Bleyer, Ismael Rodrigo; Lybeck, Lasse; Auvinen, Harri; Airaksinen, Manu; Alku, Paavo; Siltanen, Samuli

Alternating minimisation for glottal inverse filtering

Published in:
Inverse Problems

DOI:
10.1088/1361-6420/aa6eb8

Published: 17/05/2017

Document Version
Publisher's PDF, also known as Version of record

Alternating minimisation for glottal inverse filtering

OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **33** (2017) 065005 (19pp)
https://doi.org/10.1088/1361-6420/aa6eb8

Alternating minimisation for glottal inverse filtering

Ismael Rodrigo Bleyer[1], Lasse Lybeck[1], Harri Auvinen[1], Manu Airaksinen[2], Paavo Alku[2] and Samuli Siltanen[1]

[1] Department of Mathematics and Statistics, University of Helsinki, Gustaf Hällströmin Katu 2b, FI-00014 Helsinki, Finland
[2] Department of Signal Processing and Acoustics, Aalto University, Otakaari 5 A, FI-02150 Espoo, Finland

E-mail: ismael.bleyer@helsinki.fi, lybeck.lasse@gmail.com, harrikauvinen@gmail.com, manu.airaksinen@aalto.fi, paavo.alku@aalto.fi and samuli.siltanen@helsinki.fi

Received 20 July 2016, revised 28 February 2017
Accepted for publication 24 April 2017
Published 17 May 2017

Abstract

A new method is proposed for solving the glottal inverse filtering (GIF) problem. The goal of GIF is to separate an acoustical speech signal into two parts: the glottal airflow excitation and the vocal tract filter. To recover such information one has to deal with a blind deconvolution problem. This ill-posed inverse problem is solved under a deterministic setting, considering unknowns on both sides of the underlying operator equation. A stable reconstruction is obtained using a double regularization strategy, alternating between fixing either the glottal source signal or the vocal tract filter. This enables not only splitting the nonlinear and nonconvex problem into two linear and convex problems, but also allows the use of the best parameters and constraints to recover each variable at a time. This new technique, called alternating minimization glottal inverse filtering (AM-GIF), is compared with two other approaches: Markov chain Monte Carlo glottal inverse filtering (MCMC-GIF), and iterative adaptive inverse filtering (IAIF), using synthetic speech signals. The recent MCMC-GIF has good reconstruction quality but high computational cost. The state-of-the-art IAIF method is computationally fast but its accuracy deteriorates, particularly for speech signals of high fundamental frequency ($F0$). The results show the competitive performance of the new method: With high $F0$, the reconstruction quality is better than that of IAIF and close to MCMC-GIF while reducing the computational complexity by two orders of magnitude.

Keywords: ill-posed problems, glottal inverse filtering, double regularisation, alternating minimisation, glottal airflow, wavelets, deterministic

(Some figures may appear in colour only in the online journal)

1. Introduction

We study the estimation and parametrization of the voice source, the excitation waveform of speech sounds produced by the vocal folds in the human larynx. This origin of speech is called the glottal flow, named after the orifice between the vibrating vocal folds, the glottis [16]. We rely on source-filter theory [31] as the mathematical model of human voice production. Briefly, a speech signal, denoted by $m(t)$, can be represented as a convolution of three functions. The first corresponds to the glottal flow, here denoted by $g(t)$. The second is a vocal tract filter function modeling the effects caused by the oral cavity between the vocal folds and the lips, denoted in the time domain by $f(t)$. The third function corresponds to the lip radiation effect, the acoustical conversion of the air flow at the lips into a free-field pressure signal. Since the latter of the three functions can be estimated as a time derivative [31] and can be combined into the first function, the direct problem is formulated through the convolution equation

$$f(t) * g'(t) = m(t), \tag{1}$$

where t represents the time variable.

Our study focuses on the estimation of the glottal flow using a computational inversion methodology called glottal inverse filtering (GIF). The aim of GIF is to remove the effects of the vocal tract and lip radiation from the speech signal, revealing the time-domain waveform of the glottal source. GIF can be seen as the inverse problem related to (1): Given the measured speech signal waveform $m(t)$, find both $f(t)$ and $g'(t) = p(t)$ so that

$$m(t) + \delta = f(t) * p(t), \tag{2}$$

for $t \in [0, 1]$. Here δ models additive noise arising from, for instance, the recording environment. Such blind deconvolution problems are known to be *ill-posed*, or highly sensitive to modeling errors and measurement noise. Therefore, regularization is needed for successful inversion.

GIF has been used in several areas of speech science such as in studying vocal emotions and speech prosody as well as in analyzing pathological speech and singing voices (for a review of GIF, see [4]). In the past five years, GIF has also awakened increased interest in the speech technology community, particularly among scientists developing text-to-speech (TTS) synthesis technologies. Recent studies suggest that the naturalness of TTS systems can be improved by using sound excitations that have been computed from natural speech via GIF [33]. The improved naturalness of TTS in turn can help, for example, disabled people who are not capable of preserving natural vocalization due to a disease, accident, or speech disorder. An example is the world-famous physicist Stephen Hawking, who cannot speak without the help of speech synthesis software. The development of new TTS technology would benefit from computation of realistic glottal excitation signals via GIF.

GIF has been studied in the speech science community dating back to 1959; see overviews in [4, 39, 11], and further references in PhD theses [32, 38]. Within the mathematical inverse problems community, the first GIF study [35] appeared in 1970. This study was based on using Webster's horn equation to investigate the shape and the cross-sectional area of the vocal tract. In 1986, a GIF technique was suggested in [29] to study the glottal source for both

male and female voices. The topic was revived in the 2000's in [1, 2, 17, 18, 21, 22], based on the Schrödinger equation, Klein–Gordon equation and various deterministic computational approaches. Uniqueness of GIF is studied in [25]. The first Bayesian inversion approach to GIF is described in [6].

State-of-the-art GIF technologies, such as *iterative adaptive inverse filtering (IAIF)* [3], are far too simple to yield reliable glottal flow estimates particularly for speech of high $F0$ (e.g. for utterances spoken by women or children) or for voices conveying extreme emotions such as anger and hate. Bayesian statistical inversion, however, was shown to improve the estimation accuracy of the glottal flow in a recent study proposing a new GIF method called Markov chain Monte Carlo glottal inverse filtering (MCMC-GIF) [6]. However, the computational cost of MCMC-GIF is high due to extensive Markov chain Monte Carlo sampling.

In the current study, we introduce a deterministic approach called *alternating minimization glottal inverse filtering (AM-GIF)*, based on the general analysis published in [8, 9]. It requires an initial estimate $p^{(0)}$ for the derivative of the glottal flow signal, and it solves the inverse problem iteratively as follows:

(a) given $m(t)$ and a fixed $p^{(k)}$, find a regularized solution $f^{(k+1)}$ for (2);
(b) given $m(t)$ and a fixed $f^{(k+1)}$, find a regularized solution $p^{(k+1)}$ for (2).

The proposed AM-GIF method matches closely the reconstruction quality of MCMC-GIF while reducing the computational cost significantly.

We study the new method under ideal simulated conditions. The signals considered have precise periodicity, and we deal with boundary conditions in a mathematically convenient way. Natural speech signals are not so regular, and additional steps are needed for applying our results in practice. However, we expect the core properties of the alternating minimization algorithm to carry over to real-world signals.

