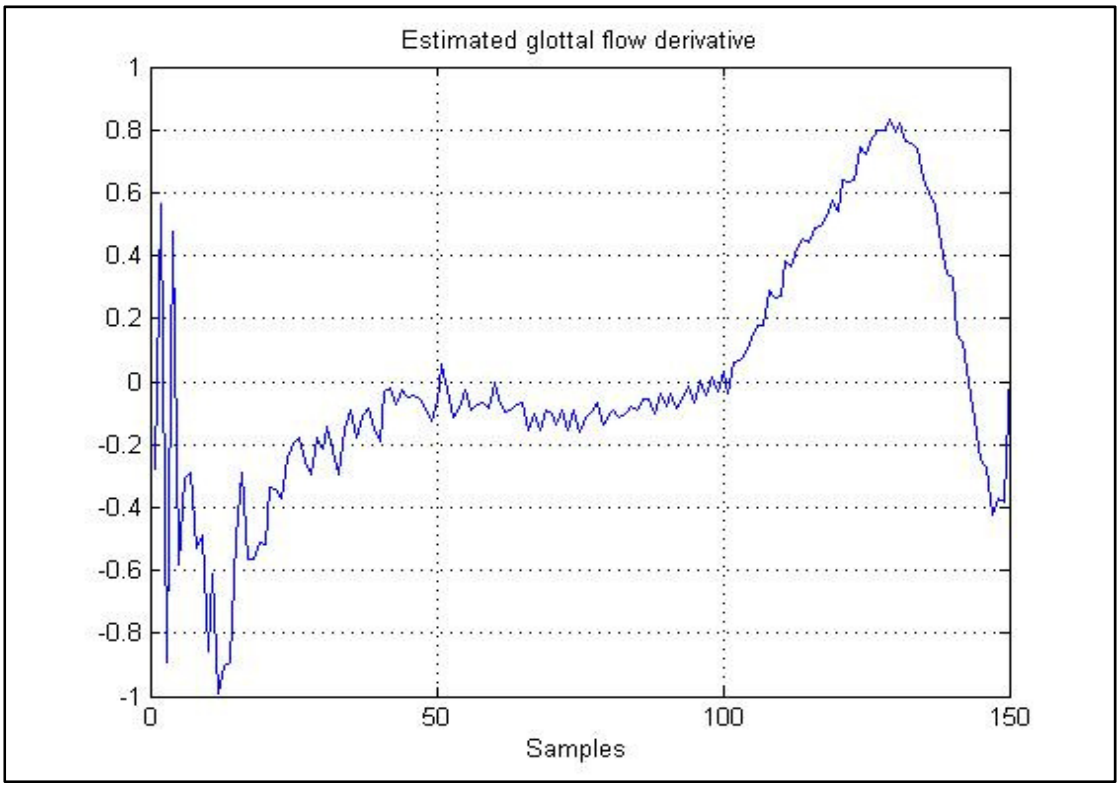

Figure 12: Glottal Flow Derivative Estimate

5. Estimating Coarse Structure of the Glottal Flow Derivative

Chapter 4 developed the techniques used to calculate the glottal flow derivative waveform from the speech signal. Now that we have the source waveform, we can estimate the parameters of a model describing the general shape of the waveform.

5.1. Formulation of the Estimation Problem

The coarse structure of the glottal flow derivative is captured using the LF model, described by the equation

$$E(t) = \frac{dU_g}{dt} = E_0 e^{at} sin\omega_g t,$$

for the period from glottal opening (T_o) to the pitch pulse (T_e), at which time the return phase starts:

$$E(t) = -\frac{E_0}{\beta T_a}\left[e^{-\beta(T-T_e)} - e^{-\beta(T_c-T_e)}\right],$$

which continues until time T_c. The figure below shows an example of the LF model:

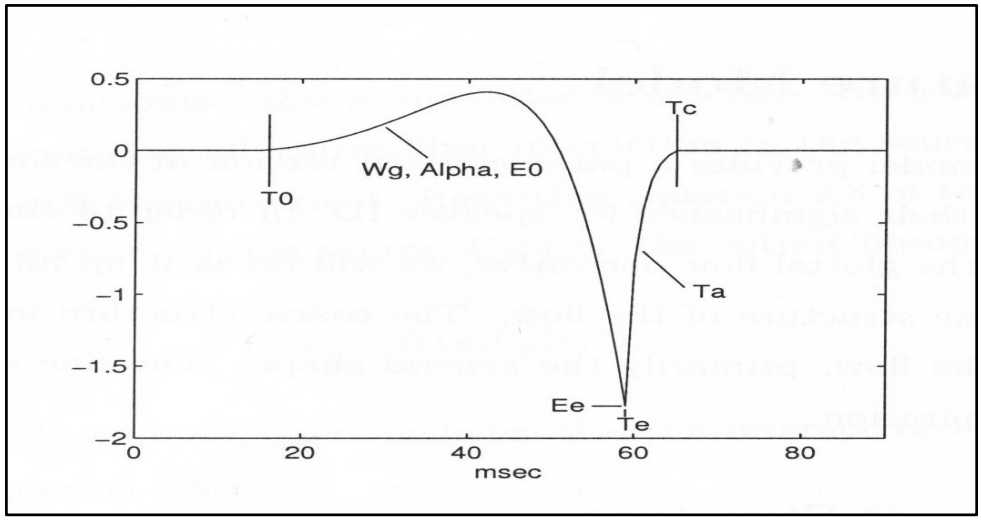

Figure 13: LF model for the glottal derivative waveform

Due to the large dependence of E_0 on a, the parameter E_e, the value of the waveform at time T_e, is estimated instead of E_0. To calculate E_0 from E_e, the equation

$$E_0 = \frac{E_e}{e^{aT_e} sin\omega_g T_e}$$

is used.

A least squares minimization problem can be set up to fit the LF model to the glottal flow derivative waveform:

$$E(\vec{x}) = \sum_{n=0}^{T_0} G^2[i] + \sum_{n=T_0+1}^{T_e} (G[n] - E_0 e^{an} sin\omega_g n)^2 + \sum_{n=T_e+1}^{T_c} (G[n] + \frac{E_0}{\beta T_a}\left[e^{-\beta(n-T_e)} - e^{-\beta(T_c-T_e)}\right])^2 + \sum_{n=T_c+1}^{N} G^2[n],$$

where the point $n = 0$ occurs after the end of the previous return phase, $n = N$ occurs before the next open phase, $\vec{x}$ is a vector of the four parameters of the LF model, and $G[n]$ is the glottal flow derivative waveform at sample n. The error E is a nonlinear function of the four model parameters, so the problem must be solved iteratively using a nonlinear least-squares algorithm. A nonlinear least-squares algorithm attempts to solve problems of the form:

$$\min_{\vec{x}} \mathrm{E}(\vec{\mathrm{x}}) = \frac{1}{2}\sum (f_i(\vec{x}, \vec{x}_0) - \vec{y}_i)^2 = \frac{1}{2}\sum r_i^2(\vec{x}) = \frac{1}{2}\vec{R}(\vec{x})^T \vec{R}(\vec{x}),$$

where $\vec{x}$ is the vector of parameters to be solved for, $\vec{y}_i$ is the data to be fitted, f_i is the value of the curve at point i using the parameters $\vec{x}$, $\vec{R}(\vec{x})$ is the residue vector, with $\vec{R}(\vec{x}) = [r_1(\vec{x}), r_2(\vec{x}), \dots, r_N(\vec{x})]$ and $\vec{x}_0$ is an initial estimate of the parameter vector.

