Joint Source-Filter Optimization for Accurate Vocal Tract Estimation using Differential Evolution

Olaf Schleusing, *Member, IEEE,* Tomi Kinnunen, Brad Story and Jean-Marc Vesin, *Member, IEEE*

***Abstract*—In this work, we present a joint source-filter optimization approach for separating voiced speech into vocal tract (VT) and voice source components. The presented method is pitch-synchronous and thereby exhibits a high robustness against vocal jitter, shimmer and other glottal variations while covering various voice qualities. The voice source is modeled using the Liljencrants-Fant (LF) model, which is integrated into a time-varying auto-regressive speech production model with exogenous input (ARX). The non-convex optimization problem of finding the optimal model parameters is addressed by a heuristic, evolutionary optimization method called differential evolution. The optimization method is first validated in a series of experiments with synthetic speech. Estimated glottal source and VT parameters are the criteria used for comparison with the iterative adaptive inverse filter (IAIF) method and the linear prediction (LP) method under varying conditions such as jitter, fundamental frequency (f_0) as well as environmental and glottal noise. The results show that the proposed method largely reduces the bias and standard deviation of estimated VT coefficients and glottal source parameters. Furthermore, the performance of the source-filter separation is evaluated in experiments using speech generated with a physical model of speech production. The proposed method reliably estimates glottal flow waveforms and lower formant frequencies. Results obtained for higher formant frequencies indicate that research on more accurate voice source models and their interaction with the VT is necessary to improve the source-filter separation. The proposed optimization approach promises to be a useful tool for future research addressing this topic.**

***Index Terms*—Global optimization, differential evolution, joint source-filter optimization, glottal inverse filtering, time-varying vocal tract estimation.**

I. Introduction

Decomposition of speech into voice source and articulation components is potentially useful in areas such as speech coding and analysis [1], parametric speech synthesis [2], remote and/or non-invasive voice disorder diagnosis [3], restoration of pathological voices [4] or as front-end processing for classification tasks such as speaker verification [5].

In conventional speech analysis schemes, voiced speech is analyzed on a frame-by-frame basis. The speech output is represented by a linear source-filter model as periodic glottal volume velocity waveforms[1] and their respective vocal tract (VT) resonances, followed by a lip radiation filter [6]. The spectral tilt of the glottal source $G(j\omega)$ is typically approximated using a second order low-pass filter having a spectral slope of -12 dB per octave [6]. The lip radiation $L(j\omega)$ is typically represented by a single-order differentiator with a spectral slope of $+6$ dB per octave. The filter $A(j\omega)$ models the resonances in the vocal tract, known as *formants*. During speech production, the filter is continuously modified by the speaker to shape the speech signal into a blended stream of speech sounds. However, $A(j\omega)$ is assumed to vary slowly enough to be considered time-invariant for the duration of an analysis frame, which is typically 20 to 30 milliseconds. Speech may then be represented by

$$S(j\omega) = G(j\omega) \cdot A(j\omega) \cdot L(j\omega). \tag{1}$$

Because of their time-invariance, $G(j\omega)$ and $L(j\omega)$ may be represented jointly by a single order filter with a net slope of -6 dB per octave. In many analysis methods, their joint effect is then cancelled by a single order *pre-emphasis* (PE) filter ($+6$ dB per octave). The PE filter effectively captures the time-average of the spectral contributions of the glottal source and the lip radiation.

Using this simple model of speech, the estimation of the glottal signal may be achieved in a straight-forward manner using *glottal inverse filtering* [7]. This involves filtering of the speech signal with the inverse of an estimate of $A(j\omega)$. The standard tool for obtaining this estimate is *linear prediction*, which assumes that the vocal tract can be represented by an all-pole filter and also that the input to the vocal tract filter after PE is spectrally white [8].

While for most vowels $A(j\omega)$ varies sufficiently slowly to be considered time-invariant during an analysis frame, this is often *not* the case for the glottal source $G(j\omega)$. According to the myoelastic-aerodynamic theory of voice production, the voice source is mainly affected by the sub-glottal air pressure, the tension of the vocal folds and the physiological configuration of the speech production organs [9]. Various combinations of these variables produce diverse glottal waveforms that are perceived as different

Manuscript received April 12, 2012; revised October 13, 2012 and February 14, 2013, accepted March 11, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Søren Jensen.

O. Schleusing was with the Department of Systems Engineering of CSEM, Neuchâtel, Switzerland and was enrolled in the EDIC Doctoral School of the Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: audio@schleusing.de).

T. Kinnunen is with the School of Computing, University of Eastern Finland, Joensuu, Finland (e-mail: tomi.kinnunen@uef.fi).

B. Story is with the Speech Acoustics Laboratory, University of Arizona, Tucson, AZ (e-mail: bstory@email.arizona.edu).

J.-M. Vesin is with the Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: jean-marc.vesin@epfl.ch).

[1]Shorter, common names for *glottal volume velocity waveform* are *glottal excitation*, *glottal source* or *glottal cycle*.

voice qualities (breathy, modal, pressed, *etc.*). For instance, variations in the period and waveform of subsequent glottal cycles are important acoustical cues often carrying prosodic and idiosyncratic information. Some voice types, in particular *pathological* voices — those produced by speakers with impaired, partially or even completely excised vocal folds as a result of laryngeal surgery — often exhibit considerable undesired inter-glottal-cycle variations. These observations imply that the glottal transfer function $G(j\omega)$ is indeed different from the residual of the conventional LP model described above and that the linear source-filter model is a simplification in several aspects.

Firstly, the glottal source is quasi-periodic in the time domain. This is reflected in the spectral domain by a sampling at multiples of the fundamental frequency, f_0. Here, $f_0 = 1/T_0$ is the rate of the vocal fold vibration, i.e. the reciprocal of the fundamental period, T_0. The formant frequencies estimated using LP tend to be biased towards the spectral peaks of nearby voice source harmonics. Thus, the vocal tract estimator is dependent on the harmonic structure of the glottal source, as can be observed in Fig. 1. This known problem of LP has motivated the development of pitch-synchronous LP (PSLP) methods [10].

Secondly, the spectral envelope of the actually observed voice source deviates largely in some voice types from the non-adaptive PE filter. This deviation may have several causes. Different voice types exhibit various degrees of spectral tilt. Also, during the glottal open phase, there exists a non-linear feedback between the pressure in the vocal tract and the glottal volume velocity waveform. As a result, the glottal source waveform is modulated by the supraglottal pressure and it exhibits ripples and a glottal formant that is not accounted for by the myoelastic-aerodynamic theory of voice production [9], [11]. Hence, if the glottal source deviates considerably from the average represented by the pre-emphasis filter, a bias is introduced into the estimated LP coefficients. To address these problems, several methods were proposed in the past. A representative example is the *iterative adaptive inverse filtering* method [12], which iteratively estimates the coefficients of two filters representing glottal source and the VT.

Thirdly, the time-invariance of the PE filter inherently poses a problem for glottal sources with high variability of waveform shape between consecutive glottal cycles. Typical LP analysis frames comprise several glottal cycles and the variation in the glottal waveform shapes is averaged throughout this duration. A longer analysis frame duration would help to reduce these variations but would also impose a reduced temporal resolution of the time-varying VT envelope. An alternative approach for reducing the glottal source variability is to restrict the LP analysis window to the zero-input closed phase (CP) of the glottal cycle, as for example in the *closed-phase covariance linear prediction* (CPLP) method [13]. However, the performance of this method depends on the duration of the closed phase. Although the covariance method of linear prediction usually outperforms the autocorrelation method for short segments, the former is not guaranteed to yield a *stable* VT filter, *i.e.* that all its poles are of magnitude less than unity [14].

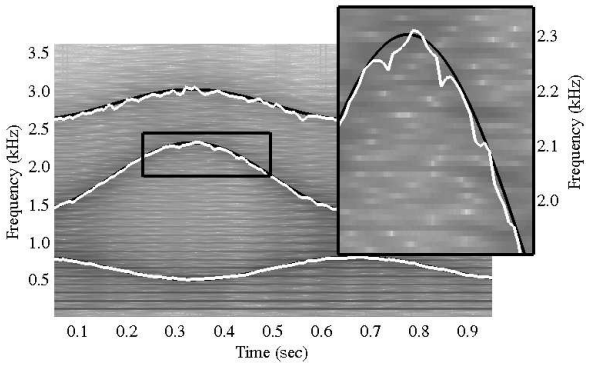

Fig. 1. Section of a spectrogram of a synthetically generated vowel transition generated using the method described in Section IV, overlaid with the true formant center frequencies (solid black lines) used for synthesis and their estimates using conventional linear prediction (solid white lines). The estimates, in particular the one of the third formant, are biased towards a lower frequency, and also exhibit a considerable variance in subsequent frames due to the underlying harmonic signal structure.

Therefore, a more complete source-filter separation may be achieved by a *joint* source-filter optimization (SFO) process, where more descriptive models of the glottal source are used to capture the glottal contribution. A variety of glottal source models has been proposed in the past. Examples are Rosenberg's model [15], which simply models the opening phase of the glottis, Klatt and Klatt's model [16] and the more complex, multi-parametric Liljencrants-Fant (LF) model [17] or different variations of these. Most glottal source models provide a parametric description of the time-domain waveform of the glottal source. They differ mainly in their complexity, *i.e.* the number of free model parameters which determine their coverage of the space of real voice source waveforms.