This paper is organized as follows. In section 2, we present an overview of the source-filter theory, including the main concepts and two mathematical models for the source signal. Section 2 is aimed at readers not familiar with the topic and can be omitted by other readers. The inverse problem is defined in section 3, with a description of the novel AM-GIF approach. Moreover, section 3 explains two baseline methods, IAIF and MCMC-GIF. Numerical experiments and comparisons are presented in section 4. Finally, the conclusions of the study are drawn in section 5.

2. Source-filter theory

Speech carries information about *phonemes*, the basic units of spoken language. Among phonemes, the current study, like almost all previous GIF investigations, focuses on non-nasalized vowels. This category of speech sounds has been prevalent in GIF studies for two reasons: non-nasalized vowels are always voiced (i.e. generated by the vibration of the vocal folds) and their vocal tract lacks coupling to the nasal tract, thereby justifying the use of all-pole type of models [27] for the vocal tract.

2.1. Source-filter theory in the frequency domain

According to source-filter theory [13, 36], speech can be interpreted as a linear combination of three processes:

$$S(z) = U(z)V(z)L(z),$$

where $S(z)$, $U(z)$, $V(z)$, and $L(z)$ denote the z-transforms of the measured speech signal, glottal flow, vocal tract, and the lip radiation effect, respectively.

This linear model has been widely used due its relative simplicity, for both speech synthesis and analysis. Source-filter theory describes, as its name suggests, speech as a two-stage process consisting of the sound *source*, which is filtered by a *filter* function.

The output speech waveform S is often described in the complex frequency domain, as stated in the previous equation. Furthermore, it is possible to combine two or more filters into a single one by, for example, representing the vocal tract filter as $V(z) = P(z)O(z)$, where P and O are the transfer functions at the pharynx and at the oral cavity respectively. In this article, the vocal tract is treated simply as a single transfer function, V.

The speech signal is in practice recorded using a free-field microphone that measures the pressure signal, not the air flow. According to [16, p 259], the acoustical conversion from the flow at the lips into a pressure signal in the free field, the so-called lip radiation effect, can be approximated at low frequencies to correspond to a fixed first-order derivative in the time domain. In the frequency domain, this derivative can be modeled by an FIR (finite impulse response) filter with a single zero

$$L(z) = 1 - \alpha z^{-1},$$

where $0.96 \leqslant \alpha < 1$. Moreover, it is often useful to combine L and U into a single filter, even though the lip radiation effect occurs physically at the lips when the flow signal has already passed through the vocal tract. This is equivalent to applying the lip radiation effect to the glottal flow by differentiating the glottal flow U [16]. Let us denote the differentiation process as $G'(z) = L(z)U(z)$ and let us refer the process output as the *glottal flow derivative*.

Finally, we can display the two-stage mathematical model as our direct problem

$$S(z) = G'(z)V(z),$$

where G' is the derivative of the glottal flow U. Note also the close relation to equation (1) introduced in the time domain.

2.2. Source models

Anatomically speaking, the lungs power an airstream outwards, pushing the air through the narrow opening between the vocal folds, called the glottis, producing puffs of air. The vocal folds are composed of twin infoldings of mucous membrane, stretched horizontally, which vibrate[3], and therefore the puffs of air are produced pseudo-periodically in time. Mathematically speaking, we are interested in a function $g : [0, 1] \to \mathbb{R}^+$ that models the volume of air traveling through the glottis at a specific time t, technically known as the *glottal flow* or the *glottal source*.

Several parametric models of the glottal flow have been proposed in the literature, and this article discusses two of them: the Rosenberg–Klatt (RK) model and the Liljencrants–Fant (LF) model.

The Rosenberg–Klatt model is a straightforward glottal flow model. It models the shape of the glottal airflow signal within one fundamental period using a cubic polynomial function that is defined by one time-domain parameter, in addition to the length of the glottal cycle, and an amplitude-domain scaling factor [23]. This model was originally proposed by Rosenberg

[3] The vocal folds usually vibrate from 100 to 300 times per second (i.e. from 100 Hz to 300 Hz), depending on gender and speaking style.

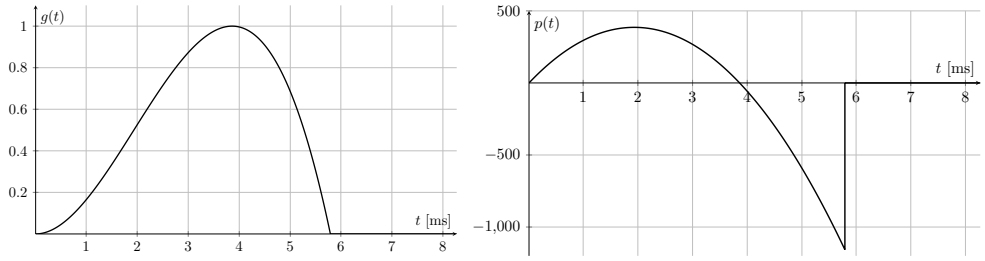

Figure 1. The glottal flow (left) and its derivative (right) generated by the RK model, with parameters $Q = 0.7$ and fundamental frequency $F0 = 120$ Hz.

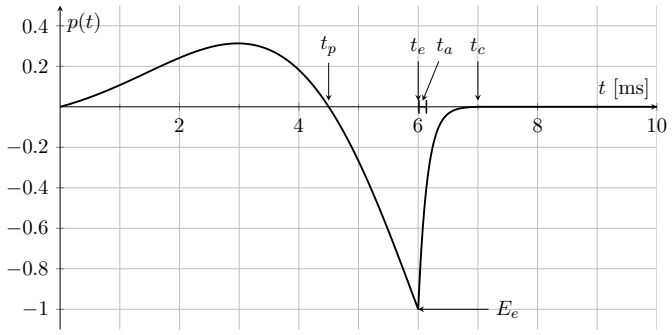

Figure 2. The glottal flow derivative generated by the LF model.

[34] and later modified by Klatt who also used the parametric waveform as an excitation in speech synthesis [23].

An example of the glottal flow and its derivative, generated by the RK model, is shown in figure 1. Note that the parametric flow pulse consists of two distinctive phases: the open phase (OP) and the closed phase (CP). The former represents the smooth increase of airflow as the membranes of the vocal folds open up (from bottom to top) and a faster decrease of the volume of air when the vocal folds close, completing one cycle. The latter represents a phase when the glottis is completely closed and there is no air passing through the vocal folds.

The Liljencrants–Fant model requires four parameters in addition to the length of the glottal cycle to create the shape of a glottal flow derivative [15]. It is one of the most widely used parametric time-domain models of the glottal source.

Compared to the RK model, the LF model describes different phases of the glottal cycle more in detail, introducing the so-called *return phase*, a phase between the instants of maximum closing discontinuity and glottal closure. Figure 2 describes the LF model in different segments: t_p is the instant of the maximum airflow (zero derivative), t_e is the instant of the maximum excitation (with amplitude E_e) or the instant when the vocal folds collide, t_a is the length of the interval $[t_e, t_e + t_a]$ that measures the effective duration of the return phase, and t_c is the instant when the vocal folds reach the maximum closure and the airflow is reduced to its minimum. The interval before t_e is the OP, between t_e and t_c is the return phase, and the section between t_c and the end of the cycle is the CP.

A detailed definition of the LF model as a time-domain function can be found in [14]. Fitting a given glottal flow with the LF model is not straightforward since it requires solving nonlinear equations.