In [10], the NL2SOL Algorithm was used. Here, due to the MATLAB environment of implementation, we used an algorithm which solves non-linear least squares problem, with similar properties of those of NL2SOL, such as the addition of bounds to enable parameters to be limited to physically reasonable values. This algorithm, called *lsqcurvefit* is a large-scale optimization algorithm, which is a subspace trust region method and is based on the interior-reflective Newton method. Each iteration

involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients (PCG).

The aforementioned algorithm makes use of the Jacobian matrix of the model function. The $(i,j)^{th}$ element of the Jacobian matrix $J(\vec{x})$ of a vector $\vec{G}(\vec{x})$ is given by

$$j_{i,l}(\vec{x}) = \frac{\partial g_i(\vec{x})}{\partial \vec{x}_l}.$$

In other words, the $(i,l)^{th}$ element of J is the partial derivative of the vector $\vec{G}(\vec{x})$ at the point i with respect to the l^{th} element of the parameter vector $\vec{x}$.

The partial derivatives of the LF model, as described in chapter 3, are given:

- $\frac{\partial f_1}{\partial a} = -\frac{E_e}{\sin\left(\frac{\pi T_e}{T_p}\right)} \mathrm{e}^{(a(t-T_e))}(t-T_e)\sin\left(\frac{\pi t}{T_p}\right)$
- $\frac{\partial f_2}{\partial a} = 0$
- $\frac{\partial f_1}{\partial T_p} = \left(t\cos\left(\frac{\pi t}{T_p}\right)\sin\left(\frac{\pi T_e}{T_p}\right) - T_e\sin\left(\frac{\pi t}{T_p}\right)\cos\left(\frac{\pi T_e}{T_p}\right)\right)\frac{E_e \pi e^{a(t-T_e)}}{\left(T_p \sin\left(\frac{\pi T_e}{T_p}\right)\right)^2}$
- $\frac{\partial f_2}{\partial T_p} = 0$
- $\frac{\partial f_1}{\partial E_e} = -\frac{e^{a(t-T_e)}\sin\left(\frac{\pi t}{T_p}\right)}{\sin\left(\frac{\pi T_e}{T_p}\right)}$
- $\frac{\partial f_2}{\partial E_e} = \left(e^{-\beta(T_0-T_e)} - e^{-\beta(t-T_e)}\right)/\left(1-e^{-\beta(t-2T_e+T_0)}\right)$
- $\frac{\partial f_1}{\partial \beta} = 0$
- $\frac{\partial f_2}{\partial \beta} = \frac{E_e\left((t-T_e)e^{-\beta(t-T_e)} - te^{-\beta(t-2T_e+T_0)} - (T_0-T_e)e^{-\beta(T_0-T_e)}\right)}{\left(1-e^{-\beta(T_0-T_e)}\right)^2}$

and the Jacobian matrix is given by:

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial a} & \frac{\partial f_2}{\partial a} \\ \frac{\partial f_1}{\partial T_p} & \frac{\partial f_2}{\partial T_p} \\ \frac{\partial f_1}{\partial E_e} & \frac{\partial f_2}{\partial E_e} \\ \frac{\partial f_1}{\partial \beta} & \frac{\partial f_2}{\partial \beta} \end{bmatrix}^T.$$

5.2. Examples

Here, we illustrate some examples of glottal flow derivatives and their corresponding fitted LF models.