For the estimation of glottal models, several joint SFO methods have been proposed in the past. Lu [18] presented a convex optimization approach for optimizing a single parameter variant of the LF model for singing voice synthesis. In [19], Fu *et al.* presented a method comprising a two stage optimization, where the initial parameters for a second stage using a more complex glottal model were found in a primer convex optimization using a simplified glottal model. Jinachitra [20] presented an iterative joint estimation approach of the glottal source and vocal tract parameters using Kalman filtering and expectation-maximization algorithm. Recently, in [21], the LF model was optimized using a combinatorial search over the entire parameter space consisting of both the glottal parameters and the VT parameters. Degottex *et al.* [22] presented a novel method that minimizes the error in the phase spectrum using a single parameter voice model.

In general, it has proven difficult to estimate a non-trivial voice source model with efficient optimization procedures. Known SFO approaches typically represent a compromise between the complexity of the voice model and the efficiency of the optimization method employed. Voice models with fewer parameters are easier to optimize, but fail to accurately describe voice types observed in real speech. On the other hand, using multi-parameter voice source models usually

prohibits the usage of classical gradient-based optimization methods due to the non-convex nature of the error surface. Instead, computationally demanding methods such as exhaustive combinatorial search of the parameter space were used [21].

In this work, we propose a novel joint SFO approach, in which the voice source is modeled using the multi-parametric LF model. The proposed method is based on a pitch-synchronous analysis-by-synthesis approach, whereby a time-varying ARX model is used to generate candidate solutions. A global, population-based, stochastic direct search method called *differential evolution* (DE) is used to optimize the voice source and the VT filter parameters [23], [24]. DE has been shown to have a computational and performance advantage in many applications scenarios [25] over similar evolutionary computation methods such as particle swarm optimization (PSO) [26]. DE has also been shown to be a robust tool in the presence of parameter dependencies and non-convex error surfaces, which makes it well-suited for the optimization of multi-parametric models such as the LF model. An objective function is constructed such that reduction of the effect of inter-glottal-cycle resonances will effectively increase the duration of the analysis window. The efficiency of the DE method allowed us to carry out extensive experiments on different speech signals. The proposed optimization method converges reliably under a variety of conditions such as environmental and glottal noise, varying fundamental frequency, jitter and vowel transitions. Finally, the method is employed in a source-filter separation experiment on signals generated using a physical model of speech.

II. Speech Production Model and Joint Optimization

A. Speech Production Model

The estimates of the speech model are updated pitch-synchronously so as to capture the inter-glottal-cycle variations of the glottal source. Eq. (1) is therefore modified. The speech signal originating from a particular glottal cycle k is modeled as

$$S_k(j\omega) = e^{-j\omega t_k} G_k(j\omega) \cdot A_k(j\omega) \cdot L_k(j\omega), \qquad (2)$$

where the temporal location of the glottal cycle is determined by the linear phase component $e^{j\omega t_k}$ with a delay of t_k in seconds.

Crucially, $G_k(j\omega)$ is mixed-phase with several zeros having a magnitude greater than unity. This implies that no stable inverse representation of $G_k(j\omega)$ exists and a direct deconvolution for obtaining the vocal tract transfer function $A_k(j\omega)$ from $S_k(j\omega)$ is impossible. Therefore, the glottal source model and vocal tract coefficients are jointly estimated using a global optimization technique in an analysis-by-synthesis framework.

As such, this model of speech production is still a simplified representation of real speech. First, errors in the model of the glottal source may potentially influence the estimator. Furthermore, the anti-formants of nasal sounds that are represented by zeros in the vocal tract transfer function are currently not included in the modeling process. These issues may be addressed in the future by extending respectively the voice and vocal tract models.

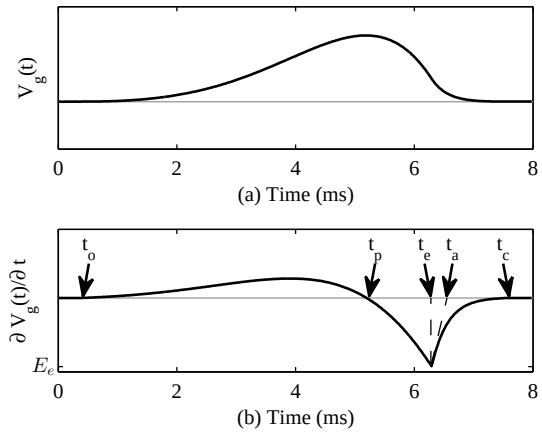

Fig. 2. Example of a glottal volume velocity waveform $V_{\mathrm{g}}(t)$ (top) and its time-derivative $v_{\mathrm{g}}(t) := \partial V_{\mathrm{g}}(t)/\partial t$ (bottom), as generated by the LF model in Eq. (3).

B. The Liljencrants-Fant Voice Source Model

The glottal excitation can be considered as a mixture of deterministic and non-deterministic components. The latter category comprises those voice source components that are not modeled deterministically, *e.g.* aspiration noise, formant ripples and other phenomena due to the non-linear coupling between vocal tract pressure and glottal volume velocity. Hereafter, we refer to the non-deterministic components as *glottal noise*.

The deterministic component originates from the periodic lateral and medial motion of the vocal folds that opens and closes the glottis. During the opening (t_{o} to t_{p}) and closing (t_{p} to t_{a}) phases of a glottal cycle, the transglottal pressure drives an air flow $V_{\mathrm{g}}(t)$ through the glottis resulting in a volume velocity waveform (see top panel of Fig. 2). Due to the vibratory dynamics of the vocal folds and the inertive properties of the lower vocal tract, the glottal volume velocity waveform typically exhibits an asymmetry in time such that the closing phase is shorter than the opening phase [27], [28].

In this study, the voice source was modeled by the *Liljencrants-Fant* (LF) model [17]. This model was chosen due to its ability to represent a wide range of natural voice variations [29]. The LF model is a piecewise-defined function serving as a parameterized representation of the *time-derivative* of the glottal flow waveform described above (see also bottom panel of Fig. 2). Its two segments are joined at the instant of the minimum of the glottal flow derivative, $t = t_{\mathrm{e}}$, which is the moment of the greatest excitation. The first segment comprises the opening phase and parts of the closing phase. It is presented as the product of a growing exponential and a low frequency sinusoid. The remaining part of the closing phase of the glottis from t_{e} until t_{c} is modeled by a decaying exponential. The time parameters t_{o}, t_{p}, t_{e} and t_{c} correspond to the instants of glottal opening, maximum glottal flow, minimum glottal flow derivative and

glottal closure, respectively. Parameter t_a is the effective return phase and is proportional to the exponential decay of the closing phase. The amplitude of the minimum glottal flow derivative is represented by E_e. The parameters $\{E_\mathrm{e}, t_\mathrm{o}, t_\mathrm{p}, t_\mathrm{e}, t_\mathrm{a}, t_\mathrm{c}\}$ can easily be identified from the glottal waveform.

For synthesis purposes it is assumed that the instantaneous fundamental frequency of the glottal cycle k is given by f_0 and the sampling rate is f_s. The number of samples representing glottal cycle k is defined by $N_\mathrm{g} = \lceil f_\mathrm{s}/f_0 \rceil$. Further, it is assumed that the glottal source signal contains no energy above the Nyquist frequency $f_\mathrm{N} = f_\mathrm{s}/2$. The timing parameters of the LF model are expressed in multiples of samples, $N_\mathrm{p} = t_\mathrm{p} \cdot f_\mathrm{s}$, $N_\mathrm{e} = t_\mathrm{e} \cdot f_\mathrm{s}$, $N_\mathrm{a} = t_\mathrm{a} \cdot f_\mathrm{s}$ and $N_\mathrm{c} = t_\mathrm{c} \cdot f_\mathrm{s}$. The shape of a single glottal cycle (see bottom panel of Fig. 2) is described by the synthesis equations, with reference to $N_\mathrm{o} = t_\mathrm{o} \cdot f_\mathrm{s} = 0$:

$$v_\mathrm{g}(n) = \begin{cases} E_0 e^{\alpha n} \sin(\omega_\mathrm{g} n), & N_\mathrm{o} = 0 \leqslant n \leqslant N_\mathrm{e} \\ -\frac{E_\mathrm{e}}{\epsilon N_\mathrm{a}} \Big[e^{-\epsilon(n-N_\mathrm{e})} & \\ \qquad -e^{-\epsilon(N_\mathrm{e}-N_\mathrm{c})} \Big], & N_\mathrm{e} \leqslant n \leqslant N_\mathrm{c} \\ 0, & N_\mathrm{c} \leqslant n \leqslant N_\mathrm{g} - 1. \end{cases} \tag{3}$$

The following relations and constraints apply:

$$\begin{aligned} \textstyle\int_0^{T_0} v_\mathrm{g}(n) dt &= 0 \\ \omega_\mathrm{g} &= \tfrac{\pi}{N_\mathrm{p}} \\ \epsilon N_\mathrm{a} &= 1 - e^{-\epsilon(N_\mathrm{c}-N_\mathrm{e})} \\ E_0 &= -\tfrac{E_\mathrm{e}}{e^{\alpha N_\mathrm{e}} \sin(\omega_\mathrm{g} N_\mathrm{e})}. \end{aligned} \tag{4}$$

The condition defined on the first line of Eq. (4) ensures that the glottal flow waveform returns to zero after each glottal cycle and is typically enforced by iteratively optimizing the damping parameter α of the exponential segment in Eq. (4) [30].