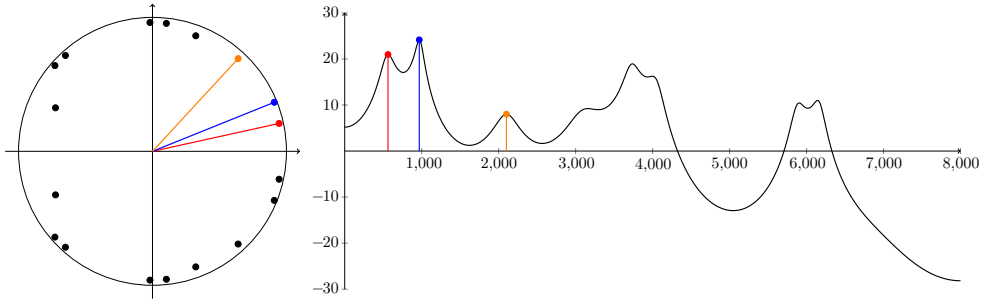

Figure 3. The transfer function: the location of the first three poles inside the unit disc (left) and the location of the first three formants $F1$, $F2$, and $F3$ (right) for the vowel [a]. The poles closer to $z = 1$ ($\omega = 0$) correspond to a low-frequency formant and the poles closer to $z = -1$ ($\omega = \pi$) correspond to a high-frequency formant.

2.3. The vocal tract as a filter

The vocal tract consists of cavities that have a strong perceptual effect on the produced speech sound. The vocal tract comprises the oral and nasal cavities that, together with the articulators—such as the tongue, soft palate, and lips—filter the produced glottal flow into its final form as an acoustical speech signal that we recognize as a certain phoneme. Mathematically, sounds are usually represented through a waveform (time-pressure) or spectrogram (time-frequency). The latter enables studying the locations of the formants, the spectral peaks of the sound, caused by the vocal tract resonances [12].

The vocal tract can be modeled as a linear time-invariant (LTI) system, defined in the complex z-domain using a transfer function $H(z)$:

$$H(z) = \frac{\sum_{n=0}^{M} b(n) z^{-n}}{\sum_{n=0}^{N} a(n) z^{-n}} = \frac{B(z)}{A(z)}.$$

Note that the roots of the numerator polynomial B are also the roots of H and they correspond to the antiformants (zeros), whereas the roots of the denominator polynomial A correspond to the formants (poles). Figure 3 shows the relation between the poles of the transfer function inside the unit disc and the corresponding formant locations.

In general, the pole-zero transfer function H is simplified in speech technology into an all-pole filter, that is, a filter for which $M = 0$ (i.e. the transfer function has only poles in the z-domain outside the origin). This is justified because (a) all-pole models give an accurate fit to most speech sounds, particularly for vowels and (b) all-pole models are typically easier to solve than pole-zero models.

3. The inverse problem: GIF

In this section we describe three approaches to GIF. The first one, the novel AM-GIF method, is based on time-frequency domain analysis using the wavelet transform and the inverse problem is solved under a deterministic setting. The second approach, the recently proposed MCMC-GIF algorithm, is based on frequency domain analysis using the z-transform and the inverse problem is solved under a stochastic setting. The third one, the state-of-the-art IAIF method, is based on linear prediction (LP) analysis and the problem is solved using a straightforward signal processing approach. Numerical experiments will be reported in section 4.

3.1. The AM-GIF approach

Tikhonov-type inversion is a common approach to solve an ill-posed inverse problem in the deterministic setting, whenever a stable solution is needed in the case of noisy data. If, however, there are uncertainties in both the operator and the measured data, the problem should be addressed with a more general approach [7].

The operator on the left-hand side of (1) will be denoted throughout this section as a linear convolution operator between Hilbert spaces $\mathcal{U} \to \mathcal{H}$, as follows:

$$p(t) * f(t) = \int_0^1 p_o(t-s) f(s) \mathrm{d}s \tag{3}$$

where $0 \leqslant t \leqslant 1$ and $p_o \in \mathcal{U}$ is called the *characterizing function* for the convolution operator.

Note that we are interested in recovering the unknown f, but at the same time the characterizing function p_o is not precisely known for real data, only mathematical models are available, as discussed in section 2.2. Therefore, solving (1) for the operator (3) should be seen as a blind deconvolution problem. Moreover, parametric models such as the RK and LF models could be good approximations for p_o, knowing *a priori* the glottal opening time. We assume the noise levels of the measurements (m_δ, p_ϵ) to be known:

$$\left\| m - m_\delta \right\|_{\mathcal{H}} \leqslant \delta, \tag{4a}$$

and

$$\left\| p_o - p_\epsilon \right\|_{\mathcal{U}} \leqslant \epsilon. \tag{4b}$$

The framework of the AM-GIF proposed here is based on the core idea of the *double regularized total least squares* (dbl-RTLS) method. In short, dbl-RTLS is a deterministic approach introduced recently in [8] to solve inverse problems with noise in both data and operator, as stated in (4). Using this method, we aim at solving the following minimization problem

$$\text{minimize}_{(p,f)} \left\{ \frac{1}{2} \left(\left\| p * f - m_\delta \right\|_{\mathcal{H}}^2 + \tau \left\| p - p_\epsilon \right\|_{\mathcal{U}}^2 \right) + \frac{\alpha}{2} \left\| Lf \right\|_{\mathcal{U}}^2 + \frac{\beta}{2} \left\| p \right\|_{\mathcal{U}}^2 \right\} \tag{5}$$

where the term inside the parentheses measures the total discrepancy, τ is a fixed scaling parameter, L is a linear bounded operator, and $\alpha > 0$ and $\beta > 0$ are regularization parameters. The amplitude of measurement noise determines the choice of α and β: Higher noise requires larger parameter values. For further properties and generalizations of the dbl-RTLS approach, see [8].

To overcome the drawback caused by the non-linearity and non-convexity in minimizing (5) with respect to the pair (p, f) at the same time, we follow the *alternating minimization* strategy, which has been successfully adopted in solving optimization problems over two variables. This strategy has been implemented for blind-deconvolution [40], and also used as standard for solving the dbl-RTLS problem [9].

The AM-GIF algorithm for equation (5) is an iteration scheme. Going from a current iterate $\left(p^{(k)}, f^{(k)}\right)$ to a new pair $\left(p^{(k+1)}, f^{(k+1)}\right)$ involves two steps. First, keeping $p^{(k)}$ fixed, define $f^{(k+1)}$ as the solution of

$$\text{minimize}_f \left\{ \frac{1}{2} \left\| p^{(k)} * f - m_\delta \right\|_{\mathcal{H}}^2 + \frac{\alpha}{2} \left\| Lf \right\|_{\mathcal{U}}^2 \right\}. \tag{6a}$$

Second, keep $f^{(k+1)}$ fixed and define $p^{(k+1)}$ as the solution of

$$\text{minimize}_p \left\{ \frac{1}{2} \left(\left\| p * f^{(k+1)} - m_\delta \right\|_{\mathcal{H}}^2 + \tau \left\| p - p_\epsilon \right\|_{\mathcal{U}}^2 \right) + \frac{\beta}{2} \left\| p \right\|_{\mathcal{U}}^2 \right\}. \tag{6b}$$

The convergence of this method is proved in [9, section 3].

Splitting the minimization (5) into two steps transforms a seemingly formidable challenge into two standard problems. Namely, step (6*a*) is just Tikhonov regularization, and (6*b*) is an example of the *regularized total least squares* method, see [19, 26].

We need a computational implementation of the convolution operator defined in (3). Assuming that p is periodic, one could follow the time-domain approach of [30] and rewrite equation (3) as matrix convolution. This yields a square circulant matrix with many desirable properties: full column rank, invertibility, and a circulant inverse [30]. One could also model (3) in the frequency domain using fast Fourier transform. However, this can be unstable if the operator is not precisely known.

