Glottal inversion with approximate vocal tract filter

Lasse Lybeck, Robert Sirviö

May 6, 2014

1 Introduction

A synthetic human vowel sound consists of a periodic signal to simulate the glottal excitation signal at the glottis and a filter to simulate the vocal tract, which the glottal signal is filtered through.[9] With a given vocal tract filter the direct problem is *given a glottal excitation signal, create the vowel sound.* The inverse problem is *given a (recorded) vowel sound, find the glottal excitation signal.* In this study we will be concentrating on the inverse problem using simulated vowel data.

The inversion from a vowel sound to the glottal signal is an important part of creating synthetic human voices and speech generators. To create a synthetic vowel both the glottal signal and the vocal tract filter are needed. However, the glottal signal cannot be directly measured, but it can be approximated with inversion of a recorded vowel. With this data models for simulating the glottal excitation signal can be created.

2 Materials and Methods

2.1 Glottal excitation signal

In this study the Rosenberg-Klatt model (RK-model) for the glottal excitation signal will be used for the generation of synthetic data and as a reference point for the obtained results. The RK-model is a simple model for the glottal signal, proposed in 1970 by Rosenberg.[8] The model is simple and easy

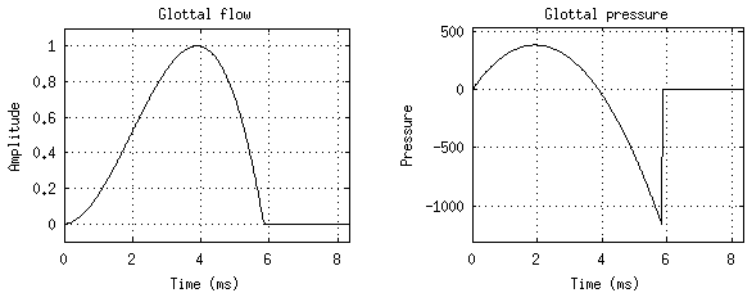

Figure 1: The airflow and pressure generated by the RK-model

to use, as it creates the signal only from two parameters, the sound frequency f and the so called Klatt-parameter Q.

The *airflow* for the glottal excitation signal created by the RK-model is defined as

$$g(t) = \begin{cases} at^2 + bt^3 & \text{if } 0 \leq t \leq QT \\ 0 & \text{if } QT < t \leq T, \end{cases} \tag{1}$$

where t is a time variable, $T = 1/f$ is the period of the pitch, $Q \in [0, 1]$ is the Klatt-parameter and a and b are variables defined in terms of $T_0 := QT$ as

$$a = \frac{27}{4T_0^2}, \quad b = -\frac{27}{4T_0^3}.$$

Here the parameter f defines the frequency of the generated signal and the Klatt-parameter Q defines the shape of the pulse.

The glottal excitation signal can be retrieved as the derivative g' of the airflow function. Here g' is the *pressure function*, and simulates the sound generated in the glottis. The pulse generated by the model can be seen in figure 1.

Another, more widely used, model for the glottal excitation signal worth mentioning is the Liljencrants-Fant model (LF-model).[3] It is regarded as more accurate than the RK-model, but it is also much more complex. It has also been shown, that the LF-model generates only marginally better approximations for the resulting vowel after the vocal tract filtering than the RK-model.[4] Due to this and the overall complexity of the LF-model we will be using the RK-model for the simulation of the glottal excitation signal in this study.

2.2 Vocal tract filter

In this study we will assume an approximate vocal tract filter to be known for the recorded vowel we want to invert. The digital filter, defined by a vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_{N_a})^T \in \mathbb{R}^{N_a}$, filters the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ as defined by the difference equation

$$\begin{cases} y_1 = x_1 \\ a_1 y_j = -\sum_{k=2}^{\min\{j-1, N_a\}} a_k y_{j-k}, \end{cases} \tag{2}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is the filtered data. We denote $\boldsymbol{y} = \varphi_{\boldsymbol{a}}(\boldsymbol{x})$.

Consider now the filter defined by $\boldsymbol{a} \in \mathbb{R}^{N_a}$ and the data $\boldsymbol{x} \in \mathbb{R}^n$ which we want to filter. Now the filter defined by (2) can be expressed by the the matrix $A \in \mathbb{R}^{n \times n}$, where

$$\begin{cases} A_{1,1} = a_1 & \\ A_{i,1} = -\sum_{k=1}^{\min\{i-1, N_a-1\}} a_{k+1} A_{i-k,1}, & 2 \leq i \leq n \\ A_{i+1,j+1} = A_{i,j}, & j \leq i \\ A_{i,j} = 0, & j > i. \end{cases} \tag{3}$$

Now $\varphi_{\boldsymbol{a}}(\boldsymbol{x}) = A\boldsymbol{x}$.

2.3 The matrix model

A vowel sound can be simulated by applying a digital filter $A \in \mathbb{R}^{n \times n}$, defined as in (3), to a sample of a glottal excitation signal $\boldsymbol{g} \in \mathbb{R}^n$ as

$$\boldsymbol{v} = A\boldsymbol{g}. \tag{4}$$

Here $\boldsymbol{v} \in \mathbb{R}^n$ is the simulated vowel.

