Spectral estimation and significance of glottal-pulse parameters

IPO, Center for Research on User-System Interaction
PO Box 513, 5600 MB Eindhoven

13.03.1997

Rapport no. 1152

Spectral estimation and significance of glottal-pulse parameters

Raymond Veldhuis

Voor akkoord: Dr. J.M.B. Terken

Spectral estimation and significance of glottal-pulse parameters

Raymond Veldhuis
March 1997

Table of Contents

1. Introduction.

During the last year research that has been done at IPO on the spectral estimation of the parameters of the Liljencrants-Fant (LF) model (Fant, Liljencrants & Lin, 1985) for the glottal pulse. This work has resulted in an estimation method that computes the glottal-pulse parameters from an estimate of the speech spectrum. This method has certain advantages over existing methods, but it sometimes produces unwanted results, such as incorrect, unexpected jumps in some of the measured glottal-pulse parameter tracks. This spectral estimation method is discussed in Section 2. The estimation errors are due to the sensitivity of the method to errors in the estimates of the speech spectrum. This sensitivity to spectral errors and the spectral significance of the R parameters of the LF model are analysed in Section 3 and Section 4. Finally, Section 5 discusses topics for further research and proposes a reduced-parameter LF model of which the parameters are better suited for spectral estimation.

2. Spectral estimation of LF parameters.

2.1. Background.

In LPC-coded synthetic diphone speech (O'Shaugnessy, 1990), each diphone is stored as a sequence of LPC frames. An LPC frame represents a speech segment of typically 10 ms and contains, among other things, the coefficients of an all-pole filter modelling the spectral envelope of the speech signal and an indication of the type of excitation signal for the filter. In the case of voiced speech the excitation signal is a sequence of mono pulses. In the case of unvoiced speech it is noise. In order to improve the segmental quality of LPC-coded diphone speech there is a plan to use another representation of voiced speech, in which each frame consists of a formant filter, modelling the resonances of the vocal tract, and parameters of the glottal pulse. The expected advantages of this representation are a better segmental quality and, via the glottal-pulse parameters, control over the voice quality. The reader is referred to Appendix A, which is extracted from Veldhuis (1997), for a more extensive description of this representation and for some notes on glottal-pulse parameters.

The improved representation requires that the glottal-pulse parameters and the coefficients of the formant filter be estimated from the speech signal. Reliable methods for the estimation of the formant filter exist, e.g. Childers & Lee (1991), as well as methods for the estimation of glottal-pulse parameters, but the latter often need manual fine tuning or solely work on stationary vowels (Strik, 1994). For the estimation of the glottal-pulse parameters from a set of approximately 1500 diphones, which may contain 15000 10 ms frames, an automatic procedure that will work on running speech is preferred.

Existing methods for the estimation of glottal-pulse parameters can be characterized as time-domain methods, which work as follows. An inverse of the formant filter is applied to a segment of speech. This is generally referred to as inverse filtering, e.g. Markel (1972), Wong, Markel, & Gray (1979), Krishnamurty & Childers (1986), Childers et al. (1991), and Childers & Wong (1994). The resulting wave form represents the time derivative of the glottal airflow, cf. Appendix A. Subsequently, a prototype glottal-pulse time derivative is matched to one pitch period of the inverse-filtered signal and the parameters of the prototype wave form are tuned until the matching is optimal in a mean-square-error sense (Strik, 1994). This parameter estimation in the time-domain has a number of disadvantages. First, it requires a procedure to align the prototype

wave form with one pitch period extracted from the inverse-filtered segment. Second, it is sensitive to phase distortion in the recording process. Third, the measurement of the return-phase time constant, see Appendix A, is error prone. Finally, fourth, this method is not robust against model deviations or errors in the inverse filter, which both are likely to occur.

As an alternative we investigated a frequency-domain glottal-pulse estimation method, which, via analysis by synthesis, optimizes a spectral match between the inverse-filtered speech and the synthetic glottal pulse time derivative. This method does not need any time alignment of the inverse-filtered speech and a prototype waveform. It is insensitive to phase errors since phase is not used in the measurement. The return-phase time constant is directly related to the spectral tilt and therefore we expected that it could be measured accurately. Finally, we expected this method to be robust against model and inverse-filtering errors because is tries to maintain the spectral content of the speech, which to a great extent determines the speech percept.

2.2. The estimation method.

We adopt the popular LF model (Fant et al., 1985) for the glottal pulse, and choose the R parameters r_0, r_a and r_k, see (44) in Appendix A, Fant, Kruckenberg, Liljencrants & Bavegard, (1994) and Fant (1995), as model parameters, but we could develop a similar estimation method for another set of LF parameters, e.g. the T parameters, or for an other glottal pulse model, e.g. the Rosenberg model (Rosenberg, 1971), the Rosenberg++ model (Veldhuis, 1997) or the model used in Cummings & Clements (1993).

We assume that a short voiced segment of speech s_k, $k = 0, \ldots, N-1$, is available that is long enough to obtain an accurate estimate of the magnitude spectrum. Generally a duration of 20-30 ms suffices. At a sample frequency f_s of 8 kHz, this comes down to a segment of 160-240 samples. In addition an estimate $f_0 = 1/t_0$ of the fundamental frequency is available, as well as an estimate $F(\Omega)$, $-\pi < \Omega \leq \pi$, of the transfer function of the formant filter. We recommend the estimation procedure for the fundamental frequency that is proposed in Childers et al. (1991). For the estimation of the transfer function of the formant we recommend a method based on cepstral analysis, e.g. Deller, Proakis & Hansen (1993), or Rabiner & Juang (1993). In such a method the cepstral coefficients are computed from a discrete Fourier transform of the (windowed) speech segment. These cepstral coefficients are truncated in order to remove the influence of the voice fundamental from the log-spectrum, after which they are converted into coefficients of an all-pole filter. The poles at frequencies below 250 Hz are removed, e.g. Childers et al. (1991). The advantage of such a cepstral method over a LPC analysis is that the influence of the voice fundamental on the formant frequencies is reduced because of the cepstral smoothing. We define the formant line spectrum by

$$F_l = \left| F\left(2\pi l \frac{f_0}{f_s}\right) \right|^2, \; l = 1, \ldots, M, \tag{1}$$

and the signal line spectrum by

$$S_l = \left| \sum_{k=0}^{N-1} w_k s_k e^{-i2\pi k l \frac{f_0}{f_s}} \right|^2, \; 1, \ldots, M, \tag{2}$$

with the number of spectral lines limited by

$$M \leq \left\lfloor \frac{f_s}{2f_0} \right\rfloor \tag{3}$$

and w_k, $k = 0, \ldots, N-1$, an appropriate window function, e.g. a Hanning window. The model glottal-pulse line spectrum is defined by

$$G_l(r_o, r_a, r_k) = \int_0^1 \dot{g}_\tau(\tau; r_o, r_a, r_k) e^{-i2\pi l\tau} d\tau, \; l = 1, \ldots, M, \tag{4}$$

with $\dot{g}_\tau(\tau; r_o, r_a, r_k)$ the time derivative $\dot{g}(t)$ of the glottal pulse, expressed in R parameters and time-scaled by $1/t_0$. The reader is referred to Appendix B for more details on the model glottal-pulse line spectrum.

The spectral glottal-pulse parameter-estimation method minimizes the mean-squared log-spectral distance (Rabiner et al., 1993)

$$Q_g(r_o, r_a, r_k) = \frac{1}{M} \sum_{l=1}^{M} \left| \log \left(\frac{G_l(r_o, r_a, r_k) / \sum_{m=1}^{M} G_m(r_o, r_a, r_k)}{(S_l/F_l) / \sum_{m=1}^{M} (S_m/F_l)} \right) \right|^2 \tag{5}$$

as a function of r_o, r_a and r_k. Equation (5) shows the log-spectral distance between the power-normalized line spectrum of the inverse-filtered signal, which is an estimate of the actual power-normalized glottal-pulse line spectrum, and the power-normalized line spectrum of a glottal pulse with parameters r_o, r_a and r_k. The power normalization is included in order to make the estimates independent of the signal or glottal-pulse power. The log-spectral distances now only depend on the spectral shape. The estimates $\hat{r}_o$, $\hat{r}_a$ and $\hat{r}_k$ are the arguments at which the mean-squared log-spectral distance attains its global minimum.

Several minimization methods have been tried on the function $Q_g(r_o, r_a, r_k)$, such as conjugate-gradient methods and iterative line searches. Unfortunately, each evaluation of $Q_g(r_o, r_a, r_k)$ requires the computation of a sequence $G_l(r_o, r_a, r_k)$, $l = 1, \ldots, M$, which is a heavy computational burden. Another problem is that the function $Q_g(r_o, r_a, r_k)$ may have local minima, which complicates the minimization tasks. Good results have obtained by first using a precalculated code book of sequences $G_l(r_o, r_a, r_k)$, $l = 1, \ldots, M$ for various values of the R parameters to find an approximation of the global minimum and the optimal R parameters. In a second step the global minimum and the corresponding values of the R parameters can be computed more accurately by using a standard minimization method in a neighbourhood of this first approximation.

2.3. Results.

Trials on sustained natural vowels show plausible results, as is illustrated in Figure 1 through Figure 6, which show results for male vowel and female vowels /a/, /i/ and /u/. These results have been obtained with the codebook-based approach described above, but without the second step that produces an accurate approximation of the global optimum. All segments were sampled at 8 kHz and contained 300 samples, which corresponds to a duration of 37.5 ms. The male vowels had an f_0 of about 110 Hz, the female vowels of 200 Hz. The number M of spectral

lines that were involved in the minimization of $Q_g(r_o, r_a, r_k)$ was 20. Synthetic waveforms were generated from the estimated parameters $\hat{r}_o$, $\hat{r}_a$ and $\hat{r}_k$ and the estimated formant filters. The figures show the original and the resynthesized speech segment, the power-normalized original and resynthesized signal line spectra, the inverse-filtered signal and the resynthesized glottal-pulse time derivative. The latter two are shown a second time with a resynthesized glottal-pulse time derivative aligned with the inverse-filtered signal. The mean-squared log-spectral distances (5) are given in the captions.

The resynthesized speech waveform often resembles the original more strongly than is the case with LPC resynthesized speech. This is not always true when we compare the inverse-filtered signal and the resynthesized glottal pulses according to the LF model. The female vowels, especially the /i/ and the /u/, show the largest deviations. There can be two reasons for this. The first is that spectral errors were introduced because a) the inverse filter did not compensate the formants or b) the estimate of the signal line spectrum (2) was biased due to the windowing or c) the f_0 was estimated incorrectly. The second reason is that the glottal-pulse model is inadequate. We exclude the influence of phase errors, since male and female vowels were recorded with the same setup and there are no large deviations in the male glottal pulses. The mean-squared log-spectral distance is between 0.08 for the male /a/ and 0.96 for the female /u/. Since we have no reference data it is, at this point, difficult to state whether this is good or bad. While analysing longer sequences of sustained vowels we observed unexpected simultaneously occurring jumps in the r_a and the r_k parameter tracks. The jumps in the r_a and the r_k parameter tracks were systematically in the opposite direction. There were not so many jumps in the r_o parameter track.

Although the spectral estimation method was designed for speech synthesis purposes rather than for the estimation of glottal-pulse parameters, we wanted to investigate to which extent it is capable of correctly estimating the LF parameters. Therefore, we tested it on synthetic vowels obtained by passing LF glottal-pulse derivatives through a formant filter. The input signals were 37.5-ms segments of a male vowel /a/ with a sampling frequency of 8 kHz and an f_0 of about 110 Hz. The number M of spectral lines that were involved in the minimization of $Q_g(r_o, r_a, r_k)$ was 20. The actual formant filter is not important, because we obtained formant spectral lines (1) directly from the formant synthesis filter. We also used the original f_0. In this way we excluded spectral errors due to inverse filtering or an erroneous f_0 estimate. We synthesized speech with various values of the parameters r_o, r_a and r_k estimated these parameters from the synthetic speech. Pairs of original and estimated R parameters are shown in Figure 7. Figure 8 shows pairs of estimation errors.

Both $\hat{r}_o$ and $\hat{r}_k$ contain errors as high as 0.2, but seem to be distributed around the lines $\hat{r}_o = r_o$ and $\hat{r}_k = r_k$, respectively. Perhaps there is a small negative bias in $\hat{r}_k$. The estimated parameter $\hat{r}_a$ has a larger positive bias. Figure 8 shows that the errors in $\hat{r}_a$ and $\hat{r}_k$ are correlated. An increase in $\hat{r}_a$ with respect to the original value leads to a decrease in $\hat{r}_k$, which corresponds to the observations made on the estimated parameter tracks of sustained vowels. Such a correlation can not be observed for the other combinations of estimation errors. The only explanation for the estimation errors in the glottal-pulse parameters must be the bias in the signal spectral lines due to the windowing in (2), because model errors are excluded as well as errors in the formant filter and in f_0.

2.4. Discussion.

The simultaneous jumps in the r_a and the r_k parameter tracks make the method in its present

form unsuited for the automatic estimation of glottal-pulse parameters as we desired for the improved representation of diphone speech. This does not mean that such a representation cannot be used, but that the glottal-pulse parameters in the diphone data base must be selected in some other way.

The errors in the glottal-pulse parameters are due to spectral errors in the signal line spectrum. They remain, even if we exclude spectral errors due to inverse filtering or an erroneous f_0 estimate. This means that the spectral errors in the signal line spectrum are caused by the windowing bias. This bias is generally accepted in spectral estimation and can be considered small. It, therefore, seems that a small change in the signal spectral lines may lead to a substantial change in the estimated glottal-pulse parameters. This is a sensitivity problem which has a relation with the spectral significance of the glottal-pulse parameters. This problem is further discussed in Section 3.

