Evaluation of Glottal Inverse Filtering Algorithms Using a Physiologically Based Articulatory Speech Synthesizer

Yu-Ren Chien, Daryush D. Mehta, *Member, IEEE*, Jón Guðnason, *Member, IEEE*, Matías Zañartu, *Member, IEEE*, and Thomas F. Quatieri, *Fellow, IEEE*

***Abstract*—Glottal inverse filtering aims to estimate the glottal airflow signal from a speech signal for applications such as speaker recognition and clinical voice assessment. Nonetheless, evaluation of inverse filtering algorithms has been challenging due to the practical difficulties of directly measuring glottal airflow. Apart from this, it is acknowledged that the performance of many methods degrade in voice conditions that are of great interest, such as breathiness, high pitch, soft voice, and running speech. This paper presents a comprehensive, objective, and comparative evaluation of state-of-the-art inverse filtering algorithms that takes advantage of speech and glottal airflow signals generated by a physiological speech synthesizer. The synthesizer provides a physics-based simulation of the voice production process and thus an adequate test bed for revealing the temporal and spectral performance characteristics of each algorithm. Included in the synthetic data are continuous speech utterances and sustained vowels, which are produced with multiple voice qualities (pressed, slightly pressed, modal, slightly breathy, and breathy), fundamental frequencies, and subglottal pressures to simulate the natural variations in real speech. In evaluating the accuracy of a glottal flow estimate, multiple error measures are used, including an error in the estimated signal that measures overall waveform deviation, as well as an error in each of several clinically relevant features extracted from the glottal flow estimate. Waveform errors calculated from glottal flow estimation experiments exhibited mean values around 30% for sustained vowels, and around 40% for continuous speech, of the amplitude of true glottal flow derivative. Closed-phase approaches showed remarkable stability across different voice qualities and subglottal pressures. The algorithms of choice, as suggested by significance tests, are closed-phase covariance analysis for the analysis of sustained vowels, and sparse linear prediction for the analysis of continuous speech. Results of data subset analysis suggest that analysis of close rounded vowels is an additional challenge in glottal flow estimation.**

***Index Terms*—Glottal excitation, glottal flow estimation, inverse filtering, performance evaluation, speech analysis, speech synthesis, voice production.**

Manuscript received December 6, 2016; revised April 11, 2017 and May 24, 2017; accepted May 26, 2017. Date of publication June 12, 2017; date of current version June 30, 2017. This work was supported in part by The Icelandic Centre for Research (RANNIS) under the project *Model-Based Speech Production Analysis and Voice Quality Assessment* under Grant 152705-051, in part by the NIH National Institute on Deafness and Other Communication Disorders under Grant R21/R33 DC011588, in part by the CONICYT under Grants BASAL FB0008 and FONDECYT 1151077. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contracts FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhen-Hua Ling. *(Corresponding author: Yu-Ren Chien.)*

Y.-R. Chien and J. Guðnason are with the Center for Analysis and Design of Intelligent Agents, Reykjavik University, Reykjavik 104, Iceland (e-mail: yrchien@ntu.edu.tw; jg@ru.is).

D. D. Mehta is with the Center for Laryngeal Surgery and Voice Rehabilitation and Institute of Health Professions, Massachusetts General Hospital, Boston, MA 02114 USA, with the Department of Surgery, Harvard Medical School, Boston, MA 02115 USA, and also with the MIT Lincoln Laboratory, Lexington, MA 02421-6426 USA (e-mail: mehta.daryush@mgh.harvard.edu).

M. Zañartu is with the Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 2390123, Chile (e-mail: matias.zanartu@usm.cl).

T. F. Quatieri is with the MIT Lincoln Laboratory, Lexington, MA 02421-6426 USA (e-mail: quatieri@ll.mit.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2714839

I. Introduction

Human voice is the result of the glottal airflow exciting the vocal tract to produce the airflow through the lips and nostrils. Since the glottal airflow is modulated by the diaphragm and the vocal folds, which are in turn coordinated by the brain through motor control, an accurate estimate of the glottal airflow from a speech signal may provide salient information related to the speaker's identity, vocal function, emotions, etc. This makes glottal flow estimation desirable for speaker identification [1], voice quality assessment [2], analysis of emotional and neurological disorders [3], and clinical voice assessment [4], [5]. Nevertheless, true glottal airflow signals have been elusive not only in ecological applications, but also in experimental settings. As a result, it has been difficult for researchers to evaluate the performance of a glottal flow estimator with confidence.

This paper presents an evaluation for a special class of glottal flow estimation methods, which we refer to as *inverse filtering* algorithms. An inverse filtering algorithm typically estimates the vocal tract filter and applies the inverse of filter estimate to the speech signal to give a glottal flow estimate. It does not constrain the waveform estimate with a glottal flow model, e.g., the Liljencrants-Fant model [6]; rather, less constrained glottal-flow assumptions are made as with a typical manual inverse filtering procedure [7], [8] where an inverse filter (with user-specified formant frequencies and formant bandwidths) is manually

2329-9290 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

adjusted to give an estimate of the glottal airflow that is ripple-free in the closed phase and has a smooth spectrum envelope. Owing to this, inverse filtering algorithms are free from a performance limitation resulting from any deviation of real glottal flow characteristics from a glottal flow model, provided that an optional glottal flow modeling procedure following inverse filtering (such as the one presented in [1]) is not performed. In addition, for the glottal flow estimation techniques that are based on a glottal flow model (and thus not considered to be inverse filtering algorithms), the objective is typically to estimate only a subset of all the parameters required for glottal flow reconstruction, leaving a glottal airflow estimate not well-defined. Consequently, among all the existing approaches to glottal flow estimation, only inverse filtering algorithms are tested in this study. In the evaluation, we aim to use synthesized glottal airflow signals as a reference, test inverse filtering algorithms on corresponding speech signals, and produce an objective assessment of the overall accuracy of each glottal airflow estimate.

In the experiments presented in this paper, both continuous speech and sustained vowels are used for performance evaluation. The specific synthesis procedures adopted to generate these test materials are physiologically based, not only simulating the voice production mechanisms at the vocal fold and vocal tract levels, but also providing the ground-truth glottal airflow signals needed for the evaluation as part of the simulation. For sustained vowels, the data set includes synthesized speech utterances for various voice qualities and subglottal pressure levels. The resulting glottal airflow estimates are compared to the simulated glottal airflow signals by measuring errors in time sample values, as well as in several types of feature values extracted from the waveform. Moreover, for the inverse filtering algorithms that make use of glottal closure instants detected from the speech signal, we evaluate the robustness to the errors in glottal closure detection with a simulation, where glottal closure instants are extracted from the synthesized glottal airflow signals, perturbed with controlled errors, and used to test these algorithms. In this paper, glottal closure instant is defined for each glottal closure event as the time sample at which the glottal-flow derivative signal starts to assume the value of zero. This definition is used in the YAGA algorithm [9].

Our contribution is presented in the subsequent sections as follows. Related works are surveyed in Section II. The tested algorithms are reviewed in Section III. In Section IV, details are provided on how the sustained-vowel and continuous-speech data sets are constructed, and the performance measures used in the evaluation are also described. In Section V, results of our glottal flow estimation experiments are documented and analyzed for the tested algorithms. These results include examples that illustrate the ground-truth and estimated glottal airflow signals, as well as performance statistics calculated at the data-set level. Concluding remarks are given in Section VI.

II. Background

Glottal flow estimation is an important task in speech analysis for which performance evaluation or literature survey has been conducted in some dedicated works. Drugman *et al.* [14] evaluated three inverse filtering algorithms on real speech data with voice quality labels, as well as on synthetic speech data. Chu *et al.* [15] tested two closely-related inverse filtering algorithms with a sound-producing instrument modeled after the glottis and vocal tract. More recently, Guðnason *et al.* [16] evaluated the performance of five inverse filtering algorithms with sustained vowels generated by an articulatory speech synthesizer, VocalTractLab [17]. Concerning literature survey, Alku [18] reviewed the literature in the topics of glottal inverse filtering, parameterization of glottal flow estimates, and applications of inverse filtering, thereby concluding that the main current limitations of most inverse filtering methods are in high-pitch, running-speech, and pathological scenarios. Drugman *et al.* [19] presented a review of works on the glottal processing of speech, covering the aspects of synchronization, estimation, parameterization, and applications.