In this study we will assume an approximation of the filter A to be known. Given the measurements $\boldsymbol{m} \in \mathbb{R}^n$ of a vowel corresponding approximately to the filter A, equation (4) can be expressed as

$$\boldsymbol{m} = A\boldsymbol{g} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ denotes the measurement noise.

2.4 The inversion method

2.4.1 Tikhonov reguralization

The classical Tikhonov regularized solution for $\boldsymbol{m} = A\boldsymbol{g} + \boldsymbol{\varepsilon}$, defined in section 2.3, is usually denoted by the vector $T_\alpha(\boldsymbol{m}) \in \mathbb{R}^n$ that minimizes

$$\|AT_\alpha(\boldsymbol{m}) - \boldsymbol{m}\|^2 + \alpha \|T_\alpha(\boldsymbol{m})\|^2 \Leftrightarrow$$

$$T_\alpha(\boldsymbol{m}) = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|A\boldsymbol{z} - \boldsymbol{m}\|^2 + \alpha \|\boldsymbol{z}\|^2 \right\},$$

where $\alpha > 0$ is called a regularization parameter. The resulting $T_\alpha(\boldsymbol{m})$ can be understood as a compromise between two conditions, namely

I. $T_\alpha(\boldsymbol{m})$ should give a small residual $AT_\alpha(\boldsymbol{m}) - \boldsymbol{m}$.

II. $\|T_\alpha(\boldsymbol{m})\|_2$ should be small.

The α parameter is used in order to tune to balance between the two conditions above.

In generalized Tikhonov regularization some prior knowledge is assumed to be known. For example, in some cases g might be known to be smooth. This information can be incorporated into the regularization by choosing

$$T_\alpha(\boldsymbol{m}) = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|A\boldsymbol{z} - m\|^2 + \alpha \|L\boldsymbol{z}\|^2 \right\}, \tag{6}$$

where L is a discretized differential operator. As shown in [5], the regularized solution satisfies

$$\left(A^TA + \alpha L^TL\right) T_\alpha(\boldsymbol{m}) = A^T\boldsymbol{m}, \tag{7}$$

which can be used to calculate the solution numerically.

In our model proposed in section 2.1 we know the airflow of the excitation signal to be smooth in the interval $[0, QT]$ and to be zero in the interval $]QT, T]$. This can be incorporated in our model by customizing the discrete differential operator matrix, described in more detail in section 2.5.

2.4.2 The conjugate gradient method

The conjugate gradient method is an iterative method for the quadratic optimization problem

$$\text{minimize } \frac{1}{2}\boldsymbol{x}^TQ\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}, \tag{8}$$

where $Q \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. We will now briefly explain the algorithm and its use to our particular problem. For a more detailed explanation, see [5].

Let $\boldsymbol{b} \in \mathbb{R}^n$ fixed, $Q \in \mathbb{R}^{n\times n}$ a symmetric positive definite matrix, $\boldsymbol{x_0} \in \mathbb{R}^n$ the initial guess and define $\boldsymbol{d_0} = -\boldsymbol{g_0} = \boldsymbol{b} - Q\boldsymbol{x_0}$. Now for $k \geq 0$ let

$$\begin{aligned} \alpha_k &= -\frac{\boldsymbol{g_k}^T \boldsymbol{d_k}}{\boldsymbol{d_k}^T Q \boldsymbol{d_k}} \\ \boldsymbol{x_{k+1}} &= \boldsymbol{x_k} + \alpha_k \boldsymbol{d_k} \\ \boldsymbol{g_{k+1}} &= Q\boldsymbol{x_{k+1}} - \boldsymbol{b} \\ \beta_k &= \frac{\boldsymbol{g_{k+1}}^T Q \boldsymbol{d_k}}{\boldsymbol{d_k}^T Q \boldsymbol{d_k}} \\ \boldsymbol{d_{k+1}} &= -\boldsymbol{g_{k+1}} + \beta_k \boldsymbol{d_k}. \end{aligned} \tag{9}$$

Now $\boldsymbol{x_k}$ converges toward the solution of (8). We now want to apply the conjugate gradient algorithm in the case of the optimization problem defined in (7) for the Tikhonov regularization.

Let $A \in \mathbb{R}^{n\times n}$, $L \in R^{n\times n}$ invertible and $\alpha > 0$. Now the square matrix $B := A^TA + \alpha L^TL$ is invertible. If we denote $\boldsymbol{f} := T_\alpha(\boldsymbol{m})$ in (7), the problem becomes to minimize the expression

$$\left\|B\boldsymbol{f} - A^T\boldsymbol{m}\right\|^2. \tag{10}$$

We see that

$$\begin{aligned} \left\|B\boldsymbol{f} - A^T\boldsymbol{m}\right\|^2 &= \langle B\boldsymbol{f}, B\boldsymbol{f}\rangle - 2\left\langle B\boldsymbol{f}, A^T\boldsymbol{m}\right\rangle + \left\langle A^T\boldsymbol{m}, A^T\boldsymbol{m}\right\rangle \\ &= \boldsymbol{f}^T B^T B\boldsymbol{f} - 2\boldsymbol{m}^T AB\boldsymbol{f} + \left\|A^T\boldsymbol{m}\right\|^2. \end{aligned} \tag{11}$$