3. Sensitivity and spectral significance.

In this section we will analyse two problems that will turn out to be related. The first concerns the observed sensitivity to spectral errors of the glottal-pulse parameter estimation methods of Section 2. This comes down to the question: 'What happens to the values of r_o, r_a, r_k at which $Q_g(r_o, r_a, r_k)$ is minimal, when we introduce a small spectral error?'. The second problem concerns the spectral significance of the glottal-pulse parameters. The question that we try to answer is: 'How do small changes in the glottal-pulse parameters affect the glottal-pulse spectral lines?'.

3.1. Sensitivity to spectral errors.

The discussion of the sensitivity problem will be a mathematical analysis of $Q_g(r_o, r_a, r_k)$ in a neighbourhood of its global minimum and we will derive results for a small perturbation of the estimated glottal line spectrum. In order to keep the notation simple, we adopt a vector notation for the glottal-pulse parameters and define

$$\underline{r} = \left[r_o \; r_a \; r_k\right]^T, \tag{6}$$

with the superscript T denoting matrix transposition. We assume that the model glottal line spectrum, which is now denoted as $G_l(\underline{r})$, $l = 1, \ldots, M$ is power normalized. Furthermore, we define the estimated glottal line spectrum, which is also power normalized, by:

$$\hat{G}_l = (S_l/F_l) / \sum_{m=1}^{M} (S_m/F_l), \; l = 1, \ldots, M. \tag{7}$$

We assume that the estimated glottal line spectrum is without spectral errors due to windowing or inverse filtering. There may be model errors, due to which the estimated glottal line spectrum will deviate somewhat from any model glottal line spectrum. Assume that the mean-square log-spectral distance

$$Q_g(\underline{r}) = \frac{1}{M}\sum_{l=1}^{M}\left|\log\left(\frac{G_l(\underline{r})}{\hat{G}_l}\right)\right|^2 \tag{8}$$

attains its global minimum at $\underline{r} = \underline{r}_{opt}$. For any $\underline{\delta}$

$$\sum_{l=1}^{M}\log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l}\right)\frac{\underline{\delta}^T\nabla G_l(\underline{r}_{opt})}{G_l(\underline{r}_{opt})} = 0, \tag{9}$$

in which the column vector $\nabla G_l(\underline{r})$ is the gradient of $G_l(\underline{r})$, defined by

$$(\nabla G_l(\underline{r}))_k = \frac{\partial}{\partial r_k}G_l(\underline{r}),\ k = 1, 2, 3. \tag{10}$$

We investigate the influence of an spectral error ξ_l in the estimated glottal line spectrum $\hat{G}_l$, on the value of $\underline{r}$ at which the mean-square log-spectral distance

$$Q_g(\underline{r};\underline{\xi}) = \frac{1}{M}\sum_{l=1}^{M}\left|\log\left(\frac{G_l(\underline{r})}{\hat{G}_l + \xi_l}\right)\right|^2, \tag{11}$$

with $\underline{\xi} = \left[\xi_1 \ \ldots \ \xi_M\right]^t$, is minimal. Because of the power normalization we have $\sum_l \xi_l = 0$. In order to be able to derive relatively simple mathematical results, we assume that $|\xi_l| \ll \hat{G}_l$, $l = 1, \ldots, M$, and that the model errors are small, i.e. $\hat{G}_l \approx G_l(\underline{r}_{opt})$, $l = 1, \ldots, M$. In the absence of model errors we would have that $\hat{G}_l = G_l(\underline{r}_{opt})$, $l = 1, \ldots, M$, and, of course, the mean-square log-spectral distance in (8) would have a zero minimum.

For $\underline{\delta}$ small enough, we can approximate $Q_g(\underline{r}_{opt} + \underline{\delta};\underline{\xi})$ by

$$Q_g(\underline{r}_{opt} + \underline{\delta};\underline{\xi}) = Q_g(\underline{r}_{opt};\underline{\xi}) + \underline{\delta}^T\nabla Q_g(\underline{r}_{opt};\underline{\xi}) + \frac{1}{2}\underline{\delta}^T\nabla^2 Q_g(\underline{r}_{opt};\underline{\xi})\underline{\delta}, \tag{12}$$

in which $\nabla^2 Q_g(\underline{r};\underline{\xi})$ is the 3×3 matrix, defined by

$$(\nabla^2 Q_g(\underline{r};\underline{\xi}))_{m,n} = \frac{\partial^2}{\partial r_m \partial r_n}Q_g(\underline{r};\underline{\xi}). \tag{13}$$

By using $|\xi_l| \ll \hat{G}_l$, $l = 1, \ldots, M$, we can write

$$Q_g(\underline{r}_{opt};\underline{\xi}) = \frac{1}{M}\sum_{l=1}^{M}\left(\log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l}\right) - \frac{\xi_l}{\hat{G}_l}\right)^2 \tag{14}$$

for the first term in (12). For the second term in (12) we can write

$$\underline{\delta}^T\nabla Q_g(\underline{r}_{opt};\underline{\xi}) = \frac{2}{M}\sum_{l=1}^{M}\log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l + \xi_l}\right)\frac{\underline{\delta}^T\nabla G_l(\underline{r}_{opt})}{G_l(\underline{r}_{opt})}. \tag{15}$$

If we use $|\xi_l| \ll \hat{G}_l$, $l = 1, \ldots, M$, and (9) this reduces to

$$\underline{\delta}^T \nabla Q_g(\underline{r}_{opt};\underline{\xi}) = -\underline{\delta}^T \frac{2}{M} \sum_{l=1}^{M} \frac{\nabla G_l(\underline{r}_{opt})}{G_l(\underline{r}_{opt})} \frac{\xi_l}{\hat{G}_l}. \tag{16}$$

The matrix $\nabla^2 Q_g(\underline{r};\underline{\xi})$ in the third term in (12) has elements

$$(\nabla^2 Q_g(\underline{r};\underline{\xi}))_{m,n} = \tag{17}$$

$$\frac{2}{M} \sum_{l=1}^{M} \frac{1 - \log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l + \xi_l}\right)}{G_l^2(\underline{r}_{opt})} \frac{\partial G_l(\underline{r}_{opt})}{\partial r_m} \frac{\partial G_l(\underline{r}_{opt})}{\partial r_n} + \frac{\log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l + \xi_l}\right)}{G_l(\underline{r}_{opt})} \frac{\partial^2 (G_l(\underline{r}_{opt}))}{\partial r_m \partial r_n}$$

If we use $|\xi_l| \ll \hat{G}_l$, $l = 1, \ldots, M$, and $\hat{G}_l \approx G_l(\underline{r}_{opt})$, $l = 1, \ldots, M$, the logarithms in (17) vanish and for the third term in (12) we can write

$$\frac{1}{2} \underline{\delta}^T \nabla^2 Q_g(\underline{r}_{opt};\underline{\xi}) \underline{\delta} = \frac{1}{M} \sum_{l=1}^{M} \left(\frac{\underline{\delta}^T \nabla G_l(\underline{r}_{opt})}{G_l(\underline{r}_{opt})} \right)^2. \tag{18}$$

We define the $M \times 3$ sensitivity matrix $\Sigma(\underline{r})$ by

$$\Sigma(\underline{r})_{l,m} = \frac{1}{G_l(\underline{r})} \frac{\partial G_l(\underline{r})}{\partial r_m} \quad , l = 1, \ldots, M, m = 1, \ldots, 3. \tag{19}$$

The element $\Sigma(\underline{r})_{l,m}$ represents the relative change in $G_l(\underline{r})$ due to a small change in r_m. The matrix therefore describes the relative spectral significance of the glottal-pulse parameters. The M vector $\underline{\xi}_r$ of relative spectral errors is defined by

$$\xi_{rl} = \xi_l / \hat{G}_l, \; l = 1, \ldots, M. \tag{20}$$

With these definitions (12) can be rewritten as

$$Q_g(\underline{r}_{opt} + \underline{\delta}) = \frac{1}{M} \left(\sum_{l=1}^{M} \left(\log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l}\right) - \frac{\xi_{rl}}{\hat{G}_l} \right)^2 - 2\underline{\delta}^T \Sigma^T(\underline{r}_{opt}) \underline{\xi}_r + \underline{\delta}^T \Sigma^T(\underline{r}_{opt}) \Sigma(\underline{r}_{opt}) \underline{\delta} \right). \tag{21}$$

Because of the differences in the value ranges of the glottal-pulse parameter r_a on the one had and the parameters r_o and r_k on the other hand, we prefer to work with relative shifts in the glottal-pulse parameters. Therefore, we define

$$\underline{\delta}_r = (\mathrm{diag}(\underline{r}_{opt}))^{-1} \underline{\delta}, \tag{22}$$

and

$$\Sigma_r(\underline{r}_{opt}) = \mathrm{diag}(\underline{r}_{opt}) \Sigma(\underline{r}_{opt}). \tag{23}$$

We then have

$$Q_g(\underline{r}_{opt} + \underline{\delta}_r) = \tag{24}$$

$$\frac{1}{M}\left(\sum_{l=1}^{M}\left(\log\left(\frac{G_l(\underline{r}_{opt})}{\hat{G}_l}\right) - \frac{\xi_{rl}}{\hat{G}_l}\right)^2 - 2\underline{\delta}_r^T\Sigma_r^T(\underline{r}_{opt})\underline{\xi}_r + \underline{\delta}_r^T\Sigma_r^T(\underline{r}_{opt})\Sigma_r(\underline{r}_{opt})\underline{\delta}_r\right)$$

We can now solve (24) for $\underline{\delta}$, which results in

$$\underline{\delta}_{r,opt} = (\Sigma_r^T(\underline{r}_{opt})\Sigma_r(\underline{r}_{opt}))^{-1}\Sigma_r^T(\underline{r}_{opt})\underline{\xi}_r. \tag{25}$$

In (24) we can identify the contributions to the mean-square log-spectral error due to a) small relative deviations $\underline{\delta}_r$ of the glottal-pulse parameters, b) small spectral errors in the estimated glottal line spectrum, reflected in $\underline{\xi}_r$ and c) small model deviations, which are quantified by $\log((G_l(\underline{r}_{opt}))/\hat{G}_l)$. Equation (25) tell us how much the optimal glottal pulse parameters change when there is a small error in the estimated glottal line spectrum. The change is linearly proportional to the relative spectral error, and this proportionality is fully determined by the relative sensitivity of the model glottal line spectrum to changes in the glottal pulse parameters, which is given by $\Sigma_r(\underline{r}_{opt})$.

We have now developed the tools to analyse the sensitivity of the estimated glottal-pulse parameters to a small spectral error. We use the singular-value decomposition of $\Sigma_r(\underline{r}_{opt})$, which is given by (Golub & Van Loan, 1986)

$$\Sigma_r(\underline{r}_{opt}) = UDV^T, \tag{26}$$

in which $U = \left[\underline{u}_1\ \underline{u}_2\ \underline{u}_3\right]$ is an $M \times 3$ matrix with orthonormal columns that span up the column space of $\Sigma_r(\underline{r}_{opt})$, $D = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3)$ is a diagonal matrix containing the singular values of $\Sigma_r(\underline{r}_{opt})$ and V is an orthonormal 3×3 matrix spanning up the row space of $\Sigma_r(\underline{r}_{opt})$. The non-negative singular values $\sigma_1, \sigma_2, \sigma_3$ can be obtained as the square roots of the eigenvalues of $\Sigma_r^T(\underline{r}_{opt})\Sigma_r(\underline{r}_{opt})$. We assume $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$. The columns of V are the eigenvectors of $\Sigma_r^T(\underline{r}_{opt})\Sigma_r(\underline{r}_{opt})$ and span up the glottal-pulse parameter space. We can write

$$\underline{\xi}_r = \alpha_1\underline{u}_1 + \alpha_2\underline{u}_2 + \alpha_3\underline{u}_3 + \alpha_\perp\underline{u}_\perp, \tag{27}$$

with $\underline{u}_\perp$ orthogonal to $\left[\underline{u}_1\ \underline{u}_2\ \underline{u}_3\right]$. It then follows that

$$\underline{\delta}_r = \frac{\alpha_1}{\sigma_1}\underline{v}_1 + \frac{\alpha_2}{\sigma_2}\underline{v}_2 + \frac{\alpha_3}{\sigma_3}\underline{v}_3, \tag{28}$$

and

$$\|\underline{\delta}_r\|^2 = \left(\frac{\alpha_1}{\sigma_1}\right)^2 + \left(\frac{\alpha_2}{\sigma_2}\right)^2 + \left(\frac{\alpha_3}{\sigma_3}\right)^2. \tag{29}$$

This means that the components of the spectral errors which are orthogonal to the column space of $\Sigma_r(\underline{r}_{opt})$ do not influence the estimation of the glottal-pulse parameters. On the other hand,

the shift in the glottal-pulse parameters will be large when the spectral error has components in the column space of $\Sigma_r(\underline{r}_{opt})$ that are associated with singular values that approach 0. For example, when $\sigma_3 \ll \alpha_3$, there will be a large shift in the glottal-pulse parameters in the direction of $\underline{v}_3$. In general, if the relative spectral errors are independent and equally distributed over all spectral lines, then the relative errors in the glottal-pulse parameters are expected to be the largest in the direction $\underline{v}_3$. They will be, on average, σ_3/σ_1 times larger than the relative errors in the least sensitive direction $\underline{v}_1$.