C. Formulation of the Proposed Joint Source-Filter Model

In the proposed method, speech production as introduced in Eq. (2), is modeled by a linear, time-varying, auto-regressive (AR) model with exogenous input (ARX) [31]. The glottal source of a particular glottal cycle k starting at t_{o_k} is provided by

$$v_{\mathrm{g}_k}(n) = v_\mathrm{g}(n) * \mathrm{sinc}(n - t_{\mathrm{o}_k}), \tag{5}$$

where sinc represents the cardinal sine function $\mathrm{sinc}(\cdot) = \sin(\pi\cdot)/(\pi\cdot)$ and $*$ stands for convolution. Note that using the cardinal sine function, rather than the conventionally used impulse train excitation [11], allows $v_{\mathrm{g}_k}(n)$ to be translated continuously and independent of the discrete sampling grid. The resulting speech of cycle k is then represented by the difference equation

$$\hat{s}_k(n) = -\sum_{i=1}^{p} a_{i,k} \hat{s}_k(n-i) + v_{\mathrm{g}_k}(n). \tag{6}$$

The parameter n is the discrete-time index defined in the range $0 \leqslant n \leqslant N_\mathrm{g}$ and p refers to the order of the VT filter. The coefficients $a_{i,k}$ of the ARX model are chosen to be real, therefore its poles always appear in complex conjugate pairs. Thus, p also corresponds to twice the number of formants and should generally be chosen to be even. Eq. (6) may be expressed with vector notation as

$$\hat{s}_k(n) = -\mathbf{a}_k^\top \hat{\mathbf{s}}_k^-(n) + v_{\mathrm{g}_k}(n) \tag{7}$$

with

$$\mathbf{a}_k = \begin{bmatrix} a_{1,k} & a_{2,k} & \dots & a_{p,k} \end{bmatrix}^\top$$

and

$$\hat{\mathbf{s}}_k^-(n) = \begin{bmatrix} \hat{s}_k(n-1) & \hat{s}_k(n-2) & \dots & \hat{s}_k(n-p) \end{bmatrix}^\top.$$

The error, or *residual*, between the observed speech $s_k(n)$ and the modeled speech $\hat{s}_k(n)$ is defined as

$$e_k(n) = s_k(n) - \hat{s}_k(n). \tag{8}$$

For convenience, the VT parameters are provided as formant frequencies $\mathbf{f}_\mathrm{F} = [f_{\mathrm{F}_1} \dots f_{\mathrm{F}_{p/2}}]$ and formant bandwidths $\mathbf{b}_\mathrm{F} = [b_{\mathrm{F}_1} \dots b_{\mathrm{F}_{p/2}}]$. For synthesis, they are transformed into the VT filter coefficients $\mathbf{a}_k$ by expanding the pairs of conjugate roots $[r_\mathrm{F}, r_\mathrm{F}^*]$ into a polynomial using the following relationships [32]:

$$\angle(r_{\mathrm{F}_m}) = \pm 2\pi f_{\mathrm{F}_m}/f_\mathrm{s} \quad ; 1 \leqslant m \leqslant p/2 \tag{9}$$

and

$$|r_{\mathrm{F}_m}| = e^{-\pi b_{\mathrm{F}_m}/f_\mathrm{s}} \quad ; 1 \leqslant m \leqslant p/2. \tag{10}$$

By defining the parameter vector $\theta_k = [E_\mathrm{e}\ t_\mathrm{p}\ t_\mathrm{e}\ t_\mathrm{c}\ t_\mathrm{a}\ \mathbf{f}_\mathrm{F}\ \mathbf{b}_\mathrm{F}]$, the optimization problem can now be formulated as

$$\begin{aligned} \min_{\theta_k} J(\theta_k) &= \min_{\theta_k} \left(\textstyle\sum_{n=0}^{N_\mathrm{g}} e_k^2(n) \right) \\ &= \min_{\theta_k} \left(\textstyle\sum_{n=0}^{N_\mathrm{g}} \left(s_k(n) + \mathbf{a}_k^\top \hat{\mathbf{s}}_k(n) - v_{\mathrm{g}_k}(n) \right)^2 \right), \end{aligned} \tag{11}$$

subject both to *inequality constraints* on the order of the temporal LF parameters,

$$0 < t_\mathrm{p} < t_\mathrm{e} < t_\mathrm{c} < T_0, \tag{12}$$

and to *bound constraints* on the temporal LF parameters, formant frequencies $\mathbf{f}_\mathrm{F}$ and formant bandwidths $\mathbf{b}_\mathrm{F}$ as listed in Table I and Table II.

The parameters of the LF model in Eq. (3) are not mutually independent [33]. Different combinations of parameters may describe very similar glottal source waveforms. As a result, the error surface defined by Eq. (3) and Eq. (11) is generally non-convex and may exhibit several local minima. Eq. (3) cannot be differentiated with respect to all of the real-valued LF model parameters. Hence, classical iterative gradient-based optimization methods cannot be applied. Instead, we chose a global optimization technique called *differential evolution* (DE) [34] as a computational tool to solve this optimization problem.

III. Methods

A. Differential Evolution

DE is a generic, population-based meta-heuristic optimization method belonging to the family of evolutionary algorithms (EA). It was first introduced in [23] and quickly gained large popularity in many engineering applications [24]. EAs iteratively explore the parameter space by using a *population*

TABLE I
LOWER AND UPPER BOUNDARY CONSTRAINTS OF THE CENTER FREQUENCIES (FREQ) AND BANDWIDTHS (BW) OF FORMANTS F1 TO F3 IN HZ.

Boundary	F1	F2	F3
Freq_{low}	450	1200	2500
Freq_{up}	860	2400	3100
BW_{low}	30	30	50
BW_{up}	70	80	200

TABLE II
BOUNDARY CONSTRAINTS OF THE LF PARAMETERS.

Boundary	t_o	t_p	t_e	t_a
lower	0.0	0.0	0.0	0.15
upper	10.0	60.0	90.0	10.0

of candidate solutions called parameter vectors or *agents*. Each agent is a concrete instantiation of a complete parameter set. A cost function J provides a criterion to determine the *fitness* of each agent. First, an initial *generation* of agents acting as parental population is populated with random values. In the case of DE, the next generation of agents is prepared using vector differences of randomly chosen agents from the previous generation. In particular, a randomly chosen base vector is *mutated* with a scaled population-derived difference vector constructed from two other agents, also randomly chosen from the previous generation. Furthermore, a trial agent is formed from the new mutation agent and the respective previous generation's agent. A crossover probability CR determines the ratio of parameters being used from either vector. A random number (j_rand in Alg. 1) ensures that at least one parameter from the mutation agent is utilized. If a parameter is outside the boundary constraints, it is reflected back from the bound by the amount of the violation [34]. The next parental generation is formed by member-wise comparison of the fitness function values of the current parental generation with the new generation. The new offspring is either discarded or it replaces the previous generation's agent. Eventually, a termination criterion is used to stop the optimization. This can be, for example, a previously specified cost function value or a maximum number of generations reached.

DE stands out from other EA algorithms in several aspects. DE is rather simple and straightforward to implement, yet its performance has been shown to be largely better than the also popular *particle swarm optimization* (PSO) and its variants over a wide variety of problems [25]. Another interesting aspect of DE is *contour matching*, which refers to the automatic adaptation of the difference vector population to the error function surface [34]. The mutation step size and its orientation are automatically adapted to the objective function landscape. Promising regions of the fitness landscape are investigated automatically once they are detected and a predetermined probability distribution for mutation, often introducing a bias, is not required. Price *et al.* [34] also highlighted that contour matching induces another important ingredient besides selection. It promotes *basin-to-basin* transfers, where search points may move from one basin of attraction, i.e., a local minimum, to another one. This considerably reduces both the necessity to initialize the population with approximately correct solutions and the probability of premature convergence to a local minimum.

Algorithm 1 Differential evolution

Step 1: Set control parameters crossover rate CR, difference scale factor F, population size NP and max. number of iterations, I_max.
Step 2: Initialize the agents $\mathbf{X}_{i,m}$ of the population number $i = 0$ with random values and subject to the constraints, where $m = [1, 2, \ldots, \text{NP}]$, $\mathbf{X}_{i,m} = [x_{1,i,m}, \ldots, x_{D,i,m}]$ and D is the dimension of the parameter vector.
Step 3:
while $i \leqslant$ I_max **do**
  for $m = 1$ to NP **do**
    Step 3.1: Mutation step
    Create a donor vector $\mathbf{V}_{i,m} = [v_{1,i,m}, \ldots, v_{D,i,m}]$:
    $\mathbf{V}_{i,m} = \mathbf{X}_{i,r_1^m} + \text{F} \cdot (\mathbf{X}_{i,r_2^m} - \mathbf{X}_{i,r_3^m})$
    using disjoint random indices r_1^m, r_2^m and r_3^m, each different from m
    Step 3.2: Crossover step
    Create a trial vector $\mathbf{U}_{i,m} = [u_{1,i,m}, \ldots, u_{D,i,m}]$:

$$u_{j,i,m} = \begin{cases} v_{j,i,m}, & \text{if rand[0,1]} \leqslant \text{CR or } j = j_\text{rand} \\ x_{j,i,m}, & \text{otherwise,} \end{cases}$$

    where $j_\text{rand} = [\text{rand[0,D]}]$.
    Step 3.3: Selection step
    Evaluate performance and select next generation member $\mathbf{X}_{i+1,m}$:

$$\mathbf{X}_{i+1,m} = \begin{cases} \mathbf{U}_{i,m}, & \text{if } J(\mathbf{U}_{i,m}) \leqslant J(\mathbf{X}_{i,m}) \\ \mathbf{X}_{i,m}, & \text{otherwise.} \end{cases}$$

  end for
  Step 3.4: Increase the generation count $i = i + 1$
end while

Another interesting aspect of DE with respect to the problem of source-filter separation is its performance in the presence of dependent parameters. Vincent [33] has shown that the parameters of the LF model are not entirely independent and several solutions describing similar voice source waveforms may exist. As already pointed out in Section II-C, the resulting error surface is not convex, but may exhibit local minima. This was one of the reasons why in the previous studies, simple, single-parameter voice source models were used for the joint SFO. In [34] and [35], it was demonstrated that choosing a high crossover rate CR in the range (0.9,1) for DE is a successful strategy for tackling the problem of parameter dependency. A large CR value ensures that the parameter space is propagated not only in parallel to the parameter axes. This reduces the likelihood to get trapped in local minima.