Further, we notice that B^TB is a positive definite symmetric matrix, since

$$\boldsymbol{v}^T(B^TB)\boldsymbol{v} = (B\boldsymbol{v})^TB\boldsymbol{v} = \|B\boldsymbol{v}\|^2 > 0$$

for any $\boldsymbol{v} \in R^n$, $\boldsymbol{v} \neq 0$, due to the fact that B is invertible. Now we define

$$Q := 2B^TB \quad \text{and} \quad \boldsymbol{b}^T := 2\boldsymbol{m}^TAB.$$

As can be seen from (11), minimizing (10) is equivalent to minimizing the expression

$$\frac{1}{2}\boldsymbol{f}^TQ\boldsymbol{f} - \boldsymbol{b}^T\boldsymbol{f}, \tag{12}$$

and thus we can use the conjugate gradient method for the optimization.

2.4.3 Morozov's discrepancy principle

The problem of finding the optimal regularization parameter is, in general, considered to be unsolved. There are, however, methods that attempt to find an optimal choice of the regularization parameter, including the Morozov discrepancy principle, which is based on the noise level in the data.

Assume that we know the size of the noise in our model defined by (5) to be $\delta > 0$. Now $T_\alpha(\boldsymbol{m})$ is an acceptable reconstruction if

$$\|AT_\alpha(\boldsymbol{m}) - \boldsymbol{m}\| \leq \delta \quad \left(A \in R^{k\times n}\right). \tag{13}$$

The idea of the Morozov's discrepancy principle is to choose $\alpha > 0$ such that

$$\|AT_\alpha(\boldsymbol{m}) - \boldsymbol{m}\| = \delta \tag{14}$$

It can be proven (see [5]) that the α that satisfies the above expression is attained by solving

$$\sum_{j=1}^{\min\{k,n\}} \left(\frac{\alpha}{d_j^2 + \alpha}\right)^2 \left(m_j'\right)^2 + \sum_{j=\min\{k,n\}+1}^{k} \left(m_j'\right)^2 - \delta^2 = 0, \tag{15}$$

where d_j are the singular values of A and $\boldsymbol{m'} = U^T\boldsymbol{m}$ where U is an orthogonal matrix acquired from the singular value decomposition $A = UDV^T$.

As previously mentioned, we will use Morozov's discrepancy principle to choose the optimal regularization parameter.

2.5 The basis and materials

In this work we will use synthetic data as our basis. We first create a synthetic glottal excitation signal using the airflow function defined in (1). To this signal we apply a previously calculated vocal tract filter, as described in section 2.2 to create a simulated vowel sound. Finally we add some normally distributed noise to the data to simulate measurement noise. Different data is created by varying the sound frequency, the value of the Klatt-parameter Q and the noise-level.

The inversion of the vowel sound is done using another filter similar to the one used in creating the synthetic data. For example we might have created the data with a filter for a male vowel /a/, and use a filter for a female vowel /a/ for the inversion. This way we avoid inverse crime. The idea is that we

can assume two different filters for the same vowel to be approximately the same. That is, given two different filters A_1 and A_2 for the same vowel and a glottal excitation signal $\boldsymbol{g}$, we assume that $A_1\boldsymbol{g} \approx A_2\boldsymbol{g}$. This is based on the fact that a vowel is defined by its two or three first *formant frequencies*, which can be assumed to be about the same for two different vowel sounds.[7]

We then attempt to solve the inverse problem by using the generalized Tikhonov regularization with a customized penalty matrix. We will assume the Klatt-parameter Q of the glottal excitation signal to be known (at least approximately), and as explained in section 2.1 we know the pressure function to be zero in the interval $]QT, T]$. This will be incorporated in the model by assigning large values to the diagonal entries $l_i \in L$ for the values of i that correspond to the previously mentioned interval, namely $i \in \{j \in \mathbb{N} : Qn < j \leq n\}$ where n is the length of our data. We also know the pressure function to be smooth in the interval $[0, Q]$. This can be incorporated by adding differential operator properties to the penalty matrix.

We can thus assign the differential operator properties

$$\begin{cases} l_{i,i} & = 1 \\ l_{i,i+1} & = -1 \end{cases} \text{ when } i \in \{j \in \mathbb{N} : 1 \leq j \leq Qn\},$$

and further, we can for example assign the larger values described above as

$$l_{i,i} = 10, \quad \text{when } i \in \{j \in \mathbb{N} : Qn < j \leq n\}$$

resulting in the penalty matrix

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 & & & \cdots & & 0 \\ 0 & 1 & -1 & 0 & & & \cdots & & 0 \\ \vdots & & \ddots & \ddots & & & & & \vdots \\ 0 & \cdots & 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 10 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 10 & 0 & \cdots & 0 \\ \vdots & & & & & & \ddots & & \vdots \\ 0 & & & \cdots & & & 0 & 10 & 0 \\ 0 & & & \cdots & & & 0 & 0 & 10 \end{pmatrix} \tag{16}$$

for a single period of the excitation signal. This procedure must of course be repeated as many times as we have periods in our measurement data.