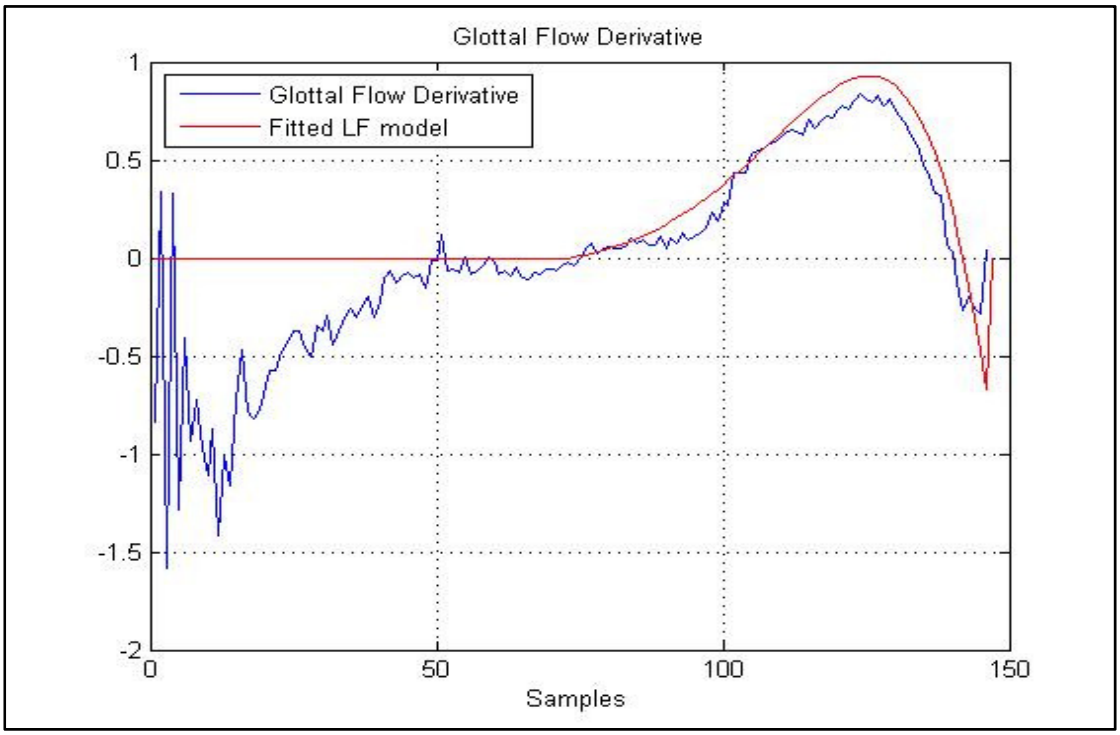

Figure 14: Glottal Flow Derivative Estimate and respective LF model

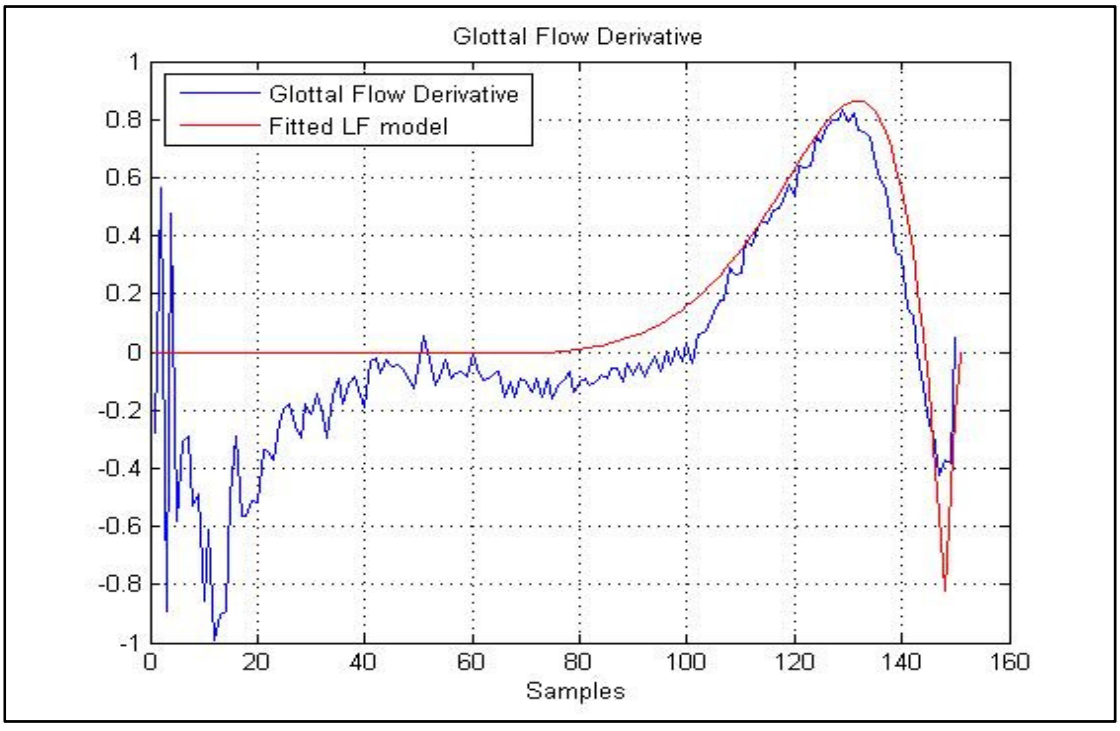

Figure 15: Glottal Flow Derivative Estimate and respective LF model

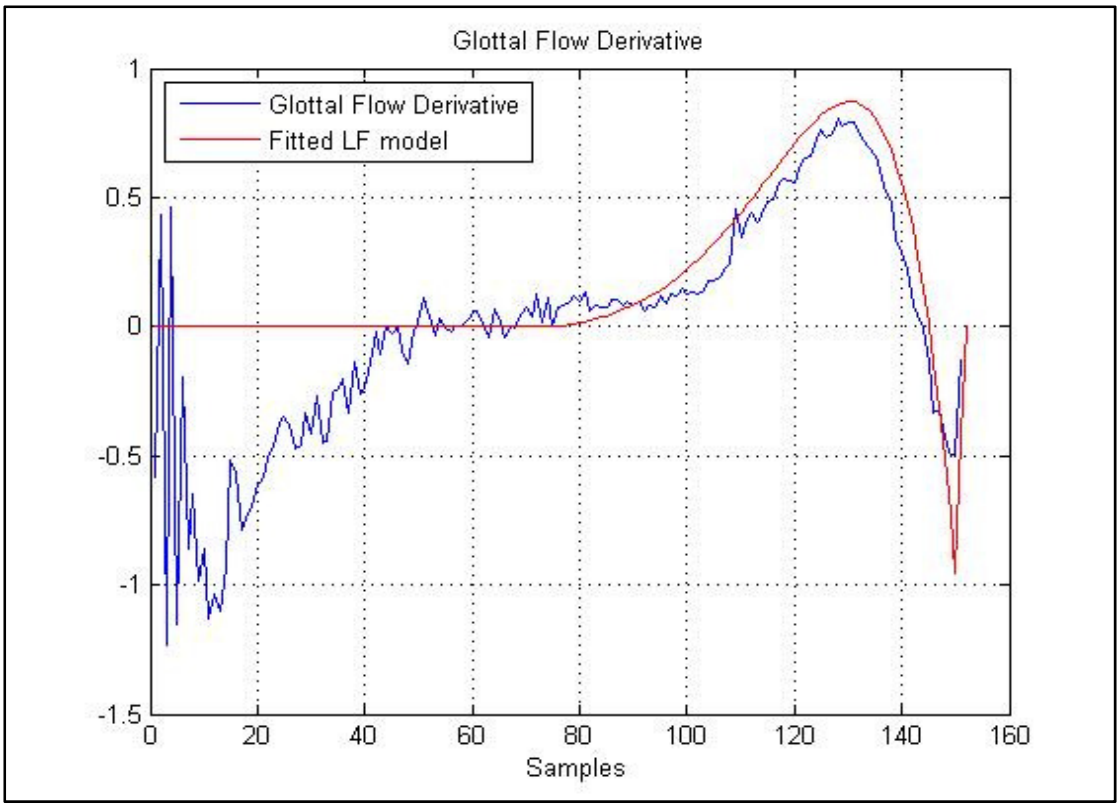

Figure 16: Glottal Flow Derivative Estimate and respective LF model

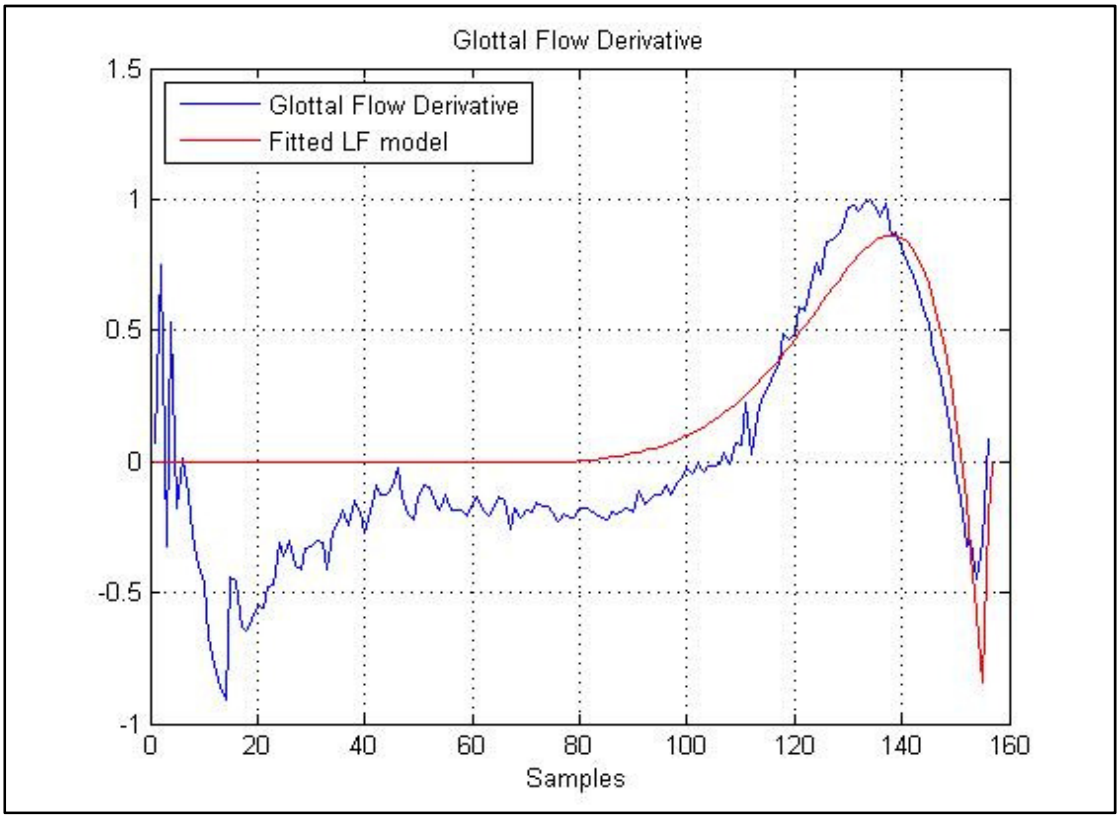

Figure 17: Glottal Flow Derivative Estimate and respective LF model

6. Spectral Representation of the Glottal Flow Derivative Waveform

The previous chapters discussed, among others, the time-domain representation of the glottal flow. This chapter deals with the spectral representation of the glottal flow. It is observed that parameter estimation seems easier in the spectral domain and the glottal flow characteristics of natural speech signals can be estimated by processing directly the spectrum, without needing time-domain parameter estimation.