In the case of inverse filtering algorithms, the glottal flow is defined with a representation more general than a parameterized waveform. Alku [20] presented a method for glottal flow estimation that is based on representing the glottal flow with a low-order linear-predictive spectrum envelope. Wong *et al.* [21] conducted linear-predictive covariance analysis in the closed phase of glottal-flow pulse, and showed that the analysis implements least-squares estimation of the vocal tract filter, and that the closed phase can be located with a normalized error energy. Alku *et al.* [22] performed a closed-phase analysis where the inverse filter is constrained in terms of DC gain and minimum phase. They carried out performance evaluation with the vowel /a/ synthesized by a physical model of voice production that allows for simulation of the interaction between glottal source and vocal tract. To achieve better robustness to the errors in closed phase detection, Airaksinen *et al.* [23] estimated the vocal tract from both closed- and open-phase time samples with more weight on the closed-phase samples, and also evaluated their approach with physical modeling. Airaksinen *et al.* [24] recently modified the traditional closed-phase analysis by introducing an additional 1-norm term in the objective function of linear prediction. Based on the assumption of a maximum-phase signal for the open phase of glottal airflow as well as minimum-phase signals for the return phase of glottal airflow and the vocal tract impulse response, Drugman *et al.* [13] were able to estimate the open-phase glottal airflow by a causal-anticausal separation in the complex-cepstrum domain that had been applied earlier to a spectrum-envelope type of speech analysis and resynthesis by Oppenheim *et al.* [25]. In a different but related approach, Zañartu *et al.* [26] presented a non-parametric scheme to remove subglottal resonances in order to obtain glottal airflow estimates from a neck surface accelerometer. This case differs from the others in that it was designed for a different sensor and sensing position, and thus could be considered in a future evaluation.

In contrast to inverse filtering algorithms, alternative approaches jointly estimate the parameters of a glottal flow model with the parameters of a vocal tract filter. In an algorithm presented by Ding *et al.* [27], parameters were estimated from speech waveforms for the Rosenberg-Klatt (RK) glottal flow model and a time-varying pole-zero-filter vocal tract model, by Kalman filtering and simulated annealing. Lu and Smith [28] estimated parameters of the KLGLOTT88 glottal flow model and an all-pole vocal tract filter by solving a convex optimization problem that depends on detected glottal closure instants. In an analysis method presented by Funaki *et al.* [29], several models are adopted, including the RK glottal flow model, a white-Gaussian random process for the aspiration noise, and a time-varying pole-zero filter for the vocal tract. They used the genetic algorithm as well as the technique of simulated annealing to fit these models to a speech signal, with phase distortion compensated by an all-pass filter. Fröhlich *et al.* [30] estimated parameters of an exponential-trigonometric (Liljencrants-Fant) glottal flow derivative model with a modified discrete all-pole modeling technique that optimizes the quality of inverse filtering. Vincent *et al.* [31] used the Liljencrants-Fant model and a time-varying all-pole-filter model for the vocal tract, with some parameters prioritized in a low-frequency analysis. Degottex *et al.* [32] used a minimum-phase vocal tract model to estimate the shape parameter and time position of the transformed Liljencrants-Fant model, and evaluated the resulting estimates with a digital vocal tract simulator. Model-based glottal flow estimation can also be achieved by fitting a glottal flow model to the glottal flow estimate given by an inverse filtering algorithm, as presented by Plumpe *et al.* [1].

In many of the above-mentioned works, glottal flow estimation experiments were conducted on synthetic audio data that is based on a shape-descriptive glottal flow model and an autoregressive vocal tract filter. Indeed, simplifications involved in such a model of voice production can result in inadequate synthesis, which in turn can give rise to a substantial performance gap between synthesized speech and real speech. This performance gap is especially relevant when many analysis approaches are actually based on the same models as the typical data synthesis procedure. In view of this, a small number of studies have drawn on physical modeling (either with numerical methods [16], [22], [23], [32]–[34] or with physical materials [15]) to fulfill realistic simulations of sustained vowels for the evaluation. In this work, we take a further step in enhancing the reality of test speech materials, by generating test data with VocalTractLab, which is capable of synthesizing continuous speech by simulating user-specified articulatory movements. Furthermore, this study also expands on [16] by 1) including multiple voice qualities and subglottal pressure levels in the test data, 2) adopting several feature-based measures in performance evaluation, and 3) performing a robustness analysis with respect to the errors in glottal closure detection.

III. Tested Algorithms

In terms of methodology, inverse filtering algorithms can be divided into three important categories, which are covariance-analysis approaches, complex-cepstrum approaches, and pitch-asynchronous approaches. In this evaluation, a small number of representative algorithms are selected from each category to provide an adequate coverage of the methodological diversity. In covariance-analysis approaches, the analysis uses a certain amount of timing information estimated for the glottal closed phase to find time samples at which a best fit of the linear prediction model is expected. At one extreme, both the estimated beginning and ending instants of the closed phase are utilized, which is the case of closed-phase covariance analysis (CPCA) [21]. At the other extreme, only an estimated glottal closure instant is utilized, and the linear prediction model is fitted in a weighted manner to the speech signal around the estimated instant to reduce the dependence on accurate timing information, which is represented in this study by the two different weighting schemes implemented in sparse linear prediction (SLP) [11] and weighted linear prediction (WLP) [12]. The above three algorithms are thus selected for covariance-analysis approaches. Complex-cepstrum approaches are completely independent of linear prediction, for which complex cepstrum decomposition (CCD) [13] is adequately representative. To the best of our knowledge, iterative adaptive inverse filtering (IAIF) [20] is the only algorithm that does not require identification of glottal closure or opening instants, which is selected for the pitch-asynchronous category.

The descriptions in this section are specific to a custom implementation of each algorithm.[1] Our implementation of CCD is based on Drugman's implementation,[2] with the estimate of glottal flow derivative post-processed by removing its DC component. All the algorithms operate at the sampling frequency of 20 kHz in our implementation, with all synthesized signals resampled from their original sampling frequency of 44.1 kHz. To ensure proper measurement of performance, the same sampling frequency is used across the input, output, and ground-truth signals. Each algorithm is applied to a uniformly spaced sequence of time frames in the analyzed utterance. Since no glottal-flow cycle exists within a non-voiced time interval in the utterance, the glottal airflow estimated at a non-voiced time frame will be ignored by a cycle-synchronous performance measure when the accuracy of glottal flow estimation is evaluated at the utterance level.

A. Closed Phase Covariance Analysis (CPCA)

At each analysis time position, say the τth position $n = n_\tau$, the vocal tract filter can be estimated by a linear-predictive covariance analysis that minimizes residual energy at closed-phase time samples [21]. Let the speech signal be denoted by $s[n]$, and let the vocal tract filter take the following form:

$$V(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \tag{1}$$

where p is set to 20 to model 10 formants below the Nyquist frequency of 10 kHz. The analysis calculates

$$a_k = -([b_{i,j}]^{+}_{N\times(p+1)} [c_i]_{N\times 1})_{k+1},\ k = 1, ..., p, \tag{2}$$

where $(\cdot)_{k+1}$ denotes the $(k+1)$th element of a vector, N is the window length (32 ms), and $(\cdot)^{+}$ denotes the pseudoinverse of a matrix. The matrix $[b_{i,j}]_{N\times(p+1)}$ is defined by

$$b_{i,j} = \begin{cases} w[n_\tau + i - 1], & \text{if } j = 1; \\ s[n_\tau + i - j] w[n_\tau + i - 1], & \text{otherwise,} \end{cases} \tag{3}$$

$$(i, j) \in \{1, ..., N\} \times \{1, ..., p + 1\}, \tag{4}$$

[1] Source code is available for each custom implementation at https://languageandvoice.files.wordpress.com/2017/03/egifa.zip.

[2] http://tcts.fpms.ac.be/˜drugman/Toolbox/GLOAT.zip.

where $w[n]$ is unity if n is within the closed phase or within a non-voiced time interval, otherwise assuming the value zero. The vector $[c_i]_{N\times 1}$ is defined by

$$c_i = s[n_\tau + i - 1]w[n_\tau + i - 1],\ i = 1, ..., N. \tag{5}$$

Once the vocal tract filter is estimated, the estimate of glottal flow derivative $\hat{\epsilon}[n]$ can be calculated by applying the inverse filter to the speech signal:

$$\hat{\epsilon}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k],\ n = n_\tau, ..., n_\tau + N - 1. \tag{6}$$

Closed-phase boundaries are derived from glottal closure and opening instants estimated with the YAGA algorithm [9]. The ending time of glottal closed phase is directly given by the glottal opening instant, which refers to the instant at which the linear-predictive residual starts to grow from zero. YAGA aims to estimate this instant along with the glottal closure instant. The starting time of each glottal closed phase is estimated by adding a guarding delay value to the glottal closure instant to ensure that linear-predictive residual is not minimized over any open-phase time samples. In the implementation, the delay value is 0.9 ms, except that when the difference between glottal opening and closure instants is less than 4.5 ms, 0.2 times the time difference is used for the delay. The delay value of 0.9 ms was chosen as 1.5 times the root-mean-square error of estimates produced by YAGA. Note that whereas the definition of glottal opening instant adopted by YAGA is based on linear prediction, the definition adopted by some other algorithms, e.g., [35], is based on the electroglottograph signal. An algorithm of the latter type can lead to substantial error in glottal flow estimation when used with CPCA.