3.2. Spectral significance of glottal-pulse parameters.

Now we turn to the analysis of the spectral significance of the glottal-pulse parameters, which is also based on $\Sigma_r(\underline{r})$. If we want to know the effect ΔG_l on $G_l(\underline{r})$ of a small change $\Delta\underline{r}$ in the glottal pulse parameters we can write for $\Delta\underline{r}$ small enough

$$\Delta G_l = (\nabla G_l(\underline{r}))^T \Delta\underline{r}, \tag{30}$$

or

$$\frac{\Delta G_l}{G_l(\underline{r})} = \frac{(\nabla G_l(\underline{r}))^T}{G_l(\underline{r})} \Delta\underline{r}. \tag{31}$$

The row vector $(\nabla G_l(\underline{r}))^T/(G_l(\underline{r}))$ in the right-hand side of this equation is the jth row of $\Sigma(\underline{r})$. Therefore, for the M vector of relative spectral deviations $\Delta_r\underline{G}$ we have, with $\Delta_r\underline{r} = (\mathrm{diag}(\underline{r}))^{-1}\Delta\underline{r}$,

$$\Delta_r\underline{G} = \Sigma_r(\underline{r})\Delta_r\underline{r}. \tag{32}$$

This shows how small changes in the glottal-pulse parameters affect the glottal-pulse line spectrum. After singular value decomposition of $\Sigma(\underline{r})$, we obtain

$$\Delta_r\underline{G} = UDV^T\Delta_r\underline{r}. \tag{33}$$

Because columns of $V = \left[\underline{v}_1\ \underline{v}_2\ \underline{v}_3\right]$ span up the glottal-pulse parameter space, we can write

$$\Delta_r\underline{r} = \beta_1\underline{v}_1 + \beta_2\underline{v}_2 + \beta_3\underline{v}_3, \tag{34}$$

and find

$$\Delta_r\underline{G} = \beta_1\sigma_1\underline{u}_1 + \beta_2\sigma_2\underline{u}_2 + \beta_3\sigma_3\underline{u}_3, \tag{35}$$

and

$$\|\Delta_r\underline{G}\|^2 = (\beta_1\sigma_1)^2 + (\beta_2\sigma_2)^2 + (\beta_3\sigma_3)^2. \tag{36}$$

The vectors $\underline{v}_1$ and $\underline{v}_3$ are the directions in the glottal-pulse parameter space of, respectively, maximum and minimum spectral significance. For instance, when σ_3 is much smaller than the other two singular values, only changes $\Delta_r\underline{r}$ in the directions $\underline{v}_1$ or $\underline{v}_2$ will affect the glottal line spectrum. The three-parameter model than essentially works as a two-parameter model.

Combining the discussions on sensitivity and spectral significance we see that when we mini-

mize the mean-square log-spectral distance $Q_g(t)$ in order to estimate the glottal-pulse parameters, the errors in the parameter vector will be largest in the direction of minimum spectral significance. In Section 4 we will illustrate these effects with numerical examples.

4. Results.

Table 6 shows R parameters and singular values $\sigma_1, \sigma_2, \sigma_3$ of $\Sigma_r(t)$ for glottal pulses of various voice types. The entries 1-9 were obtained from Childers et al. (1991), the entries 10-27 are

Table 6: R parameters and singular values $\sigma_1, \sigma_2, \sigma_3$ of $\Sigma_r(t)$ for glottal pulses of various voice types.

nr.	voice type	r_a	r_k	r_o	σ_1	σ_2	σ_3
1	male, modal /i/	0.021	0.31	0.64	9.7503	4.1131	1.2103
2	male, modal /i/	0.025	0.34	0.71	11.3122	5.9142	1.2540
3	male, modal /a/	0.015	0.33	0.68	9.3728	3.7541	1.3269
4	male, slight vocal fry /a/	0.008	0.28	0.63	7.8381	1.8001	0.9388
5	male, vocal fry /a/	0.005	0.25	0.25	7.3763	1.6511	0.3701
6	male, falsetto /a/	0.133	0.35	0.77	16.0980	1.2916	1.1507
7	male, falsetto /a/	0.043	0.44	0.89	20.1658	5.6871	2.0611
8	male, breathy /i/	0.068	0.42	0.68	41.8595	14.3189	0.9217
9	male, breathy /i/	0.100	0.45	0.84	26.8449	6.7635	1.9537
10	male, normal	0.020	0.38	0.54	11.4197	7.6993	1.3301
11	female, normal	0.050	0.52	0.71	60.6060	17.0320	1.0535
12	male, low F_0	0.026	0.43	0.61	16.7893	10.7277	1.3701
13	female, low F_0	0.051	0.42	0.76	28.4513	12.7848	1.0284
14	male, medium F_0	0.015	0.45	0.56	13.1106	8.3405	1.5098
15	female, medium F_0	0.042	0.48	0.71	37.9375	14.3674	1.0335
16	male, high F_0	0.098	0.31	0.87	65.7867	8.4822	1.4049
17	female, high F_0	0.030	0.49	0.65	28.3418	13.5025	1.3557
18	male, low level	0.027	0.41	0.69	15.0627	10.5274	1.2931
19	female, low level	0.110	0.57	0.81	99.4240	23.8867	2.2549
20	male, medium level	0.019	0.45	0.57	15.7155	10.2345	1.5197

Table 6: R parameters and singular values $\sigma_1, \sigma_2, \sigma_3$ of $\Sigma_r(\underline{r})$ for glottal pulses of various voice types.

nr.	voice type	r_a	r_k	r_o	σ_1	σ_2	σ_3
21	female, medium level	0.037	0.51	0.68	36.6113	15.4317	1.2459
22	male, high level	0.016	0.38	0.49	11.8488	7.6708	1.2408
23	female, high level	0.019	0.52	0.64	21.5118	13.1364	1.5834
24	male, breathy	0.046	0.51	0.65	55.6362	16.4460	1.2002
25	female, breathy	0.081	0.48	0.79	49.3108	15.4530	1.6230
26	male, pressed	0.013	0.40	0.41	11.2541	6.4124	1.2206
27	female, pressed	0.032	0.50	0.71	31.4372	13.9071	1.2377

from Karlsson & Liljencrants (1996). For all entries we see that σ_3 is at least a factor of 7 smaller than σ_1. The average ratio σ_3/σ_1 equals 0.07, the average ratio σ_2/σ_1 equals 0.4. A scatter plot of the orthogonal projection of the vectors $\underline{v}_3$ on the $\Delta r_k/r_k$-$\Delta r_a/r_a$ plane is shown in Figure 9. Nearly all points in this plot lie on the unit circle. This means that the vectors $\underline{v}_3$ have a negligible component in the $\Delta r_o/r_o$ direction and that the largest relative estimation errors in the R parameters are expected to be in the $\Delta r_k/r_k$-$\Delta r_a/r_a$ plane. We also see that the orthogonal projection of the vectors $\underline{v}_3$ on the $\Delta r_k/r_k$-$\Delta r_a/r_a$ plane are all in more or less the same direction, even though the voice types have been labelled as very different. This means that for all these cases we can expect the same type of errors, which are described by

$$\frac{dr_k}{r_k}\sin(\varphi) - \frac{dr_a}{r_a}\cos(\varphi) = 0, \tag{37}$$

with $\left[\cos(\varphi)\ \sin(\varphi)\right]^T$ the normalized orthogonal projection of the vector $\underline{v}_3$ on the $\Delta r_k/r_k$-$\Delta r_a/r_a$ plane. From this we can derive that due to spectral errors, the parameters r_a and r_k will move along a curve

$$r_a = C r_k^{\tan(\varphi)}, \tag{38}$$

with C an arbitrary constant. The vectors $\underline{v}_1$ and $\underline{v}_2$ do not show such a systematic behaviour, except that the first element of $\underline{v}_1$ is large and often (in about 50% of the cases) close to 1. This means that $\underline{v}_1$ is substantially, or sometimes mainly, in the direction of r_o which is therefore relatively insensitive to spectral errors.

With regard to the spectral significance of the R parameters we can state that r_o usually has a strong (or the strongest) spectral significance and that there is a certain covariation of r_a and r_k, given by (38), which will only lead to small or minimal spectral changes.

As an example we will present the mean-squared log-spectral distance

$$Q_g(\underline{r} + \mathrm{diag}(\underline{r})\underline{\delta}_r) = \frac{1}{M}\sum_{l=1}^{M}\left|\log\left(\frac{G_l(\underline{r} + \mathrm{diag}(\underline{r})\underline{\delta}_r)}{G_l(\underline{r})}\right)\right|^2 \tag{39}$$

as a function of the relative deviation $\underline{\delta}_r = [\Delta r_o/r_o, \Delta r_a/r_a, \Delta r_k/r_k]^T$ for the R parameters $\underline{r} = [r_o, r_a, r_k]^T = [0.614, 0.029, 0.400]^T$, which is close to entry 12 of Table 6. For the matrix D of singular values of $\Sigma_r(\underline{r}_{opt})$, we find

$$D = \begin{bmatrix} 16.01 & 0 & 0 \\ 0 & 10.24 & 0 \\ 0 & 0 & 1.25 \end{bmatrix},$$

and for the corresponding basis of the glottal-pulse parameter space

$$V = \begin{bmatrix} 0.9276 & -0.3734 & 0.0095 \\ 0.2959 & 0.7499 & 0.5916 \\ 0.2280 & 0.5460 & -0.8061 \end{bmatrix}.$$

Figure 10 shows three-dimensional plots of the mean-squared log-spectral distance for $\Delta r_o = 0$ (top left), $\Delta r_k = 0$ (top right) and $\Delta r_o = 0$ (bottom). Figure 11 shows the corresponding contour plots.

If the relative spectral errors are independent and equally distributed over all spectral lines, then the smallest relative estimation errors can be expected in the direction

$$\begin{bmatrix} \Delta r_o/r_o \\ \Delta r_a/r_a \\ \Delta r_k/r_k \end{bmatrix} = \lambda \begin{bmatrix} 0.9276 \\ 0.2959 \\ 0.2280 \end{bmatrix},$$

which is mainly in the direction of r_o. This also means that changes of the glottal-pulse parameters in this direction are spectrally the most significant. The largest relative estimation errors can be expected in the direction

$$\begin{bmatrix} \Delta r_o/r_o \\ \Delta r_a/r_a \\ \Delta r_k/r_k \end{bmatrix} = \lambda \begin{bmatrix} 0.0095 \\ 0.5916 \\ -0.8061 \end{bmatrix},$$

which is almost completely in the $\Delta r_k/r_k$-$\Delta r_a/r_a$ plane and corresponds to the narrow valley shown in the top left picture of Figure 10. This relative error will on average be nearly $\sigma_1/\sigma_3 = 13$ times larger than the relative error in r_o. An relative error of 5% in r_o would then mean that the estimates of the other two R parameters are useless. The other pictures show that the estimation errors in r_o are more or less independent of estimation errors in the other parameters.

Figure 12 shows the columns of UD (top panel), which span the column space of $\Sigma_r(\underline{r}_{opt})$, and are weighted with the corresponding singular values. The relative spectral changes due to relative modifications of r_o, r_k and r_a parameters in the directions $\underline{v}_1$, $\underline{v}_2$ and $\underline{v}_3$ are indicated by solid lines, dashed lines and dash-dotted lines, respectively. The bottom panel of this figure shows the columns of $\Sigma_r(\underline{r}_{opt})$. The relative spectral changes due to relative perturbations of the r_o, r_k and r_a parameters presented as solid lines, dashed lines and dash-dotted lines, respec-

tively. Note the similarity between the first columns of UD and $\Sigma_r(\underline{r}_{opt})$.

5. Topics for further research.

5.1. Perceptual relevance of LF parameters.

The columns of the matrix V obtained via the singular value decomposition $\Sigma_r(\underline{r}) = UDV^T$ contain the coefficients of linear combinations of glottal-pulse parameters that have largest, medium and smallest spectral significance. A perceptual experiment can show to which extent this spectral relevance, quantified by the singular values, can predict the perceptual relevance of these linear combinations. A possible approach would be to estimate the just-noticeable differences in the directions of the columns of V, and compare those with the singular values. A topic for further discussion would be the question whether a possible lack of correlation between the measured just-noticeable differences and the singular values shows the importance of phase to speech quality.

5.2. A two-parameter LF model for spectral parameter estimation.

We have seen in Section 4 that in many cases the three-parameter LF model is too rich to allow robust spectral estimation of its parameters. If in certain applications spectral estimation is useful, for instance because it is easier, then a reduced parameter set would be required. One parameter in this set could be r_0, because changes $\Delta r_0/r_0$ have a significant or even dominant effect on the glottal-pulse line spectrum. Moreover, r_0 has a familiar interpretation as the open quotient. It was seen in Section 4 that estimation errors in r_0 were nearly independent to those in the other R parameters. Therefore, the second parameter in the two-parameter LF model should control the parameters r_k and r_a. Evolutions of r_k and r_a in the direction $\underline{v}_3$ must be avoided, because these are the most sensitive to spectral errors. Fortunately, the projection of $\underline{v}_3$ on the $\Delta r_k/r_k$-$\Delta r_a/r_a$ plane seems more or less constant for a large variety of glottal pulses. Therefore, we use the average projection, which we denote by $[\cos(\varphi)\ \sin(\varphi)]^T$. We require

$$\frac{dr_k}{r_k}\cos(\varphi) + \frac{dr_a}{r_a}\sin(\varphi) = 0, \tag{40}$$

or

$$\log(r_k)\cos(\varphi) + \log(r_a)\sin(\varphi) = C, \tag{41}$$

with C an arbitrary constant. A possible parameter is ρ in

$$\begin{bmatrix}\log(r_k)\\ \log(r_a)\end{bmatrix} = C\begin{bmatrix}\cos(\varphi)\\ \sin(\varphi)\end{bmatrix} + \rho\begin{bmatrix}\sin(\varphi)\\ -\cos(\varphi)\end{bmatrix}. \tag{42}$$

The constant C can be obtained by fitting (42) to data. This two-parameter model can be used to estimate the parameters r_0 and ρ, which are relevant for the signal's spectrum. If we use (42) to compute estimates $\hat{r}_k$ and $\hat{r}_a$ we may find incorrect values due to the parameter reduction. The true estimates are to be found for some 'spectrally invisible' value of λ along the curve

$$\begin{bmatrix} \log(r_k) \\ \log(r_a) \end{bmatrix} = \lambda \begin{bmatrix} \cos(\varphi) \\ \sin(\varphi) \end{bmatrix} + \begin{bmatrix} \log(\hat{r}_k) \\ \log(\hat{r}_a) \end{bmatrix}. \tag{43}$$

If the estimates $\hat{r}_k$ and $\hat{r}_a$ are to be used in speech synthesis, and we require a strong similarity between the original segment and the resynthesized one, then it is important that the parameter λ is of little perceptual relevance. This emphasizes the need for the investigation proposed in Subsection 5.1.

