Utilizing Markov Chain Monte Carlo (MCMC) Method for Improved Glottal Inverse Filtering

Harri Auvinen[1]*, Tuomo Raitio*[2]*, Samuli Siltanen*[1]*, Paavo Alku*[2]

[1]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

harri.auvinen@helsinki.fi, tuomo.raitio@aalto.fi

Abstract

This paper presents a new glottal inverse filtering (GIF) method that utilizes Markov chain Monte Carlo (MCMC) algorithm. First, initial estimates of the vocal tract and glottal flow are evaluated by an existing GIF method, the iterative adaptive inverse filtering (IAIF). Simultaneously, the initially estimated glottal flow is synthesized using the Klatt model and filtered with the estimated vocal tract filter. In the MCMC estimation process, the first few poles of the initial vocal tract model and the Klatt parameter are refined in order to minimize the error between the original and the synthetic signals. MCMC converges to the optimal result, and the final estimate of the vocal tract is found by averaging the parameter values of the Markov chain. Experiments show that the MCMC-based GIF method gives more accurate results compared to the original IAIF method.

Index Terms: glottal inverse filtering, Markov chain Monte Carlo, MCMC

1. Introduction

Glottal inverse filtering (GIF) [1] is a technique for estimating the glottal volume velocity waveform (i.e. the glottal source) from a speech signal. Glottal inverse filtering involves first estimating the vocal tract filter, which is then used to cancel the effect of the vocal tract by filtering the signal by the inverse model of the tract. The resulting signal is the time-domain waveform of the estimated glottal source. Glottal inverse filtering is an inverse problem that is difficult to solve accurately. Moreover, it is not possible to observe non-invasively the glottal source, and thus it is difficult to validate the accuracy of GIF.

During the past decades, many GIF methods have been developed (for a review on GIF methods, see e.g. [2]). This paper addresses GIF capable of estimating automatically the glottal source from a speech pressure signal recorded outside the lips. For sustained non-nasalized vowels produced with normal phonation of low pitch, the existing GIF methods are, in general, capable of estimating the glottal source with tolerable accuracy. However, the performance of current GIF methods typically deteriorates in the analysis of high-pitch voices. This accuracy degradation is explained by the prominent harmonic structure in high-pitch voices which gives rise to biasing of the formant estimates [3]. Biased formant estimates, in turn, result in insufficient cancellation of the true resonances of the vocal tract which distorts the estimated voice source.

In this paper, computational inversion methods are utilized in order to improve the accuracy of GIF. More specifically, we take advantage of Markov chain Monte Carlo (MCMC) methods in order to determine new vocal tract models for GIF. The method is based on first finding an initial estimate of the vocal tract filter, and then refining the GIF model parameters within the MCMC method in order to get optimal inverse filtering result. The goodness of the results is evaluated by comparing the original speech signal to a synthetic one created by filtering the Klatt model [4, 5] based excitation signal with the estimated vocal tract filter. The GIF model parameters include a few first poles of the vocal tract filter and the Klatt parameter. The final tract estimate is given by the mean of the MCMC chain, excluding a burn-in period.

2. MCMC-based glottal inverse filtering

Let us write the blind deconvolution problem in the form

$$m = p * f + \varepsilon, \tag{1}$$

where $p = p(z_1, z_2, \ldots, z_N)$ is the all-pole filter parameterized by N complex numbers, $f = f(t) = f(t; \theta_1, \ldots, \theta_M)$ is the pressure signal at the glottis parameterized by M real numbers, $m = m(t)$ is the measurement recorded using a microphone, and ε is random noise. The inverse problem is to recover the parameter vector $\vec{\theta} := [r_1, \alpha_1, r_2, \alpha_2 \ldots, r_N, \alpha_N, \theta_1, \ldots, \theta_M]^T$ from a given measurement m. Here $z_j = r_j \exp(i\alpha_j)$.