2.5.1 Noise estimation

As previously mentioned, we will use Morozov's discrepancy principle for optimization of the regularization parameter α. We assume the noise to originate from a scaled standard multivariate normal distribution, i.e. if

$$\boldsymbol{\varepsilon}' = (Z_1, \ldots, Z_n), \quad Z_i \sim N(0,1) \perp\!\!\!\perp \text{ for all } i \in \{1, \ldots, n\} \tag{17}$$

then our assumed noise vector is a sample from $\boldsymbol{\varepsilon}$, defined as

$$\boldsymbol{\varepsilon} := c\boldsymbol{\varepsilon}' = (cZ_1, \ldots, cZ_n), \quad c > 0. \tag{18}$$

Note that

$$cZ_i \sim N\left(0, c^2\right) \text{ for all } i \in \{1, \ldots, n\}. \tag{19}$$

The scaling is done to make the error relative to the data, which again is preferable in order to generate consistent data. We want to estimate the magnitude of the noise. Let $X \sim \chi_n^2$ and $Y \sim \chi_n$ where χ_n^2 denotes the chi-squared distribution and χ_n the chi distribution with n degrees of freedom respectively. We now see ([1], [2]) that

$$\begin{aligned} E\left(\|\boldsymbol{\varepsilon}\|\right) &= E\left(\|c\boldsymbol{\varepsilon}'\|\right) = E\left(c\,\|\boldsymbol{\varepsilon}'\|\right) = cE\left(\|\boldsymbol{\varepsilon}'\|\right) \\ &= cE\left(\sqrt{Z_1^2 + \ldots + Z_n^2}\right) = cE\left(\sqrt{X}\right) \\ &= cE\left(Y\right) = c\sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \end{aligned}$$

so the natural choice of δ is

$$\delta = c\sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \tag{20}$$

This is numerically impossible to compute in a straightforward manner due to fast growth of the gamma function, so in practice we will use

$$\delta = c\sqrt{2}\exp\left[\log\Gamma\left(\frac{n+1}{2}\right) - \log\Gamma\left(\frac{n}{2}\right)\right] \tag{21}$$

where the $\log\Gamma$ function is, in our case, more numerically stable. Note that expressions (20) and (21) are equivalent.

In our case we know our approximated vocal tract filter to be at least to some extent erroneous. Therefore the measurement noise must be estimated somewhat differently to take the errors attained from the differences between the filters into account.

Let $\boldsymbol{m} \in \mathbb{R}^n$ be the given measurement data. As we know the sound frequency and the value of the Klatt-parameter we can generate a glottal excitation signal, and filter it with the approximated filter used in the inversion to create a similar vowel sound as our measurements. Let us denote it with $\boldsymbol{v} \in \mathbb{R}^n$. Now we can calculate an upper bound for the measurement noise as

$$\delta_{\max} = \|\boldsymbol{v} - \boldsymbol{m}\| . \tag{22}$$

However, we know our filter to be only an approximation of the filter used in generating the measurement data, not the exact same filter. Therefore it is reasonable to assume that the approximations done in the selection of the filter give rise to some differences in the vowel sound that cannot be regarded as measurement errors, but rather systematic errors attained by the approximations. Therefore we will use the estimation $\delta = \delta_{\max}/2$ for the noise level of the measurement data. This estimation of the measurement noise is not at all a trivial task, and should require more careful examination. This is however beyond the scope of this study.

When we have acquired an estimation for the measurement noise the solution of equation (15) is approximated with Newton's method [6].

3 Results

In this section we will present some of the results of our study. For obvious reasons we will not include sound data from our results in this paper. Note that in all the following graphs of the target glottal pressure and the reconstructed glottal pressure are represented by the colors green and blue, respectively. We will also denote the Klatt-parameter with Q. The relative errors of the reconstructions are calculated with the formula

$$\delta_{rel} = \frac{\|\boldsymbol{g} - T_\alpha(\boldsymbol{m})\|_2}{\|\boldsymbol{g}\|_2} \cdot 100\%, \tag{23}$$

where $\boldsymbol{g} \in \mathbb{R}^k$ is the original glottal excitation signal and $T_\alpha(\boldsymbol{m})$ is the reconstruction calculated from the noisy data m with the regularization parameter α.

The inversion was done, in addition to the Tikhonov regularization strategy, also with the naïve inversion method in order to illustrate the ill-posedness of the inverse problem. A comparison of the methods is presented below in figures 2 and 3.