In summary, DE has only a few control parameters, namely the crossover rate CR, the difference weight F and the population size NP, which makes its application straightforward and easy. Furthermore, its algorithmic nature qualifies DE to

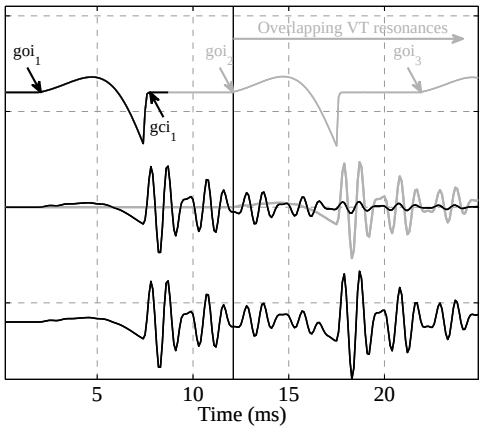

Fig. 3. Synthetic modal glottal excitations (upper graph) and their respective (middle graph) and joint (bottom graph) vocal tract resonances of a vowel */a:/*. The decaying VT resonances of the first glottal excitation (black solid line), depicted in the middle graph, clearly overlap with the subsequent glottal excitation, resulting in the commonly observed speech waveform shown in the bottom graph. The abbreviations goi_k and gci_k refer to the kth glottal opening and glottal closing instances respectively.

benefit well from the current massive trend in hardware development towards parallel computing environments [36]. The values used for the joint SFO in this paper were determined by empirical observations and set to CR $= 0.9$, F $= 0.3$ and NP $= 120$. The termination criterion used was a maximum number of iterations of I_max $= 600$. A summary of the glottal cycle optimization procedure is given in Algorithm 1.

B. Pitch-Synchronous Optimization

The aim is to use an analysis-by-synthesis approach to find the set of model parameters θ_k^* that minimizes (11) for a particular glottal cycle k. The approximate solution found for the previous glottal cycle, (θ_{k-1}^*), is used to reduce the effect of overlapping resonances.

The speech signal is first segmented into analysis frames, $s_k(n)$, the length of which correspond to the period between successive glottal opening instants (t_o in Fig. 2). It is assumed that the fundamental frequency and the location of each glottal cycle is known *a priori*. Numerous methods exist that may assist in finding these values (e.g. [37], [38]).

For the first iteration $i = 0$, an initial population of M candidate solutions $\theta_{k,i=0}^m$ with $m = [1 \,..\, M]$ is populated with random values. The temporal LF model parameters in $\theta_{k,i=0}^m$ adhere to the inequality constraints defined in Eq. (12). The boundary constraints for the parameters are listed in Table I and Table II. The values for the formant frequencies were derived from [30] and the values for the formant bandwidths were taken from [39].

Each iteration starts by calculating the cost of all population members m. Therefore, the parameter set $\theta_{k,i}^m$ is used to synthesize $\hat{s}_k^m(n)$ as defined in Eq. (6). Note that the vocal tract, represented by the first term on the right hand side of Eq. (6), is an auto-regressive structure. Vector $\mathbf{a}_k$ represents the coefficients of a recursive all-pole filter using its past output as its input. Depending on the bandwidths of the formants, the decay times of this filter are often found to be considerably longer than the fundamental period of the voice. This results in an overlapping of the resonances across subsequent glottal cycles, as illustrated in Fig. 3. A method is therefore devised that helps to decrease the influence of the resonances of previous cycles. First $\hat{s}_{k-1}^*(n+l)$ is defined to be the synthetic speech generated by approximate solution θ_{k-1}^* found for glottal cycle $k-1$. Here, l corresponds to the number of samples between the beginnings of cycles $k-1$ and k. $\hat{s}_{k-1}^*(n+l)$ is then subtracted from $s_k(n)$ before the optimization of glottal cycle k starts. Eq. (11) thus is rewritten as

$$\begin{aligned}\min_{\theta_k} J(\theta_k) &= \min_{\theta_k} \left(\textstyle\sum_{n=0}^{N_\text{g}} \left(e'(n)\right)^2 \right) \\ &= \min_{\theta_k} \big(\textstyle\sum_{n=0}^{N_\text{g}} \big(s_k(n) - \hat{s}_{k-1}^*(n+l) \\ &\quad + \mathbf{a}_k^\top \hat{\mathbf{s}}_k^-(n) - v_{\text{g}_k}(n) \big)^2 \big),\end{aligned} \tag{13}$$

where $e'(n)$ stands for the modified residual shown in (13).

Subsequently, the DE algorithm heuristics and iterations are applied until a fixed number of iterations is reached (see Section III-A). Fig. 4 provides an example of the optimization process.

IV. Experiments

A proper evaluation of source-filter separation methods is a difficult task due to the uncertainty regarding the *correct* glottal source and VT. In fact, there exists no method that allows measuring the glottal excitation directly from the human larynx while preserving natural voice production. Therefore, often synthetic speech is used in the evaluation of the performance of estimation methods. This approach may be considered problematic though, if both the synthesized samples and the evaluated method are based on the same hypothesis regarding the mechanisms of human speech production. Hence, while such synthetic vowels may be used for the validation of methodology, they are in principle insufficient for assessing the accuracy of a method that uses the same source-filter model as the one used for generating the experimental data. As an alternative to such synthetic vowels, physical models of voice production for the generation of synthetic speech samples were used in [40]. For the experiments presented in this work we chose such a model, which is described in Section II-B.

Following the above discussion, the proposed optimization method is first validated in a series of experiments using synthetic speech samples (Section IV-A). These experiments aim at investigating the behavior of the proposed method under varying environmental noise, fundamental frequency and glottal jitter. In another experiment, the effect of mismodeling the glottal source is investigated (Section IV-B). Eventually, the performance of the proposed joint SFO is evaluated using speech signals generated by the above mentioned physical model of speech.

The proposed method is compared to two other widely used methods for inverse filtering.

1) *Iterative adaptive inverse filtering (IAIF)*: IAIF was first introduced by Alku in [12] and has since found many applications, for example in speech synthesis [2]. IAIF uses an autoregressive error minimization method such

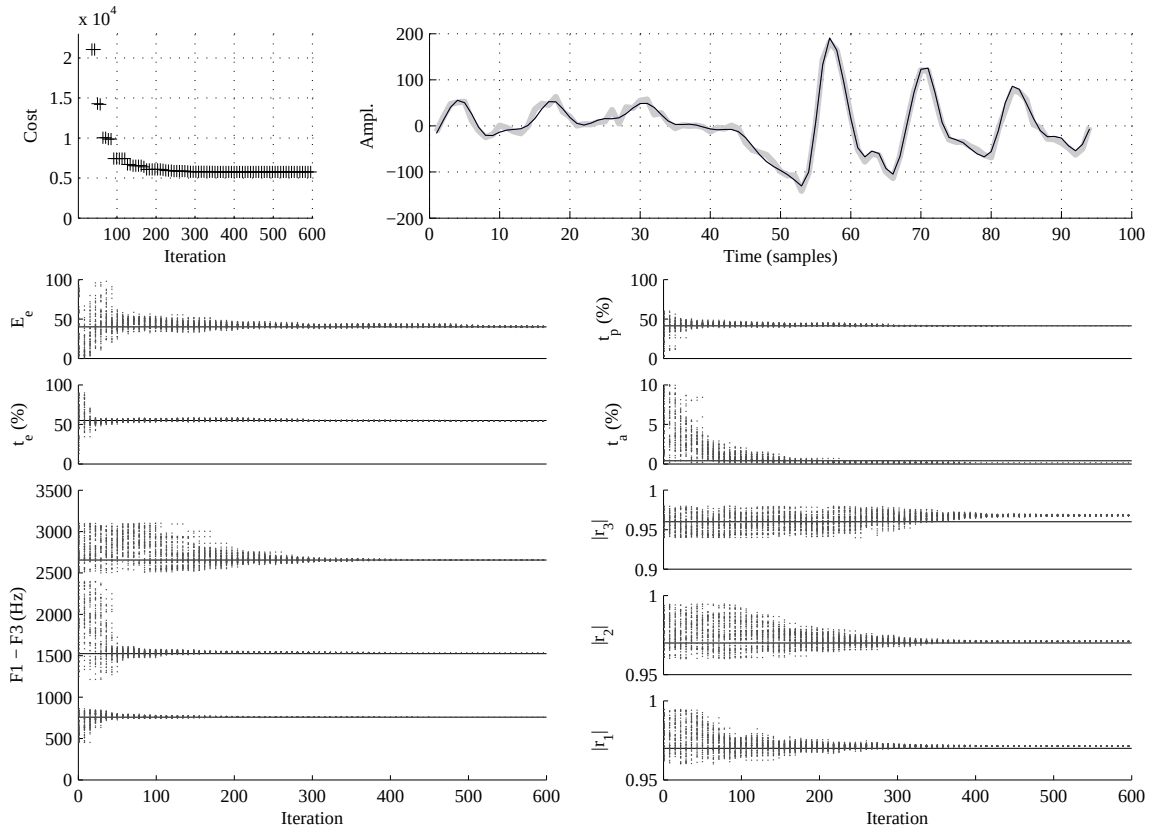

Fig. 4. Optimization of a glottal cycle of synthetic speech generated by Eq. (16), embedded in additive environmental noise of 15 dB SNR level. The thick, grey line in the right, top panel represents the original speech signal. The thin black line is the signal generated with the parameter set found by the optimization, θ_k^*. One may observe the resonances of the previous glottal cycle during the first 50 samples, which are canceled in the optimization cost function (see Section III-B, Eq. (3)). The top left panel illustrates the respective minimum cost found in each iteration. The remaining panels display scatterplots illustrating the evolution of the parameter set throughout the optimization process. The dots represent a subset of the NP population members as they converge towards the ground truth values (solid lines). F1-F3 are the formant frequencies corresponding to parameters $\mathbf{f}_\mathrm{F}$ and $|r_1|$ to $|r_3|$ are the pole radii determined by the bandwidth parameters b_1 to b_3 (see Eq. (10)). The temporal parameters are given in per cent of the glottal cycle duration.