For CPCA (and for SLP and WLP as well), analysis time positions are spaced with a hop size of 16 ms. The hop size used in [21] was unspecified. In [21], the setting for the filter order was $p = 8$, with the sampling frequency unspecified. Since a filter of order 8 is typically used to model 4 formants for frequencies below 4 kHz, the sampling frequency there could have been 8 kHz. The window length used in [21] was 4.75 ms if a sampling frequency of 8 kHz was used. This ensured a time resolution that was sufficiently high for identifying the closed phase from linear-predictive residuals.

B. Sparse Linear Prediction (SLP)

As with CPCA, SLP estimates the vocal tract filter by a linear-predictive covariance analysis. However, this analysis minimizes a weighted sum of residual energy at all the time samples, with higher weights allocated to time samples farther from glottal closure instants [11]. Also using the (2), (3), and (5), the analysis defines its own weighting as follows:

$$w[n] = 1 - \kappa \cdot \sum_{l=1}^{L} \exp\frac{-(n-\gamma_l)^2}{2(\sigma f_s)^2}, \tag{7}$$

where γ_l denotes the lth of a total of L glottal closure instants detected from the speech signal [9], f_s denotes the sampling frequency in Hz, and κ and σ are parameters fixed to predefined constants (0.9 and 0.25 ms, respectively). The value of σ used in [11] was 4.42 ms. Note that glottal closure instants were detected in [11] by the algorithm of Drugman and Dutoit [10].

C. Weighted Linear Prediction (WLP)

The WLP algorithm differs from SLP only in that its weighting is defined by a piecewise-linear function [12], rather than by a sum of upside-down, shifted Gaussian functions. The weighting is characterized by two distinct levels of weight (1.0 and 0.05), with the higher-level value taken by all the time samples that are at a distance from glottal closure instants. Shortly before each glottal closure instant, the weight begins to ramp down, reaching the lower-level value before the glottal closure instant. After retaining the low value (for 0.4 times the fundamental period) past the glottal closure instant, the weight starts to ramp up (for 0.45 ms), going back to the higher level shortly after the glottal closure instant. Ramping down takes 0.45 ms, and the lower level is reached 0.32 times the fundamental period before the glottal closure instant. The value used in [12] for the lower level of weight was 0.01, determined from a synthetic development data set with true glottal closure instants.

D. Iterative Adaptive Inverse Filtering (IAIF)

Prior to estimating the vocal tract filter, the spectral contribution of glottal flow derivative can be estimated and removed from the speech signal with a low-order linear predictive analysis [20]. IAIF is a two-pass procedure based on this concept. In the first pass, a first-order linear predictive autocorrelation analysis is applied to the speech signal to give an estimate of the glottal-flow spectrum envelope. After applying an inverse filter of this envelope to the speech signal, a 20th-order linear predictive autocorrelation analysis is applied to the filtered signal to give an estimate of the vocal tract filter, according to which a second inverse filtering procedure yields the estimated glottal flow derivative for the first pass. In the second pass, low-order (4th-order) linear predictive analysis is again used to estimate the source contribution, but applied to the glottal flow estimated in the first pass. Similarly to the first pass, two inverse filtering steps follow to give the final estimate of glottal flow derivative. All the linear predictive analyses in IAIF are carried out with a window length of 32 ms and a hop size of 16 ms. In [20], the higher order of linear prediction was set to 10 with a sampling frequency of 8 kHz.

E. Complex Cepstrum Decomposition (CCD)

At each analysis time position, say the lth position $n = \gamma_l$ which coincides with the lth glottal closure instant detected from the speech signal (by the algorithm of Drugman and Dutoit [10]), the glottal flow can be estimated directly by separating a maximum-phase component from the speech signal, without first estimating a vocal tract filter [13]. The CCD algorithm approaches the separation by calculating the complex cepstrum of the speech signal:

$$\hat{\mathbf{x}} = \mathrm{DFT}^{-1}\{\log|\mathrm{DFT}\{\mathbf{x}\}| + j\angle\mathrm{DFT}\{\mathbf{x}\}\}, \tag{8}$$

where $\mathrm{DFT}\{\cdot\}$ denotes the discrete Fourier transform, $\angle(\cdot)$ denotes the unwrapped phase of a complex number, and $\mathbf{x}$ denotes a time frame of the speech signal $s[n]$ centered at $n = \gamma_l$, spanning 1.8 cycles, multiplied by a Blackman window, and zero-padded to 102.4 ms (a default setting in Drugman's implementation that ensures a sufficiently high spectral resolution needed for phase unwrapping). The maximum-phase component is represented by the anti-causal component $\hat{\mathbf{x}}'$ in the complex cepstrum:

$$(\hat{\mathbf{x}}')_i = \begin{cases} \frac{1}{2}(\hat{\mathbf{x}})_1, & \text{if } i = 1; \\ 0, & \text{if } 2 \le i \le N_0/2; \\ (\hat{\mathbf{x}})_i, & \text{if } N_0/2 < i \le N_0, \end{cases} \tag{9}$$

where N_0 denotes the length of $\mathbf{x}$. The time-domain representation of the glottal flow estimate is then given by inverting the complex-cepstral calculation:

$$\mathbf{x}' = \mathrm{DFT}^{-1}\{\exp(\mathrm{DFT}\{\hat{\mathbf{x}}'\})\}, \tag{10}$$

from which an estimate of the glottal flow derivative can be calculated by taking the differences between adjacent elements.

IV. Experimental Procedure

A. Data Sets

All the utterances used in our experiments are generated by the software VocalTractLab 2.1 [17]. The synthesis of vowels performs time-domain, finite-difference simulation of acoustic wave motion for a two-mass, triangular-glottis model of the vocal folds [36] and a transmission-line model of the vocal tract. Despite the fact that the glottal area waveforms simulated from the vocal-fold model may deviate to a certain degree from the waveforms measured with, e.g., high-speed digital imaging [37], the simulation reproduces the nonlinear, time-varying coupling between glottal source and vocal tract [38] by coupling an external force in the vocal-fold model to the vocal-tract acoustics through the supraglottal pressure. Another physiological advantage of this glottis model is its capability of simulating a continuum of voice qualities from pressed voice to breathy voice. Voice quality concerns the degree of glottal closure within each glottal-flow cycle, which can vary both within the same utterance and among different speakers. The pressed voice is characterized by a relatively long phase for closed vocal folds, whereas the vocal folds can lack a complete closure in the case of breathy voice. By being self-oscillating, the model promises more realistic glottal flow simulations than geometric approaches. The synthesizer includes a subglottal system, where the trachea is modeled by 23 tube sections up to 24 cm below the glottis, with a cross-sectional area around 2.5 square cm for most of the sections. A short-circuit termination impedance is used to simulate the bronchi and lungs [39]. The output sampling frequency of the synthesizer is 44.1 kHz. The approach is not currently capable of simulating pathological voices; therefore, we limit our analysis to the conditions currently included in VocalTractLab 2.1.

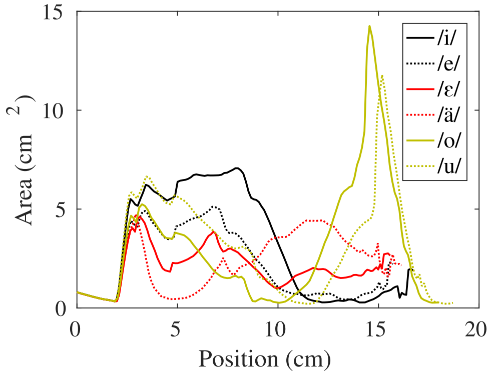


Fig. 1. Area functions realizing the 6 sustained vowel types in the experiments. Each function represents the position-varying cross-sectional area of vocal tract, with the glottis being the positional origin.

To evaluate the performance of inverse filtering algorithms under various controlled conditions, we carried out sustained-vowel time-domain simulation of voice production with VocalTractLab, giving a structured set of 750 speech utterances along with a corresponding set of glottal flow signals.[3] These samples consisted of all the combinations of 5 target fundamental frequencies (for controlling vocal-fold tension; 90 Hz, 120 Hz, 150 Hz, 180 Hz, and 210 Hz), 5 subglottal pressure levels (500 Pa, 708 Pa, 1,000 Pa, 1,414 Pa, and 2,000 Pa), 5 voice qualities (pressed, slightly pressed, modal, slightly breathy, and breathy), and 6 vowel types (/i/, /e/, /ϵ/, /ä/, /o/, and /u/; see Fig. 1). Each sample is a sustained-vowel utterance that lasts for 0.6 seconds.