Accurate processing of the vocal flow characteristics is needed for dealing with voice quality in high quality speech synthesis. In the context of synthesis, a frequency-domain approach appears desirable, because voice quality is better described by spectral parameters. The main spectral parameters found for synthesizing voices with different qualities are: 1/ spectral tilt; 2/ amplitude of the first few harmonics; 3/ increase of the first formant bandwidth; 4/ noise in the voice source. We will consider only the first two of these in this chapter.

In most of the studies, the spectrum is obtained by Fourier transform of the glottal waveform. Therefore, little insight is brought on the role played by each individual component of the waveform in the spectral domain, no analytic formulas are provided for the spectrum, and no spectral model of the glottal glow are proposed.

In this text, using the results from [4], we show the spectral correlates of the LF-model. The analytic formula of the spectrum of the LF-model is presented. Then, formulas are given for computing spectral tilt and amplitudes of the first of the first harmonics as functions of the LF-model parameters.

6.1. R_k, R_g, R_a parameter transformations of the LF model

The LF-model is considered here as a five parameter model of the glottal flow derivative. The five parameters commonly used to describe the LF-model are:

$$T_0, E_e, R_g, R_k, R_a.$$

T_0 is the fundamental period; it will only change the harmonic frequencies. E_e is the maximum flow declination rate; it will only change the overall harmonic amplitudes. R_g is the ratio of T_0 over twice the peak flow time T_e. It behaves much like the open quotient O_q. The spectral effect of an increased R_g is to expand the frequency scale, resulting in shifting energy from low frequency harmonics to medium frequency harmonics. R_k is the inverse of the speed quotient: $R_k = (T_e - T_p)/T_p$; it will change the waveform skewness, and will essentially affect the first harmonics amplitude. R_a measures the duration of the return phase: $R_a = T_a/T_0$; it will change the spectral tilt adding a -6db/oct above a frequency which depends on R_a, R_g, and R_k, and then will essentially affect high order harmonics amplitudes. The open quotient is related to both $R_g$and R_k: $O_q = (1 + R_k)/2R_g$. See [7] for details on LF-model parameters. The LF-model can produce a great variety of waveforms with the different parameter settings. But a given set of parameters does not ensure to give a plausible speech waveform. In order to do so, the parameters must satisfy their theoretical ranges: $E_e > 0, T_0 > 0, R_g > 0.5, 1 > R_k > 0, R_a > 0$. But they must also verify the following equations: $R_k < 2R_g - 1$, which ensures that the closing time is inside the period, and $R_a < 1 - \frac{1+R_k}{2R_g}$, which ensures that the return phase is a decreasing exponential. Furthermore, if $R_k > 0.5$ then the negative maximum of the flow derivative is no longer E_e. Thus, to keep the meaning of E_e as the maximum flow declination rate, one must force $R_k < 0.5$.

6.2. Spectrum of the LF-model

In [4], the derivative spectrum of the LF-model is computed. The result is given below:

$$\begin{aligned}\acute{U}_g(v) = E_0 \frac{1}{(a - j2\pi v)^2 + \omega_g^2}\Big[&\omega_g \\ &+ \exp\big((a - j2\pi v)T_e\big)\Big((a - j2\pi v)\sin\big(\omega_g T_e\big) - \omega_g \cos\big(\omega_g T_e\big)\Big)\Big] \\ &+ E_e \frac{\exp\left(-j2\pi v T_e\right)}{\beta T_a j2\pi v(\beta + j2\pi v)}\big[\beta(1 - \beta T_a)\big(1 - \exp\big(-j2\pi v(T_0 - T_e)\big)\big) \\ &- \beta T_a j2\pi v\big].\end{aligned}$$

The variables E_0, ω_g, T_e, T_a, and β are functions of the model parameters and variable a is obtained solving an implicit equation. The reason for an implicit equation is the condition of zero net gain of flow during a fundamental period which implies area balance in the flow derivative: $\int_0^{T_0} u_{LF}(t)dt = 0$.

6.3. Spectral Correlates of the LF-model Parameters

With the help of the above analytic expression of the LF-model spectrum, one can obtain the following results on the spectral correlates of the LF-model:

6.3.1. Spectral Tilt

The spectral tilt is an important parameter of voice quality, especially for female voices. It is related to the spectrum behavior when the frequency tends towards $+\infty$. If the parameter R_a is set to 0, then $\left|\acute{U}_g(v)\right| \sim E_e/2\pi v$ when $v \to \infty$, which corresponds to a spectral slope of -6 dB/oct. If R_a is not equal to 0, then an extra -6 dB/oct is added to the spectrum, leading to a -12dB/oct spectral slope, above a cutoff frequency which can be computed as $f_c = F_a + \frac{a}{2\pi} + \frac{R_g}{T_0}\cot\big(\pi(1+R_k)\big)$, where $F_a = \frac{1}{2\pi R_a}$. In comparison to the predicted cutoff frequency value of F_a given by Fant [7], this analytically calculated value gives a correction term that is not negligible: for instance, with $R_g = 1.3, R_k = 0.3, R_a = 0.1$, then $F_a = 160\ Hz$ although the cutoff frequency is equal to $f_c = 290\ Hz$; in this case, taking F_a instead of f_c leads to a more than 5 dB error in the determination of the spectral tilt. Notice that the amplitude of the first harmonics is also affected by this parameter. In conclusion, the spectral tilt depends mostly on the parameterR_a. This parameter is responsible for an extra -6 dB/oct attenuation above frequencyf_c. However, f_c depends also on $R_g$and R_k according to the analytic expression $f_c = F_a + \frac{a}{2\pi} + \frac{R_g}{T_0}\cot\big(\pi(1+R_k)\big)$. Thus, f_c cannot always be approximated by F_a.

6.3.2. First Harmonics

In a similar manner, one can study the low frequency harmonic amplitudes. Of particular interest is the ratio H1-H2, where H1 and H2 are the amplitudes of the first two harmonics (in dB). We will see in the next section some examples of the variation of this ratio as a function of R_k. As can be seen, H1-H2 has a range of about 10 dB for common parameter ranges $0.3 < R_k < 0.6$ and $1.0 < R_g < 1.3$. The amplitude ratio of the two first harmonics depends mostly on the open quotient and the speed quotient (or equivalently to R_g and R_k). Changes in spectral tilt are also noticeable. This ratio increases with the open quotient and its range increases with R_k as shows the 1dB approximation: $H1 - H2 = 12\left(\frac{O_q}{0.7}\right)^2\left(1 - \left(1 - \frac{R_k}{0.7}\right)^2\right) - 6$.

6.4. Examples

Here, we illustrate some examples of the LF spectrum and its properties for different values of the parameters.