as *discrete all-pole model* (DAP) [40] or *LP* to obtain estimates of AR models of the voice source and the VT. First, the voice source is estimated from a windowed signal segment spanning several glottal cycles using a low-order (order 1) all-pole model. After canceling the estimated effect of the source, a preliminary, higher order (order p) estimate of the vocal tract is obtained. In a second iteration, a refined estimate of each model is obtained by repetition of the first steps. The voice source model is refined by using a higher order (order g) AR model for its representation and by estimating it after cancellation of the preliminary estimate of the VT resonances from the first iteration. Again, the effect of the voice source is canceled before estimating a refined version of the VT model. In this paper, the choice of parameters was based on the values used in [40]. In particular, we used the DAP estimation method, a window length of 200 ms, $g = 2$ and $p = 10$. The windowed segments are positioned centric with respect to a glottal cycle and shifted pitch-synchronously.

2) *Linear prediction*: Linear prediction is probably the most widely used method for the estimation of vocal tract coefficients [11]. In this paper, a pre-emphasis ($b_1 = -0.98$) filter is applied. The LP window length is chosen to be 51.2 ms and the LP model order is $p = 10$. As with the IAIF method, the windowed segments are positioned in time so as to be centered on a glottal cycle and shifted pitch-synchronously.

In the following, all signals are sampled at 10 kHz.

A. Methodology Validation using Synthetic Speech

A glottal source signal is controlled by the glottal opening instant t_{o_k}, the LF model parameters contained in θ_k defined in Section II-C and a glottal noise $w^{\sigma_\mathrm{g}}(n)$ with standard deviation σ_g added to the glottal source $g(n)$:

$$g(n) = \sum_{k=0}^{K} v_{\mathrm{g}_k}(n) + w^{\sigma_\mathrm{g}}(n). \tag{14}$$

A glottal cycle $v_{\mathrm{g}_k}(n)$ is generated using Eq. (5). The aspiration noise $w^{\sigma_\mathrm{g}}(n)$ is produced by a high-pass filtered (f_c=2 kHz) white Gaussian noise that was pitch-synchronously amplitude modulated in order to create a perceptionally coherent aspiration noise, as proposed in [41]. A clean speech signal $s_\mathrm{c}(n)$ is then generated using

$$s_\mathrm{c}(n) = -\sum_{i=1}^{p} a_i(n) s_\mathrm{c}(n-i) + g(n). \tag{15}$$

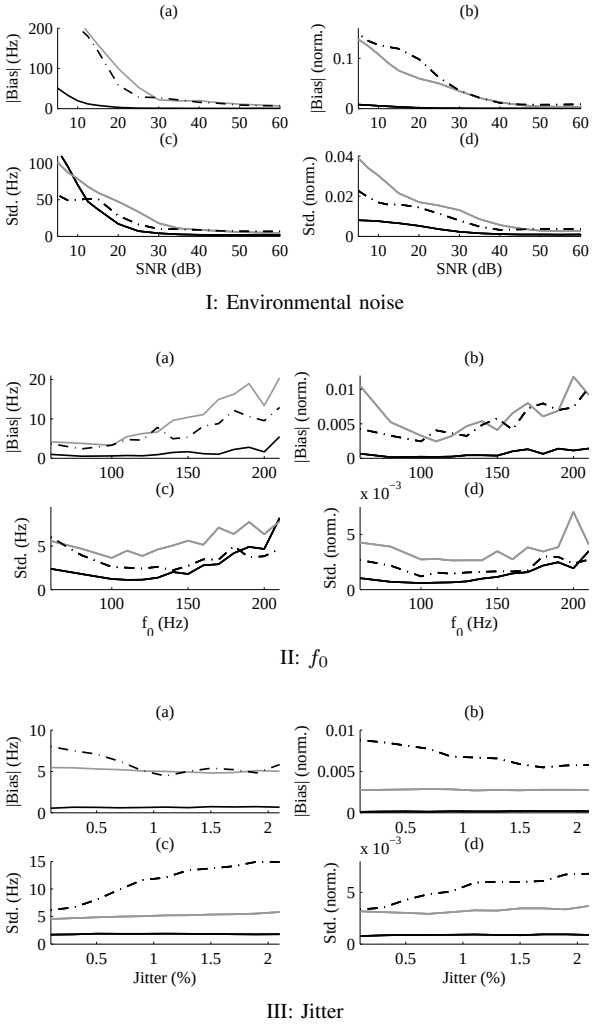

Fig. 5. Absolute value of bias (top) and variance (bottom) of the estimation errors regarding the formant frequencies (left) and the formant radii (right) as measured over a range of environmental noise levels (I), fundamental frequency (II) and jitter (III). The method based on LP (gray solid line) is most affected by noise, whereas the bias and variance of the proposed method (black solid line) outperforms the IAIF method (black dash-dot line).

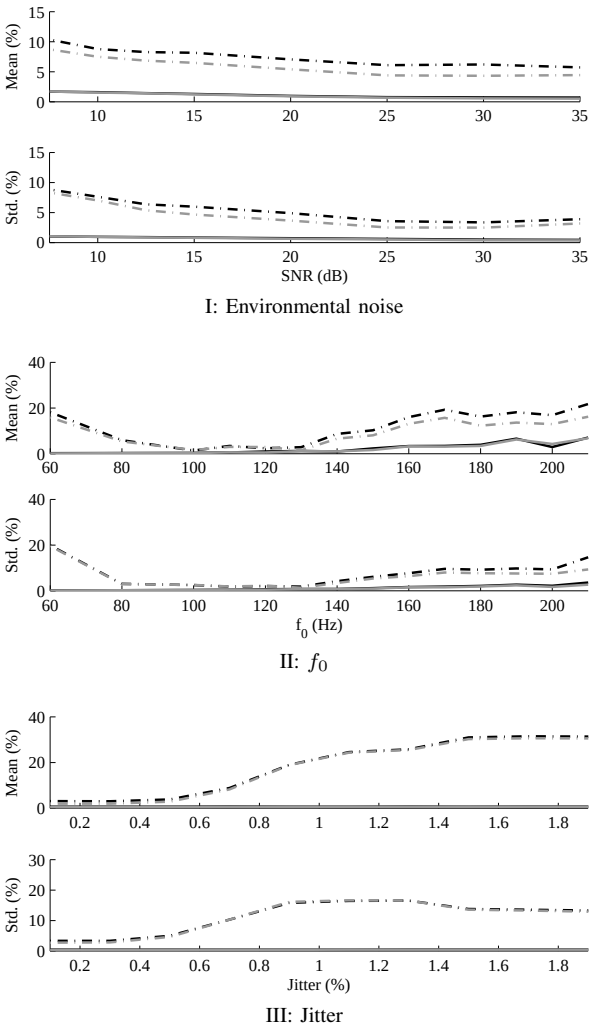

Fig. 6. Relative error of the estimated LF model parameters t_p and t_e with respect to the instantaneous glottal period. The estimates of the IAIF method are displayed using a dashed line, estimates of the proposed method are displayed in a solid lines. The black color represents the error in t_p, whereas the gray color refers to the error in t_e.

Eventually, environmental noise $w^{\sigma_\mathrm{e}}(n)$ is added to $s_\mathrm{c}(n)$ in order to emulate a real world speech recording environment. The final synthetic speech signal is then represented by

$$s(n) = s_\mathrm{c}(n) + w^{\sigma_\mathrm{e}}(n). \tag{16}$$

The noise $w^{\sigma_\mathrm{e}}(n)$ has standard deviation σ_e and was chosen to be a white Gaussian noise for mathematical convenience. The energy of either Gaussian noise source was chosen so as to obtain a particular signal-to-noise ratio ($\mathrm{SNR_g}$ and $\mathrm{SNR_e}$). The VT coefficients $a_i(n)$ are obtained by expanding the polynomial roots determined by the formant frequencies $\mathbf{f}_\mathrm{F}$ and formant bandwidths $\mathbf{b}_\mathrm{F}$, contained in θ_k, and interpolation to generate a set of coefficients at each sample, n.