We chose to implement the convolution operator in the wavelet domain [10, 28, 37, 41].

3.1.1. Convolution and the wavelet transform. Let us denote the wavelet family by $\{\varphi_\lambda\}_{\lambda\in\Lambda}$, which constitutes an orthonormal basis for a Hilbert space, where the index set Λ is defined by

$$\Lambda = \{1\} \cup \left\{ (j,l) \mid j \in \mathbb{N}_0,\ 0 \leqslant l \leqslant 2^j - 1 \right\}.$$

Furthermore, we denote

$$\varphi_\lambda = \begin{cases} \phi & \text{if} \quad \lambda = 1 \\ \psi_{j,l} & \text{if} \quad \lambda = (j,l). \end{cases}$$

Here ϕ is the *scaling function* and $\psi_{0,0}$ is the *mother wavelet*. Any function $f \in \mathcal{U}$ can be decomposed as

$$f = \sum_{\lambda\in\Lambda} \langle f\ ,\ \varphi_\lambda \rangle \varphi_\lambda, \quad \text{where} \quad \langle f\ ,\ \varphi_\lambda \rangle = \int_0^1 \varphi_\lambda(s) f(s) \mathrm{d}s.$$

We denote the coefficient sequence of $f \in \mathcal{U}$ by $F(f)$:

$$\begin{array}{rccl} F & : & \mathcal{U} & \longrightarrow & \ell_2 \\ & & f & \longmapsto & F(f) := \{\langle f\ ,\ \varphi_\lambda \rangle\}_{\lambda\in\Lambda}. \end{array} \tag{7}$$

In other words, we represent the signal f by the sequence $x = F(f)$ which belongs to ℓ_2. For numerical reasons we have to truncate the summation of j at a certain fixed index J, called the *maximal level*.

The next goal is to express convolution equation (3) in the wavelet domain to obtain the desired data function m by applying the direct operator entirely in the wavelet context. We first compute an operator C that depends only on J and on the choice of the wavelet family.

Let $y := F(p)$ and $x := F(f)$ be the coefficients from the characterizing and input function respectively. Then the coefficient of $d := F(m)$ is determined via $C(y, x)$ as

$$\begin{array}{rccl} C & : & \ell_2 \times \ell_2 & \longrightarrow & \ell_2 \\ & & (y,x) & \longmapsto & d = C(y,x). \end{array}$$

The major work is to compute the operator C. Once accomplished, for a fixed interval, the maximal wavelet level J, and the number of samples N, one could vary the characterizing function or the input function. So the computation of the convolution operator remains straightforward via a few matrix-vector multiplications. More precisely, the operator C is a sequence of square matrices $\{C_\mu\}_\mu$, where $1 \leqslant \mu \leqslant 2^{J+1}$, in such a way that the sequence coefficient $\{d_\mu\}_\mu$ is computed by

$$y^t C_\mu x = d_\mu.$$

The detailed computation of the sequence of matrices $\{C_\mu\}_\mu$ is explained in appendix.

For a fixed characterizing function p, the sequence of matrices can be combined into a unique square matrix A whose μth row $A(\mu,:)$ is given by

$$A(\mu,:) = y^t C_\mu, \tag{8}$$

where $1 \leqslant \mu \leqslant 2^{J+1}$. Now the computation of the wavelet coefficient sequence $d = F(m)$ takes the form of a single matrix-vector multiplication

$$Ax = d.$$

Analogously, for a fixed input function f one can define the matrix

$$B(\mu,:) = C_\mu x \tag{9}$$

and so

$$By = d$$

solves the forward convolution problem for a fixed function f.

3.1.2. The AM-GIF algorithm. Now we can rewrite algorithm (6) by taking advantage of the wavelet decomposition. We now have convenient matrix machinery for the appropriate operators for fixed p and f:

$$\begin{array}{rcccl} A & : & \ell_2 & \longrightarrow & \ell_2 \\ & & x & \longmapsto & Ax = d, \end{array}$$

and

$$\begin{array}{rcccl} B & : & \ell_2 & \longrightarrow & \ell_2 \\ & & y & \longmapsto & By = d. \end{array}$$

As seen in figure 4, both A and B are sparse matrices due to the compact support of the Haar wavelet basis.

Note that the wavelet coefficients d_δ of the measured data m_δ may contain errors arising from measurement noise.

The first step of the AM algorithm (6*a*) in the time domain, for a fixed p, is equivalent to the minimization of the following functional:

$$\text{minimize}_x \left\{ \frac{1}{2}\left\|Ax - d_\delta\right\|_2^2 + \frac{\alpha}{2}\left\|Lx\right\|_2^2 \right\}, \tag{10a}$$

that is, recovering the wavelet coefficients.

Whereas the second step of the alternating minimization algorithm (6*b*) in the time domain, for a fixed f, is equivalent to the minimization of the following functional:

$$\text{minimize}_y \left\{ \frac{1}{2}\left(\left\|By - d_\delta\right\|_2^2 + \tau\left\|y - y_\epsilon\right\|_2^2 \right) + \frac{\beta}{2}\left\|y\right\|_2^2 \right\} \tag{10b}$$

where $y_\epsilon = F(p_\epsilon)$.

For the first step it is well-known that the solution of the above problem (10*a*) satisfies

$$(A^*A + \alpha L^*L)\,\bar{x} = (A^* d_\delta)$$

where the adjoint operator is the transpose matrix.

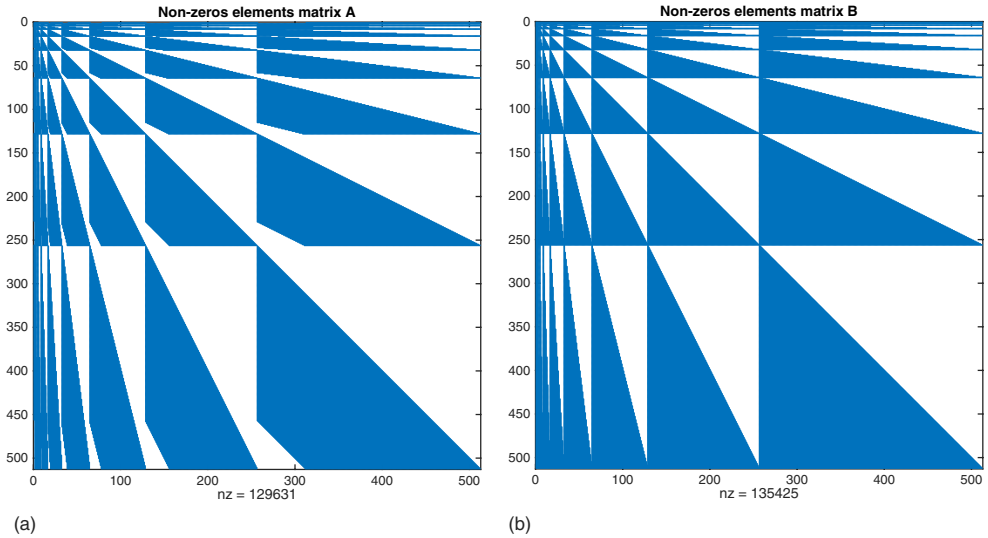

Figure 4. The sparse structure of the matrices for one cycle of the AM strategy, with the maximal level $J = 8$ for the Haar wavelet basis. (a) Matrix A (characterizing function). (b) Matrix B (filter function).