5.3. Comparison of spectral estimation of LF parameters with other methods.

The sensitivity to signal-spectral-line estimation errors of this method is undesirable, because these errors are likely to occur. It is an interesting question whether such a sensitivity is also present in time-domain glottal-pulse parameter estimation methods. We therefore propose to do a similar analysis as has been done in Section 3 and Section 4 for a time-domain method based on minimizing a mean squared error.

A. The LF model.

For analysis and synthesis purposes, speech production is often modelled by a source-filter model. Figure 13 shows two versions of a source-filter model. On the left we see a model consisting of a source producing a signal $g(t)$ which models the air flow passing the vocal cords, a filter with a transfer function $H(j\omega)$ which models the spectral shaping by the vocal tract and an operator R which models the conversion of the air flow to a pressure wave $s(t)$ as it takes place at the lips and which is called lip radiation. The operator R is essentially a differentiation operator. On the right we see a simplified version of this model, in which the differentiation operator has been combined with the source, which now produces the time derivative $\dot{g}(t)$ of the air flow passing the vocal cords. The opening between the vocal cords is called glottis, therefore the source is referred to as the glottal source. In voiced speech the signal $g(t)$ is periodic and one period is called a glottal pulse. The glottal pulse or, more often, its time derivative has been the topic of many studies because it is expected to determine the voice quality and to be related to the production of prosody, e.g. Childers et al. (1991), Cummings et al. (1993), Gobl (1989), Klatt & Klatt (1990), Pierrehumbert (1989), Rosenberg (1971), Strik (1994). The time derivative of the glottal pulse is studied rather than the glottal pulse because it is easier to obtain it from the speech signal and to derive some of the glottal-source parameters from it.

The LF model (Fant et al., 1985) has become a reference model for glottal-pulse analysis. Unfortunately, its use in speech synthesizers is limited because of its computational complexity. This computational complexity is due to the difference between the specification parameters and the generation parameters of the LF model. The computation of the generation parameters from the specification parameters is computationally complex, because it involves solving a nonlinear equation. Figure 14 shows typical examples of $g(t)$ and $\dot{g}(t)$ and introduces the specification parameters t_0, t_p, t_e, t_a and U_o or E_e (Fant et al., 1985). The length of a pitch period is t_0. The maximum air flow U_o occurs at t_p and the maximum excitation with amplitude E_e, which corresponds to the instant when the vocal cords collide, occurs at t_e. The interval with approximate length $t_a = E_e/\ddot{g}(t_e)$ just after the instant of maximum excitation is called the return phase. During this phase the vocal folds reach maximum closure and the air

flow reduces to its minimum. The minimum air flow is often referred to as leakage. Here we assume that there is no leakage, therefore $g(0) = g(t_0) = 0$. The air flow in the return phase is of perceptual importance, because it determines the spectral slope. The parameters t_0, t_p, t_e, t_a are called the T parameters. Instead of the T parameters, sometimes the R parameters (Fant et al., 1985) are used, which are defined as follows:

$$r_o = t_e/t_0,\ r_a = t_a/t_0,\ r_k = (t_e - t_p)/t_0. \tag{44}$$

The parameters r_o and r_a denote the relative duration of the open phase and the return phase, respectively. The parameter r_k quantifies the symmetry of the glottal pulse. The parameter r_o is usually referred to as the open quotient (OQ).

The following expression is a general description of the glottal air flow derivative $\dot{g}(t)$ with an exponential decay modelling the return phase

$$\dot{g}(t) = \begin{cases} f(t) & \text{for } 0 \le t < t_e \\ f(t_e)\dfrac{\exp\left(-\dfrac{(t-t_e)}{t_a}\right) - \exp\left(-\dfrac{(t_0-t_e)}{t_a}\right)}{1 - \exp\left(-\dfrac{(t_0-t_e)}{t_a}\right)} & \text{for } t_e \le t < t_0 \end{cases}. \tag{45}$$

We require $f(0) = 0$. In addition we have that $f(t_0) = 0$. Integration leads to the following expression for the glottal air flow:

$$g(t) = \begin{cases} \int_0^t f(\tau)\mathrm{d}\tau & \text{for } 0 \le t < t_e \\ \int_0^{t_e} f(\tau)\mathrm{d}\tau + t_a f(t_e)\dfrac{1 - \exp\left(-\dfrac{t-t_e}{t_a}\right) - \dfrac{t-t_e}{t_a}\exp\left(-\dfrac{t_0-t_e}{t_a}\right)}{1 - \exp\left(-\dfrac{t_0-t_e}{t_a}\right)} & \text{for } t_e \le t < t_0 \end{cases}. \tag{46}$$

Since there is no leakage we require $g(t) \ge 0$ and $g(0) = g(t_0) = 0$, from which one can derive the following continuity condition

$$\int_0^{t_e} f(\tau)\mathrm{d}\tau + t_a f(t_e) D(t_0, t_e, t_a) = 0, \tag{47}$$

with

$$D(t_0, t_e, t_a) = \frac{1 - \dfrac{t_0 - t_e + t_a}{t_a}\exp\left(-\dfrac{t_0-t_e}{t_a}\right)}{1 - \exp\left(-\dfrac{t_0-t_e}{t_a}\right)}. \tag{48}$$

Any parameters of $f(t)$ must be chosen such that condition (47) is satisfied.

The parameter t_a in the above definitions for the glottal air flow $g(t)$ and its derivative $\dot{g}(t)$ is the time constant of the exponential decay in the return phase. This is slightly different from the situation in Figure 14, where $t_a = E_e/\dot{g}(t_e)$. For $t_a \ll t_0 - t_e$, which is usually the case, both definitions are equivalent. If this is not the case then there exists a simple relation between both t_a parameters.

The LF model presented in Fant et al. (1985), but with the modified definition of t_a, follows from (45) and the choice

$$f(t) = B\sin\left(\pi\frac{t}{t_p}\right)\exp(\alpha t), \tag{49}$$

with B the amplitude of the glottal-pulse derivative. The generation parameter α can only be solved numerically from the continuity equation (47), which in this case reads:

$$\frac{\pi - \exp(\alpha t_e)\left(\pi\cos\left(\pi\frac{t_e}{t_p}\right) - \alpha t_p\sin\left(\pi\frac{t_e}{t_p}\right)\right)}{\pi^2 + (\alpha t_p)^2} + \frac{t_a}{t_p}\exp(\alpha t_e)\sin\left(\pi\frac{t_e}{t_p}\right)D(t_0, t_e, t_a) = 0. \tag{50}$$

Solving (50) for α is a heavy computational load in a speech-synthesizer, where the T parameters may vary typically every 10 ms.

We now derive expressions for the LF model in terms of the R parameters (44). We consider a time scaled time derivative of the glottal pulse, defined by

$$\dot{g}_r(\tau; r_o, r_a, r_k) = \dot{g}(\tau t_0),\ 0 \le \tau < 1, \tag{51}$$

with the R parameters defined in (44). Expression (45) can than be rewritten as

$$\dot{g}_r(\tau) = \begin{cases} f_r(\tau) & \text{for } 0 \le \tau < r_o \\ f_r(r_o)\dfrac{\exp\left(-\dfrac{(\tau - r_o)}{r_a}\right) - \exp\left(-\dfrac{(1 - r_o)}{r_a}\right)}{1 - \exp\left(-\dfrac{(1 - r_o)}{r_a}\right)} & \text{for } r_o \le \tau < 1 \end{cases}, \tag{52}$$

in which $f_r(\tau)$ is given by

$$f_r(\tau) = B\sin\left(\pi\tau\frac{r_k + 1}{r_o}\right)\exp(\beta\tau), \tag{53}$$

with

$$\beta = \alpha t_0. \tag{54}$$

Definition (48) is replaced by

$$D_r(r_o, r_a) = \frac{1 - \frac{1 - r_o + r_a}{r_a} e^{-\frac{1 - r_o}{r_a}}}{1 - e^{-\frac{1 - r_o}{r_a}}}. \qquad (55)$$

Finally, β is solved from

$$\frac{\pi + e^{\beta r_o}\left(\pi\cos(\pi r_k) - \frac{\beta r_o}{r_k + 1}\sin(\pi r_k)\right)}{\pi^2 + \left(\frac{\beta r_o}{r_k + 1}\right)^2} - \frac{r_a(r_k + 1)}{r_o} e^{\beta r_o}\sin(\pi r_k) D_r(r_o, r_a) = 0. \qquad (56)$$

B. The glottal-pulse line spectrum.

We derive an expression for $G_l(r_o, r_a, r_k)$ in (4). For the integral on the right-hand side of (4) we can write

$$\int_0^1 g_r(\tau; r_o, r_a, r_k) e^{-i2\pi l\tau} d\tau = \qquad (57)$$

$$B\left(\int_0^{r_o} \sin\left(\pi\tau\frac{r_k + 1}{r_o}\right) e^{\beta\tau} e^{-i2\pi l\tau} d\tau - \sin(\pi r_k) e^{\beta r_o} \int_{r_o}^1 \left(\frac{e^{-\frac{\tau - r_o}{r_a}} - e^{-\frac{1 - r_o}{r_a}}}{1 - e^{-\frac{1 - r_o}{r_a}}}\right) e^{-i2\pi l\tau} d\tau\right) =$$

$$B(I_1 - I_2)$$

For I_1 we find

$$I_1 = \frac{\left(\left(\pi\frac{r_k + 1}{r_o}\right)\cos(\pi r_k) - (\beta - i2\pi l)\sin(\pi r_k)\right) e^{\beta r_o} + \pi\frac{r_k + 1}{r_o}}{(\beta - i2\pi l)^2 + \left(\pi\frac{r_k + 1}{r_o}\right)^2}, \qquad (58)$$

and for I_2

$$I_2 = \sin(\pi r_k) e^{\beta r_a} \frac{e^{-\frac{1-r_a}{r_a}}}{1 - e^{-\frac{1-r_a}{r_a}}} \left(\frac{r_a \left(e^{\frac{1-r_a}{r_a}} e^{-i2\pi l r_a} - 1 \right)}{1 + i2\pi l r_a} - \frac{e^{-i2\pi l r_a} - 1}{i2\pi l} \right). \quad (59)$$

References

Childers, D.G. & Lee, C.K. (1991). Voice quality factors: Analysis synthesis and perception. *Journal of the Acoustical Society of America, 90*, 2394-2410.

Childers, D.G. & Wong, C.-F. (1994). Measuring and modelling vocal source-tract interaction. *IEEE Transactions on Biomedical Engineering, 41*, 663-671.

Cummings. K.E. & Clements, M.A. (1993) Application of the analysis of glottal excitation of stressed speech to speaking style modification. *Proceedings ICASSP-93*, 207-210, Minneapolis.

Deller. J.R., Proakis, J.G. & Hansen, J.H.L. (1993). *Discrete-Time Processing of Speech Signals.* Macmillan. New York.

Fant. G. (1995). The LF model revisited. Transformations and frequency domain analysis. *Speech Transmission Laboratory Quarterly Progress Report 2-3/95*, KTH.

Fant, G.. Kruckenberg, A., Liljencrants, J. & Bavegard, M. (1994). Voice source parameters in continuous speech. Transformation of LF parameters. *Proceedings of the ICSLP-94*, 1451-1454, Yokohama.

Fant. G.. Liljencrants, J. & Lin. Q. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress Report 4/85*, KTH.

Gobl. C. (1989). A preliminary study of acoustic voice quality correlates. *Speech Transmission Laboratory Quarterly Progress Report 4/89*, KTH.

Golub. G.H. & Van Loan, C.F. (1986). *Matrix Computations.* North Oxford Academic, London.

Karlsson. I. & Liljencrants, J. (1996). Diverse voice qualities: models and data *TMH-QPSR 2/1996*, KTH.

Klatt. D.H. & Klatt. L.C. (1990). Analysis synthesis. and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820-856.

Krishnamurty, A.K. & Childers, D.G. (1986). Two-channel speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34, 730-743.

Markel. J.D. (1972). Digital inverse filtering-A new tool for formant trajectory estimation. *IEEE Transactions on Audio Electroacoustics*, 20, 129-137.

O'Shaugnessy. D. (1990). *Speech Communication.* Addison Wesley, Reading.

Pierrehumbert, J.B. (1989). A preliminary study of the consequences of intonation for the voice source. *Speech Transmission Laboratory Quarterly Progress Report 4/89*, KTH.