We use the Bayesian inversion approach where $\vec{\theta}, m$ and ε are modelled as random variables. The posterior distribution

$$\pi(\vec{\theta}\,|\,m) = \frac{\pi(\vec{\theta})\pi(m\,|\,\vec{\theta})}{\pi(m)} \tag{2}$$

is considered to be the complete answer to the inverse problem. Here $\pi(m)$ is a normalizing constant, $\pi(\vec{\theta})$ is the *prior model* expressing all the knowledge we have about the parameters apart from the measurement data, and $\pi(m\,|\,\vec{\theta})$ is the *likelihood model*.

The prior model should attach small or zero probability to parameter values that are unexpected in light of *a priori* information and high probability otherwise. In the case of glottal inverse filtering, we use the following *a priori* information. The poles of the vocal tract filter are located in the open unit disc. The angles are always ordered as $\alpha_1, < \alpha_2 < \cdots < \alpha_N$, and the radii r_j satisfy certain lower and upper bounds. Also, the parameters $\theta_1, \theta_2, \ldots, \theta_M$ can only vary within an *a priori* known interval. We express the above facts as probability density functions in our algorithms. In this paper we use the Klatt [4, 5] model $f(t) = at^2 - bt^3$ that contains only one parameter to control the shape of glottal excitation, the ratio a/b, therefore here $M = 1$.

Under the assumption of independent white noise ε with standard deviation $\sigma > 0$, the likelihood model takes the form

$$\pi(m \,|\, \vec{\theta}) = \pi_\varepsilon(p * f - m) \quad (3)$$
$$= \exp(-\frac{1}{2\sigma^2}\|p(\vec{\theta}) * f(t, \vec{\theta}) - m(t)\|_2^2).$$

In the inversion process, the role of the likelihood model is to keep data discrepancy low. In this work we found it important to measure data discrepancy simultaneously in the time-domain and in the frequency-domain, resulting in the following likelihood function:

$$\exp(-\alpha\|p * f - m\|_2^2 - \beta\||\text{FFT}(p * f)| - |\text{FFT}(m)|\|_2^2); \quad (4)$$

here $\alpha > 0$ and $\beta > 0$ are parameters that need to be evaluated experimentally.

To get a useful answer to our inverse problem, we need to draw a representative estimate from the posterior probability distribution (2). We study the *conditional mean estimate* defined as the integral

$$\vec{\theta}_{\text{CM}} := \int_{\mathbb{R}^{N+M}} \pi(\vec{\theta} \,|\, m)\, d\vec{\theta}. \quad (5)$$

However, the integration in (5) is over a high-dimensional space, and standard numerical integration quadratures are ineffective. We resort instead to Markov chain Monte Carlo (MCMC) methods, whose basic idea is to generate a random sequence $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \ldots, \vec{\theta}^{(K)}$ of samples with the property that

$$\vec{\theta}_{\text{CM}} \approx \frac{1}{K}\sum_{k=1}^{K} \vec{\theta}^{(k)}. \quad (6)$$

The Metropolis-Hastings method [6] is the most classical algorithm for computing the samples $\vec{\theta}^{(k)}$. The idea is to pick a random candidate $\vec{\theta}'$. The posterior probability of the candidate is compared to that of the most recent member $\vec{\theta}^{(k)}$ in the chain, and we set $\vec{\theta}^{(k+1)} = \vec{\theta}'$ if the candidate is more probable. However, even if $\pi(\vec{\theta}' \,|\, m) < \pi(\vec{\theta}^{(k)} \,|\, m)$ we might accept the candidate: draw a uniformly distributed random number $r \in [0, 1]$ and set $\vec{\theta}^{(k+1)} = \vec{\theta}'$ if $\pi(\vec{\theta}' \,|\, m)/\pi(\vec{\theta}^{(k)} \,|\, m) > r$. Otherwise set $\vec{\theta}^{(k+1)} = \vec{\theta}^{(k)}$. In this work, a modern variant of the Metropolis-Hastings algorithm called DRAM [7] is used. The implementation of the DRAM method is found in the MCMC Matlab package [8].

The aforementioned glottal inverse filtering method is hence cited as the MCMC-GIF method.