Using the synthesized speech as described above, the accuracy of the proposed method is first assessed with respect to variations in *(a)* environmental noise w^{σ_e}, the *(b)* fundamental frequency and *(c)* glottal jitter. While focusing on the effect of varying one particular variable, the respective *other* variables were fixed to the following default values: $f_0 = 108\,\mathrm{Hz}$, $\mathrm{SNR_g} = 80\,\mathrm{dB}$, $\mathrm{SNR_e} = 80\,\mathrm{dB}$ and jitter $= 0.3\%$ of the fundamental period $T_0 = 1/f_0$. This jitter value was reported to be commonly found in normal phonation [42]. As test material, six samples of 2 s in duration were generated. These samples cover the range of the combinations of three vowel configurations (see Table III) and two voice types (see Table IV). For an example of a vowel transition, see Fig. 1. The LF parameters used for generating the glottal source are specified in per cent of the duration of the glottal cycles as listed in Table IV. In addition, the LF parameters obey a normal distribution with standard deviation of $2\,\%$ around these nominal values, varying from glottal cycle to glottal cycle, as described in [30]. The results for each experiment and each test configuration were averaged from 100 glottal cycles.

For an objective comparison, two types of errors related to the VT and to the glottal source are reported. First, the error on each formant frequency and formant radius relative to the ground truth is computed and averaged over all voice

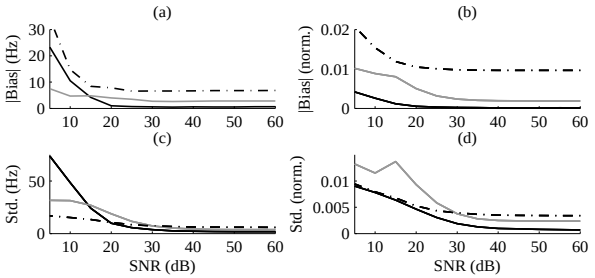

Fig. 7. Absolute value of bias (top) and variance (bottom) of the estimation errors regarding the formant frequencies (left) and the formant radii (right) as measured over a range of glottal noise levels. Down to a certain SNR level, the bias and variance of the proposed method (black solid line) show a good performance compared to the method based on LP (gray solid line) and the IAIF method (black dash-dot line) with respect to formant frequencies.

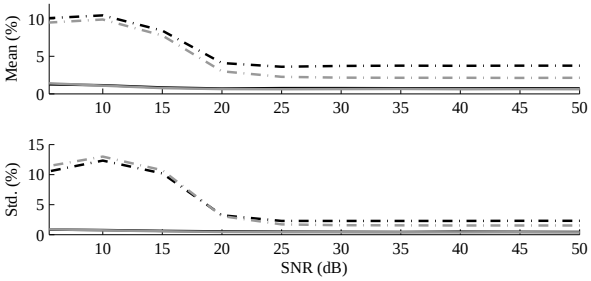

Fig. 8. Relative error of the estimated LF model parameters t_p and t_e with respect to the instantaneous glottal period. The estimates of the IAIF method are displayed using a dashed line; estimates of the proposed method are displayed using solid lines. The black color represents the error in t_p, whereas the gray color refers to the error in t_e.

types, vowel configurations and the three formants. The glottal source error is related to the shape of the extracted glottal waveform. In particular, the errors related to the temporal instants of the maximum of the glottal flow waveform and the minimum of the glottal flow derivative waveform are reported, *i.e.* t_p and t_e. Similarly to the formant errors, the voice source related errors are also averaged over all voice types and vowel configurations. The LP method is excluded from this second result, since the residual of the LP method is optimized to be spectrally white and therefore is not meant to extract the glottal waveform. The voice source-related values extracted by the IAIF method were obtained using methods found in the *Aparat* toolbox [43], [44].

(a) Environmental noise: In the first test, we assess the influence of the presence of background noise on the reliability of the proposed method. In Fig. 5 I, the absolute value of the bias (upper panels) and standard deviation (lower panels) of the estimated formant frequencies (left panels) and radii (right panels) are displayed. As expected, the formant frequencies and radii estimated by the proposed method exhibit a reduced bias compared to the other two methods. Notably, it was observed that the value of the lower formant frequencies estimated using the proposed method exhibited a high accuracy at all SNRs. For low SNR values, the estimate of the highest formant occasionally got trapped in a local minimum, which could not be prevented by an increased population size NP. This resulted in sporadic outliers of the estimated third formant. This explains the largely increased standard deviation of the average formant frequency estimates for SNR values below $15\,\mathrm{dB}$. The estimated formant radii exhibited a lower bias and standard deviation throughout all SNR values compared to the other methods.

The errors related to the glottal source temporal parameters are displayed in Fig. 6 I. The error of the proposed method is relatively small at high SNR values and steadily increases for lower SNR values. In comparison, the error of the IAIF method is generally higher and appears to be more affected by the increasing noise level.

(b) Fundamental frequency (f_0): For this experiment, synthetic vowels with different fundamental frequencies were generated. As pointed out in Section I, frame-based analysis methods may be influenced by the harmonics of the fundamental frequency of the voice source. At lower values of f_0, the estimated poles form a well-defined spectral envelope over the densely distributed f_0-harmonics. At higher values of f_0, the harmonics are sparser and thus represent single points of attraction for the poles. Thus, with rising f_0, it becomes more likely that a pole models a harmonic instead of a formant. This is what can be observed in Fig. 5 II. The error in the estimated formant frequency is increasing with higher values of f_0 for the LP and the IAIF methods, while the proposed method is unaffected up to a certain value. Above $f_0 = 200\,\mathrm{Hz}$, the error of the proposed method rises sharply due to the considerably shortened analysis window.

TABLE III
FORMANT FREQUENCIES AND BANDWIDTHS (BW) IN HZ USED FOR SYNTHESIZING THE TEST MATERIAL FOR THE FIRST TWO EXPERIMENTS.

Vowel	**F1** (Bw)	**F2** (Bw)	**F3** (Bw)
/a:/	800 (65)	1400 (68)	2600 (128)
/i/	500 (63)	2300 (78)	3000 (129)
/a:/ /i/	repeated transition through above vowels		

The results with respect to the glottal source timing parameters are displayed in Fig. 6 II. Notably, the error of the proposed method is less affected across different f_0 values and is also smaller compared to the error of the IAIF estimates.

(c) Glottal jitter: This experiment investigates the error induced by different values of jitter in the fundamental period of the voice source. Jitter is a measure of deviation from perfect harmony, *i.e.* how much a particular glottal cycle deviates from an averaged, instantaneous glottal period, T_0. Jitter is measured in percent, relative to T_0.

The results are displayed in Fig. 5 III and Fig. 6 III. An increased value of jitter does not influence the estimates of the VT nor of the glottal parameters measured by the proposed method. An increase of the standard deviation of the VT measures of the IAIF method can be observed, whereas the LP method also is not affected by jitter. The voice source related errors (t_p and t_e) are very similar in both methods (IAIF and the proposed method), but IAIF is affected by higher values of jitter. This is to be expected from a frame-based analysis method.

TABLE V
FORMANT FREQUENCY ESTIMATION RESULTS USING THE SPEECH SYNTHESIZED BY THE PHYSICAL MODEL. THE VALUES REPRESENT THE ABSOLUTE VALUE OF THE BIAS (IN Hz) FOLLOWED BY THE ERROR STANDARD DEVIATION IN PARENTHESES.

Vowel	Method	pressed			modal			breathy		
		F1	**F2**	**F3**	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**
	LP	48.4 (0.7)	74.5 (1.2)	56.4 (4.9)	0.5 (1.4)	10.0 (4.8)	39.4 (24.9)	8.3 (14.7)	84.4 (53.5)	88.9 (58.4)
/a/	IAIF	21.7 (0.5)	31.3 (0.6)	18.1 (1.5)	10.0 (0.8)	10.2 (5.5)	46.0 (11.6)	39.5 (5.9)	43.7 (27.5)	124.5 (23.7)
	DE	**3.6** (0.4)	16.2 (1.2)	9.1 (9.2)	**0.6** (0.5)	29.5 (4.7)	26.7 (19.7)	9.6 (4.9)	48.1 (22.7)	36.1 (29.7)
	LP	41.0 (0.3)	45.9 (1.9)	74.1 (3.0)	6.7 (0.4)	28.5 (7.1)	12.0 (20.1)	23.6 (1.4)	70.0 (46.0)	100.1 (67.0)
/i/	IAIF	7.3 (0.4)	6.2 (0.7)	14.8 (0.8)	5.1 (0.9)	7.8 (2.9)	18.4 (11.7)	14.3 (5.2)	10.5 (13.7)	178.8 (39.1)
	DE	**1.0** (1.0)	14.9 (13.3)	25.9 (19.1)	**3.7** (2.0)	20.8 (16.7)	40.6 (30.2)	2.3 (3.7)	53.6 (262.9)	94.9 (0.2)
	LP	43.4 (5.0)	59.8 (13.0)	64.0 (8.9)	2.1 (3.7)	14.5 (11.3)	17.3 (28.7)	14.6 (12.4)	63.1 (57.1)	113.3 (67.8)
trans. */a/*	IAIF	10.1 (13.3)	10.0 (25.6)	15.6 (8.2)	6.5 (9.3)	8.0 (23.5)	32.4 (10.6)	11.6 (23.7)	58.1 (45.6)	187.2 (56.7)
to */i/*	DE	**0.3** (5.1)	24.1 (16.2)	9.3 (22.5)	**1.5** (2.3)	26.9 (15.6)	39.5 (33.3)	9.3 (9.3)	39.9 (32.1)	80.4 (49.7)

TABLE IV
LF PARAMETERS USED FOR SYNTHESIZING THE TEST MATERIAL FOR THE FIRST TWO EXPERIMENTS IN PER CENT OF THE GLOTTAL CYCLE DURATION.