Rabiner. L.R. & Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey.

Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America. 49*, 583-590.

Strik. H (1994). *Physiological Control and Behaviour of the Voice Source in the Production of Prosody*. PhD. Thesis, University of Nijmegen.

Veldhuis, R.N.J. (1997). An alternative for the LF model. *IPO Annual Progress Report.* IPO Eindhoven.

Wong, D.Y., Markel, J.D. & Gray, A.H. (1979). Least squares glottal inverse filtering from the acoustic speech signal. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 4, 350-355.

Figures

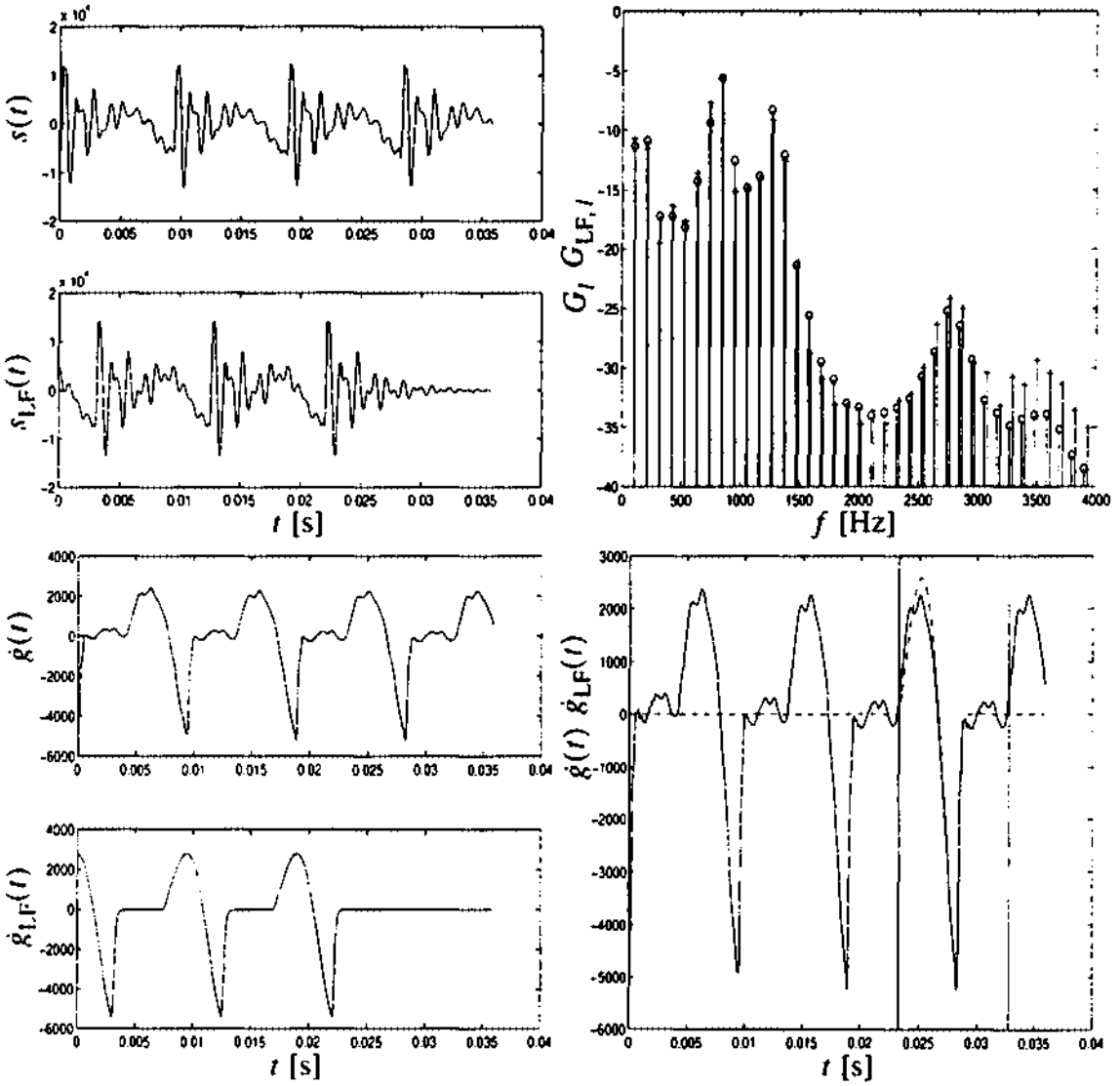

Figure 1: Results from spectral glottal-pulse parameter estimation on a male vowel /a/. Top left: original segment $s(t)$ and resynthesized segment $s_{LF}(t)$, arbitrary units. Top right: power-normalized original signal line spectrum G_l (+) and resynthesized signal line spectrum $G_{LF,l}$ (o) in dB.Bottom right: inverse-filtered segment $\dot{g}(t)$ and resynthesized glottal pulse time derivative $\dot{g}_{LF}(t)$, arbitrary units. Bottom left: inverse filtered segment $\dot{g}(t)$ and (solid) resynthesized glottal pulse time derivative $\dot{g}_{LF}(t)$ (dashed), arbitrary units. Mean-squared log-spectral distance: 0.08.

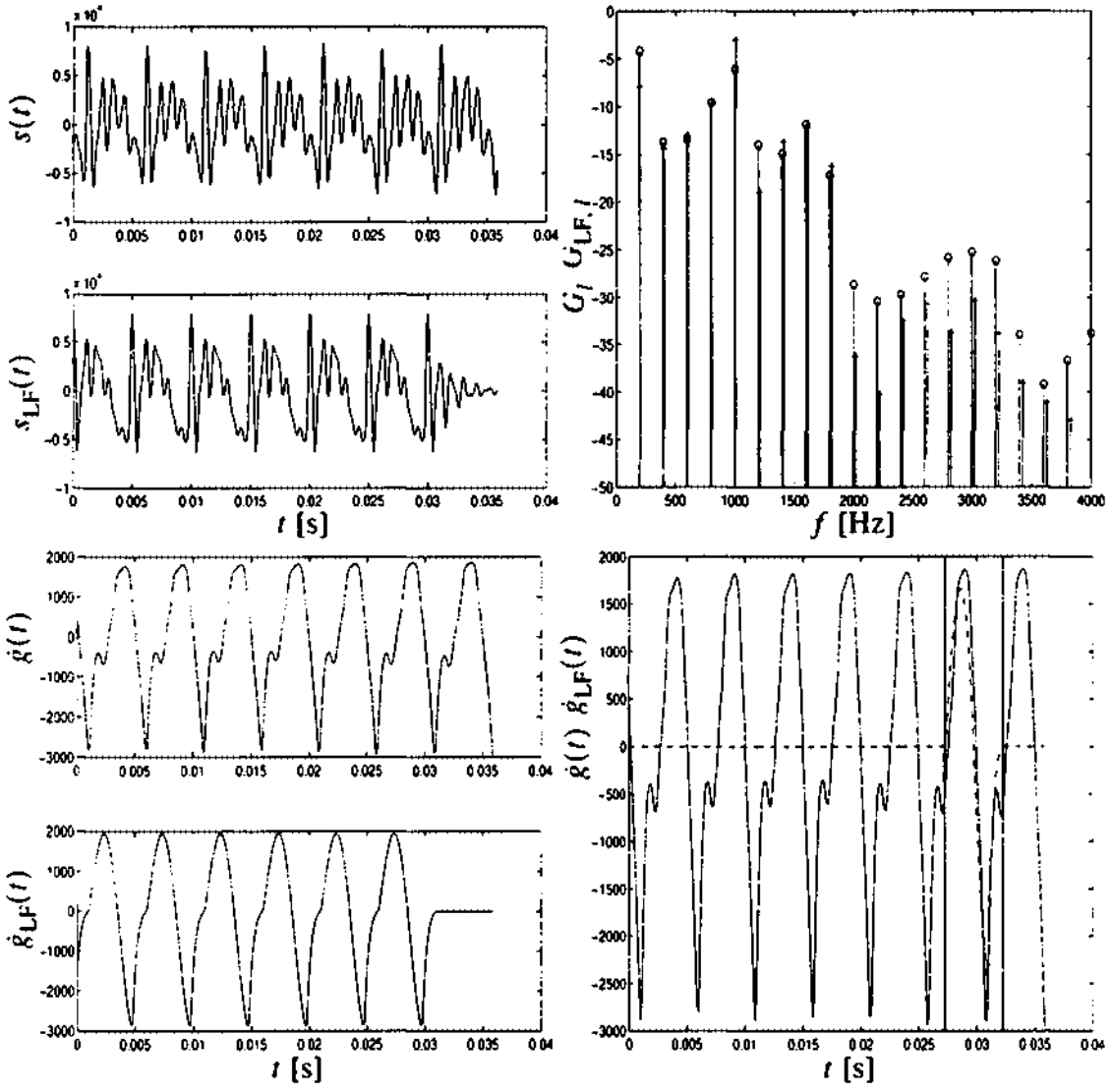

Figure 2: Results from spectral glottal-pulse parameter estimation on a female vowel /a/. Top left: original segment $s(t)$ and resynthesized segment $s_{\mathrm{LF}}(t)$, arbitrary units. Top right: power-normalized original signal line spectrum G_l (+) and resynthesized signal line spectrum $G_{\mathrm{LF},l}$ (o) in dB. Bottom right: inverse-filtered segment $\dot{g}(t)$ and resynthesized glottal pulse time derivative $\dot{g}_{\mathrm{LF}}(t)$, arbitrary units. Bottom left: inverse filtered segment $\dot{g}(t)$ and (solid) resynthesized glottal pulse time derivative $\dot{g}_{\mathrm{LF}}(t)$ (dashed), arbitrary units. Mean-squared log-spectral distance: 0.67.

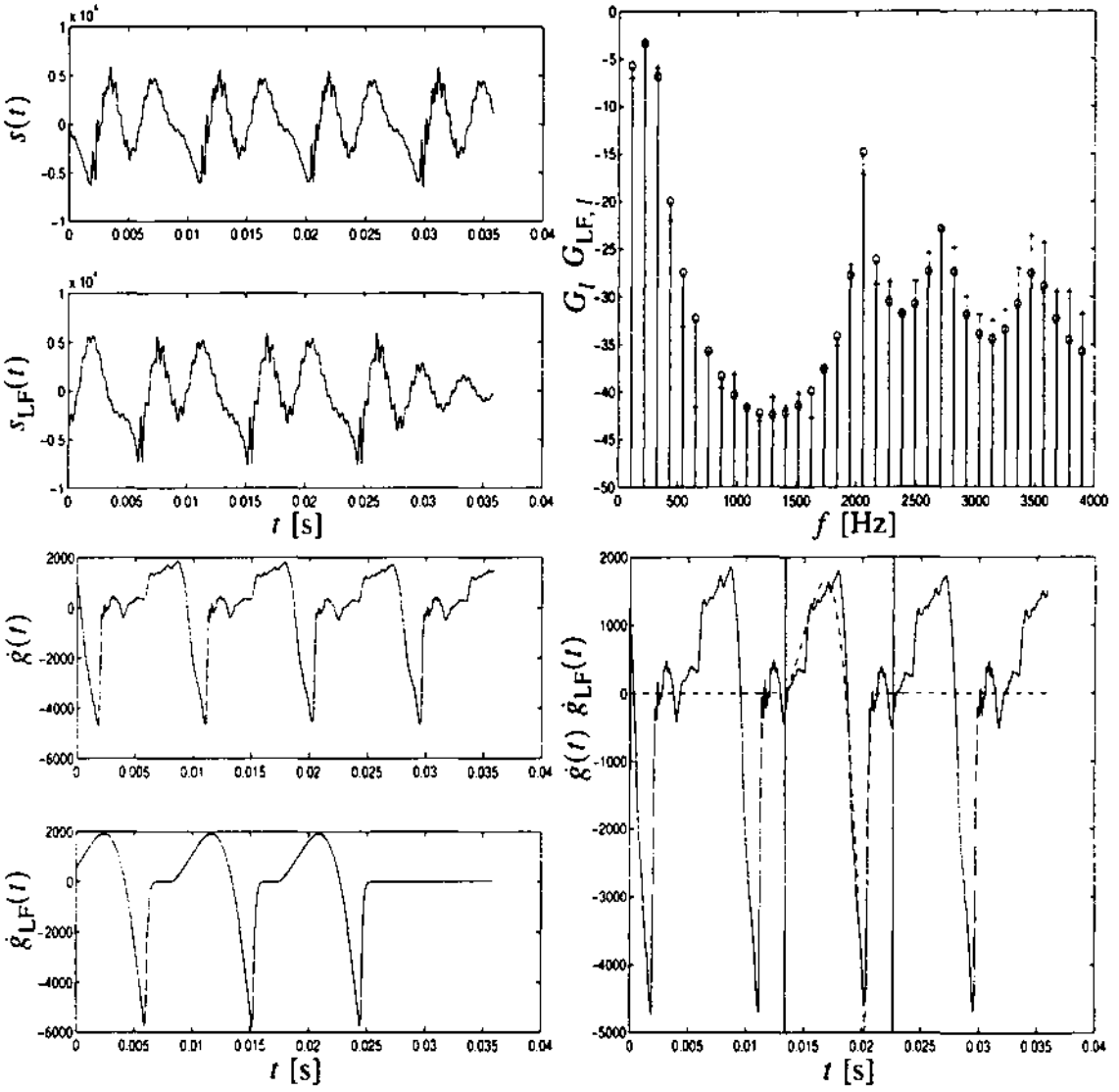

Figure 3: Results from spectral glottal-pulse parameter estimation on a male vowel /i/. Top left: original segment $s(t)$ and resynthesized segment $s_{\mathrm{LF}}(t)$, arbitrary units. Top right: power-normalized original signal line spectrum G_l (+) and resynthesized signal line spectrum $G_{\mathrm{LF},l}$ (o) in dB. Bottom right: inverse-filtered segment $\dot{g}(t)$ and resynthesized glottal pulse time derivative $\dot{g}_{\mathrm{LF}}(t)$, arbitrary units. Bottom left: inverse filtered segment $\dot{g}(t)$ and (solid) resynthesized glottal pulse time derivative $\dot{g}_{\mathrm{LF}}(t)$ (dashed), arbitrary units. Mean-squared log-spectral distance: 0.36.