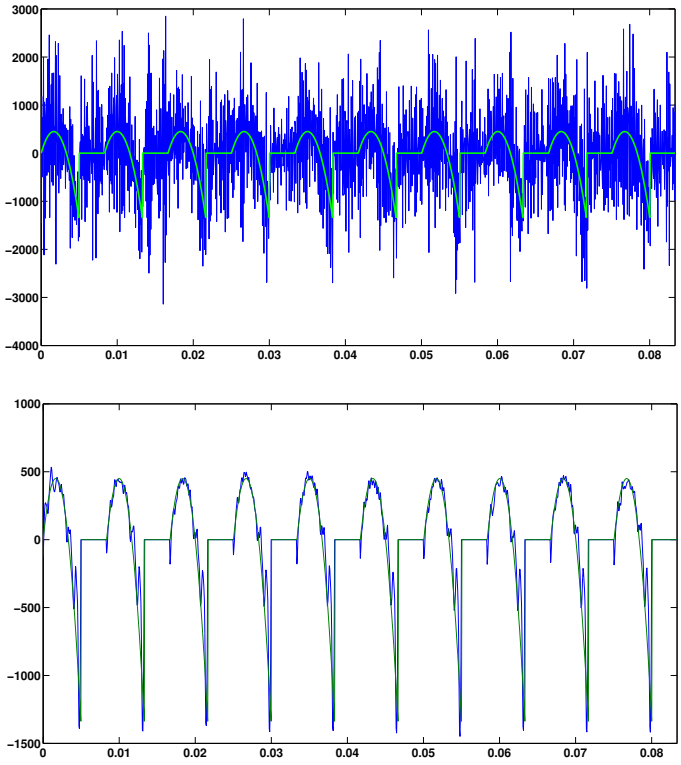

Figure 2: Comparison between the naïve inversion method and the Tikhonov regularization strategy with inverse crime. Sound frequency = 90 Hz, Q = 0.7, relative error of the reconstruction $\approx 43.9\%$, $\alpha \approx 63.7$.

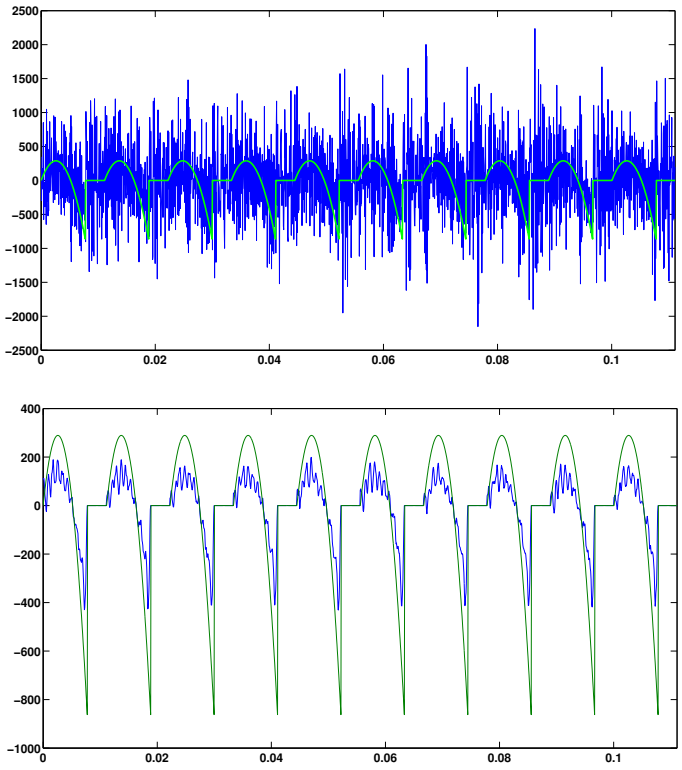

Figure 3: Comparison between the naïve inversion method and the Tikhonov regularization strategy without inverse crime. Sound frequency = 90 Hz, Q = 0.7, relative error of the reconstruction $\approx 63.1\%$, $\alpha \approx 63.2$.

In order to evaluate the accuracy and correctness (in the sense of the relative error of the glottal impulse) of the α-value chosen according to Morozov's discrepancy principle the relative errors were calculated with different values of the parameters α. This was then compared with corresponding reconstruction with an α-value chosen according to Morozov's discrepancy principle. The results can be seen in figure 4.