A second data set is constructed for the continuous-speech experiments, which is generated by simulating manually planned movements in vocal-tract and vocal-fold configurations with VocalTractLab.[4] All the utterances in this data set are derived from a prototype score of glottal and articulatory movements, which was composed by the author of VocalTractLab for the German sentence “Lea und Doreen mögen Bananen.” The score describes 8 types of vocal movements, each of which is defined by a sequence of target configurations. Among the 8 movement types, three concern glottal movements (the other five types all concerning vocal-tract movements), i.e., target fundamental frequency (continuous-valued), subglottal pressure (continuous-valued), and voice quality (pressed, slightly pressed, modal, slightly breathy, or breathy). To generate utterances that exhibit different conditions of phonation, we adapted this prototype score by introducing various translations to the three glottal configuration sequences, such that each translated glottal configuration sequence has a new median value. The resulting adaptations consist of the 125 combinations of 5 median target fundamental frequencies, 5 median pressure levels, and 5 median voice qualities, which share specifications with the sustained-vowel data. The 125 new movement scores were used to synthesize 125 speech utterances, which make up our continuous-speech data set. In the adaptation, a translation by δ

[3] The sustained-vowel data set is available at https://languageandvoice.files.wordpress.com/2017/03/vowel.zip.

[4] The continuous-speech data set is available at https://languageandvoice.files.wordpress.com/2017/03/speech.zip.

is introduced to the sequence of M voice quality values on the linear scale (with the 5 possible voice qualities encoded by the integers 1, ..., 5):

$$\phi_m^{(\delta)} = \phi_m^{(0)} + \delta,\; m = 1, ..., M, \tag{11}$$

where $\phi_m^{(0)}$ and $\phi_m^{(\delta)}$ denote the mth prototype and translated voice quality values, respectively, such that the new sequence of voice quality values $\{\phi_m^{(\delta)}\}_{m=1}^{M}$ has one of the five desired median values while preserving the sequential variations in the prototype. Target fundamental frequencies (in Hz) and subglottal pressures (in Pa) are similarly adapted, except that these are adapted on the logarithmic scale.

B. Performance Measures

Consider an utterance for which a glottal airflow estimate has been produced by an inverse filtering algorithm. We assess the accuracy of the estimate in a cycle-synchronous fashion, accumulating cycle-wise error measurements over the whole utterance to give an overall error measurement for the utterance. The utterance is segmented automatically into cycles according to a glottal area signal derived from the synthesis process. At each time point, the area between the upper (superior) vocal-fold sections, and that between the lower (inferior) vocal-fold sections, are available from the speech synthesizer as part of the simulation. With the glottal area defined as the smaller of these two areas, the utterance is segmented whenever the glottal area waveform drops below a threshold value that indicates glottal closure. The threshold value is set to an area that is 10^{-6} m^2 larger than the minimum area over the utterance. Note that the instant when the glottal area goes to zero does not typically coincide exactly with the instant when the negative peak of glottal flow derivative occurs [8]. The glottal area signals exhibit simple trends without impulse-like events, lending themselves to reliable detection of glottal closure events.

1) Waveform Errors: To determine the extent to which the estimated waveform deviates from the true glottal flow derivative, we calculate the *normalized median absolute waveform error* (MAE-Wave). The first step in this calculation is time-alignment of the ground-truth waveform with the estimated waveform. Although the acoustic propagation delay in the voice transmission through the vocal tract can ideally be canceled by the inverse filter, the acoustic propagation delay in voice radiation cannot be modeled by an inverse filtering algorithm in general, which leads to a time delay between the estimated and ground-truth glottal flow signals that needs to be compensated with an alignment. This alignment is implemented by a 0.65-ms delay of the ground-truth waveform relative to the estimated waveform, which corresponds to a 22-cm radiation distance. Within a particular cycle, let the true and estimated glottal flow derivative signals be denoted by $\epsilon_c[n]$ and $\hat{\epsilon}_c[n]$, respectively. For pulse shape comparison, we calculate a scaled version of the estimate whose amplitude is aligned with the true signal, with a scaling factor that minimizes the Euclidean distance between the scaled version and the true signal (i.e., by an orthogonal projection):

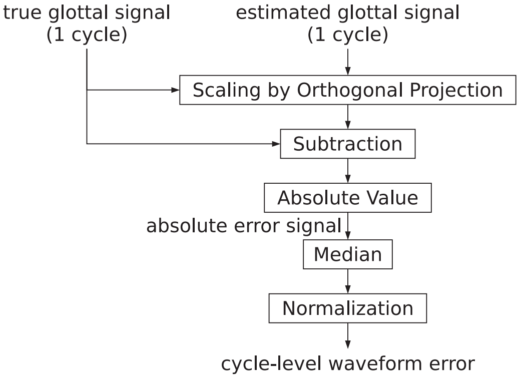


Fig. 2. Waveform error evaluation for a particular cycle identified from the synthesized glottal area signals. The median of all cycle-level waveform errors in an utterance is calculated to give an MAE-Wave.

$$\tilde{\epsilon}_c[n] = \frac{\sum_{i=0}^{N_c-1} \epsilon_c[i]\hat{\epsilon}_c[i]}{\sum_{i=0}^{N_c-1} \hat{\epsilon}_c^2[i]} \cdot \hat{\epsilon}_c[n],\; n = 0, ..., N_c - 1, \tag{12}$$

where N_c denotes the length of this cycle. As shown in Fig. 2, a cycle-level waveform error is calculated by taking the error magnitude of $\tilde{\epsilon}_c[n]$ with respect to $\epsilon_c[n]$ for each time sample, taking the median of error magnitudes over all time samples in the cycle, and normalizing the median value by the utterance-wide root-mean-square amplitude of the true signal. The utterance-level waveform error, i.e., the MAE-Wave error measure, is calculated by taking the median of all cycle-level errors. The utterance-level waveform error is not equivalent to a median calculated over all time samples in an utterance because the number of time samples within each cycle can vary from one cycle to another. Here the median-based measurement ensures that the resulting error accounts for a majority of its components, both on the cycle level and on the utterance level.

In the early days of voice production studies, inverse filtering used to be performed with dedicated hardware that came with no capability of optimization or matrix computation for formant frequency estimation, but allowed the user to assess glottal flow waveforms that resulted from various (user-specified) formant frequency settings [7]. The analysis implemented on a legacy inverse filtering device is typically limited to a bandwidth that only accounts for the first formant of vocal-tract frequency response. In the present study, to evaluate the accuracy of an estimated waveform in terms of what would have been given by single-formant processing, a variant of the aforementioned waveform error is calculated by applying the same error evaluation procedure to a low-pass filtered version of the true signal and a low-pass filtered version of the estimated signal. The low-pass filter is a 10th-order digital Butterworth filter with a cut-off frequency of 1 kHz [7]. On the other hand, to measure the higher-formant error component that could not be observed from single-formant processing, another variant of MAE-Wave is similarly calculated with a high-pass filter cut off at 1 kHz.

2) Feature Errors: The accuracy of a glottal flow estimate can also be assessed in terms of important waveform features

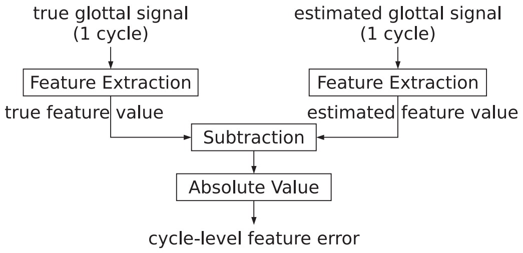


Fig. 3. Error evaluation for a particular cycle (identified from the synthesized glottal area signals) and each of the features NAQ, H1-H2, and HRF. For each feature, the median of all cycle-level errors in an utterance is calculated to give a median absolute feature error.

that traditionally represent voice quality. To that end, we use the normalized amplitude quotient (NAQ) [40], the H1-H2 feature [41], and the harmonic richness factor (HRF) [42], calculating the *median absolute NAQ, H1-H2, and HRF errors* (MAE-NAQ, MAE-H1H2, and MAE-HRF). For each cycle of the true signal $\epsilon_c[n]$, an NAQ is evaluated by dividing the peak-to-peak glottal flow amplitude by the product of fundamental period and maximum flow declination rate. The maximum flow declination rate refers to the maximum magnitude of negative slope on the pulse shape of glottal flow (i.e., magnitude of the lowest point in the derivative pulse shape), which apparently varies with the fundamental frequency and signal amplitude. The NAQ feature eliminates this variance by normalizing the maximum rate by the diagonal slope of the rectangle spanned by the single-cycle pulse shape of the glottal flow. The features H1-H2 and HRF are also extracted from the true glottal flow as spectral descriptors of the single-cycle pulse shape. H1-H2 subtracts the amplitude of the second harmonic (in decibels) from the amplitude of the first harmonic. HRF measures the total power (in decibels) of overtones, i.e., the harmonics with an order greater than one, relative to the power of the fundamental. Here the harmonic amplitudes of the true glottal flow (integral of $\epsilon_c[n]$) are calculated by taking the absolute value of its discrete Fourier transform (without zero-padding before the transform) and extracting the consecutive frequency bins that correspond to harmonic frequencies greater than 0 Hz and less than 3 kHz. Both NAQ and H1-H2 could be regarded as a measure of voice breathiness, while HRF is negatively correlated with breathiness [40]–[42]. The same features are also extracted from the glottal flow derivative estimate $\hat{\epsilon}_c[n]$. As shown in Fig. 3, to evaluate the error in glottal flow estimation, three error magnitudes are calculated respectively for the three features for each cycle, and an utterance-level error is calculated by taking the median of all cycle-level errors for each of the three features.