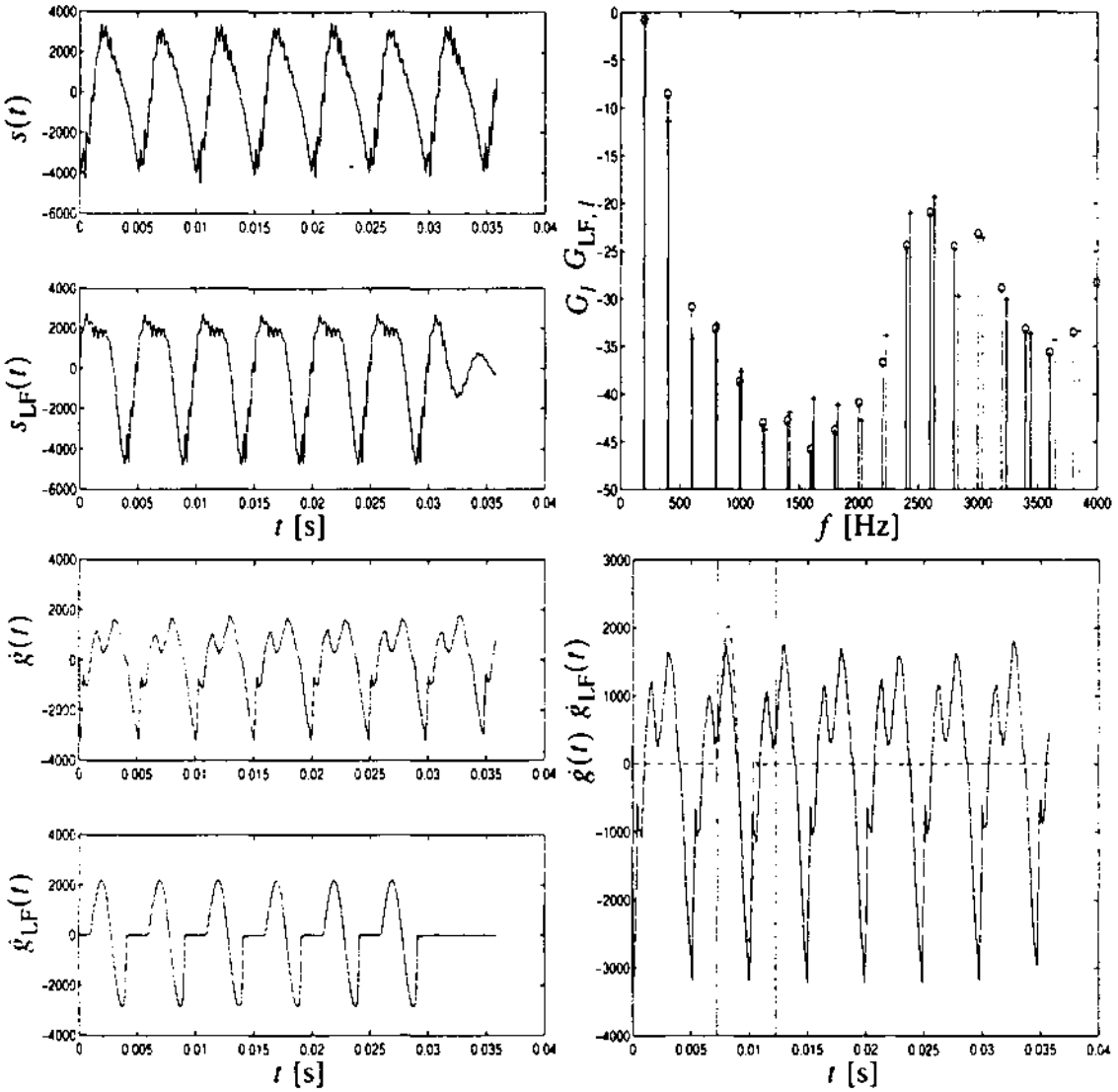

Figure 4: Results from spectral glottal-pulse parameter estimation on a female vowel /i/. Top left: original segment $s(t)$ and resynthesized segment $s_{\mathrm{LF}}(t)$, arbitrary units. Top right: power-normalized original signal line spectrum G_l (+) and resynthesized signal line spectrum $G_{\mathrm{LF},l}$ (o) in dB.Bottom right: inverse-filtered segment $\dot{g}(t)$ and resynthesized glottal pulse time derivative $\dot{g}_{\mathrm{LF}}(t)$, arbitrary units. Bottom left: inverse filtered segment $\dot{g}(t)$ and (solid) resynthesized glottal pulse time derivative $\dot{g}_{\mathrm{LF}}(t)$ (dashed), arbitrary units. Mean-squared log-spectral distance: 0.31.

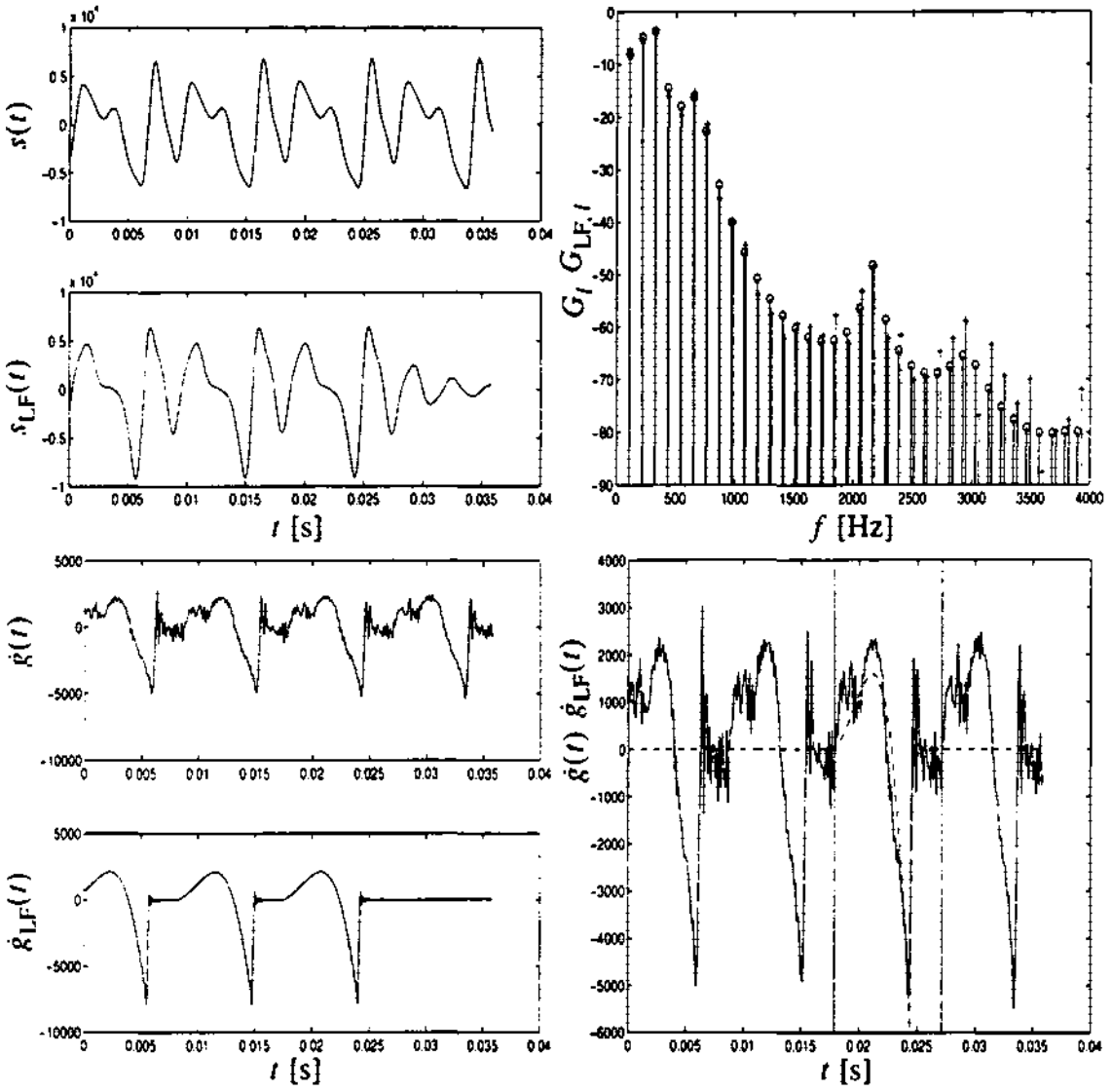

Figure 5: Results from spectral glottal-pulse parameter estimation on a male vowel /u/. Top left: original segment $s(t)$ and resynthesized segment $s_{LF}(t)$, arbitrary units. Top right: power-normalized original signal line spectrum G_l (+) and resynthesized signal line spectrum $G_{LF,l}$ (o) in dB. Bottom right: inverse-filtered segment $\dot{g}(t)$ and resynthesized glottal pulse time derivative $\dot{g}_{LF}(t)$, arbitrary units. Bottom left: inverse filtered segment $\dot{g}(t)$ and (solid) resynthesized glottal pulse time derivative $\dot{g}_{LF}(t)$ (dashed), arbitrary units. Mean-squared log-spectral distance: 0.24.

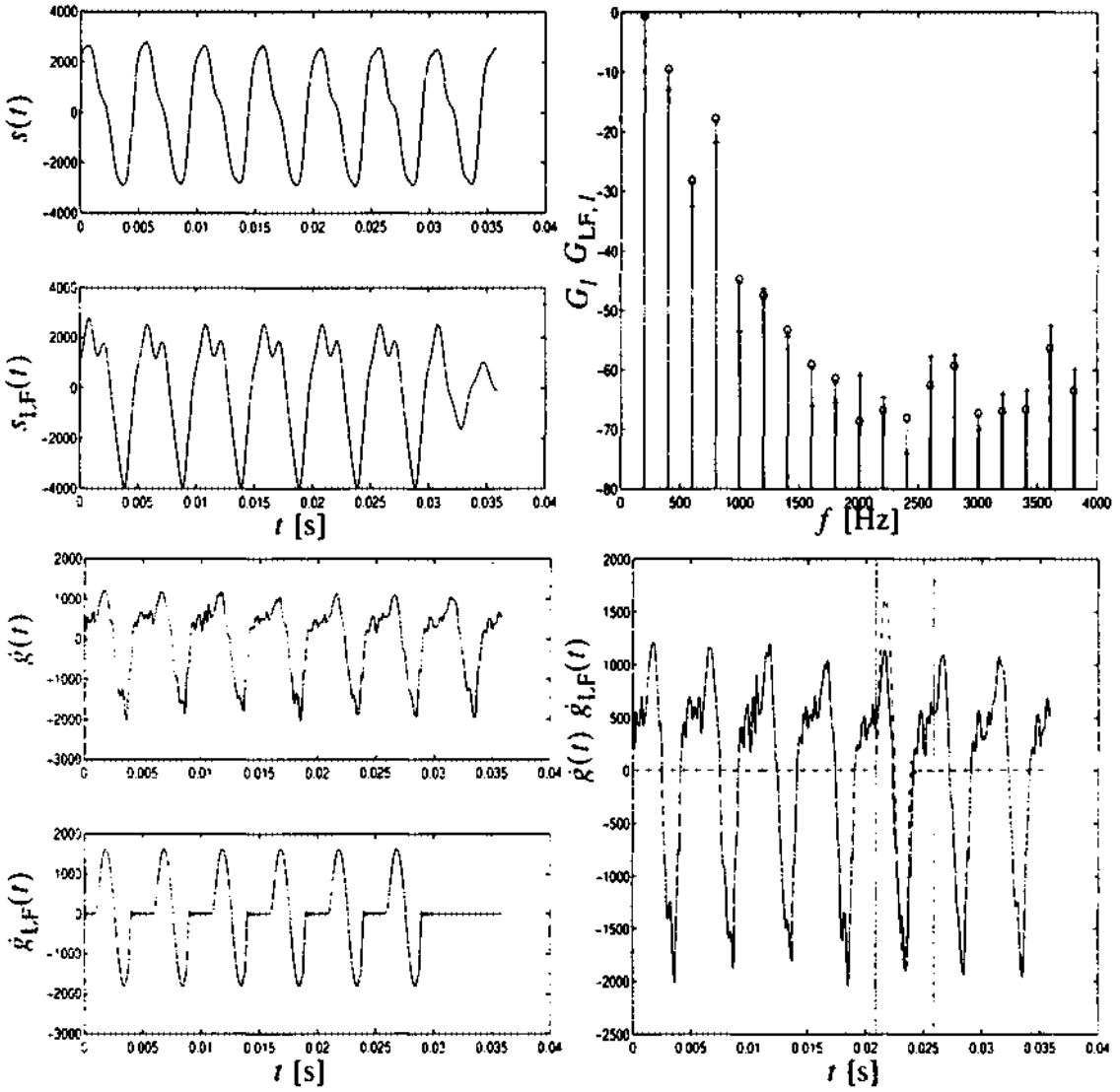

Figure 6: Results from spectral glottal-pulse parameter estimation on a female vowel /u/. Top left: original segment $s(t)$ and resynthesized segment $s_{LF}(t)$, arbitrary units. Top right: power-normalized original signal line spectrum G_l (+) and resynthesized signal line spectrum $G_{LF,l}$ (o) in dB.Bottom right: inverse-filtered segment $\dot{g}(t)$ and resynthesized glottal pulse time derivative $\dot{g}_{LF}(t)$, arbitrary units. Bottom left: inverse filtered segment $\dot{g}(t)$ and (solid) resynthesized glottal pulse time derivative $\dot{g}_{LF}(t)$ (dashed), arbitrary units. Mean-squared log-spectral distance: 0.96.