Also, the spectrum of the derivative of the LF model is illustrated below. That's because the general slope for $F\{u'_{LF}(t)\}(f)$ is flat (0 dB/oct) when there is no return phase, and is decreasing at -6 dB/oct after the cutoff frequency f_c that is controlled by the return phase parameter, T_a, when there is a return phase. The difference between those 2 cases (with and without a return phase) is better seen on the $u''_{LF}(t)'s$ spectrum than on the $u'_{LF}(t)'s$, as can be seen on figures below.

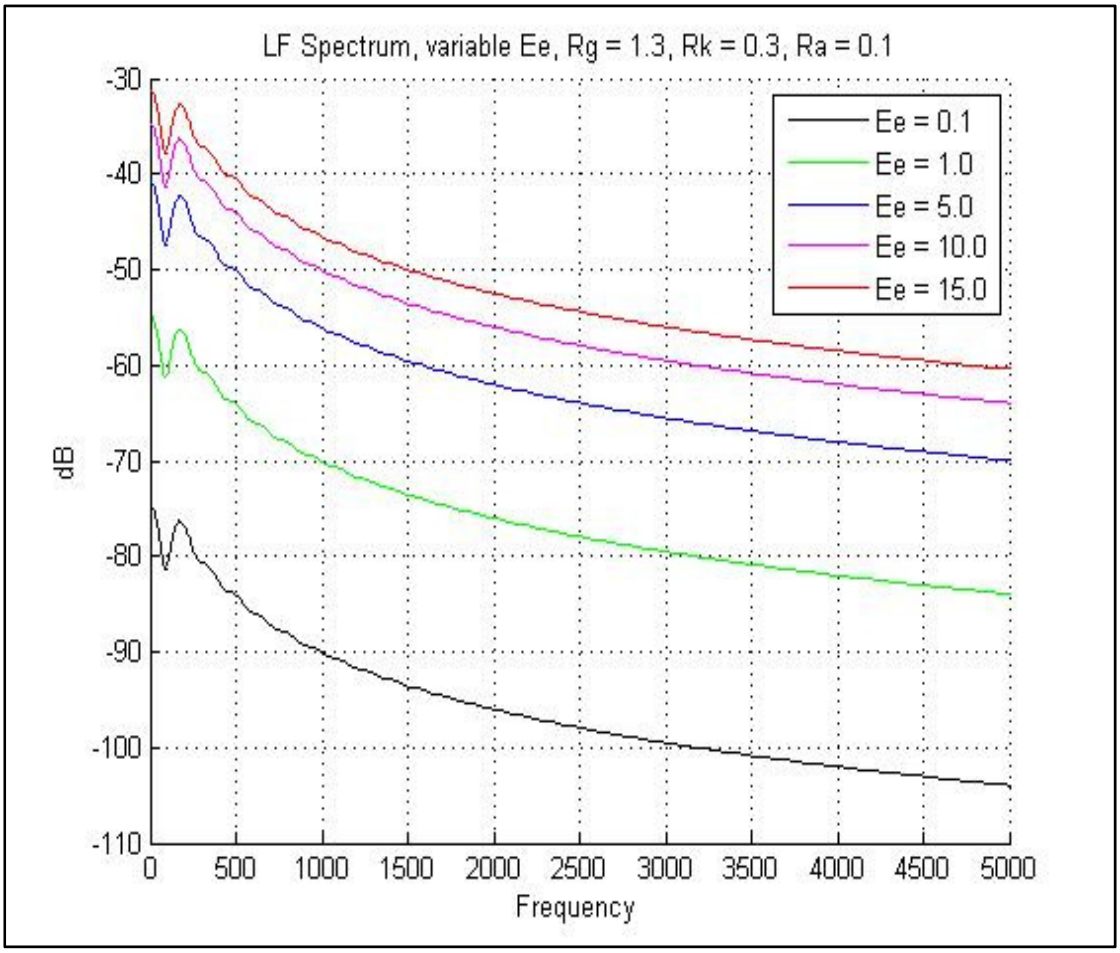

Figure 18: LF Spectrum, variable Ee

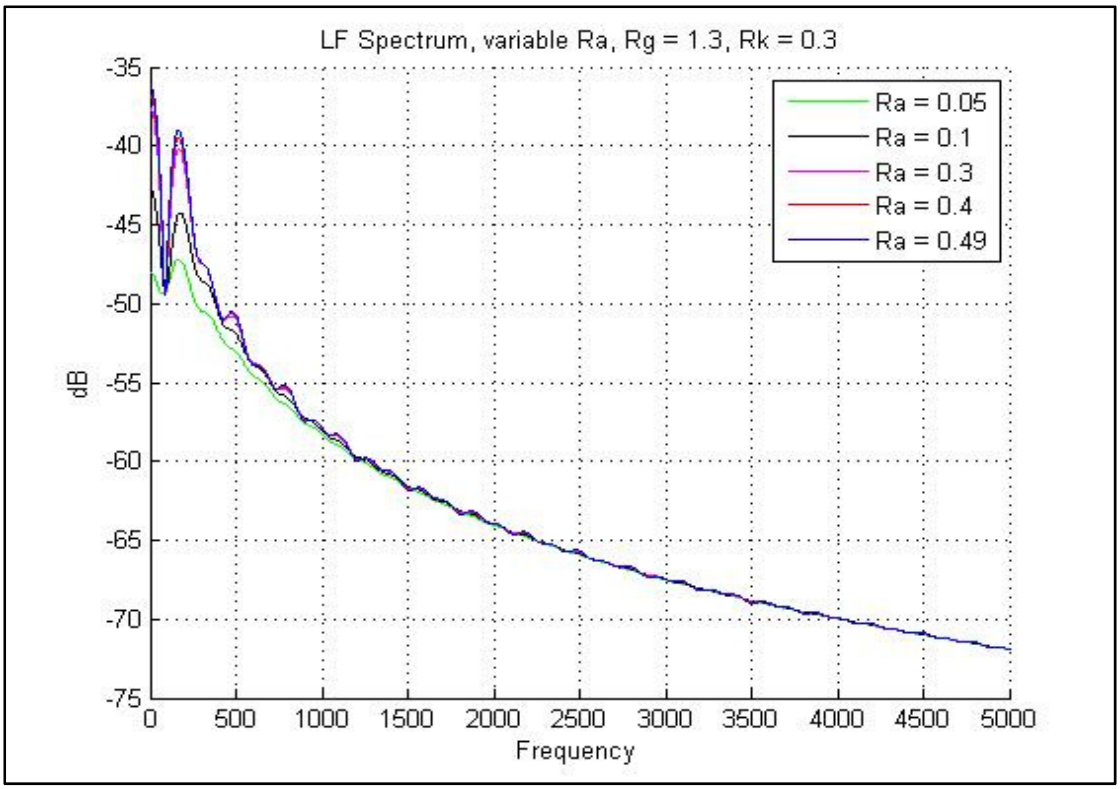

Figure 19: LF Spectrum, variable Ra

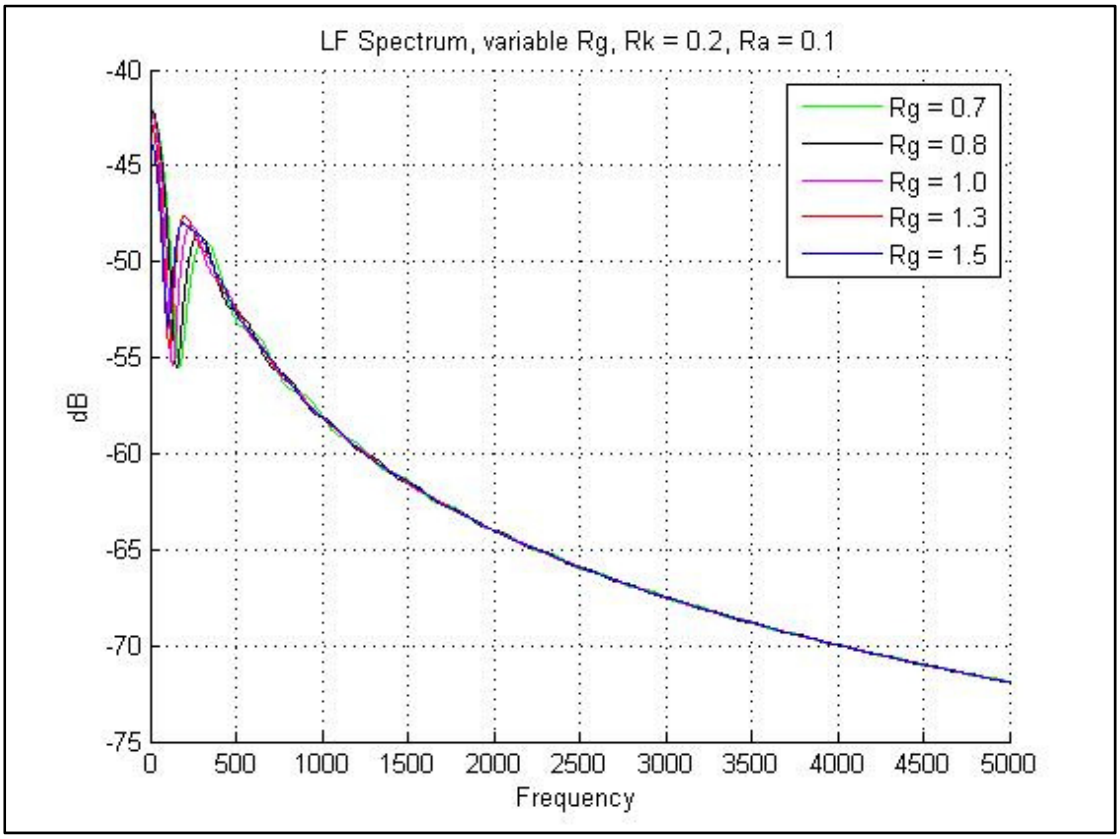

Figure 20: LF Spectrum, variable Rg

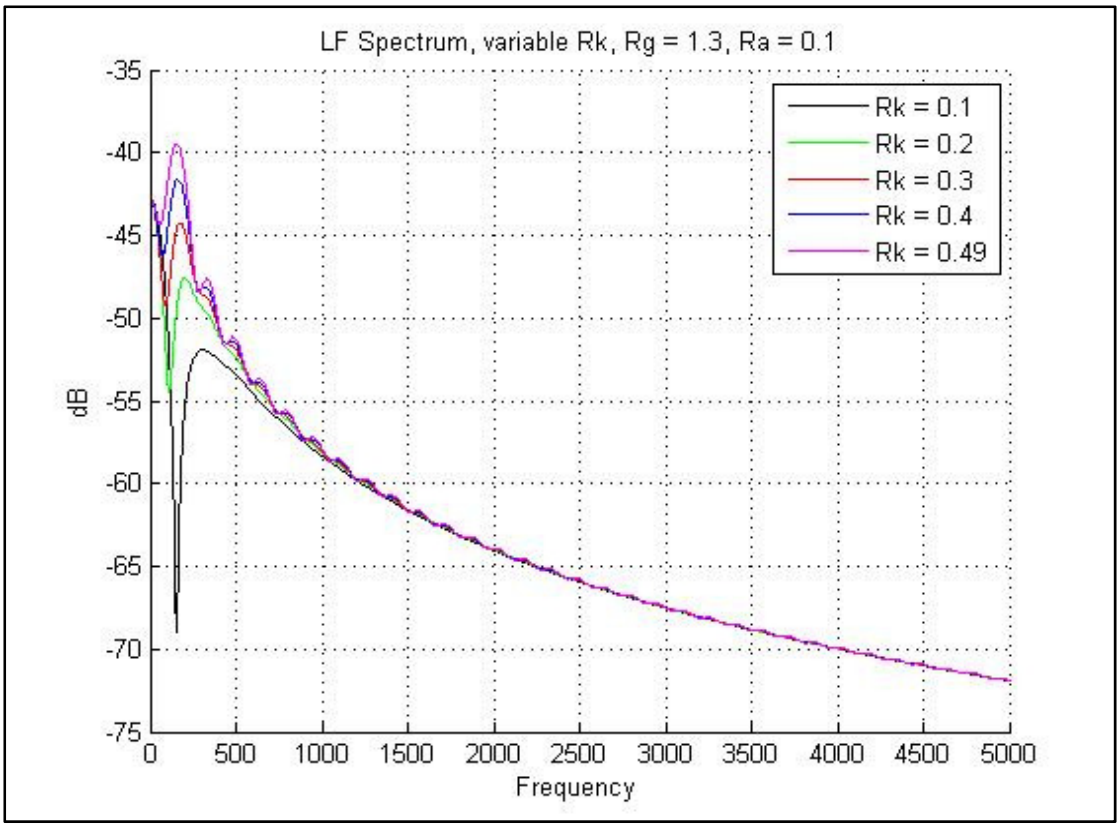

Figure 21: LF Spectrum, variable Rk

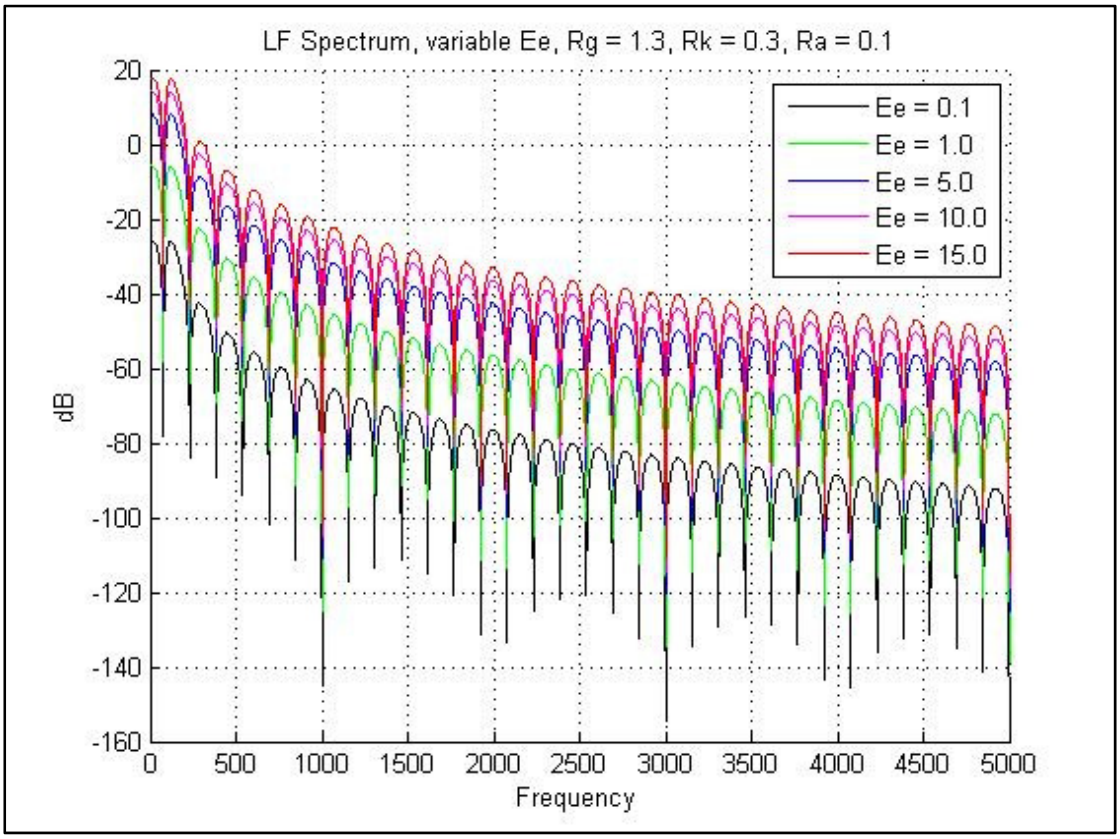

Figure 22: LF derivative Spectrum with variable Ee

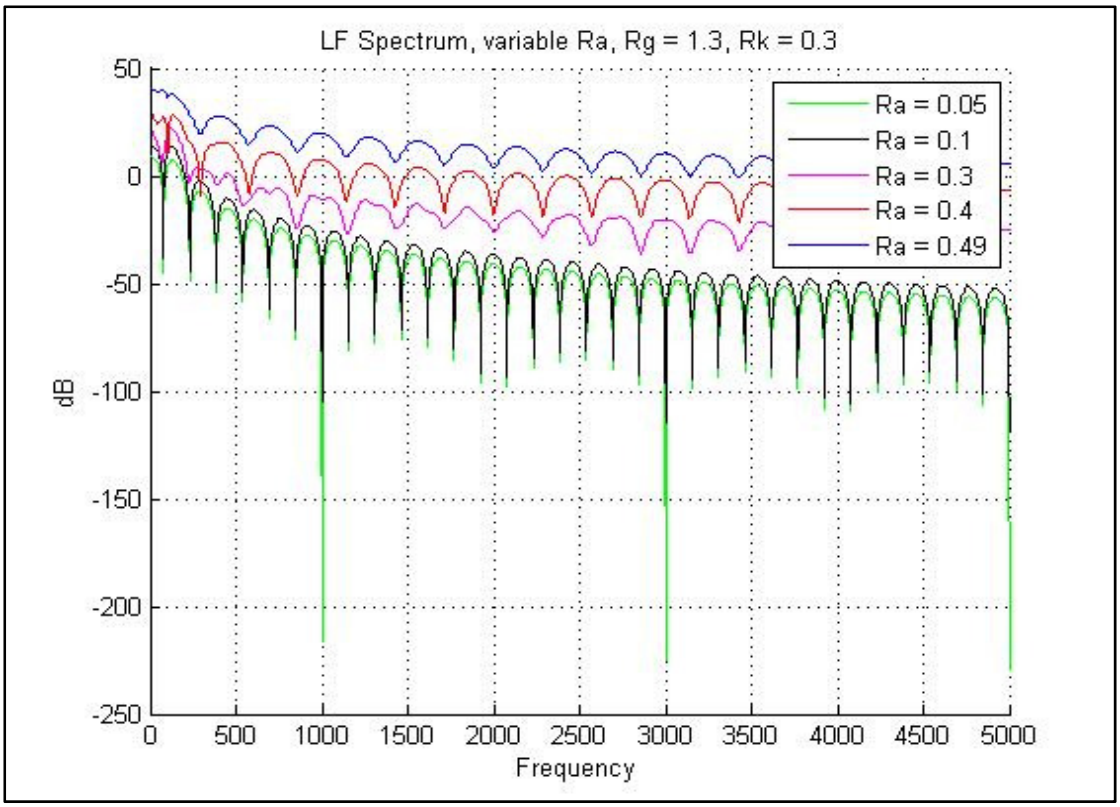

Figure 23: LF derivative Spectrum with variable Ra

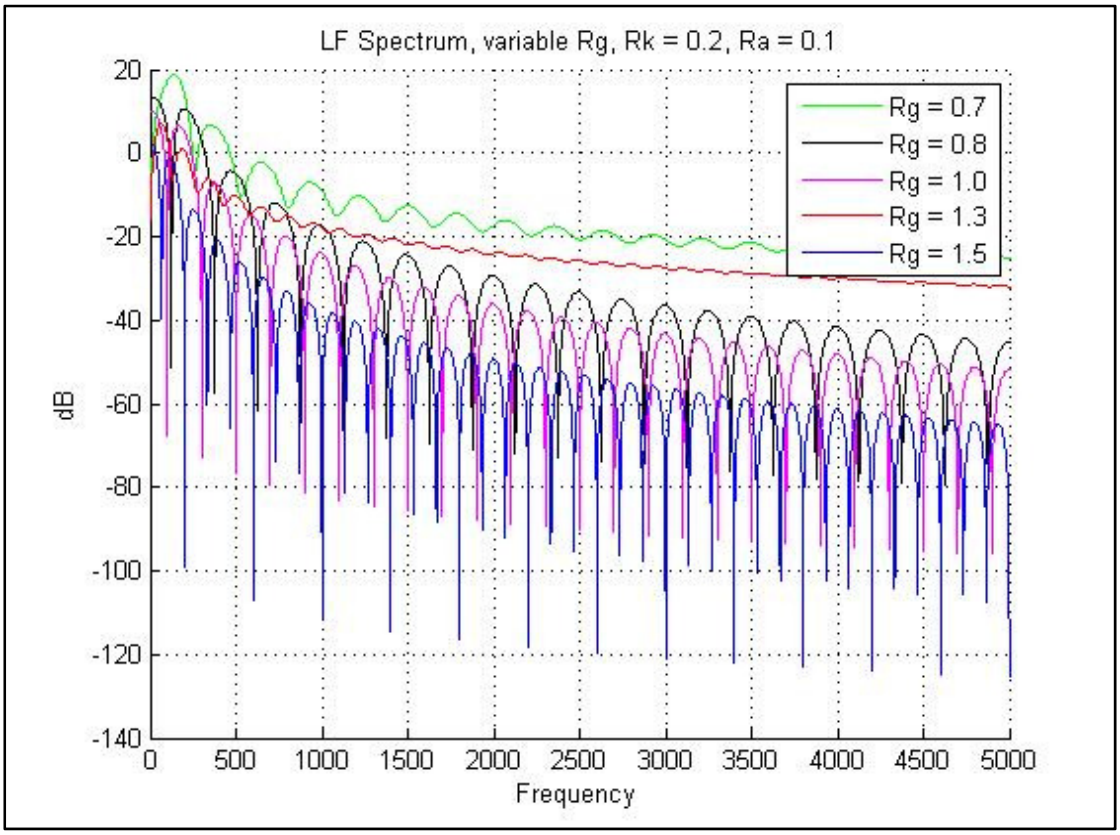

Figure 24: LF derivative Spectrum with variable Rg

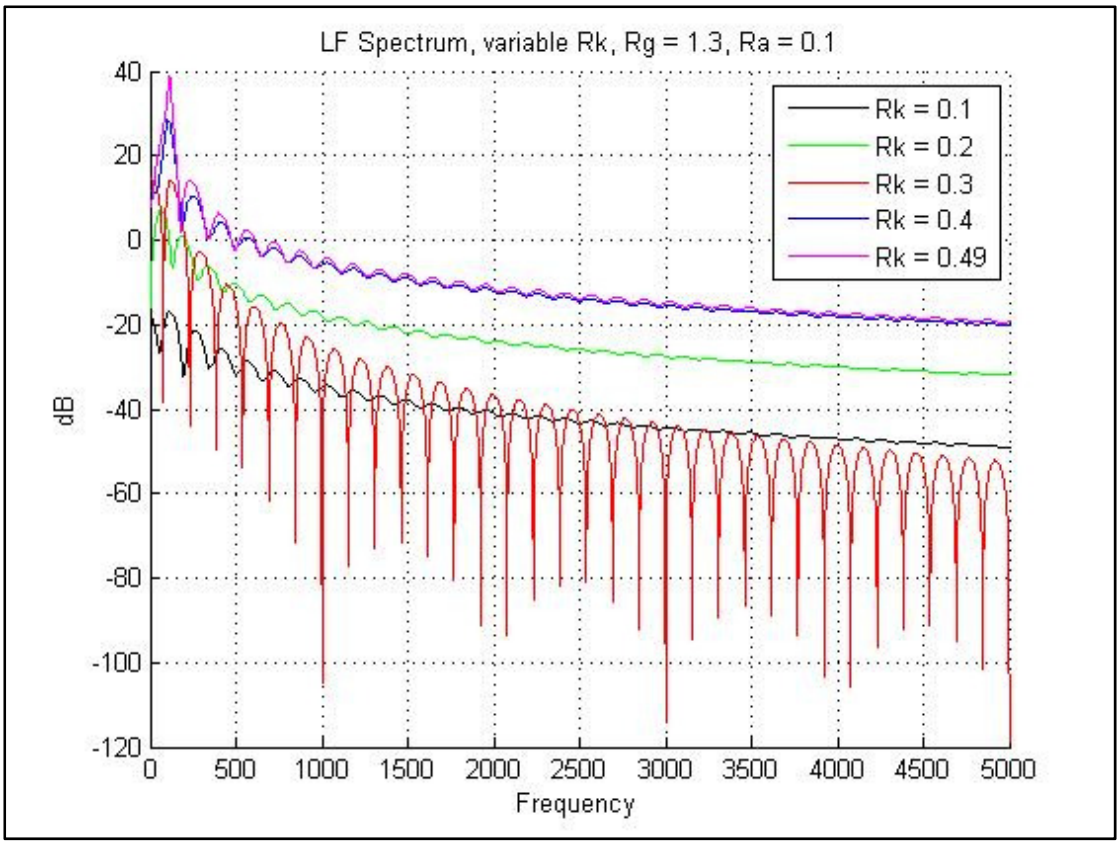

Figure 25: LF derivative Spectrum with variable Rk

7. Discussion and Future Work

7.1. Summary

In this text, we have discussed the glottal flow derivative waveform of the speech production system, an algorithm for extracting it from speech waveform, and a mathematical model for representing it in both time and frequency domain.