It is easy to see that the minimization of (10*b*) and

$$\text{minimize}_y \left\{ \frac{1}{2}\left\|By - d_\delta\right\|_2^2 + \frac{(\tau+\beta)}{2}\left\|y\right\|_2^2 - \tau\left\langle y\ ,\ y_\epsilon\right\rangle \right\}$$

have the same solution, satisfying the equation

$$(B^*B + (\beta+\tau)I)\ \bar{y} = (B^*d_\delta + \tau y_\epsilon)\,,$$

as one reads optimality condition for the previous minimization problem. A summary of this procedure can be seen in Algorithm 1.

Algorithm 1. The AM-GIF algorithm

Require: Load matrix C; Set scaling parameter τ;
1: Compute the wavelet coefficients $d_\delta = F(m_\delta)$ and $y_\epsilon = F(p_\epsilon)$;
2: Set starting point, e.g. $\bar{y} = y_\epsilon$
3: **repeat**
4: Set $y \mapsto \bar{y}$
5: Compute matrix A as $A(\mu,:) = y^t C_\mu$
6: Find regularization parameter α
7: Solve $(A^*A + \alpha L^*L)\ \bar{x} = (A^*d_\delta)$
8: Set $x \mapsto \bar{x}$
9: Compute matrix B as $B(\mu,:) = C_\mu x$
10: Find regularization parameter β
11: Solve $(B^*B + (\beta+\tau)I)\ \bar{y} = (B^*d_\delta + \tau y_\epsilon)$
12: **until** Convergence
13: **return** coefficients $\bar{x}$ and $\bar{y}$; reconstruct functions $\bar{f}$ and $\bar{p}$

3.2. The MCMC-GIF approach

The MCMC-GIF method was introduced in [6]. In Bayesian inversion, robustness against modeling errors and measurement noise is achieved by complementing measurement data by *a priori* information. The measurement process is described probabilistically in the form of a likelihood model. Further, any *a priori* information about the unknown quantities is represented as prior probability distribution. The product of the likelihood and prior yields the *posterior distribution*. The solution of the ill-posed inverse problem takes the form of stable computation of the mean of the posterior distribution by Monte Carlo integration.

In the MCMC-GIF approach, the vocal tract model is assumed to be an all-pole filter. Each jth complex pole of the filter (see figure 3, left side) is parametrized by a pair (r_j, θ_j) from its complex representation $z_j = r_j \exp(i\theta_j)$ where r_j and θ_j are, respectively, the radius and angle of the pole in the z-domain. The glottal airflow is parametrized with the open quotient parameter of the RK model [34], denoted by Q in the following. Therefore, the inverse problem can be expressed as the recovery of the (combined) parameters vector

$$\vec{v} := [r_1, \theta_1, r_2, \theta_2 \ldots, r_N, \theta_N, Q]^T$$

from a given measurement m, where N denotes the number of poles.

The computational procedure is based on an initial estimate of the vocal tract filter, here given by the IAIF method. Next, the algorithm refines the vocal tract model parameters within the MCMC-GIF method in order to obtain a more accurate glottal flow estimate. This GIF method is known for its good performance in the estimation of the glottal flow from high-pitched signals.

3.3. The IAIF Approach

The IAIF method was proposed more than 20 years ago [3] and it is still in use due its simplicity and fast computation.

The method relies on the assumption that the glottal flow is obtained by canceling the effects of the vocal tract and lip radiation from the speech measurement with an iterative structure. The theory of speech production via a chain of filters was introduced in section 2.1. According to [3] the vocal tract transfer function is modeled after eliminating the average glottal contribution. Then the glottal excitation is obtained by canceling the effects of the vocal tract and lip radiation by inverse filtering. This approach is based on LP analysis [27] in order to estimate the vocal tract filter function.

4. Numerical experiments

In this section, we will examine two vowels segments produced by adult speakers: the vowel [i], similar to the vowel sound in the English word *meet*, and the vowel [a], similar to the vowel sound in *bath*. These two vowels were chosen because they represent very different articulations: the former has a low first formant ($F1$) and high second formant ($F2$), whereas the latter has a high $F1$ and low $F2$.

All the computations were performed using Matlab version 2014b.

4.1. Simulation of the data

The forward problem (1) was computed with the following two functions, the *synthetic glottal flow*: created with sampling frequency $f_s = 16\,\text{kHz}$, computed using the LF model with the following parameters $t_p = 0.47$, $t_e = 0.65$, $t_a = 0.01$, and $t_c = 1$; and the *vocal tract transfer*