Voice Type	$\mathbf{t_p}(\%)$	$\mathbf{t_e}(\%)$	$\mathbf{t_a}(\%)$	$\mathbf{E_e}$
Modal	41.21	54.93	0.42	40.03
Harsh	25.01	29.89	0.99	39.98

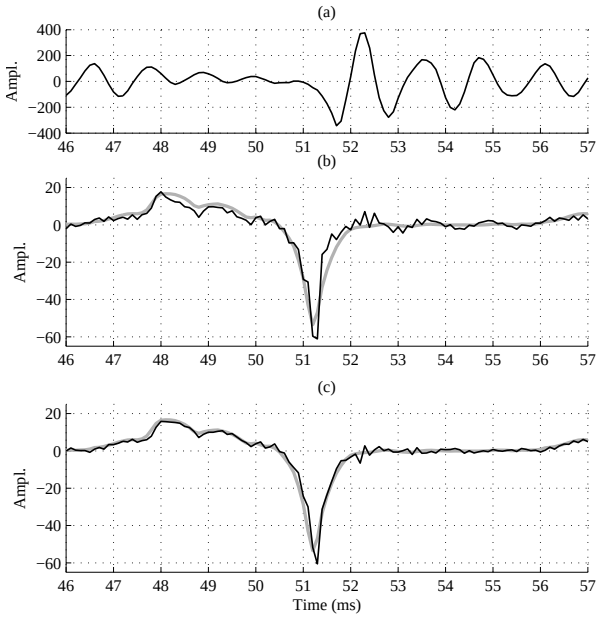

Fig. 9. Example of a speech segment of a vowel */a/* (top) synthesized with the physical model of speech. In the middle and bottom panel, the true glottal flow derivative (gray line) is shown and the inverse filter residual (black line) of the IAIF method is shown in the middle panel and the respective residual of the proposed method is shown in the bottom panel.

B. Glottal Source Distortion

This experiment addresses two issues. On one hand, glottal noise is largely composed of aspiration noise carrying idiosyncratic and semantic cues. On the other hand, glottal noise represents a distortion in the glottal source, because it is not captured in the LF model that represents only the deterministic voice source components (see Section II-B). Hence, this experiment can be seen as validation against glottal noise and voice source miss-modeling.

The results were computed in the same manner as in the previous experiment. The errors of the estimated VT envelopes and LF model parameters across a range of glottal noises w^{σ_g} are displayed in Figs. 7 and 8, respectively. For all three methods, the influence of the glottal distortion is negligible up to $\mathrm{SNR_g} = 20\,\mathrm{dB}$. The LP method is the most affected method by a further increase of the glottal noise, although mainly with respect to bias. Here, the proposed method shows a similar degree of degradation. As in the case of environmental noise, the proposed method performs well also in the presence of glottal noise with respect to the estimated formant frequencies in terms of bias, but less well in terms of the standard deviation of the estimates.

As in the previous experiments, the LF model parameters estimated by the proposed method exhibit a smaller bias and a smaller standard deviation compared with IAIF.

C. Physical Model of Speech

In our final experiment, we assess the performance of the proposed method on synthetic vowel samples, representative of an adult male speaker, generated using a physical, computational model of the speech production system. The voice source component of the model consists of a kinematic representation of the medial surfaces of the vocal folds ([45], [46]; and specifically [47]) for which surface bulging, adduction, length, and thickness are control parameters, as well as fundamental frequency. Vocal fold length and thickness are set to be 1.6 cm and 0.3 cm, respectively. As the vocal fold surfaces are driven in vibration the model produces a time-varying glottal area that is coupled to the acoustic pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations [48]. The resulting glottal volume velocity is determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis.

The vocal tract shape, which extends from glottis to lips, was specified by area functions representative of */i/* and */a/* vowels, or as a transition from */i/* to */a/*, and were based

on data reported by Story [49]. The tracheal shape was also specified by an area function that extended from the glottis to bronchi [50]. Acoustic wave propagation in the subglottal and supraglottal airspaces was computed with a wave-reflection model [50], [51] that included energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips [50]. This form of the computational model was similarly used to generate synthetic speech samples for [47]; a more extensive description of the model can be found there.

The test material consisted of nine speech samples, each $0.7\,\mathrm{s}$ long. Three vowel configurations were used (*/a/*, */i/*, transition */i/* to */a/*). Of each vowel, three different realizations were synthesized using three different voice types (*pressed*, *modal* and *breathy*) and a constant fundamental frequency of $f_0 = 105\,\mathrm{Hz}$. Along with the synthesized speech, a true glottal flow signal generated by interaction with trachea and VT, as well as true formant frequencies are available. All speech samples were low-pass filtered ($f_\mathrm{c} = 4\,\mathrm{kHz}$) and downsampled to a sampling rate of $f_\mathrm{s} = 10\,\mathrm{kHz}$.

An example of an inverse filtered glottal derivative waveform is shown in Fig. 9. Both, IAIF and the proposed method, are able to retain the general waveform of the glottal source including low frequency glottal distortions (observable in the first $4\,\mathrm{ms}$ of the example). From visual inspection it is also observable that there remains slightly more high-frequency noise in the IAIF residual. In this particular example, the IAIF method did not capture all the VT resonance components in the estimated VT filter. These remaining spectral components were thus not removed by inverse filtering. A possible explanation is temporal averaging in the IAIF method. The glottal VT coefficients estimated for a speech segment may well represent the average spectra of the observed respective components, but individual glottal cycles may diverge considerably from this average. No parametric representation of the true glottal source is available, thus no objective results are reported.

In Table V, the errors related to the estimated formant frequencies are presented. In virtually all examples, the bias of the first formant (F1) estimated using the proposed method is smaller compared to that of the other two methods. The standard deviation of the F1 estimate varies from example to example but compares similar to the other methods.

In general, the proposed method performs best for pressed voice and worst for breathy voice. This is expected, since in pressed voice the instant of greatest excitation (t_e) occurs relatively early in the glottal cycle and thus a longer duration of the analysis window contains the VT resonances. Furthermore, it is well known that the duration of the return phase of the glottal source, t_a, is strongly correlated with the spectral tilt of the voice source [29]. A small value of t_a, as found in pressed voice, yields a low spectral tilt and results in the higher glottal energy in higher frequency bands. As a result, higher formants exhibit a larger SNR and are more likely to be estimated correctly.

An interesting observation concerns the error found for *higher formant* estimates. The performance of the proposed method appears to deteriorate when compared to the other methods for some configurations (*e.g.* pressed and modal voice of vowel */i/*). By inspection of the results of individual glottal cycles it was discovered that these errors were mostly introduced by outliers in formant estimation for some glottal cycles. Further inspection of glottal cycles exhibiting estimation outliers revealed that their spectra in frequency ranges corresponding to higher formants (above $2\,\mathrm{kHz}$) show considerable, cycle-specific, attenuations and amplifications in relatively narrow frequency bands. In other words, the high frequency spectral characteristics of some glottal cycles show large frequency-dependent deviations from the constant spectral decay assumed by the LF model. The LF model used in the proposed method is not capable of describing such fine details due to the constant decay in high frequencies. It may be argued that the LF model, despite its relatively high degrees of freedom, lacks in its ability to represent the details of high frequency components of the glottal source. Therefore, errors in the estimated formant frequencies are introduced.

V. Discussion and Conclusion

A novel method for robust joint source-filter optimization was proposed. The focus of our work was to combine multi-parametric voice source models with efficient, global optimization methods. In our approach, the LF model is used to represent the voice source and an ARX process models speech production. The respective optimal model parameters were found using an evolutionary algorithm named differential evolution.

A first series of experiments showed the accuracy and robustness of the proposed optimization method against a variety of changing parameters such as fundamental frequency, jitter, glottal and environmental noise. The proposed method outperformed the comparative LP and IAIF methods at formant estimation and voice source parameter estimation accuracy. In particular, the bias of the estimated parameters was shown to be largely reduced. Therefore, a promising direction for future research is the consideration of *a priori* information, for example by using probabilistic tracking schemes, to further reduce the standard deviation of the estimated parameters.

A second experiment using speech generated with a physical model of speech production revealed that the accuracy of lower formant estimates was mostly improved. In higher spectral bands, where the glottal source deviated from the spectral characteristics of the LF model, the accuracy of the formant estimation deteriorated. Nevertheless, the experiments have shown that for voices within the boundaries of the used voice source model, the proposed method is a reliable and efficient method for source-filter separation.

The proposed method has been designed with source-filter decomposition, rather than computational efficiency, in mind. The real-time factor of our mixed Matlab/C++ implementation was measured as 1:200, averaged over all the experiments. This means that, on average, it takes approximately 200 seconds to analyze 1 second of speech. Following the above considerations, the method is best suited for applications with moderate amounts of data and non-real-time requirements, such as clinical speech analysis or as a research tool in acoustic phonetics. The proposed method also represents a promising approach for research aiming at improved models of speech

production. Voice restoration, voice transformation, parametric speech coding and speaker verification may also benefit from such improved models. To be applicable for large-scale speech or speaker recognition systems, further speed optimization is necessary.

ACKNOWLEDGEMENT

The authors would like to thank *Recherche Suisse Contre le Cancer* (KFS 02681-08-2010) and the *Centre Suisse d' Electronique et de Microtechnique* for their help in enabling this research. This work was supported by the Academy of Finland (projects 132129 and 253120). Further thanks go to the reviewers whose valuable comments have significantly improved this article.