In particular, the estimation of the glottal flow derivative is automatic and requires only information which can be directly calculated from the speech signal. An innovative technique is used: identifying the closed phase through formant modulation calculated by a sliding covariance analysis. By identifying statistically significant variations in the frequency of the estimated first formant, we are able to identify when the glottis finishes closing and when it begins opening. The formant motions are predicted by the theory of interaction between the glottal flow and the vocal tract.

Next, a nonlinear least-squares algorithm is used to fit the LF model to the glottal flow derivative waveform for each pitch period. Steps must be taken in order to ensure that the curve fitting is performed in a manner that yields meaningful results. This is done by setting bounds to the estimated parameters so they can take only physically reasonable values.

Finally, the spectrum of the LF model is studied. In [4], an analytic formula of the LF model spectrum is derived. It is shown that it is possible to model in the spectral domain accurate description of the glottal flow characteristics. It is also possible to switch from frequency to time domain or from time to frequency domain with the help of exact formulas. This formulation allows for spectral modeling. These results are challenging the more traditional time-domain approaches of glottal modeling. It opens a new way for glottal parameters estimation in speech.

7.2. Future Work

The LF model fitted to the glottal flow derivative waveform calculated using the formant modulation technique can be extended from four parameter model to

seven parameter model, so as to include the glottal timings. As can be seen in [10], this could be useful for SID purposes.

Also, the identification of glottal opening and closing is done by whitening the speech waveform. Several other techniques can be used to provide more accurate identification.

Furthermore, a high fundamental frequency pose a problem in linear prediction analysis and formant tracking estimation. A two-window covariance based linear prediction analysis could be used to help minimize difficulty with high pitched speakers.

A useful application of the time domain part of this text could be the comparison of the closed phase speech samples and the glottal flow derivative in ‘Speech in Noise’ and in ‘Speech without Noise’. Speech in noise is high quality recorded speech when background noise is placed into the speaker’s headphones. Speech without noise is normal recorded speech in silence.

Finally, the spectral representation of the LF model can be studied in more depth so as to provide a method for spectral modeling of the glottal flow.

8. Bibliography

[1] T.V. Ananthapadmanabha and G. Fant. “Calculation of true glottal flow and its components”. *Speech Communications*, pages 167-184, 1982.

[2] T.V. Ananthapadmanabha and B. Yegnanarayana. “Epoch extraction from linear prediction residual for identification of closed glottis interval”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):309-319, August 1979.

[3] Kathleen E. Cummings and Mark A. Clements. “Analysis of glottal waveforms across stress styles”. In *ICASSP*, pages 369-372, 1990.