C. Simulation of Glottal Closure Instants

To evaluate the susceptibility of inverse filtering algorithms to the errors in glottal closure detection, we extract all the glottal closure instants from each true glottal flow signal in the data set, use these true instants to simulate estimated instants of a certain accuracy, and substitute these simulated estimates for the real detector-produced estimates in a glottal flow estimation experiment.

To extract glottal closure instants from a true glottal flow signal and its derivative, the signals are first segmented into cycles with the same area-based procedure as described in Section IV-B. For each cycle, in order to identify closed-phase time samples, the maximum value of glottal flow is calculated. Time samples with a glottal-flow value below 0.1 times the maximum value are considered to be within the closed phase. Among the closed-phase time samples, the one with the minimum derivative value is extracted as a true glottal closure instant. In case that no closed-phase time sample can be found (which can sometimes occur for breathy voice), the minimum-flow time sample is taken as a true glottal closure instant.

The error in an estimated glottal closure instant can be measured in relation to the instantaneous fundamental period, as a phase error in the quasi-periodic structure of glottal closure instants. To see the effect that this phase error has on the performance of glottal flow estimation, we simulate estimates of glottal closure instants that have a constant phase error of θ radians throughout an utterance:

$$\tilde{\gamma}_l = \left\lfloor \bar{\gamma}_l + (\bar{\gamma}_{l+1} - \bar{\gamma}_l) \cdot \frac{\theta}{2\pi} + 0.00065 f_s + 0.5 \right\rfloor, \tag{13}$$

$$l = 1, \ldots, L-1, \tag{14}$$

$$\tilde{\gamma}_L = \left\lfloor \bar{\gamma}_L + (\bar{\gamma}_L - \bar{\gamma}_{L-1}) \cdot \frac{\theta}{2\pi} + 0.00065 f_s + 0.5 \right\rfloor, \tag{15}$$

where $\tilde{\gamma}_l$ denotes the lth simulated glottal closure instant in samples, $\bar{\gamma}_l$ denotes the lth true glottal closure instant in samples, f_s denotes the sampling frequency in Hz, and a rounding to the nearest integer and a 0.65-ms delay give the simulated estimate.

To test in this simulation an algorithm that also uses glottal opening instants, such as CPCA, the instants are derived from the simulated glottal closure instants without a separate simulation procedure. To that end, the YAGA algorithm is used to generate candidates for the glottal opening instants, from which a sequence of glottal opening instants can be chosen with reference to the simulated sequence of glottal closure instants.

V. Results

A. Results on Sustained Vowel and Continuous Speech Data

Results of the sustained-vowel and continuous-speech glottal flow estimation experiments are presented in Table I. For sustained vowels, all the five algorithms gave normalized waveform errors around 0.3, with standard deviations around 0.2, which shows no substantial performance difference among the algorithms. Listed on the row titled "MAE-Wave-S" are results obtained with a signed variant of the waveform error, where a signed error is calculated in place of an error magnitude for each time sample to reveal any systematic bias in the signal estimate. This shows that CCD tends more to overestimate glottal flow derivative values than to underestimate them, whereas there is a slight tendency for IAIF to underestimate glottal flow derivative values. Still, even for these two algorithms the bias does not predominantly account for the unsigned waveform error.

The similarity between the low-pass filtered and unfiltered waveform errors (measured as described in Section IV-B1)

TABLE I
ERROR (MEAN ± STANDARD DEVIATION) OF GLOTTAL FLOW ESTIMATES ACROSS THE SUSTAINED-VOWEL AND CONTINUOUS-SPEECH DATA SETS

Measure	Data	CPCA	SLP	WLP	IAIF	CCD
MAE-Wave	vowel	**0.27** ± 0.19	0.29 ± 0.18	0.29 ± 0.17	0.32 ± 0.20	0.34 ± 0.24
	speech	0.40 ± 0.11	**0.39** ± 0.12	**0.39** ± 0.12	0.43 ± 0.11	0.41 ± 0.21
MAE-Wave-S	vowel	**0.000** ± 0.05	**0.000** ± 0.06	0.001 ± 0.07	−0.008 ± 0.07	0.041 ± 0.11
	speech	−0.016 ± 0.03	−0.017 ± 0.03	−0.018 ± 0.03	−0.022 ± 0.03	−**0.012** ± 0.13
MAE-Wave-LP	vowel	**0.24** ± 0.19	0.26 ± 0.18	0.26 ± 0.17	0.29 ± 0.20	0.34 ± 0.22
	speech	0.34 ± 0.10	**0.33** ± 0.10	0.34 ± 0.10	0.38 ± 0.10	0.42 ± 0.18
MAE-Wave-LP-S	vowel	0.014 ± 0.06	0.014 ± 0.06	0.012 ± 0.07	**0.002** ± 0.07	−0.009 ± 0.10
	speech	**0.000** ± 0.03	−0.003 ± 0.03	−0.002 ± 0.03	−0.011 ± 0.03	−0.095 ± 0.15
MAE-Wave-HP	vowel	**0.09** ± 0.054	0.10 ± 0.055	0.10 ± 0.056	0.10 ± 0.057	0.10 ± 0.056
	speech	0.16 ± 0.089	0.16 ± 0.087	0.16 ± 0.089	0.16 ± 0.089	**0.15** ± 0.090
MAE-Wave-HP-S	vowel	**0.000** ± 0.004	**0.000** ± 0.005	**0.000** ± 0.005	**0.000** ± 0.005	−0.002 ± 0.005
	speech	**0.000** ± 0.002	**0.000** ± 0.002	**0.000** ± 0.002	**0.000** ± 0.002	−0.001 ± 0.002
MAE-NAQ	vowel	0.035 ± 0.027	0.031 ± 0.023	0.032 ± 0.024	**0.029** ± 0.023	0.049 ± 0.045
	speech	0.035 ± 0.017	0.034 ± 0.017	0.035 ± 0.016	**0.033** ± 0.015	0.045 ± 0.030
MAE-NAQ-S	vowel	0.030 ± 0.032	0.024 ± 0.029	0.026 ± 0.030	**0.020** ± 0.030	−0.045 ± 0.048
	speech	0.026 ± 0.026	0.026 ± 0.026	0.025 ± 0.027	**0.024** ± 0.024	−0.039 ± 0.033
MAE-H1H2	vowel	3.3 ± 4.0	3.4 ± 4.1	**3.2** ± 3.9	4.2 ± 4.9	5.8 ± 5.0
	speech	3.1 ± 1.7	**3.0** ± 1.6	3.1 ± 1.7	3.6 ± 1.7	5.6 ± 3.8
MAE-H1H2-S	vowel	−**0.9** ± 5.0	−2.5 ± 4.7	−2.1 ± 4.6	−3.2 ± 5.6	−5.6 ± 5.2
	speech	−0.6 ± 1.6	−**0.5** ± 1.6	−**0.5** ± 1.5	−1.5 ± 1.6	−5.2 ± 3.9
MAE-HRF	vowel	3.0 ± 2.9	3.0 ± 3.0	**2.7** ± 2.8	3.8 ± 4.3	5.7 ± 4.9
	speech	**2.7** ± 1.6	**2.7** ± 1.6	**2.7** ± 1.6	3.3 ± 1.7	6.7 ± 4.6
MAE-HRF-S	vowel	**0.5** ± 4.1	2.2 ± 3.6	1.4 ± 3.6	2.9 ± 4.9	5.6 ± 5.0
	speech	0.4 ± 1.7	0.4 ± 1.7	**0.3** ± 1.7	1.6 ± 1.6	6.6 ± 4.6

The suffix S represents the signed variant of an error measure. The suffixes LP and HP refer to low- and high-pass filtered variants of MAE-Wave. The error given by the best-performing algorithm is shown in boldface for each combination of data set and measure. As defined in Section IV-B, the measures MAE-H1H2 and MAE-HRF (and their variants) are in dB, and the other measures are unit-less.