REFERENCES

[1] M. Schröder, *Affective Information Processing*. London: Springer, 2009, ch. Expressive Speech Synthesis: Past, Present, and Possible Futures, pp. 111–126.

[2] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, and P. Vainio, M. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Language Processing*, vol. 19(1), pp. 153–165, Jan. 2011.

[3] D. Hartl, S. Hans, L. Crevier Buchman, O. Laccourreye, J. Vaissiere, and D. Brasnu, "Dysphonia: current methods of evaluation," *Ann Otolaryngol Chir Cervicofac.*, vol. 122(4), pp. 163–172, 2005.

[4] O. Schleusing, R. Vetter, P. Renevey, J.-M. Vesin, and V. Schweizer, *CCIS: Biomedical Engineering Systems and Technologies*. Springer, 2011, vol. 127, ch. Prosodic Speech Restoration Device: Glottal Excitation Restoration using a Multi-Resolution Approach, pp. 177–188.

[5] M. Plumpe, T. Quatieri, and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 7(5), pp. 569–586, Sep. 1999.

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

[7] P. Alku, "Glottal inverse filtering analysis of human voice production a review of estimation and parameterization methods of the glottal excitation and their applications," *Sādhana - Academy Proceedings in Engineering Sciences*, vol. 36(5), pp. 623–650, Oct. 2011.

[8] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[9] I. Titze, *Principles of voice production*. National Center for Voice and Speech, 2000.

[10] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[11] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2001.

[12] P. Alku, E. Vilkman, and U. Laine, "Analysis of glottal waveform in different phonation types using the new IAIF-method," in *Proc. 12th lnternat. Congress Phonetic Sciences*, vol. 4, Aug. 1991, pp. 362–365.

[13] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27(4), pp. 350–355, Aug. 1979.

[14] P. Vaidyanathan, *The Theory of Linear Prediction*. Morgan & Claypool, 2008.

[15] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels,," *J. Acoust. Soc. Am.*, vol. 49, pp. 583–590, 1971.

[16] D. Klatt and L. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[17] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL–QPSR, Tech. Rep., 1985.

[18] H. Lu, "Toward a high-quality singing synthesizer with vocal texture control," Ph.D. dissertation, CCRMA, Stanford University, 2002.

[19] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14(2), pp. 492–501, Mar. 2006.

[20] P. Jinachitra, "Robust structured voice extraction for flexible expressive resynthesis," Ph.D. dissertation, CCRMA, Stanford University, 2007.

[21] P. Ghosh and S. Narayanan, "Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter," *Elsevier Speech Communication*, vol. 53(1), pp. 98–109, Jan. 2011.

[22] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 19(5), pp. 1080–1090, Jul. 2011.

[23] R. Storn and K. Price, "Differential evolution a simple and efficient adaptive scheme for global optimization over continuous spaces," ICSI, Tech. Rep., 1995.

[24] S. Das and P. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evol. Comput.*, vol. 15(1), pp. 4–31, Feb. 2011.

[25] J. Vesterstrøm and R. Thomson, "Comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems," in *Proc. IEEE Congr. Evol. Comput.*, 2004, pp. 1980–1987.

[26] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Intern. Conf. Neural Networks*, vol. IV, 1995, pp. 1942–1948.

[27] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Am.*, vol. 53(6), pp. 1632–1645, 1973.

[28] I. Titze, "The physics of small amplitude oscillation of the vocal folds," *J. Acoust. Soc. Am.*, vol. 83(4), pp. 1536–53, Apr. 1988.

[29] B. Doval and C. d'Alessandro, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, pp. 1026–1046, 2006.

[30] D. Childers, *Speech Processing and Synthesis Toolboxes*. New York: J. Wiley & Sons, Inc., 2000.

[31] L. Ljung, *System Identifcation*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[32] J. Smith, *Introduction to Digital Filters with Audio Applications*. http://www.w3k.org/books/: W3K Publishing, 2007.

[33] D. Vincent, "Analyse et controle du signal glottique en synthese de la parole (in french)," Ph.D. dissertation, ENST, Paris, France, 2007.

[34] K. Price, R. Storn, and J. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization*. Springer, 2005.

[35] J. Ronkkonen, S. Kukkonen, and K. Price, "Real parameter optimization with differential evolution," in *Proc. IEEE CEC*, 2005, pp. 506–513.

[36] H. Sutter, "Welcome to the parallel jungle!" *Dr. Dobb's Journal*, Jan. 2012. [Online]. Available: http://drdobbs.com/parallel/232400273

[37] M. Thomas, J. Gudnason, and P. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20(1), pp. 82–91, January 2012.

[38] A. Kounoudes, P. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE ICASSP*, Orlando, FL, 2002, pp. 349–352.

[39] G. Fant, "Formant bandwidth data," *Quarterly Status Report, KTH*, vol. 3(1), pp. 1–2, 1962.

[40] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production," *Folia Phoniatrica et Logopaedica*, vol. 58(2), pp. 102–113, 2006.

[41] D. Hermes, "Synthesis of breathy vowels: some research methods," *Speech Communication, Special issue on speaker characterization in speech terminology*, vol. 10(5–6), pp. 497–502, Dec. 1991.

[42] M. Brockmann, C. Storck, P. Carding, and M. Drinnan, "Voice loudness and gender effects on jitter and shimmer in healthy adults," *Journal of Speech, Language and Hearing Research*, vol. 51, pp. 1152–1160, Oct 2008.

[43] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proc. of INTERSPEECH*, Lisbon, Portugal, Sep. 2005, pp. 2145–2148.

[44] Helsinki University of Technology (HUT), TKK Laboratory of Acoustics and Audio Signal Processing, "TKK Aparat," Online Resource: http://www.acoustics.hut.fi/software/aparat/.

[45] I. Titze, *The myoelastic aerodynamic theory of phonation*. Iowa City: National Center for Voice and Speech, 2006.

[46] ——, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc. Am.*, vol. 75(2), pp. 570–580, Feb. 1984.

[47] R. Samlan and B. Story, "Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling," *J. Speech, Lang and Hear. Res.*, vol. 54(5), pp. 1267–1283, Oct. 2011.

[48] I. Titze, "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," *J. Acoust. Soc. Am.*, vol. 111(1), pp. 367–376, Jan. 2002.

[49] B. Story, "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," *J. Acoust. Soc. Am.*, vol. 123(1), pp. 327–335, Jan. 2008.
[50] ——, "Speech simulation with an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, University of Iowa, Iowa City, 1995.
[51] J. Liljencrants, "Speech synthesis with a reflection-type line analog," Ph.D. dissertation, Royal Inst. of Tech., Stockholm, Sweden, 1985.

Olaf Schleusing received the degree of Diplom-Ingenieur of Mediatechnology from the Technical University of Ilmenau in 2002. He has worked in R/D at BDTi in Berkeley, CA, at Fraunhofer IDMT in Ilmenau, Germany and at Studer Professional Audio AG in Zurich, Switzerland. He returned to academia in 2007 as assistant researcher in the Computer Science Department of the National University of Singapore (NUS). Since 2008, he was employed at CSEM in Neuchâtel, Switzerland and obtained his Ph.D. from EPFL in Lausanne, Switzerland in 2012. In that year he started working on 3D audio algorithms at SonicEmotion in Zurich, Switzerland.

His research interests are in audio signal processing, numerical optimization and algorithms. Furthermore he is interested in aspects of efficient implementations, in programming languages and lean methodologies.

Tomi Kinnunen received the M.Sc., Ph.Lic. and Ph.D. degrees in computer science from the University of Joensuu (now Univ. of Eastern Finland, UEF), Finland, in 1999, 2004 and 2005, respectively. From 2005 to 2007, he worked as an associate scientist at the Institute for Infocomm Research (I^2R), Singapore. Since 2007, he has been with UEF. From 2010 to 2012, he was funded by a post-doc grant from Academy of Finland and he currently holds position of university researcher. He serves as an associate editor in *Digital Signal Processing* (Elsevier) and he is the chair of forecoming *Odyssey 2014: the Speaker and Language Recognition Workshop*. His primary research interests include speaker recognition, speech signal processing and voice conversion.

His research interests cover speaker recognition, speech signal processing, pattern recognition and biometric person authentication.

Brad Story is an Associate Professor of Speech, Language, and Hearing Sciences. After receiving his Bachelors degree in Applied Physics from the University of Northern Iowa, Dr. Story was employed in industry as an acoustical engineer. He received his Ph.D. in Speech Science from the University of Iowa in 1995, and then conducted postdoctoral research in speech and singing at the Denver Center for the Performing Arts. Dr. Storys research is focused on the use of computer models to aid in understanding how the shapes, sizes, and movements of the larynx and vocal tract contribute to the sounds of speech and song. He and colleagues have recently begun a long term project using a computational model to study the development of speech production in children.

Jean-Marc Vesin graduated from the Ecole Nationale Supérieure d'Ingénieurs Electriciens de Grenoble (ENSIEG, Grenoble, France) in 1980. He received his M. Sc. from Laval University, Québec city, Canada, in 1984, where he spent four years on research projects. After two years in the industry, he joined the Signal Processing Institute of the Swiss Federal Institute of Technology, Lausanne, Switzerland (EPFL), where he obtained his Ph. D. in 1992. He currently heads the Applied Signal Processing Group (ASPG) at EPFL. His main interests are biomedical signal processing, adaptive signal analysis, nonlinear signal modeling and analysis, and computer modeling of heart electrical activity. He has authored or co-authored more than fifty publications in peer-reviewed journals, as well as several book chapters.