[4] Boris Doval and Christophe D’Alessandro. “Spectral Correlates of glottal waveform models: an analytic study”. In *Proceedings ICASSP-97*, Munich, 1997.

[5] G. Fant. “The LF model revisited: Transformations and frequency domain analysis”. *STL-QPSR*, 2-3/95, KTH, 1996.

[6] G. Fant. “Some Problems in Voice Source Analysis”. *Speech Communications*, 13:7-22, 1993.

[7] G. Fant, J. Liljencrants, and Q. Lin. “A four parameter model of glottal flow”. *STL-QPSR*, 4/85, pages 1-13, KTH, 1985.

[8] John Makhoul. “Linear Prediction: A tutorial review”. In *Proceedings of the IEEE*, volume 63, pages 561-580, April 1975.

[9] R. J. McAulay and T.F. Quatieri. “Pitch Estimation and Voicing Detection based on a sinusoidal model”. In *ICASSP*, pages 249-252, 1990.

[10] Michael D. Plumpe, T.F. Quatieri, and Douglas A. Reynolds. “Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification”. *IEEE Transactions on Speech and Audio Processing*, 7(5):569-586, September 1999.

[11] Lawrence R. Rabiner and Ronald W. Schafer. “Digital Processing of Speech Signals”, *Prentice Hall*, Inc. 1978.

[12] Raymond N.J. Veldhuis. “The Spectral Relevance of Glottal-Pulse Parameters”. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, USA; 1998:873-876.

[13] David Y. Wong, John D. Markel, and Augustine H. Gray Jr. “Least Squares glottal inverse filtering from the acoustic speech waveform”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):350-355, August 1979.

[14] B. Yegnanarayana and Raymond N.J. Veldhuis. “Extraction of Vocal-Tract System Characteristics from Speech Signals”. *IEEE Transactions on Speech and Audio Processing*, 6(4):313-327, July 1998.