TABLE II
MATRIX OF p-VALUES FOR SUSTAINED VOWELS

	CPCA	SLP	WLP	IAIF	CCD
CPCA	1.00	**0.00**	**0.00**	**0.00**	**0.00**
SLP	1.00	1.00	1.00	**0.00**	**0.00**
WLP	1.00	**0.00**	1.00	**0.00**	**0.00**
IAIF	1.00	1.00	1.00	1.00	0.12
CCD	1.00	1.00	1.00	0.88	1.00

Each p-value was given by a paired left-tailed Wilcoxon signed rank test conducted between two inverse filtering algorithms on the MAE-Wave error. The row and column labels identify the first and second sample data, respectively. A p-value less than 0.05 (shown in boldface) indicates that the row algorithm tends to give a lower MAE-Wave than the column algorithm at the 5% significance level.

suggests a consistency of the present performance measurement with earlier research. Although large signal value errors could occur in the return phase (because of the typically abrupt change in glottal flow derivative) and thus be captured by the high-pass filtered measure, such errors would be confined within a small number of time samples in each cycle and have no substantial impact on the median-based high-pass measure. This explains why the low-pass error component dominates the waveform errors.

The NAQ results again show a similarity of performance among the algorithms, but reveal that errors in NAQ are overwhelmingly either underestimations (with a large, negative signed error for CCD) or overestimations (with a large, positive signed error for the other algorithms) within an algorithm. This suggests the possibility of improving NAQ estimates given by a specific algorithm by canceling the bias observed here. The results for the spectral features H1-H2 and HRF show relatively poor performance for CCD with average errors around 6 dB, and substantial biases (underestimations of H1-H2 and overestimations of HRF) for all the algorithms except CPCA.

Regarding the continuous-speech results, mean MAE-Wave was again similar (approximately 0.40) across all the algorithms, and comparison of the MAE-Wave results with those obtained with the variant measures exhibits an absence of substantial bias, as well as a consistency of unfiltered results with low-pass filtered results. The NAQ results reveal the biasedness of all NAQ estimates. CCD produced H1-H2 and HRF estimates with a bias that resulted in an average error around 6 dB.

Although the five algorithms exhibited similar performance in terms of MAE-Wave, a statistically significant performance difference between any two algorithms may be detected by a hypothesis test. To ascertain the best-performing algorithm, the paired, left-tailed Wilcoxon signed rank test was applied to the pairs (750 for sustained vowels, or 125 for continuous speech) of MAE-Wave values produced by each pair of algorithms. The test operates on a pair of sample data, producing a p-value for the null hypothesis that the median difference between the first sample data and the second sample data is zero, against the alternative hypothesis that the median difference is negative. Results of the significance tests are presented in Tables II and III, from which we can conclude 1) that CPCA performs the best in terms of MAE-Wave and sustained vowels, and 2) that SLP performs the best in terms of MAE-Wave and continuous speech if CCD is excluded from the comparison.

TABLE III
MATRIX OF p-VALUES FOR CONTINUOUS SPEECH

	CPCA	SLP	WLP	IAIF	CCD
CPCA	1.00	1.00	1.00	**0.00**	0.65
SLP	**0.00**	1.00	**0.00**	**0.00**	0.35
WLP	**0.00**	1.00	1.00	**0.00**	0.43
IAIF	1.00	1.00	1.00	1.00	0.96
CCD	0.35	0.65	0.57	**0.04**	1.00

Each p-value was given by a paired left-tailed Wilcoxon signed rank test conducted between two inverse filtering algorithms on the MAE-Wave error. The row and column labels identify the first and second sample data, respectively. A p-value less than 0.05 (shown in boldface) indicates that the row algorithm tends to give a lower MAE-Wave than the column algorithm at the 5% significance level.

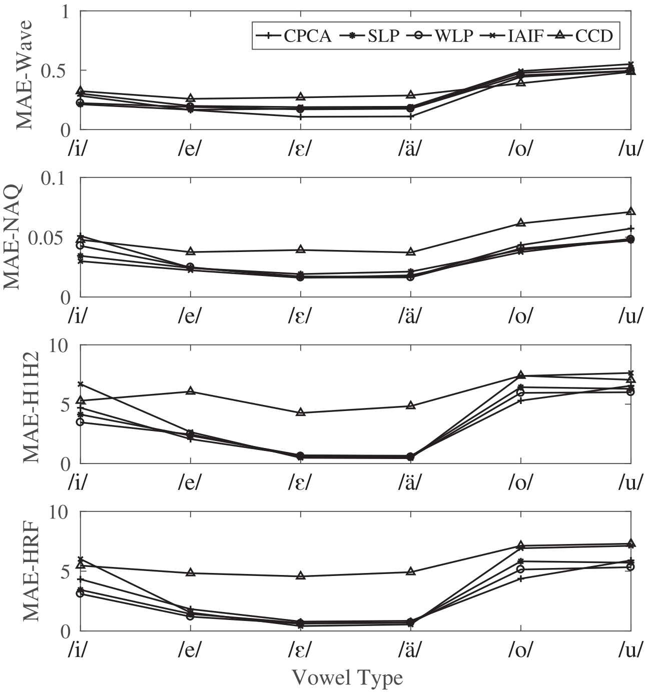


Fig. 4. Subset error averages for vowel types.

B. Results on Data Subsets

1) Vowel Types: It has been observed by some researchers that some vowels with a low first formant frequency cannot be adequately analyzed by an inverse filtering algorithm, whereas the vowel /ä/ has a first-formant frequency that is sufficiently high to avoid interference with the primarily low-frequency energy distribution of glottal source [20]. To see the impact of vowel type on the performance of algorithms, we took a separate average of errors for each vowel-specific subset of the sustained-vowel data. As shown in Fig. 4, the close rounded vowels /o/ and /u/ are associated with substantially higher errors than other vowels. This confirms that the analysis of close rounded vowels remains difficult as far as inverse filtering algorithms are concerned. Throughout the rest of this paper, we will move on to explore some other factors that could also have an effect on algorithm performance, while factoring out the effect of vowel types by testing the algorithms on utterances of the vowel /ä/ only.

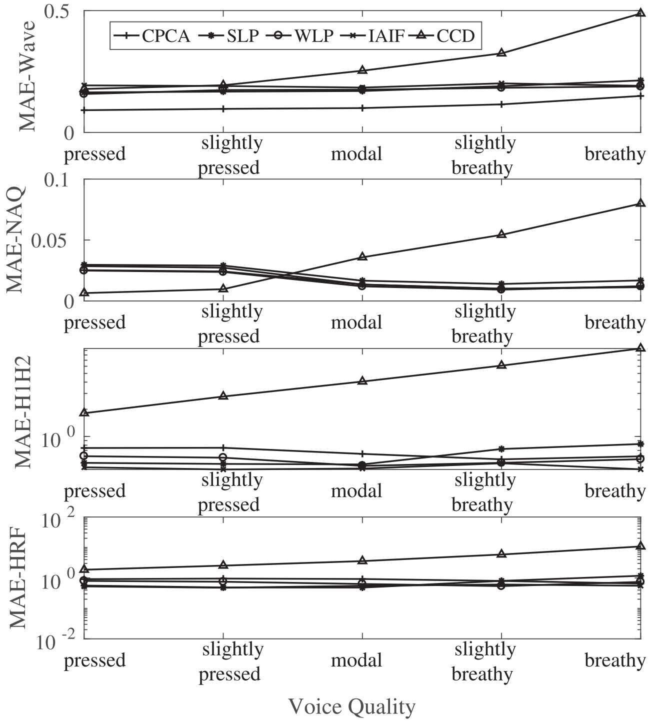


Fig. 5. Subset error averages for voice qualities. Only utterances of vowel /ä/ in the sustained-vowel data set are used.

2) Voice Qualities: The performance of algorithms on utterances of different voice qualities is examined in Fig. 5. For CCD, the breathy voice quality is associated with a substantially higher average error than the pressed voice quality with respect to every performance measure, which suggests that the maximum phase property assumed for the glottal-flow open phase may not be as valid for breathy voice as for pressed voice. All the other algorithms demonstrate roughly constant performance over the voice qualities with respect to several measures. This is remarkable for the closed-phase approaches in particular, for which only a small number of time samples are available in each analysis time frame for the estimation of vocal tract filter in the case of breathy voice. An exception to this constant performance is the NAQ error, for which the pressed and slightly pressed voice qualities have slightly higher errors. This resulted from the narrow negative peaks in pressed glottal flow derivative waveforms, which are not represented accurately by the 20-kHz signal sampling in our experiments. Accurate performance evaluation in terms of the NAQ feature would require a sampling frequency higher than 44.1 kHz because even the un-resampled derivative waveforms from the synthesizer for pressed voice, exhibit maximum flow declination rates that vary substantially between adjacent cycles.