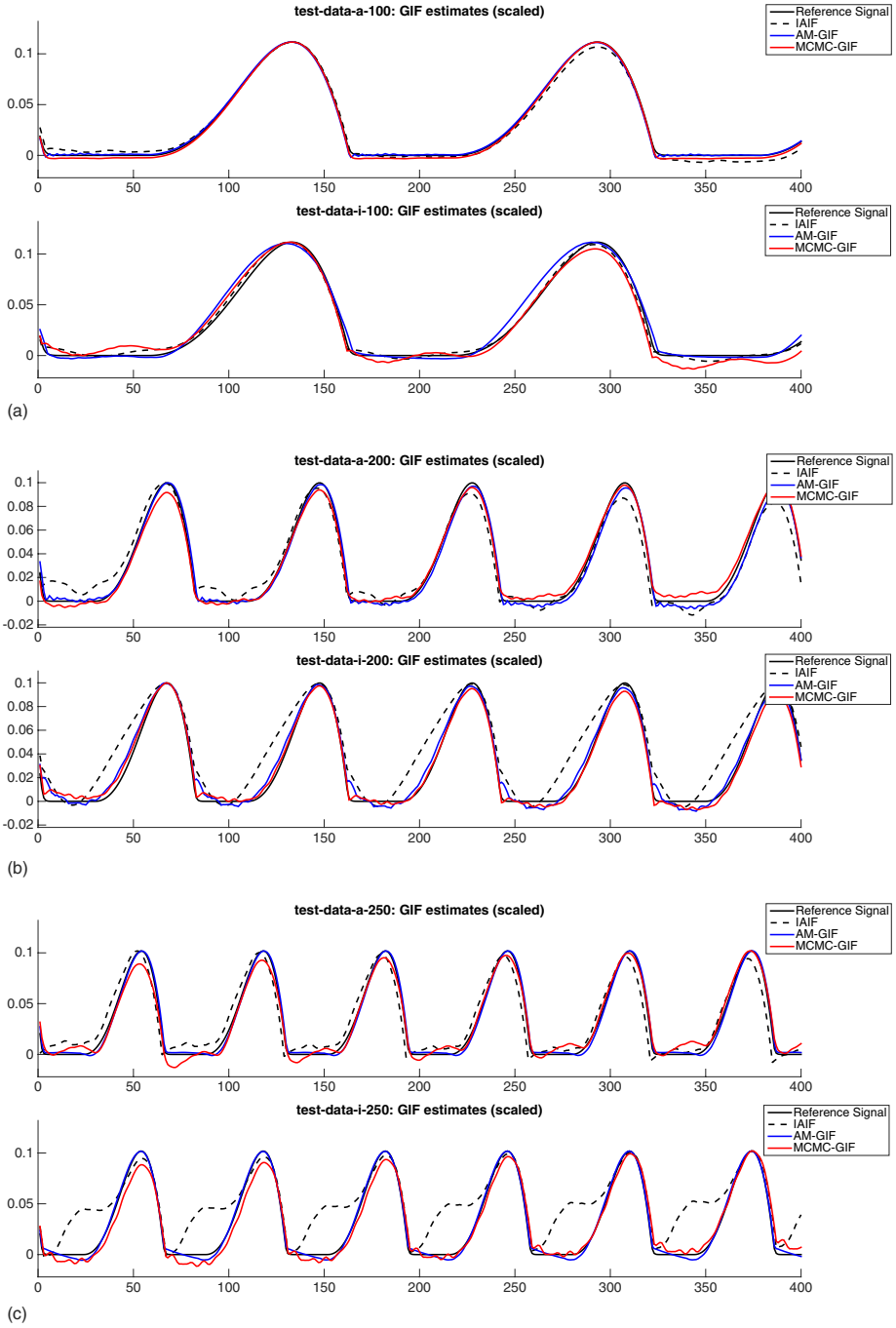

Figure 5. Reference glottal flow and three estimated glottal flows computed by IAIF, MCMC-GIF and AM-GIF for three values of $F0$ (100 Hz, 200 Hz, 250 Hz) and for two vowels ([a] and [i]). (a) (a) Vowel [a] (top) and vowel [i] (bottom) with $F0 = 100$ Hz. (b) (b) Vowel [a] (top) and vowel [i] (bottom) with $F0 = 200$ Hz. (c) (c) Vowel [a] (top) and vowel [i] (bottom) with $F0 = 250$ Hz.

Table 1. Estimation errors computed with three measures—the L_2 relative error norm (top), H1–H2 (middle), and the NAQ (bottom)—comparing the three GIF methods (IAIF, AM-GIF, MCMC-GIF) using two vowels ([a], [i]) and three values of $F0$ (100 Hz, 200 Hz, 250 Hz).

	(a) L_2 relative error norm for vowel [a] (left) and vowel [i] (right)					
	100 Hz	200 Hz	250 Hz	100 Hz	200 Hz	250 Hz
IAIF	0.007 50	0.094 13	0.079 39	0.008 43	0.035 71	0.258 36
AM-GIF	0.004 25	0.005 44	0.000 33	0.031 89	0.010 98	0.021 28
MCMC-GIF	0.041 40	0.059 80	0.061 30	0.006 48	0.037 65	0.058 04
	(b) H1–H2 error for vowel [a] (left) and vowel [i] (right)					
	100 Hz	200 Hz	250 Hz	100 Hz	200 Hz	250 Hz
IAIF	0.06 dB	0.76 dB	0.79 dB	0.10 dB	5.27 dB	6.14 dB
AM-GIF	0.08 dB	0.11 dB	0.32 dB	2.41 dB	3.09 dB	1.53 dB
MCMC-GIF	0.03 dB	0.01 dB	1.25 dB	0.34 dB	1.09 dB	1.26 dB
	(c) NAQ error for vowel [a] (left) and vowel [i] (right)					
	100 Hz	200 Hz	250 Hz	100 Hz	200 Hz	250 Hz
IAIF	0.8%	0.5%	8.3%	2.2%	25.9%	7.7%
AM-GIF	1.4%	4.5%	4.8%	0.6%	12.9%	4.9%
MCMC-GIF	0.8%	0.01%	4.9%	2.0%	3.2%	5.0%

function, computed using 20 poles for the all-poles vocal tract filter (see section 2.3), provided by the MCMC-GIF method from a male voice sample where $F0 = 100$ Hz. We repeat the latter process for both the vowel [a] and [i]. For each fixed articulation, we evaluate the forward problem with glottal airflow pulses of three different $F0$ values: 100 Hz, 200 Hz, and 250 Hz. These three $F0$ values correspond roughly to average pitch in the speech of men, women, and children, respectively.

Finally, note that the forward problem is computed via the LF model, which has more degrees of freedom than the RK model. Therefore, the LF pulse is able to better capture the dynamics of the natural glottal airflow. In doing so, we also avoid the 'inverse crime', since the inversion does not utilize the LF model in any form. In this article, the computations were done using the noise level $\epsilon = 0.2873$ for $F0 = 100$ Hz, $\epsilon = 0.3404$ for $F0 = 200$ Hz and $\epsilon = 0.3753$ for $F0 = 250$ Hz, and the noise level $\delta = 0.0001$ for all frequencies, as white noise.

4.2. Inversion by AM-GIF

Algorithm [alg:amgif]1 is relative easy to implement and fast. Most of its CPU time is required to pre-compute the C matrix, but it can be stored and loaded for future experiments, with the same wavelet maximal level J. In the numerical tests, we set $J = 8$ for the Haar wavelet basis. Some additional sub-routines (available in the Matlab wavelet toolbox as `cwtft` and `icwtft`) are required to compute the wavelet coefficients and the inverse transformation.

Once matrix C is available, matrices A and B are computed for each step respectively through (8) and (9). Each step requires solving a linear system in order to recover the wavelet coefficients of both the filter and characterizing function. For the former, we need an operator L to enforce the shape of the filter function (fading out to zero in the time domain). For the latter reconstruction we use the classical regularization term with the identity operator. Moreover, for the minimization problem (10*b*), the required approximation $y_\epsilon = F(p_\epsilon)$ is indeed fundamental. In our experiments, the best reconstruction quality was obtained by a

hybrid approximation. The hybrid approximation was obtained by splitting the signal into two parts: the CP and the OP. In the CP, we used the RK model. In the OP, we used the structure given by our first reconstruction of the filter function.

To solve both linear systems we used `ldl` Matlab's built-in factorization. Overall, the algorithm also required scaling and regularization parameters, which were heuristically chosen and fixed for all cycles.

4.3. Inversion by MCMC-GIF

The MCMC-GIF algorithm is based on a modern variant of the Metropolis-Hastings algorithm called DRAM [20]. The implementation used in the present study was taken from the MCMC Matlab package [24]. MCMC-GIF requires an initial guess for the vocal tract filter, which is provided by the IAIF method. MCMC-GIF has shown good performance, particularly for voices of high pitch. However, the computational cost is high: running the MCMC-GIF algorithm typically takes several hours using a single core processor. During each experiment the MCMC-GIF algorithm produced 10^5 samples. The CPU time could be reduced with parallel coding and by reducing the number of samples.

4.4. Inversion by IAIF

The IAIF algorithm [3] can be easily implemented, mainly with two build-in Matlab functions: the `lpc` and `filter`. IAIF is the fastest of the three GIF methods compared in the current study and it requires just a few parameter settings (for further details, see REFN). The estimation quality of IAIF is generally good for speech signals of low pitch but the method suffers from poor performance in the estimation of the glottal flow of high-pitched speech.

4.5. Comparison of the GIF approaches

Figure 5 displays the numerical reconstructions of the glottal flow for both [a] and [i] and for all three $F0$ values. The numerical errors are displayed in table 1.

In order to quantify the obtained glottal airflows, two special measurements that are widely used in the speech processing community were adopted in addition to the standard error measured according to the norm of the space (table 1(a)). The first one, H1–H2, measures the spectral tilt of the glottal flow. It is defined in the dB scale as the difference between the amplitude of the first and second harmonic of the glottal flow spectrum. The second one, the normalized amplitude quotient (NAQ), was introduced in [5]. It measures the relative time duration of the glottal closing phase from the ratio of the peak flow and the negative peak amplitude of the glottal flow derivative, normalized with respect to the length of the fundamental period. Estimation errors measured by H1–H2 and NAQ are displayed in tables 1(b) and (c).

5. Conclusion

A non-invasive inversion method, called AM-GIF, was proposed for the estimation of the glottal flow and vocal tract of a given speech waveform.