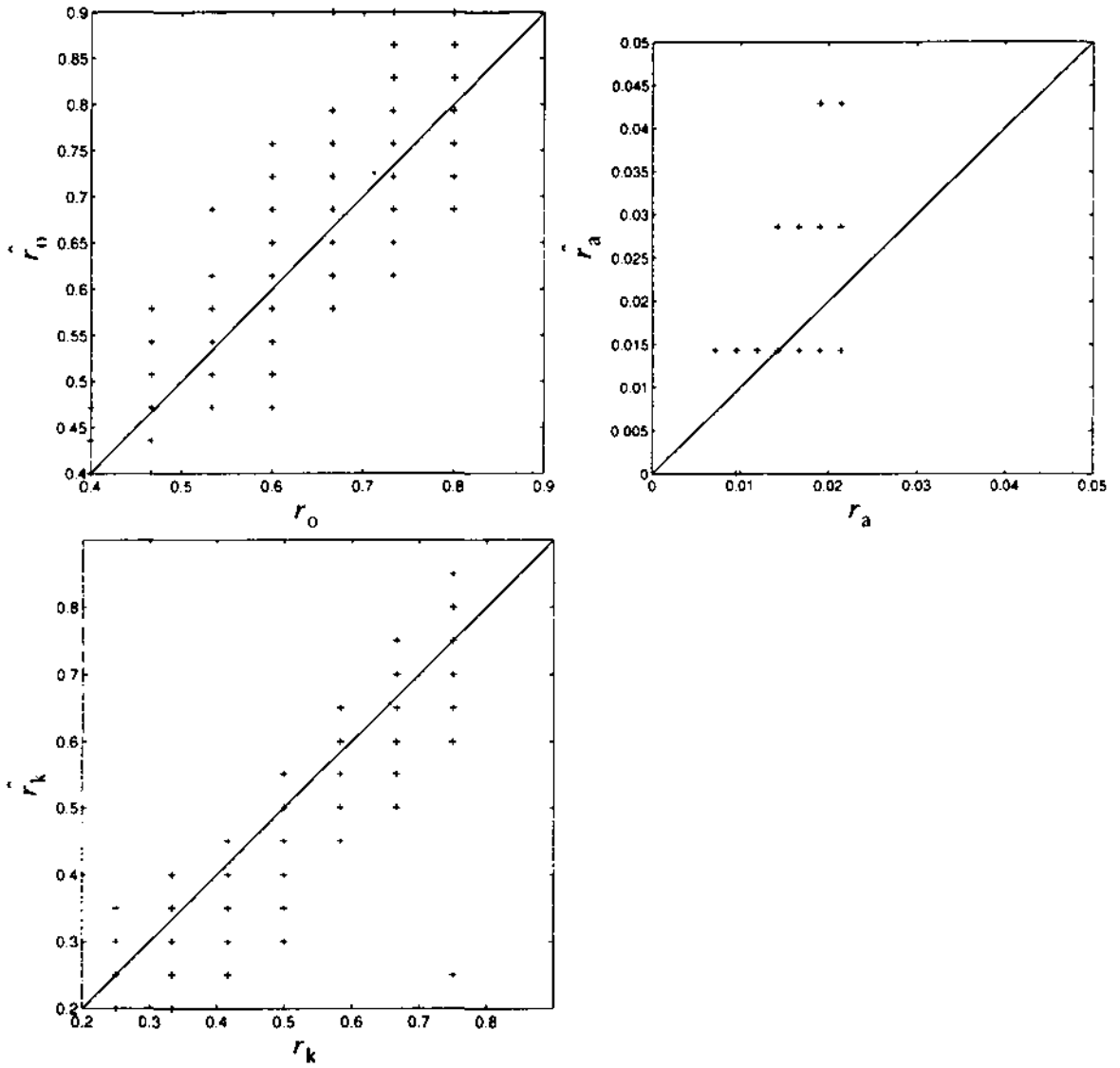

Figure 7: Estimated glottal parameters r_o, r_a and r_k paired with parameters r_o, r_a and r_k of a synthetic vowel /a/. The solid line indicates equality between original and estimated parameters.

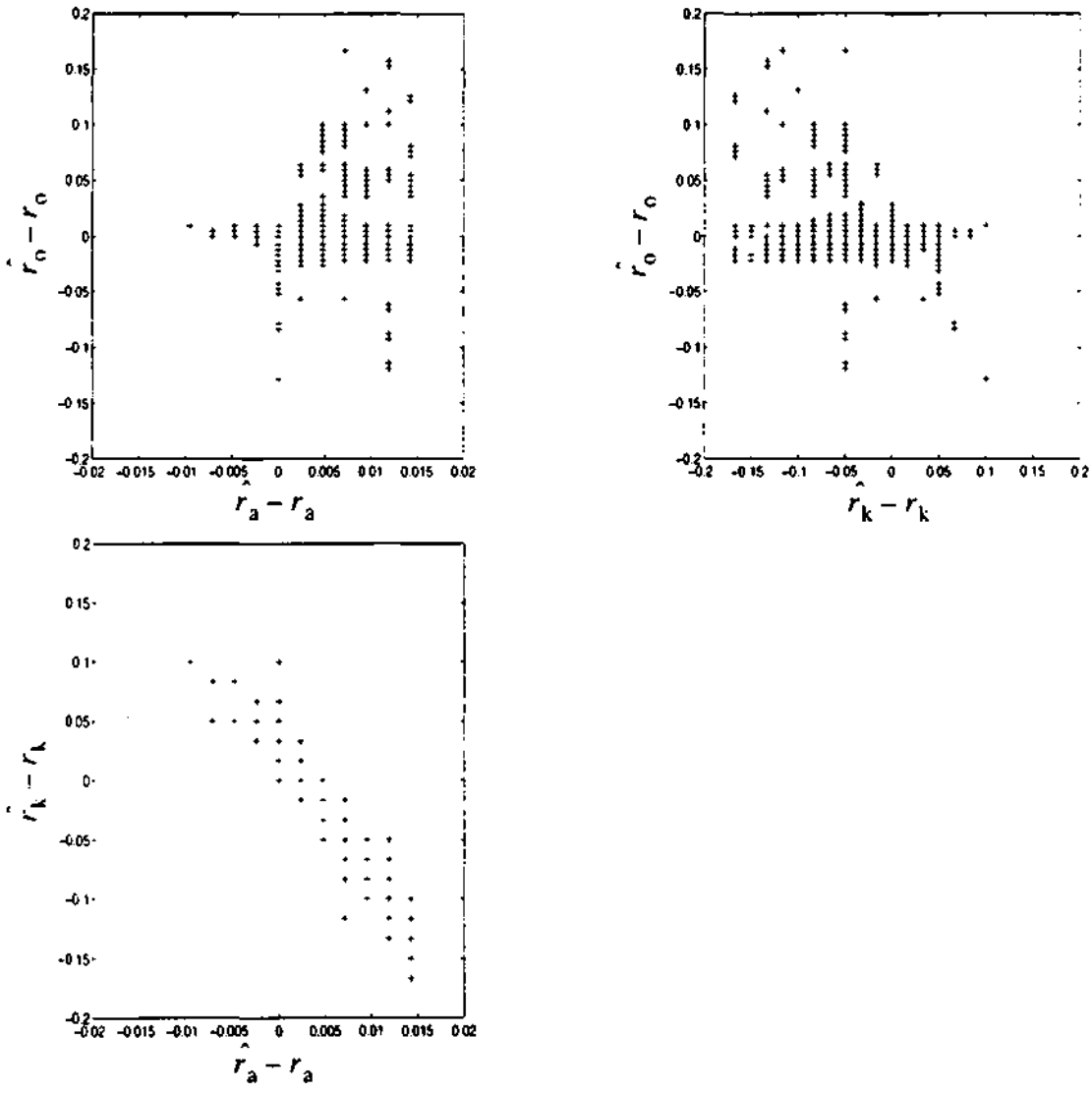

Figure 8: Pairs of various estimation errors in parameters r_o, r_a and r_k.

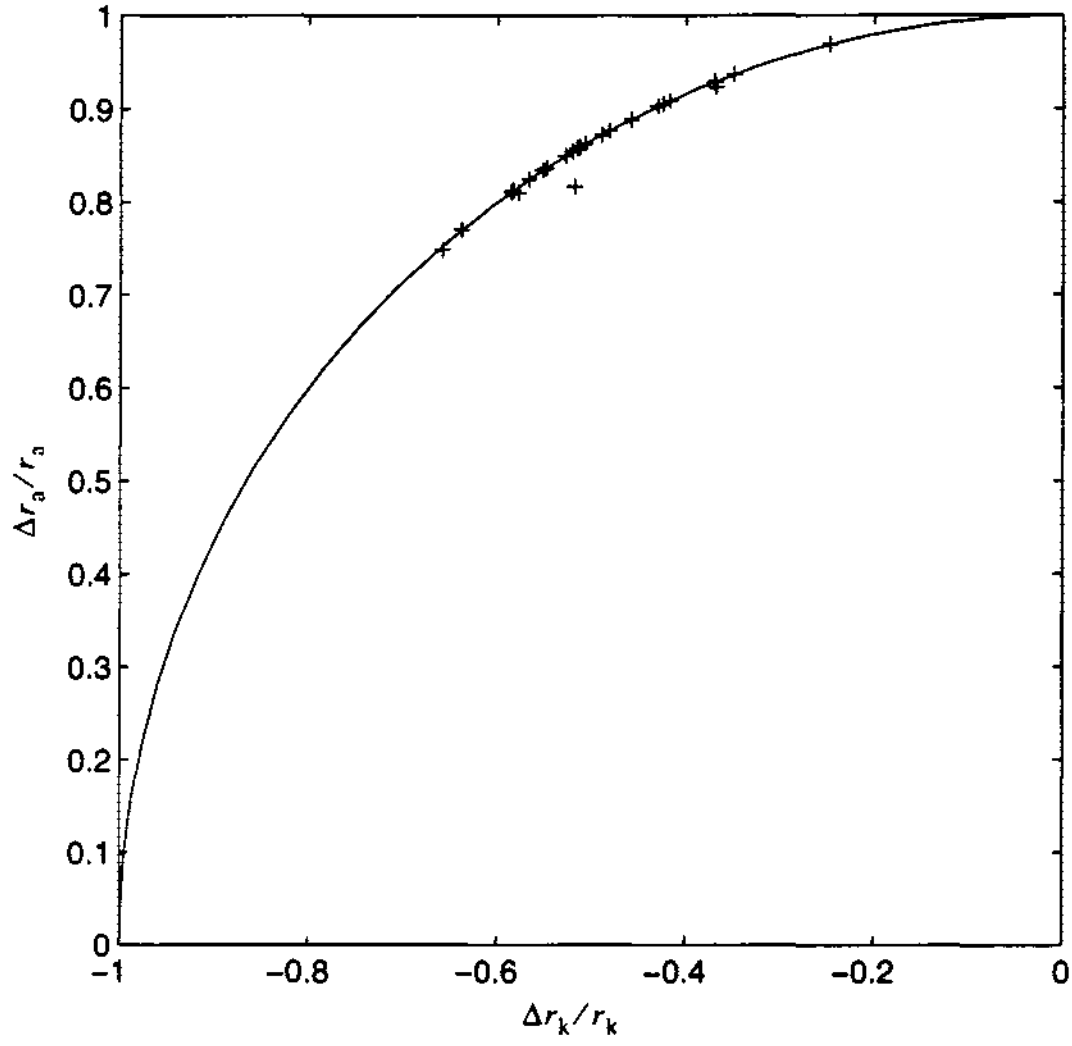

Figure 9: Scatter plot of the components of the basis vectors $\underline{v}_3$ associated to the smallest singular value in the directions $\Delta r_k / r_k$ and $\Delta r_a / r_a$.

