Glottal inversion with an approximate vocal tract filter

Lasse Lybeck, Robert Sirviö

April 1, 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.[4] 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, starting with both a simulated vowel and a real recording.

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.[6] The model is simple and easy

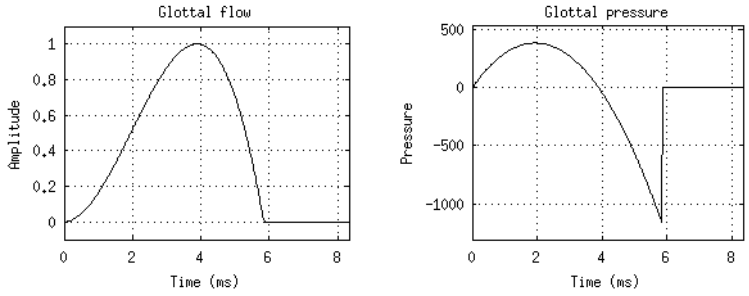

Figure 1: The airflow and pressure generated by the RK-model

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

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

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).[1] 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.[2] 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 $a \in \mathbb{R}^{N_a}$, filters the data $x \in \mathbb{R}^n$ 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 $y \in \mathbb{R}^n$ is the filtered data. We denote $y = \varphi_a(x)$.

Consider now the filter defined by $a \in \mathbb{R}^{N_a}$ and the data $x \in \mathbb{R}_n$ 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 $\varphi_a(x) = Ax$.

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 $g \in \mathbb{R}^n$ as

$$v = Ag. \tag{4}$$

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

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

$$m = Ag + \varepsilon, \tag{5}$$

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

2.4 The inversion method

2.4.1 Tikhonov reguralization

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

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

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

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

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

II. $\|T_\alpha(m)\|_2$ 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

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

where L is a discretized differential operator. As shown in [3], the regularized solution satisfies

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

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

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

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 [3].

Let $b \in \mathbb{R}^n$ fixed, $Q \in \mathbb{R}^{n\times n}$ a symmetric positive definite matrix, $x_0 \in \mathbb{R}^n$ the initial guess and define $d_0 = -g_0 = b - Qx_0$. Now for $k \geq 0$ let

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

Now 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 $B := A^T A + \alpha L^T L$ is invertible. If we denote $f := T_\alpha(m)$ in (7), the problem becomes to minimize the expression

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

We see that

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

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

$$v^T (B^T B) v = (Bv)^T Bv = \|Bv\|^2 > 0$$

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

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

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

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

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 Mozorov 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 $T_\alpha(m)$ is an acceptable reconstruction if

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

If we assume that

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n, \quad \varepsilon_k \sim N(0,1) \text{ for all } k \in \{1, \ldots, n\}, \tag{14}$$

we can choose $\delta = \sqrt{n}$ since $E(\|\varepsilon\|) = \sqrt{n}$.

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

$$\|AT_\alpha(m) - m\| = \delta \tag{15}$$

It can be proven (see [3]) that the α that satisfies the above expression is attained by solving

$$\sum_{j=1}^{\min\{k,n\}} \left(\frac{\alpha}{d_j^2 + \alpha}\right)^2 \left(m_j'\right)^2 + \sum_{j=\min\{k,n\}+1}^{k} \left(m_j'\right)^2 - \delta^2 = 0, \tag{16}$$

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

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 g, we assume that $A_1 g \approx A_2 g$. 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.[5]

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

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 & & & \cdots & & 0 \\ 0 & 1 & -1 & 0 & & & \cdots & & 0 \\ \vdots & & \ddots & \ddots & & & & & \vdots \\ 0 & \cdots & 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 10 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 10 & 0 & \cdots & 0 \\ \vdots & & & & & & \ddots & & \vdots \\ 0 & & & \cdots & & & 0 & 10 & 0 \\ 0 & & & \cdots & & & 0 & 0 & 10 \end{pmatrix} \tag{17}$$

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.

3 Results

4 Discussion

References

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

[2] 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.

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

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

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

[6] 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.