The complexity of solving the blind deconvolution problem involves recovering two variables through a nonlinear and nonconvex minimization, which is a nonlinear ill-posed inverse problem. A double regularization strategy was successfully applied, solving instead two linear and convex problems. Additionally, in order to reconstruct the functions with more desirable

properties, the alternating minimization technique gave broader control in each step, tailoring constraints and specific regularization parameters. We also remark on the crucial role of the approximation p_ϵ and the insight to create a hybrid signal, reflecting the reconstruction quality of the glottal flow.

One of the main benefits of the developed alternating minimization technique is to preserve the good quality of the MCMC-GIF method for high-pitched signals using a fraction of the computation time. With high $F0$, the new AM-GIF method was found to yield reconstruction quality close to MCMC-GIF and better than IAIF while reducing the computational burden of MCMC-GIF by two orders of magnitude. The new method clearly showed a much better fit to the reference in the closed phase of the glottal cycle. This happened particularly for the vowel [i] when the pitch was high, which, importantly, represents difficult material for all GIF methods (due to combination of low $F1$ and high $F0$). The objective measures, the NAQ and H1–H2, respectively focus mainly on the temporal and spectral properties of the glottal closing phase. Hence, they are unable to take into account the behavior of the flow waveform in the glottal open phase where GIF-AM shows, as demonstrated visually by figure 5, the best fit to the LF pulse.

The GIF algorithm proposed in the current study can be developed further in several directions. For instance, changing the wavelet family could improve the accuracy of the inverse problem solution. In addition, studying more complex regularization terms could help when searching for improvements in the overall quality.

Finally, we hope that the current study is capable of awakening the general interest of the mathematical inverse problems community in the highly interesting, yet difficult topic of human speech production.

Acknowledgments

The authors would like to thank Ahmed Geneid, MD, hospital district of Helsinki and Uusimaa, for his time, discussion, and insights that enriched this research topic. The work of I.R.B. was supported by CNPq (Brazilian National Council for Scientific and Technological Development, grant 249845/2013-0). The work of H.A. was supported by the Academy of Finland (LASTU program on computational science, project 134868). Three of the authors (I.R.B., L.L., and S.S.) were supported by the Academy of Finland (Finnish Centre of Excellence in Inverse Problems Research 2012–2017, decision 250215). The work of the authors M A and P A was supported by Academy of Finland (projects 256961 and 284671).

Appendix. Computation of the operator *C*

Decompose

$$p = \sum_{\eta\in\Lambda} p_\eta \varphi_\eta \quad \text{and} \quad f = \sum_{\lambda\in\Lambda} f_\lambda \varphi_\lambda,$$

where $p_\eta = \langle p,\ \varphi_\eta \rangle$ and $f_\lambda = \langle f,\ \varphi_\lambda \rangle$, computed with the same maximal wavelet level J so that they have equal size.

Combining the previous decomposition with equation (3),

$$\begin{aligned}(Kf)(t) &= \int_0^1 p(t-s)f(s)\mathrm{d}s \\ &= \int_0^1 p(t-s)\sum_{\lambda\in\Lambda} f_\lambda\varphi_\lambda(s)\mathrm{d}s \\ &= \int_0^1 \sum_{\eta\in\Lambda} p_\eta\varphi_\eta(t-s)\sum_{\lambda\in\Lambda} f_\lambda\varphi_\lambda(s)\mathrm{d}s \\ &= \sum_{\eta\in\Lambda} p_\eta \sum_{\lambda\in\Lambda} f_\lambda \int_0^1 \varphi_\eta(t-s)\varphi_\lambda(s)\mathrm{d}s\end{aligned}$$

we get

$$Kf = \sum_{\eta\in\Lambda} p_\eta \sum_{\lambda\in\Lambda} f_\lambda\left(\varphi_\eta * \varphi_\lambda\right).$$

We are interested to find the coefficients of the function $m = Kf$, denoted by d. Therefore we apply the operator F defined in (7)

$$d = \{\langle Kf\ ,\ \varphi_\mu\rangle\}_{\mu\in\Lambda}.$$

Let's give a close look at each component of d

$$\begin{aligned}d_\mu &= \langle Kf\ ,\ \varphi_\mu\rangle \\ &= \left\langle \sum_{\eta\in\Lambda} p_\eta \sum_{\lambda\in\Lambda} f_\lambda\left(\varphi_\eta * \varphi_\lambda\right)\ ,\ \varphi_\mu \right\rangle \\ &= \sum_{\eta\in\Lambda} p_\eta \sum_{\lambda\in\Lambda} f_\lambda\langle \varphi_\eta * \varphi_\lambda\ ,\ \varphi_\mu\rangle.\end{aligned}$$

Notice that the summations above are finite, both on η and λ, namely $1 \leqslant \eta \leqslant 2^{J+1}$ and $1 \leqslant \lambda \leqslant 2^{J+1}$. Identically $1 \leqslant \mu \leqslant 2^{J+1}$.

For a fixed μ we expand the summation on η in order to get better representation. We denote $n := 2^{J+1}$.

$$\begin{aligned}d_\mu = p_1 \sum_{\lambda\in\Lambda} f_\lambda\langle \varphi_1 * \varphi_\lambda\ ,\ \varphi_\mu\rangle + p_2 \sum_{\lambda\in\Lambda} f_\lambda\langle \varphi_2 * \varphi_\lambda\ ,\ \varphi_\mu\rangle \\ + \cdots + p_n \sum_{\lambda\in\Lambda} f_\lambda\langle \varphi_n * \varphi_\lambda\ ,\ \varphi_\mu\rangle\end{aligned}$$

Expanding on λ

$$\begin{aligned}d_\mu = p_1\left(f_1\langle\varphi_1 * \varphi_1\ ,\ \varphi_\mu\rangle + f_2\langle\varphi_1 * \varphi_2\ ,\ \varphi_\mu\rangle + \cdots + f_n\langle\varphi_1 * \varphi_n\ ,\ \varphi_\mu\rangle\right) \\ + p_2\left(f_1\langle\varphi_2 * \varphi_1\ ,\ \varphi_\mu\rangle + f_2\langle\varphi_2 * \varphi_2\ ,\ \varphi_\mu\rangle + \cdots + f_n\langle\varphi_2 * \varphi_n\ ,\ \varphi_\mu\rangle\right) \\ \vdots \\ + p_n\left(f_1\langle\varphi_n * \varphi_1\ ,\ \varphi_\mu\rangle + f_2\langle\varphi_n * \varphi_2\ ,\ \varphi_\mu\rangle + \cdots + f_n\langle\varphi_n * \varphi_n\ ,\ \varphi_\mu\rangle\right)\end{aligned}$$

This summation is easily understood as a product vector-matrix-vector

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}^t \begin{pmatrix} \langle \varphi_1 * \varphi_1 \ , \ \varphi_\mu \rangle & \langle \varphi_1 * \varphi_2 \ , \ \varphi_\mu \rangle & \cdots & \langle \varphi_1 * \varphi_n \ , \ \varphi_\mu \rangle \\ \langle \varphi_2 * \varphi_1 \ , \ \varphi_\mu \rangle & \langle \varphi_2 * \varphi_2 \ , \ \varphi_\mu \rangle & \cdots & \langle \varphi_2 * \varphi_n \ , \ \varphi_\mu \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \varphi_n * \varphi_1 \ , \ \varphi_\mu \rangle & \langle \varphi_n * \varphi_2 \ , \ \varphi_\mu \rangle & \cdots & \langle \varphi_n * \varphi_n \ , \ \varphi_\mu \rangle \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}.$$

Defining the coefficient vectors

$$y = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}, x = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}$$

and the matrix C_μ, where the element on the row η and column λ is defined as

$$[C_\mu]_{\eta,\lambda} = \langle \varphi_\eta * \varphi_\lambda \ , \ \varphi_\mu \rangle,$$

where $*$ represents the convolution operator over the whole interval [0,1] and the inner product is the standard L_2 inner product.