3) Subglottal Pressure Levels: Sustained-vowel utterances of a particular subglottal pressure are also isolated to give an average error specific to the pressure level. These errors are plotted in Fig. 6, where the only remarkable effect of the pressure level occurs with the waveform and NAQ errors given by

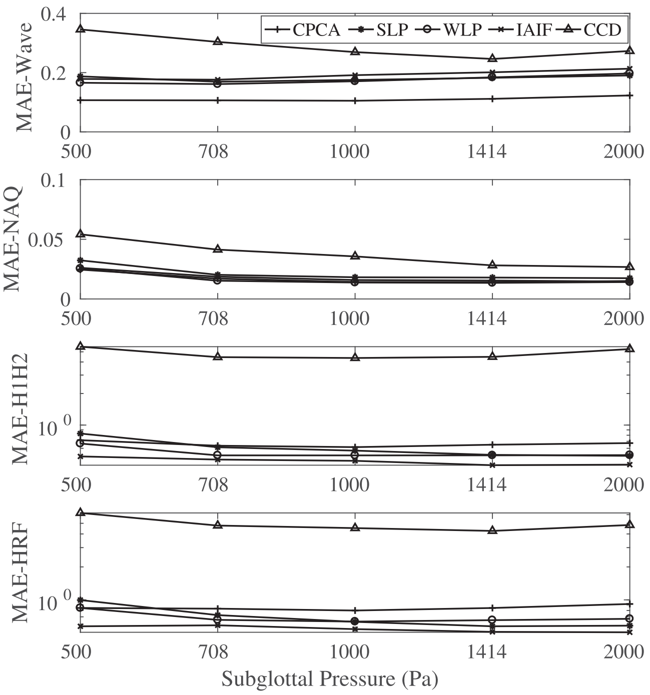


Fig. 6. Subset error averages for subglottal pressure levels. Only utterances of vowel /ä/ in the sustained-vowel data set are used.

the CCD algorithm. The raised error for low pressure could be an effect similar to that of breathiness observed for CCD in Section V-B2; i.e., low subglottal pressure tends to result in a glottal-flow pulse shape typical of a breathy voice.

4) Target Fundamental Frequencies: Inverse filtering algorithms typically involve the estimation of vocal tract filter, which explicitly or implicitly relies on the harmonic amplitudes of speech signal as observable samples of the spectrum envelope. As the fundamental frequency increases, the observable harmonics become sparser in the spectrum, which can gradually turn the envelope estimation problem into an under-determined one. The degradation of glottal flow estimation performance under increasing fundamental frequency has been well documented and discussed in the literature, which is also observed on our data set as a general trend of MAE-Wave in Fig. 7. In comparison to the other algorithms, the evidently inferior performance of CCD presumably results from the limited validity of its assumption on the maximum-phase open-phase glottal flow, given that none of the others is based on the assumption.

C. Examples

To demonstrate the performance of each inverse filtering algorithm, consider the utterance for which median performance was observed among all the utterances concerned. The median-performance utterance is determined in terms of the MAE-Wave measure and the CPCA algorithm. The utterance is selected such that its error is the 63rd lowest (0.099) among all the 125 utterances of vowel /ä/. For this example utterance, results can be examined not only in terms of MAE-Wave and CPCA, but also in terms of other measures and algorithms. Cycle-level

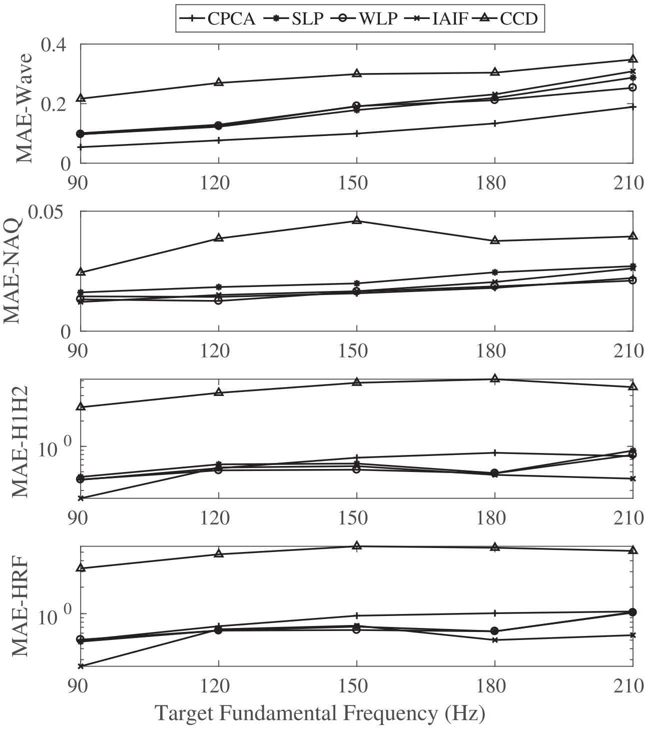


Fig. 7. Subset error averages for target fundamental frequencies. Only utterances of vowel /ä/ in the sustained-vowel data set are used.

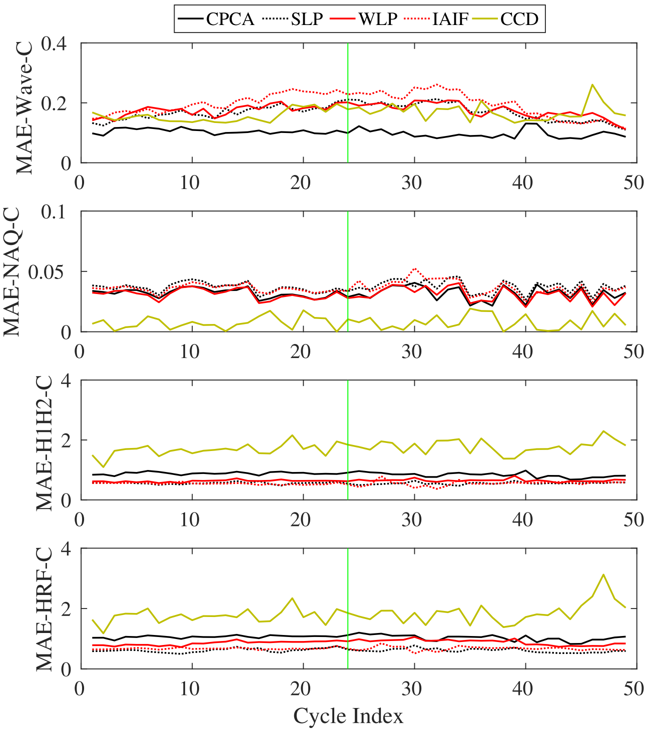


Fig. 8. Cycle-level errors in a sustained-vowel example utterance, which has a slightly pressed glottal flow, a vowel type of /ä/, a target fundamental frequency of 150 Hz, and a subglottal pressure of 500 Pa. The suffix C refers to the cycle-level errors underlying an utterance-level measure. Marked with a vertical green line is the cycle with the 25th lowest error among the 49 cycles in terms of the CPCA algorithm and the cycle-level components of MAE-Wave.

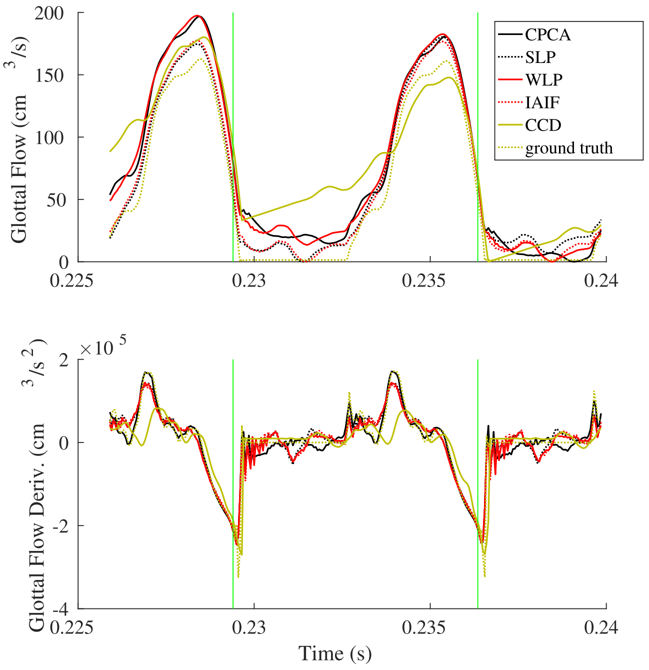


Fig. 9. Glottal flow and derivative ($\tilde{\epsilon}_c[n]$) estimates generated by the tested algorithms at the median cycle marked in Fig. 8. The endpoints of the cycle are marked with vertical green lines.