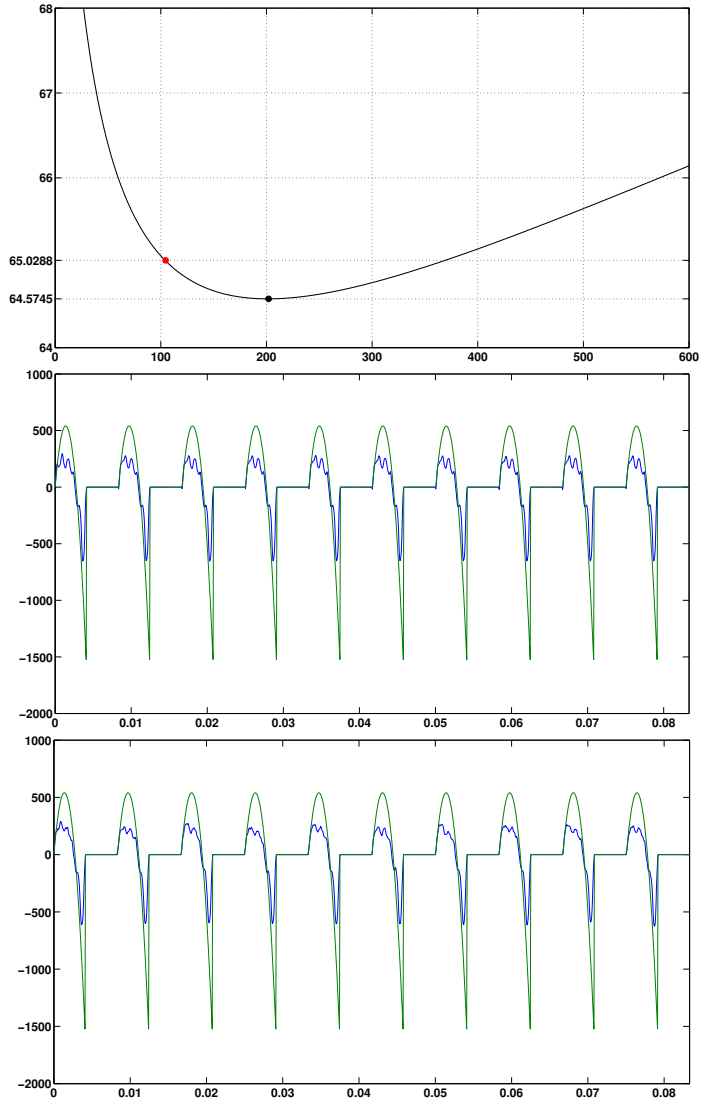

Figure 4: The uppermost picture plots the relative error values vs. α-values of the iterative calculation of the inversion. The red dot denotes the α-value acquired from Morozov's discrepancy principle and the black dot denotes the α-value corresponding to the least relative error in the reconstruction (from the iterative calculation). The middle picture graphs the reconstruction with the α-value acquired from Morozov's discrepancy principle and the bottom picture graphs the reconstruction α-value corresponding to the least relative error in the reconstruction.

Since the number of (a priori known) variables is quite high a large amount of data needed to be harvested in order to properly analyze the inversion method. The following figures demonstrate some of the data that was collected.

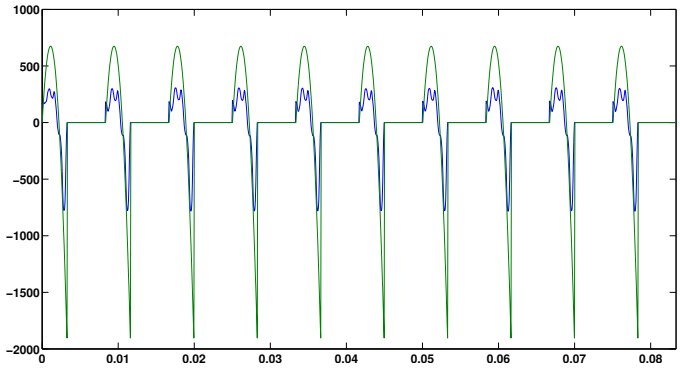

Figure 5: Sound frequency = 120 Hz, Q = 0.4, relative error of the reconstruction $\approx 69.1\%$, $\alpha \approx 123.9$.

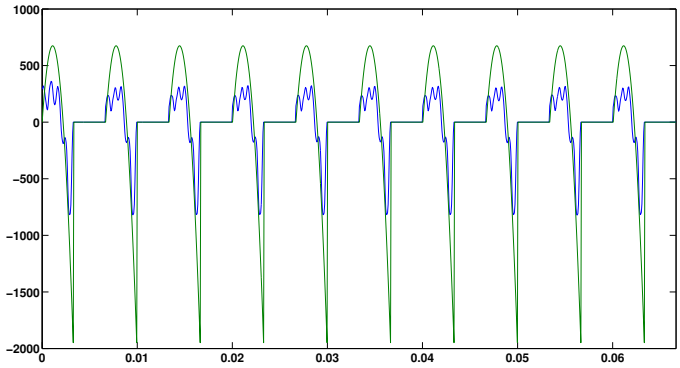

Figure 6: Sound frequency = 150 Hz, Q = 0.5, relative error of the reconstruction $\approx 70.1\%$, $\alpha \approx 86.0$.

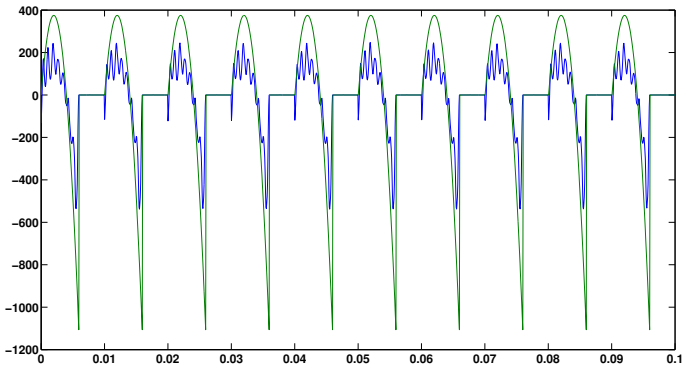

Figure 7: Sound frequency = 100 Hz, Q = 0.6, relative error of the reconstruction $\approx 64.1\%$, $\alpha \approx 66.6$.