3. Experiments

This section describes the preliminary experiments performed with the MCMC-GIF method. First, a set of synthetic vowels were used where the glottal excitation signal is known. Secondly, natural vowels were used in order to test the performance of the method with natural speech. The iterative adaptive inverse filtering (IAIF) method was used for estimating the initial model of the vocal tact filter, which is described below.

3.1. Iterative adaptive inverse filtering

IAIF [9, 10] is a method that automatically decomposes speech into the vocal tract transfer function and the glottal source signal. The method is applied pitch-asynchronously over several fundamental periods. IAIF is based on the assumption that the contribution of the glottal excitation can be estimated by a low-order all-pole filter. Thus, by canceling this effect, an estimate of the vocal tract filter can be computed with an all-pole model. The method has two phases. First, a preliminary estimate of the contribution of the glottal flow is estimated by a first-order all-pole model. In the second phase, a higher-order all-pole model is computed, yielding a more accurate estimate of the glottal contribution.

Since the IAIF method is based on all-pole modeling, the harmonic structure of the excitation may introduce bias to the vocal tract estimate, especially with high-pitched voices [3]. To reduce such bias, the initial estimates given by the IAIF are refined by the MCMC-GIF method.

3.2. Speech data

The fundamental problem in evaluating GIF methods with natural speech is that the real glottal excitation cannot be observed. If synthetic speech signals are used, the glottal excitation is known, and evaluation is possible. However, the problem with synthetic speech signals is that they are based on the same principle of speech production [11] as the glottal inverse filtering methods, and thus the evaluation is not truly objective.

In this study, synthetic speech generated by physical modeling of the vocal folds and the vocal tract [12] is used in order to overcome this problem. The glottal excitation is generated by a three-mass model, and the resulting excitation signal is coupled with a physical model of the trachea and vocal tract using a wave-reflection approach [12]. Synthetic [a] vowels of normal phonation type with fundamental frequencies (F_0) from 100 Hz to 400 Hz were produced by using the physical modeling. The synthetic speech samples were sampled at 16 kHz.

Secondly, recorded speech was used in order to study the performance of the MCMC-GIF method with natural speech. The natural speech data consists of sustained [a] vowels produced by a male speaker with F_0 ranging from 110 Hz to 292 Hz. The natural speech samples were sampled at 8 kHz.

3.3. Experimental setup

The MCMC-GIF method consists of first estimating the initial model of the vocal tract filter and the glottal source signal with the IAIF method from a 25-ms pitch-asynchronous window. Simultaneously, the glottal closure instants (GCIs) of the estimated glottal excitation are estimated with a simple method based on peak picking at fundamental period intervals, after which preliminary Klatt pulses are synthesized according to the GCIs. Thus, the goodness of the parameter estimates given by the Markov chain can be evaluated by comparing the original speech signal to the synthetic speech signal created by filtering the Klatt based excitation signal with the estimated vocal tract filter. The comparison is performed both in time and frequency domains according to the likelihood model in Equation 4. The time domain error is the mean square error between the signals, and the frequency domain error is defined by the mean square error of the magnitude spectra of the signals.

The parameters to be estimated by the MCMC method are the poles of the first few formants, defined as a shift of the angle in radians from the IAIF estimate and as the norm of the poles, and the shape parameter of the Klatt pulse. In order to efficiently start the MCMC-GIF method, the vocal tract pole estimates of the IAIF method are used as starting values for the Markov chain. The shape parameter of the Klatt model is purely estimated by MCMC-GIF method. The total length of the Markov chain is set to 40 000. Thus a single run of

Table 1: *Comparison of the GIF methods with H1-H2 and NAQ quantities using synthetic speech with four different F_0.*

H1-H2	100 Hz	200 Hz	300 Hz	400 Hz
IAIF	3.9 dB	2.1 dB	11.6 dB	-2.1 dB
MCMC-GIF	4.0 dB	5.1 dB	3.5 dB	5.7 dB
MCMC-Klatt	4.0 dB	2.5 dB	4.7 dB	3.1 dB
Reference	4.0 dB	3.8 dB	3.7 dB	3.7 dB
NAQ	100 Hz	200 Hz	300 Hz	400 Hz
IAIF	0.043	0.024	0.019	0.009
MCMC-GIF	0.044	0.031	0.030	0.022
MCMC-Klatt	0.067	0.061	0.079	0.076
Reference	0.108	0.104	0.101	0.100

