Glottal inversion with an approximate vocal tract filter

Lasse Lybeck, Robert Sirviö

March 28, 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.[citation needed] 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.[4] As the model creates the signal

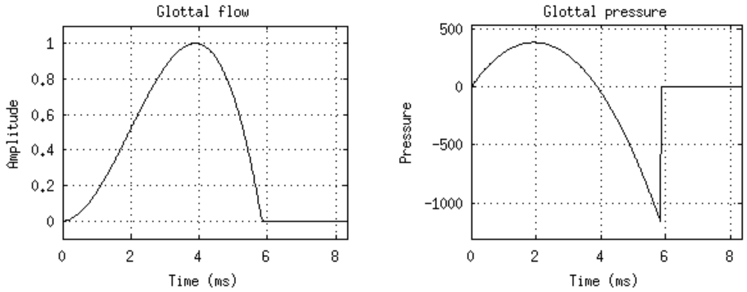

Figure 1: The airflow and pressure generated by the RK-model

only from two parameters, the sound frequency f and the so called Klatt-parameter Q, it is easy to use and is therefore selected as the model to be used in this study.

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 not be using it 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 $x \in \mathbb{R}^{N_x}$ as defined by the difference equation

$$y_1 = x_1 \tag{2}$$

$$a_1 y_n = - \sum_{k=2}^{\min\{n-1,N_a\}} a_k y_{n-k} \tag{3}$$

jatkuu...

2.3 The matrix model

A vowel $v \in \mathbb{R}^n$ can be simulated by applying a digital filter $A \in \mathbb{R}^{n \times n}$ to a sample of a glottal excitation signal $g \in \mathbb{R}^n$ as

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

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

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

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

jatkuu...

2.4 The inversion method

2.4.1 Tikhonov reguralization

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

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

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

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. In some cases f 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. In our model proposed in [citation needed] we know the glottal impulse to be zero in an interval [mera kama hit när modellen är skriven]

2.4.2 The conjugate gradient method

3 Results

4 Discussion

References

[1] Fant, G., Liljencrants, J., Lin, Q., (1985). *A four-parameter model of glottal flow.* STL-QPSR 26 (4), 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. pp. 1605–1608.

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

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