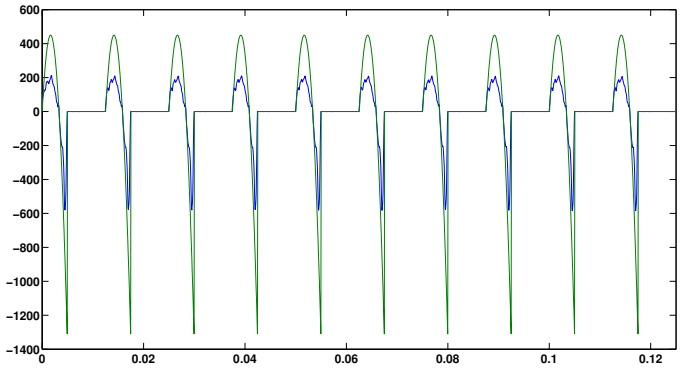

Figure 8: Sound frequency = 80 Hz, Q = 0.3, relative error of the reconstruction $\approx 64.6\%$, $\alpha \approx 84.5$.

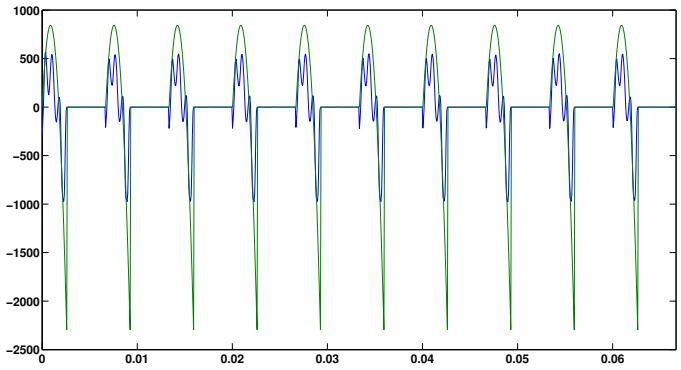

Figure 9: Sound frequency = 150 Hz, Q = 0.4, relative error of the reconstruction $\approx 74.0\%$, $\alpha \approx 48.5$.

4 Discussion

4.1 Ill-posedness

As it can be seen from figures 2 and 3 the inverse problem to determine the glottal impulse is clearly *ill-posed*. It can be seen that naïve inversion fails completely to achieve the shape of the glottal excitation signal, and thus a regularization method is required.

In this work the Tikhonov regularization strategy was chosen for solving the inverse problem. It can further be seen from figures 2 and 3 that when the same problem that failed with naïve inversion is solved with Tikhonov regularization the results are significantly better.

It must be noted that the situation displayed in figure 2 is done with inverse crime, which means that the same filter was used both in creating the data and in solving the inverse problem. This is the reason why the error of the reconstruction is so low; such low values could not be achieved with real data. The situation in figure 3 is a much better example of how reconstructions with real data could look like.

4.2 Tikhonov regularization

Apart from the naïve inversions displayed in figures 2 and 3 discussed above, all the reconstructions in figures 2 to 9 are done with the Tikhonov regu-

larization strategy. For all reconstructions the relative error was well below 75%, which can be regarded as a good result.

As seen from the figures, the shape of the reconstruction is in all the cases clearly the same as in the original data, although the reconstruction is somewhat flattened. This is due to the high values of the regularization parameter used to compensate for the noise. The problem here is that although the *sound* generated by the reconstruction is close to the original data the relative error of the reconstruction can become quite high; as the reconstruction is flattened it only affects the *amplitude* of the generated sound as long as the *shape* is unaltered. This is further discussed in section 4.4.

Although the Tikhonov regularization strategy was found to work well for the glottal inversion problem, it relies heavily on many pieces of a-priori information. The information needed for the discussed method to work include the sound frequency and the Klatt-parameter of the glottal excitation signal used to generate the vowel sound. The sound frequency is an easy task to find from the measurement data, but the Klatt-parameter can be a harder task to find precisely.

If the Klatt-parameter is approximated too low (a too large part of a period of the reconstruction is suppressed by the penalty matrix, see section 2.5) the peaks of the excitation signal would also be suppressed, thus ruining the reconstruction. If the value of the Klatt-parameter is approximated too high (smaller part of a period suppressed) a part of the reconstruction that should have the value 0 will become oscillatory (see example in figure 10). The latter would probably yield a better sound, as the characteristic shape of the glottal excitation signal would be established, although with some added noise. From this reconstruction the parameter could however be approximated again, as it will probably show in the reconstruction if the parameter value is estimated wrong. However, this means that the inversion process cannot be automated, as the Klatt-parameter must be found by hand.

Another way of tackling the problem of not knowing the Klatt-parameter exactly would be to alter the penalty matrix, for example by "smoothening" the transition from the discrete first derivative penalty to the suppressing part of the matrix. In this way the exact value would not be as important, but a good approximation would suffice. This is however beyond the scope of this study.