Table 2: *Relative errors of H1-H2 and NAQ quantities of the GIF methods for synthetic speech with four different F_0.*

H1-H2	100 Hz	200 Hz	300 Hz	400 Hz
IAIF	1.3 %	43.4 %	214.9 %	156.2 %
MCMC-GIF	0.3 %	35.4 %	4.9 %	53.5 %
MCMC-Klatt	1.0 %	34.6 %	26.8 %	16.1 %
NAQ	100 Hz	200 Hz	300 Hz	400 Hz
IAIF	60.2 %	76.9 %	81.2 %	90.7 %
MCMC-GIF	59.3 %	70.2 %	70.3 %	78.2 %
MCMC-Klatt	38.0 %	41.3 %	21.8 %	24.2 %

the MCMC-GIF method takes roughly 3–4 hours in a standard workstation. However, MCMC codes can be easily parallelized for faster computation of the GIF estimates. The final parameter estimates are computed as a mean of the Markov chain, excluding the burn-in period (10 000 first members of the chain).

The GIF results are evaluated both with time and frequency domain features. First, the magnitude difference between the first and the second harmonics of the glottal flow is measured. This feature, usually denoted by H1-H2 [13], is widely used as a measure of vocal quality. Normalized amplitude quotient (NAQ) [14] is used as a time domain feature, which is also commonly used as a measure of vocal quality.

3.4. Results for synthetic vowels

In the case of the synthetic vowels, four poles of the vocal tract (describing the first four formants) and the Klatt parameter were refined using the MCMC-GIF method. Thus the total number of estimated parameters was nine. The weights of time and frequency domain errors were set equal.

Table 1 contains the H1-H2 and NAQ quantities of the glottal flow estimates given by the GIF methods, and Table 2 contains their relative errors. Here MCMC-Klatt refers to the glottal excitation generated using the Klatt model inside MCMC-GIF method. MCMC-GIF refers to the glottal inverse filtering estimate achieved using the data and the initial vocal tract estimate. The MCMC-GIF method provided more accurate estimates of the glottal flow than the IAIF in terms of both H1-H2 and NAQ. The improvement is clear especially with high fundamental frequencies. The average relative error for IAIF, MCMC-GIF, and MCMC-Klatt methods were 103.9%, 23.5%, and 19.6% for H1-H2, and 77.3%, 69.5%, and 31.3% for NAQ, respectively.

Figure 1 shows an example of the glottal flow estimates of different methods. In this example, MCMC-GIF and MCMC-

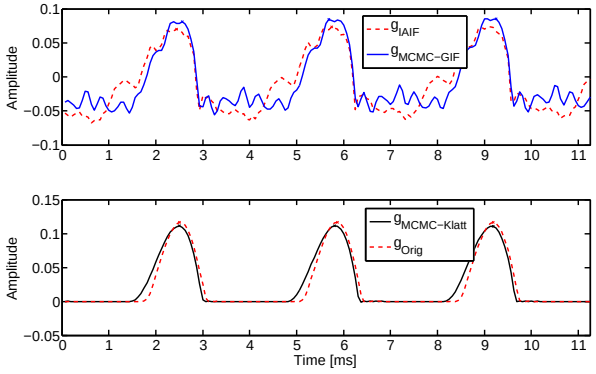

Figure 1: *Glottal flow estimates from a synthetic vowel with $F_0 = 300$ Hz. Upper panel: IAIF and MCMC-GIF estimates. Lower panel: original excitation and corresponding MCMC-Klatt estimate.*