In summary, the sequence $d = \{d_\mu\}_\mu$ is computed by

$$y^t C_\mu x = d_\mu.$$

Notice that, depending on the wavelet choice, the matrix C_μ will be symmetric for each μ, since the convolution operator is commutative $\varphi_\mu * \varphi_\lambda = \varphi_\lambda * \varphi_\mu$, and therefore only half of the computation will be required.

References

[1] Aktosun T 2005 Inverse scattering for vowel articulation with frequency-domain data *Inverse Problems* **21** 899–914

[2] Aktosun T 2007 Inverse scattering to determine the shape of a vocal tract *The Extended Field of Operator Theory* (*Operator theory, advances and applications* vol 171) (Basel: Birkhäuser) pp 1–16

[3] Alku P 1992 Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering *Speech Commun.* **11** 109–18

[4] Alku P 2011 Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications *Sadhana* **36** 623–50

[5] Alku P, Bäckström T and Vilkman E 2002 Normalized amplitude quotient for parameterization of the glottal flow *J. Acoust. Soc. Am.* **112** 701–10

[6] Auvinen H, Raitio T, Airaksinen M, Siltanen S, Story B H and Alku P 2014 Automatic glottal inverse filtering with the markov chain monte carlo method *Comput. Speech Lang.* **28** 1139–55

[7] Bleyer I R and Leitão A 2015 *Novel Regularization Methods for ill-posed problems in Hilbert and Banach spaces* (*30 Colóquio Brasileiro de Matemática* vol 121) (Rio de Janeiro: Publicações Matemáticas do IMPA)

[8] Bleyer I R and Ramlau R 2013 A double regularization approach for inverse problems with noisy data and inexact operator *Inverse Problems* **29** 025004

[9] Bleyer I R and Ramlau R 2015 An alternating iterative minimisation algorithm for the double-regularised total least square functional *Inverse Problems* **31** 075004

[10] Cheikhrouhou I, Atitallah R B, Ouni K, Hamida A B, Mamoudi N and Ellouze N 2004 Speech analysis using wavelet transforms dedicated to cochlear prosthesis stimulation strategy *First Int. Symp. on Control, Communications and Signal Processing* pp 639–42

[11] Drugman T, Bozkurt B and Dutoit T 2012 A comparative study of glottal source estimation techniques *Comput. Speech Lang.* **26** 20–34

[12] Fant G 1970 *Acoustic Theory of Speech Production* (Mouton: The Hague)
[13] Fant G 1971 *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations* (*Description and Analysis of Contemporary Standard Russian*) (Boston: De Gruyter)
[14] Fant G 1995 The LF-model revisited. Transformations and frequency domain analysis *STL-QPSR* **36** 119–56
[15] Fant G, Liljencrants J and Lin Q 1985 A four-parameter model of glottal flow *STL-QPSR* **26** 1–13
[16] Flanagan J L 1972 *Speech Analysis, Synthesis and Perception* (*Kommunikation und Kybernetik in Einzeldarstellungen*) (New York: Springer)
[17] Forbes B J, Pike E R, Sharp D B and Aktosun T 2006 Inverse potential scattering in duct acoustics *J. Acoust. Soc. Am.* **119** 65–73
[18] Fu Q and Murphy P 2006 Robust glottal source estimation based on joint source-filter model optimization *IEEE Trans. Audio Speech Lang. Process.* **14** 492–501
[19] Golub G H, Hansen P C and O'leary D P 1999 Tikhonov regularization and total least squares *SIAM J. Matrix Anal. Appl* **21** 185–94
[20] Haario H, Laine M, Mira A and Saksman E 2006 DRAM: efficient adaptive MCMC *Stat. Comput.* **16** 339–54
[21] Kako T and Touda K 2005 Numerical approximation of Dirichlet-to-Neumann mapping, its application to voice generation problem *Domain Decomposition Methods in Science, Engineering* (*Lecture Notes in Computational Science and Engineering* vol 40) (Berlin: Springer) pp 51–65
[22] Kako T and Touda K 2006 Numerical method for voice generation problem based on finite element method *J. Comput. Acoust.* **14** 45–56
[23] Klatt D H and Klatt L C 1990 Analysis, synthesis, and perception of voice quality variations among female and male talkers *J. Acoust. Soc. Am.* **87** 820–57
[24] Laine M 2013 MCMC toolbox for Matlab http://helios.fmi.fi/~lainema/mcmc/
[25] Leonov A S and Sorokin V N 2012 On the uniqueness of determination of a vocal source from a speech signal and formant frequencies *Doklady Mathematics* vol 85 (Berlin: Springer) pp 432–5
[26] Lu S, Pereverzev S V and Tautenhahn U 2008 Dual regularized total least squares and multi-parameter regularization *Comput. Methods Appl. Math.* **8** 253–62
[27] Makhoul J 1975 Linear prediction: a tutorial review *Proc. IEEE* 63 561–80
[28] Mallat S 1998 *A Wavelet Tour of Signal Processing* (San Diego: Academic)
[29] Milenkovic P 1986 Glottal inverse filtering by joint estimation of an ar system with a linear input model *IEEE Trans. Acoust. Speech Signal Process.* **34** 28–42
[30] Ng M K, Plemmons R J and Qiao S 1997 Regularized blind deconvolution using recursive inverse filtering *IEEE Trans. Image Process.* **9** 1130–4
[31] Rabiner L and Schafer R 1978 *Digital Processing of Speech Signals* (Englewood Cliffs: Prentice-Hall)
[32] Raitio T 2015 Voice source modelling techniques for statistical parametric speech synthesis *Doctoral Dissertation* Aalto University Finland
[33] Raitio T, Suni A, Yamagishi J, Pulakka H, Nurminen J, Vainio M and Alku P 2011 HMM-based speech synthesis utilizing glottal inverse filtering *IEEE Trans. Audio Speech Lang. Process.* **19** 153–65
[34] Rosenberg A 1971 Effect of glottal pulse shape on the quality of natural vowels *J. Acoust. Soc. Am.* **49** 583–90
[35] Sondhi M M and Gopinath B 1971 Determination of vocal tract shape from impulse response at the lips *J. Acoust. Soc. Am.* **49** 1867–73
[36] Stevens K N 2000 *Acoustic Phonetics* (*Current studies in linguistics series*) (London: MIT Press)
[37] Tan B T, Lang R, Schröder H, Spray A and Dermody P 1994 Applying wavelet analysis to speech segmentation, classification *Wavelet Applications and volume Proc. SPIE 2242* pp 750–61
[38] Touda K 2007 Study on numerical method for voice generation problem *PhD Thesis* The University of Electro-Communications Tokyo, Japan
[39] Walker J and Murphy P J 2005 Advanced methods for glottal wave extraction *Nonlinear Analyses, Algorithms for Speech Processing, Int. Conf. on Non-Linear Speech Processing (Barcelona, Spain, 19–22 April 2005)* pp 139–49 (Revised selected papers)

[40] You Y-L and Kaveh M 1996 A regularization approach to joint blur identification and image restoration *IEEE Trans. Image Process* **5** 416–28
[41] Ziolko M, Galka J, Ziolko B, Jadczyk T, Skurzok D and Wicijowski J 2010 Automatic speech recognition system based on wavelet analysis *IEEE Fourth Int. Conf. on Semantic Computing* pp 450–1