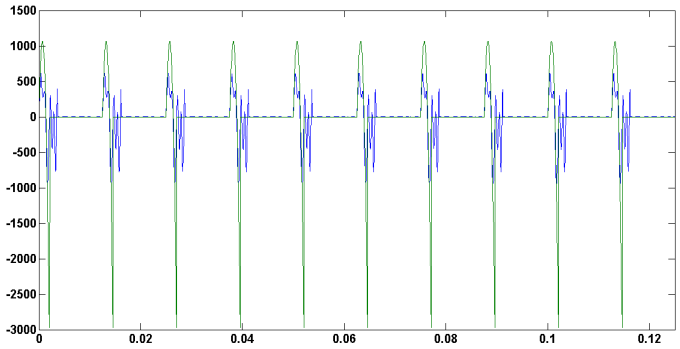

Figure 10: Example of a reconstruction done with a too big Klatt-parameter value.

4.3 Morozov's discrepancy principle

The regularization parameters acquired with Morozov's discrepancy principle yielded good results. As can be seen from figure 4, the reconstruction with a regularization parameter acquired with Morozov's discrepancy principle gave almost as good a reconstruction as with the best possible value of the parameter (in the sense of the values in the iteration) when using the relative error of the reconstructions as a reference point. The difference in the audio data was indistinguishable to the human ear.

Considering that the selection of the regularization parameter with Morozov's discrepancy principle works completely without the need of human interaction, the method selection can be regarded as a success for this particular problem. With an optimization of the noise estimation process (see 2.5.1) potentially more accurate results can be acquired.

4.4 Reconstruction quality

As mentioned earlier in section 4.2, the relative error of the reconstruction compared with the original data might not be the perfect quantity to measure the quality of the reconstructions. As the values of the reconstruction signal are flattened, only the amplitude of sound generated by the reconstruction is affected, as long as the *shape* of the signal is the same as for the original data. Therefore another way of measuring the error might be used.

The shape of the reconstruction signal could be compared by scaling the reconstruction and finding the smallest possible value for the relative error. This way only the shape, not the amplitude, of the reconstruction would matter. The shape error could thus be calculated as

$$\delta_{\text{shape}} = \min_{c \in \mathbb{R}} \left\{ \frac{\|\boldsymbol{g} - cT_\alpha(\boldsymbol{m})\|_2}{\|\boldsymbol{g}\|_2} \right\} \cdot 100\%, \tag{24}$$

where $\boldsymbol{g} \in \mathbb{R}^k$ is the original glottal excitation signal and $T_\alpha(\boldsymbol{m})$ is the reconstruction calculated from the noisy data m with the regularization parameter α.

However, it turned out that this way of comparing the reconstruction with the original data yielded similar results to the relative error. Therefore only the relative error was reported for the reconstructions.

4.5 Further work

In this work the inversion was done solely with simulated data. The same method might however be succesfully applied even to real recorded data. The vocal tract filter used in the inversion with the synthetic data is selected to simulate only an approximation of the real vocal tract filter, as the inversion uses another filter for the same vowel as what the data was created with. The situation with real data is not far from this starting point.

With the Tikhonov regularization strategy the reconstruction easily becomes oscillatory. As it was seen from the results in this work, to even out the oscillations the regularization parameter had to have a relatively high value. This in turn resulted in suppressed reconstructions. Also, the penalty matrix had to be heavily customized with Tikhonov regularization. Therefore, the inversion might work with the *total variation* regularization strategy.[5] This could also be a good method of choise because it is planned to work well for piecewise constant functions, and a big part of the glottal excitation signal has the constant value zero.

References

[1] Abramowitz, M & Stegun, I.A. (1965) *Handbook of mathematical functions with formulas, graphs and mathenatical tables* New York, NY: Dover, p. 940

[2] Evans, M., Hastings, N. & Peacock, B. (2000) *Statistical distributions*. New York: Wiley, p. 57

[3] Fant, G., Liljencrants, J., Lin, Q., (1985). *A four-parameter model of glottal flow*. STL-QPSR 26 (4), p. 1-13

[4] Fujisaki, H., Ljungqvist, M., 1986. Proposal and evaluation of models for the glottal source waveform. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 11. p. 1605–1608.

[5] Mueller, Jennifer L. & Siltanen Samuli, (2012). *Linear and Nonlinear Inverse Problems with Practical Applications*. SIAM, 1st edition.

[6] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (2007) *Numerical Recipes: The Art of Scientific Computing*. New York: Cambridge University Press, 3rd edition, p. 362

[7] Rabiner, L. R., Schafer, R. W., (1987). *Digital processing of speech signals*. Englewood Cliffs: Prentice-Hall, p. 38-107.

[8] Rosenberg, A., (1971). *Effect of glottal pulse shape on the quality of natural vowels*. Journal of the Acoustical Society of America 49 (2B), p. 583–590.

[9] Touda, K., (2007) *Study on numerical method for voice generation problem*. PhD thesis. The University of Electro-Communications.