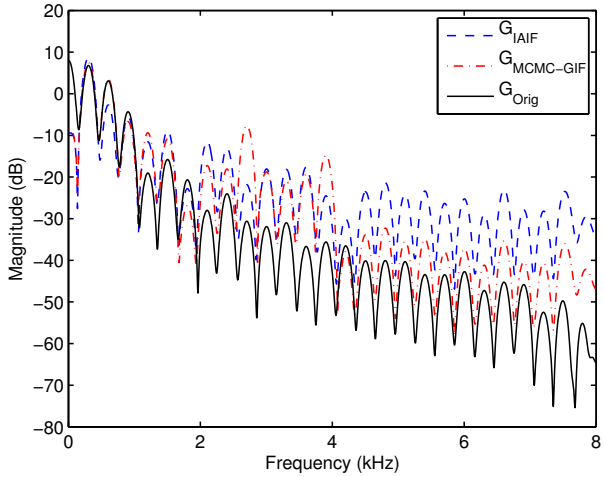

Figure 2: *Spectra of the glottal flow estimates given by the GIF methods. The spectrum of the original synthetic excitation with $F_0 = 300$ Hz is shown for reference.*

Klatt methods provide more realistic and accurate estimate of the glottal flow than the IAIF method. Figure 2 shows an example of the spectra of the glottal excitation estimates of the different methods. The original excitation of the synthetic vowel serves here as a reference. The MCMC-GIF estimate is closer to the original excitation than the IAIF estimate. This can be noticed especially at the higher frequencies.

3.5. Results for natural vowels

For natural vowels, four poles of the vocal tract and the Klatt parameter were estimated with the MCMC-GIF method, and only the time-domain error was used for defining the fit. In the case of natural vowels, the actual glottal excitation is unknown. Thus, the MCMC-GIF and MCMC-Klatt results are compared against the IAIF estimates. Since no objective evaluation can be performed, the results were only observed visually. The results suggest that the MCMC-GIF yields similar or better results than the IAIF method.

Figure 3 shows an example of the vocal tract estimates of the IAIF and MCMC-GIF methods for natural vowel [a] with

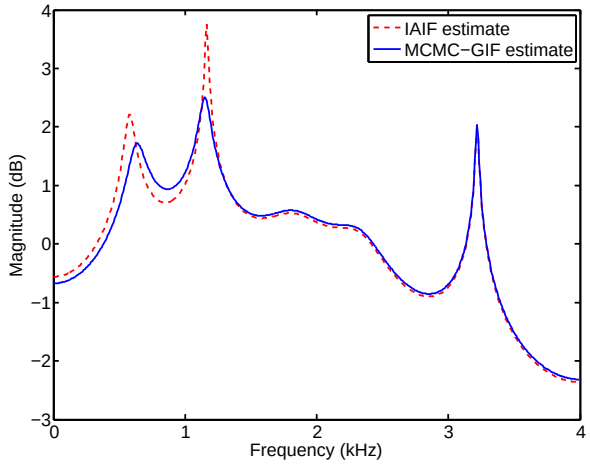

Figure 3: *Estimates of the vocal tract of the IAIF and the MCMC-GIF methods for natural vowel [a] with $F_0 = 292$ Hz. MCMC-GIF method has shifted the first two formants from the initial IAIF estimate in order to get a better GIF result.*

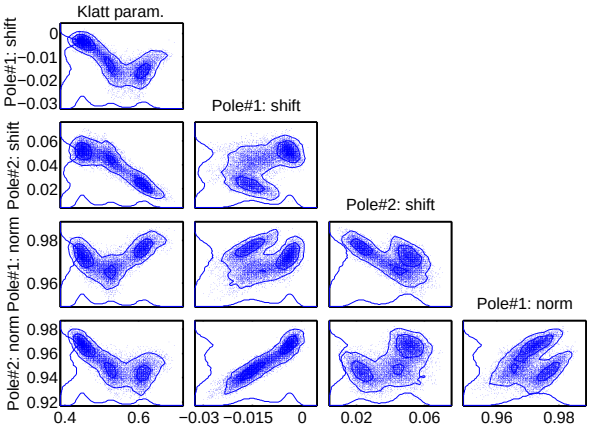

Figure 4: *Estimated marginal posterior distributions for a natural vowel [a] with $F_0 = 292$ Hz, showing a few modes of possible solutions. The total number of estimated parameters is five, consisting of the Klatt parameter and two poles, defined as a shift of the angle in radians from the IAIF estimate and as the norm of the pole.*

$F_0 = 292$ Hz. MCMC-GIF method has shifted especially the estimate of the first formant and lowered the levels of the two first formants, correcting the IAIF estimate possibly biased by the harmonics of the excitation.