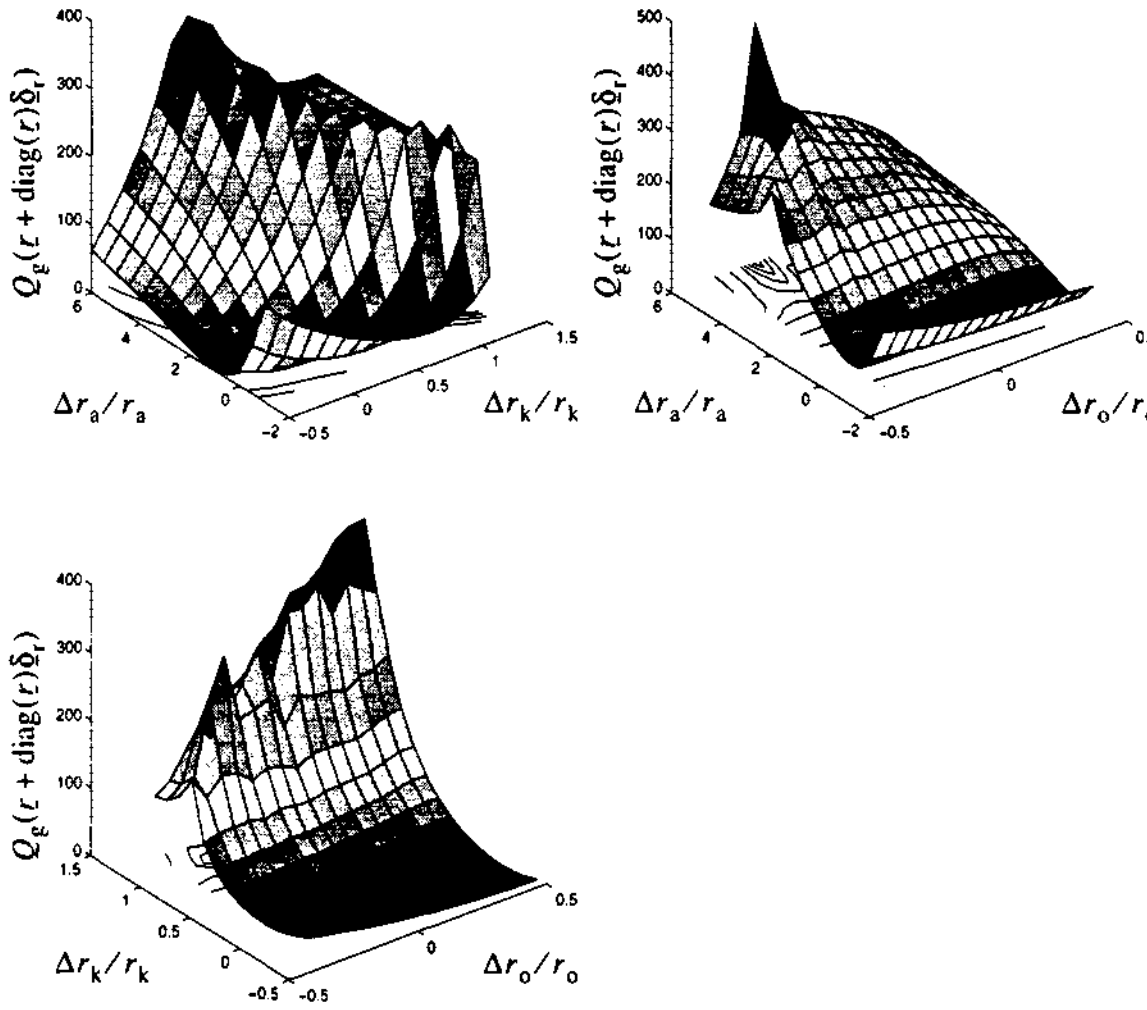

Figure 10: Mean-squared log-spectral distance $Q_g(\underline{r} + \mathrm{diag}(\underline{r})\underline{\delta}_r)$, to an LF glottal pulse with parameters $[r_o, r_a, r_k]^T = [0.614, 0.029, 0.400]^T$ as a function of the relative deviation in the parameters. Top left: $\Delta r_o = 0$. Top right: $\Delta r_k = 0$. Bottom left: $\Delta r_a = 0$.

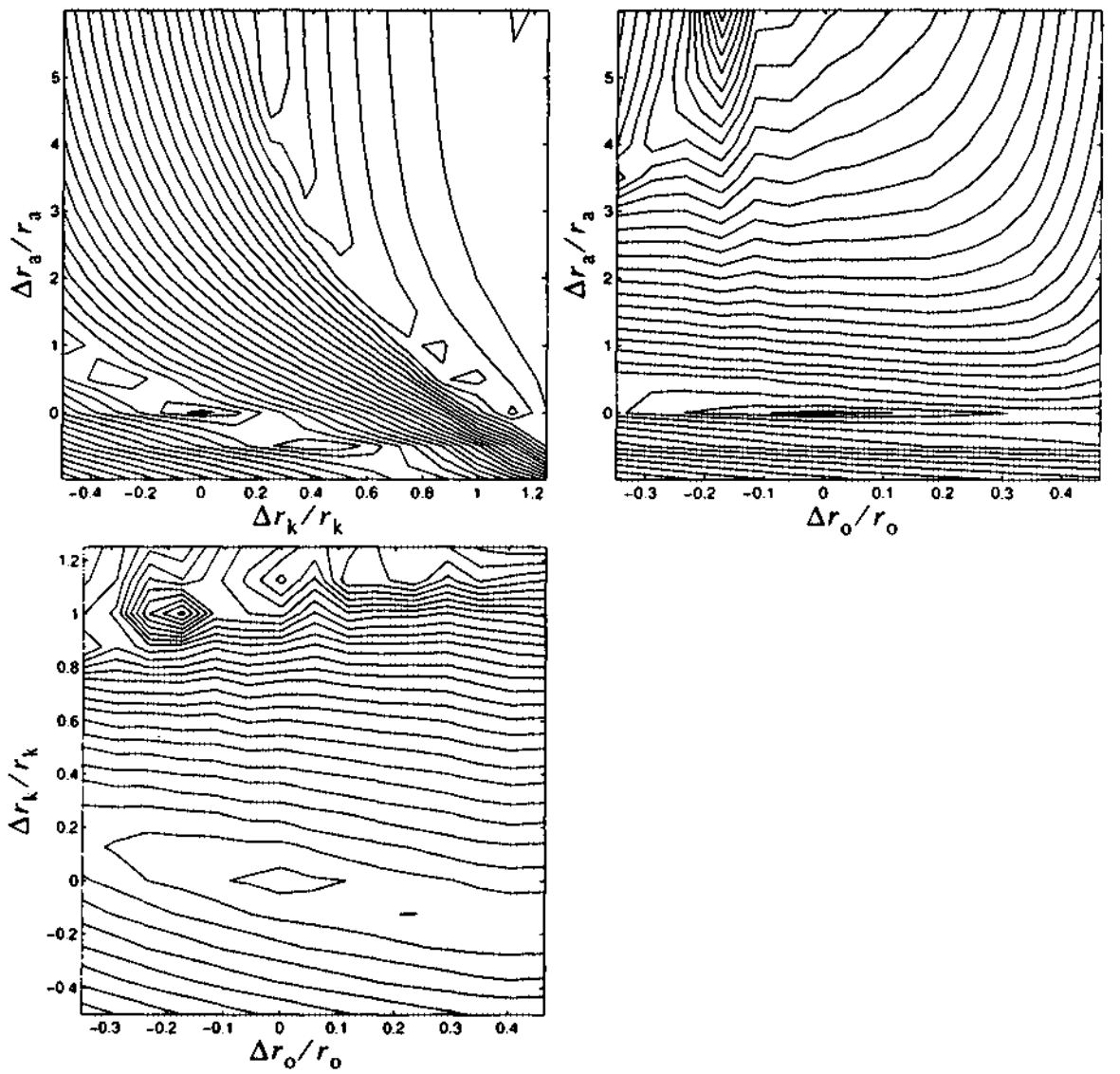

Figure 11: Contour plots of the mean-squared log-spectral distance $Q_g(\underline{r} + \mathrm{diag}(\underline{r})\underline{\delta}_r)$, *to an LF glottal pulse with parameters* $[r_o, r_a, r_k]^T = [0.614, 0.029, 0.400]^T$ *as a function of the relative deviation in the parameters. The contours are drawn at levels* $0.5 \times (1, 2, 4, 9, 16, \ldots)$. *Top left:* $\Delta r_o = 0$. *Top right:* $\Delta r_k = 0$. *Bottom left:* $\Delta r_a = 0$.

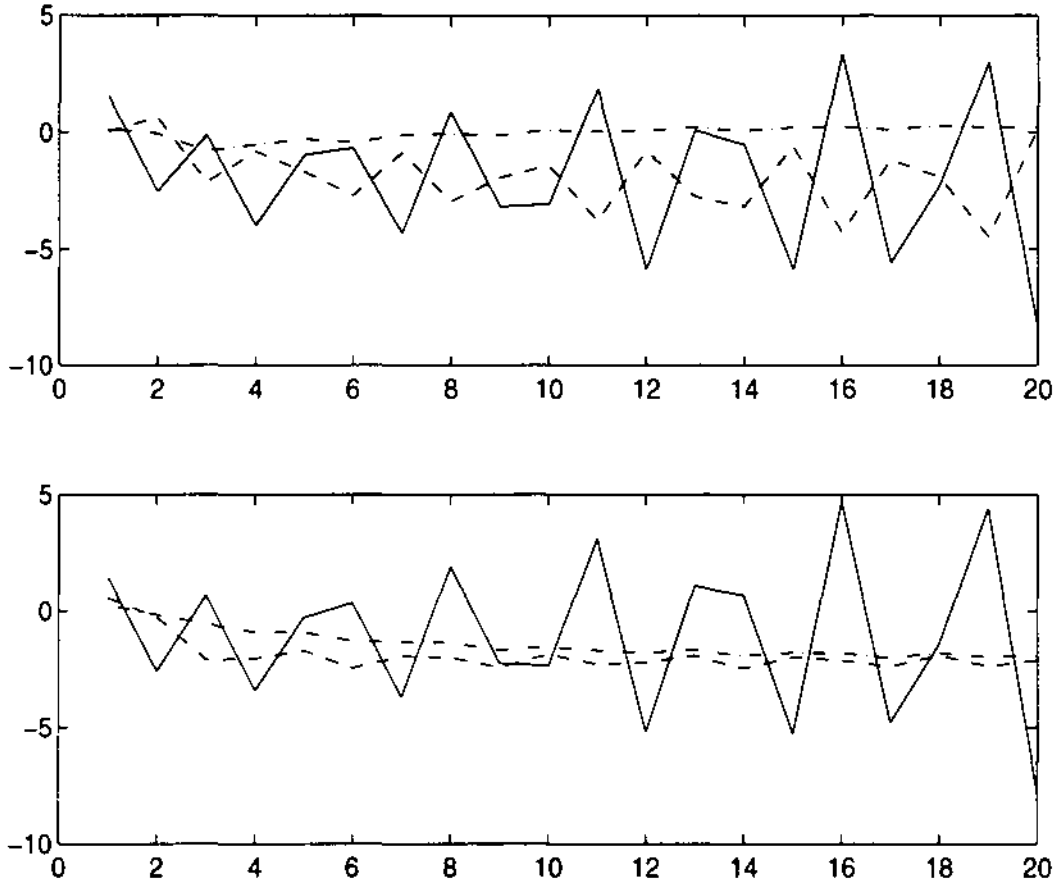

Figure 12: Top: Columns of UD, which span the column space of $\Sigma_r(\underline{r}_{opt})$, and are weighted with the corresponding singular values. The relative spectral changes due to relative modifications of r_o, r_k and r_a parameters in the directions $\underline{v}_1$, $\underline{v}_2$ and $\underline{v}_3$ are indicated by solid lines, dashed lines and dash-dotted lines, respectively. Bottom: Columns of $\Sigma_r(\underline{r}_{opt})$. The relative spectral changes due to relative perturbations of the r_o, r_k and r_a parameters presented as solid lines, dashed lines and dash-dotted lines, respectively.

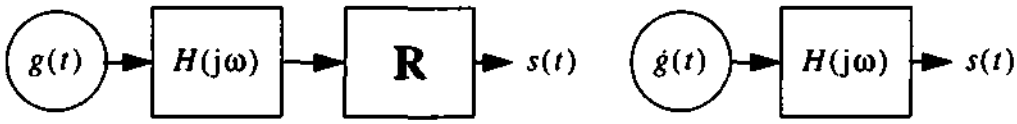

Figure 13: Left: Source-filter model with glottal source, vocal-tract filter and lip radiation. Right: Simplified source-filter model.

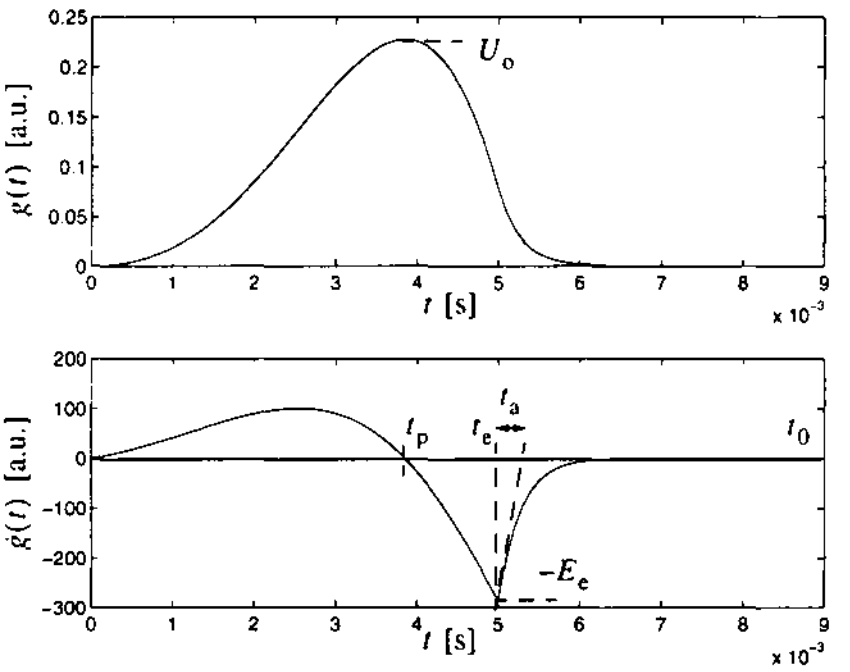

Figure 14: Glottal pulse (top) and its time derivative (bottom).