Figure 4 illustrates the marginal posterior distribution estimated by the MCMC-GIF method. Each blue dot in the figure indicates a possible solution of the GIF inversion problem. The concentration of the dots correspond to the probability of the solution. The shape of the posterior distribution shows that the inversion problem contains only few modes of possible solutions, and justifies the choice of MCMC method as a solver.

4. Conclusions

Accurate and automatic estimation of the voice source from speech pressure signals is known to be difficult with current glottal inverse filtering (GIF) techniques especially in the case of high-pitch speech. In order to tackle this problem, the present study proposes the use of the Bayesian inversion approach in GIF. The proposed method takes advantage of the Markov chain Monte Carlo (MCMC) modeling in defining the parameters of the vocal tract inverse filter. The new technique, MCMC-GIF, enables applying detailed *a priori* distributions of the estimated parameters. Furthermore, the MCMC-GIF method provides an estimate of the whole *a posteriori* distribution in addition to a single maximum likelihood estimate.

The novel MCMC-GIF method was preliminary tested in the current study by using test vowels synthesized with a physical modeling approach. The results are encouraging in showing that the MCMC-GIF yields more accurate glottal flow estimates than a known reference method. The improvement was most prominent for the vowel of the highest fundamental frequency hence indicating that the approach suggested holds promises in the automatic voice source analysis of high-pitch speech. In the future, more experiments will be conducted by involving more advanced *a priori* distributions of the estimated parameters.

5. Acknowledgements

This study was supported by the Academy of Finland (LASTU programme, decision no 134868) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678.

6. References

[1] Miller, R.L., "Nature of the vocal cord wave", J. Acoust. Soc. Am., 31(6):667–677, 1959.

[2] Alku, P., Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications, Sadhana Vol. 36, Part 5, pp. 623–650, 2011.

[3] El-Jaroudi, A. and Makhoul, J., "Discrete all-pole modeling", IEEE Trans. Signal Proc., 39(2):411–423, 1991.

[4] Klatt, D., "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. 67(3):971–995, 1980.

[5] Klatt, D.H., "Review of text-to-speech conversion for English", J. Acoust. Soc. Am., 82(3):737–793, 1987.

[6] Metropolis, N., Rosenbluth, A.W, Rosenbluth, M.N., Teller, A.H., and Teller, E., "Equations of state calculations by fast computing machine", J. Chem. Phys., 21:1087–1091, 1953.

[7] Haario, H., Laine, M., Mira, A. and Saksman, E., "DRAM: Efficient adaptive MCMC", Stat. Comput. 16:339–354, 2006.

[8] Laine, M., MCMC toolbox for Matlab, online: http://www.helsinki.fi/∼mjlaine/mcmc/

[9] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Commun., 11(2–3):109–118, 1992.

[10] Alku, P., Tiitinen, H. and Näätänen, R., "A method for generating a natural-sounding speech stimuli for cognitive brain research", Clinical Neurophysiology, 110:1329–1333, 1999.

[11] Fant, G., "Acoustic Theory of Speech Production", The Hague: Mouton, 1960.

[12] Alku, P., Magi, C., Yrttiaho, S., Bäckström, T. and Story, B., "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering", J. Acoust. Soc. Am., 125:3289–3305, 2009.

[13] Titze, I. and Sundberg, J., "Vocal intensity in speakers and singers", J. Acoust. Soc. Am., 91(5):2936–2946, 1992.

[14] Alku, P., Bäckström, T. and Vilkman, E., "Normalized amplitude quotient for parametrization of the glottal flow", J. Acoust. Soc. Am., 112(2):701–710, 2002.