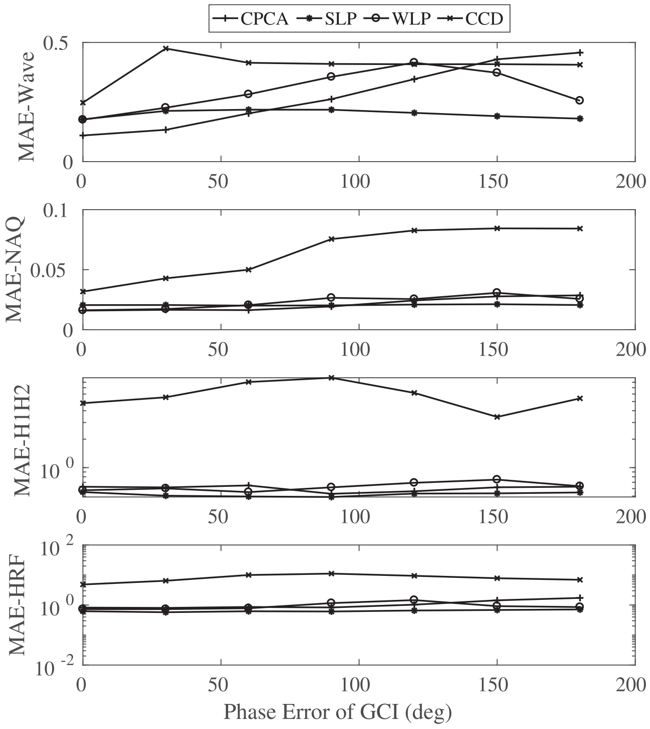


Fig. 10. Performance of four algorithms under various amounts of error in each simulated glottal closure instant (GCI) estimate used by the algorithms. Only utterances of vowel /ä/ in the sustained-vowel data set are used in these experiments.

errors are plotted for this utterance in Fig. 8, which shows that the best-performing algorithm varies on the utterance level, depending on the error measure used: When the waveform error is used, CPCA gave the lowest error. When the NAQ error is used, the lowest error was given by CCD. When either of the two spectral-feature errors is used, IAIF and SLP performed the best.

Physiologically based speech synthesis could simulate a “ripple effect” in the glottal airflow that is beyond the representation of a typical glottal flow model. It consists in some ripples in the open-phase glottal flow derivative waveform that result from the nonlinear coupling between vocal tract and glottis [38]. To see how well these ripples can be captured by an inverse filtering algorithm, we assess the accuracy of glottal flow estimation also at the cycle level. To that end, we apply the same median selection strategy to the cycles in the example utterance, illustrating with the median-performance cycle determined in terms of the cycle-level components of MAE-Wave and the CPCA algorithm. The example cycle is selected such that its error is the 25th lowest (0.099) among all the 49 cycles in the example utterance. The estimates given by the five algorithms for the selected cycle are shown in Fig. 9. In the derivative plot, CPCA slightly deviates from the ground truth during the closed phase, but closely matches the ground truth during the open phase, where the ripples are evident. In contrast, CCD deviates considerably from the ground truth during the open phase. The latter deviation is so severe that spectral-feature errors reach 2 dB for CCD. Given a ground-truth value of 0.07 for this cycle, NAQ is underestimated by CCD at 0.06 and overestimated by the other algorithms (at 0.11 by IAIF and at 0.10 by the 3 covariance-analysis algorithms).

D. Robustness to Errors in Glottal Closure Detection

Results for the simulated glottal closure detection are presented in Fig. 10. As intended by the weighted minimization of residual energy, the dependence of performance on the accuracy of glottal closure detection is minimal for SLP under every performance measure. The strong dependence for the other three algorithms is evident in terms of the waveform error. Despite this, the zero-phase-error MAE-Wave values in Fig. 10 (resulting from the use of true glottal closure instants) are fairly close to the MAE-Wave values in Fig. 4 for the vowel /ä/. This implies that the errors in (non-simulated) glottal closure detection do not constitute a primary factor that limited the evaluated performance of analyzing this vowel, leaving high target fundamental frequencies as the only important limiting factor. Note that IAIF does not rely on glottal closure detection.

VI. Conclusion

In this paper, the performance of several inverse filtering algorithms has been evaluated with synthesized test data. These algorithms aim to provide accurate glottal flow estimates without assuming a glottal flow model. With the test data generated with a physiologically relevant, articulatory speech synthesizer that simulates articulatory movement as well as voice production, the resulting evaluation serves to predict the performance of these algorithms in analyzing real speech.

The fundamental techniques that underlie the tested methods include linear-predictive covariance analysis, linear-predictive autocorrelation analysis, and the complex cepstrum. The experiments showed that each method gives an average MAE-Wave around 0.3 over the sustained-vowel data, and an average error of the same type around 0.4 over the continuous-speech data. Significance tests identified CPCA as the algorithm that gives the lowest MAE-Wave in sustained-vowel analysis. SLP was shown by significance tests to outperform CPCA, WLP, and IAIF in the case of continuous speech analysis. The average waveform errors evaluated over the close rounded vowel subsets of the sustained-vowel data are above 0.4 for all the methods, which confirmed that the methods are not as effective for close rounded vowels as for open vowels. Comparison among data subsets of an open vowel and of different voice qualities revealed that CCD does not produce glottal flow estimates as accurately for breathy voice as for pressed voice, which suggests that the validity of the maximum-phase assumption on open-phase glottal flow is questionable in the case of breathy voice. According to the robustness analysis performed with respect to the errors in glottal closure detection, the algorithm of choice for the analysis of vowel /ä/ is IAIF or SLP when accurate glottal closure instants are not available.

Results of the experiments suggest that the difficulty in analyzing close rounded vowels remains a major factor that limits the applicability of inverse filtering algorithms to accurate glottal flow estimation from continuous speech. This difficulty could have resulted from the first-formant resonance in close rounded vowels coinciding with the frequency band where glottal source energy is primarily distributed. It would be an important direction for future research to inquire models of voice production that are effective for the analysis of close rounded vowels. Other challenges in glottal flow estimation also merit further investigation, including high-pitched phonation, disordered speech, and estimation from non-audio signals such as oral airflow and neck-surface accelerometry. Regarding biometric and clinical applications, it will be of great interest to evaluate the impact of current limitations of inverse filtering algorithms in a specific application, as well as to explore how the application should be approached to make the most of the information revealed by an inverse filtering algorithm. For instance, a relevant clinical application is the discrimination between normal and hyperfunctional voices. Espinoza *et al.* [43] presented an approach to this type of discrimination, which is based on a set of glottal-flow measures extracted from the output of an inverse filtering algorithm. Future efforts can thus look into the accuracy of clinical discrimination achievable with the best-performing algorithm identified in this study.

Acknowledgment

The authors would like to thank Dr. Peter Birkholz for sharing a handcrafted score of vocal movements, and Dr. Mark R. P. Thomas for sharing his implementation of IAIF.

References

[1] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

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

[3] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[4] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, “Objective assessment of vocal hyperfunction: An experimental framework and initial results,” *J. Speech, Lang. Hearing Res.*, vol. 32, pp. 373–392, 1989.

[5] D. D. Mehta *et al.*, “Using ambulatory voice monitoring to investigate common voice disorders: Research update,” *Frontiers Bioeng. Biotechnol.*, vol. 3, no. 155, 2015, DOI: 10.3389/fbioe.2015.00155.

[6] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 26, pp. 1–13, 1985.

[7] J. S. Perkell, E. B. Holmberg, and R. E. Hillman, “A system for signal processing and data extraction from aerodynamic, acoustic, and electroglottographic signals in the study of voice production,” *J. Acoust. Soc. Amer.*, vol. 89, no. 4, pp. 1777–1781, 1991.

[8] S. Granqvist, S. Hertegård, H. Larsson, and J. Sundberg, “Simultaneous analysis of vocal fold vibration and transglottal airflow; exploring a new experimental set-up,” *TMH-QPSR*, vol. 45, no. 1, pp. 35–46, 2003.

[9] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, “Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 82–91, Jan. 2012.

[10] T. Drugman and T. Dutoit, “Glottal closure and opening instant detection from speech signals,” in *Proc. INTERSPEECH*, 2009, pp. 2891–2894.

[11] V. Khanagha and K. Daoudi, “An efficient solution to sparse linear prediction analysis of speech,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 3, 2013, DOI: 10.1186/1687-4722-2013-3.

[12] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, “Formant frequency estimation of high-pitched vowels using weighted linear prediction,” *J. Acoust. Soc. Amer.*, vol. 134, no. 2, pp. 1295–1313, 2013.

[13] T. Drugman, B. Bozkurt, and T. Dutoit, “Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation,” *Speech Commun.*, vol. 53, pp. 855–866, 2011.

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

[15] D. T. W. Chu, K. Li, J. Epps, J. Smith, and J. Wolfe, “Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics,” *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. EL358–EL362, 2013.

[16] J. Guðnason, D. D. Mehta, and T. F. Quatieri, “Evaluation of speech inverse filtering techniques using a physiologically based synthesizer,” in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 920–924.

[17] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PLoS ONE*, vol. 8, no. 4, 2013, Art. no. e60603, DOI: 10.1371/journal.pone.0060603.

[18] P. Alku, “Glottal inverse filtering analysis of human voice production—A review of estimation and parameterization methods of the glottal excitation and their applications,” *Sādhanā*, vol. 36, no. 5, pp. 623–650, 2011.

[19] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, “Glottal source processing: From analysis to applications,” *Comput. Speech Lang.*, vol. 28, pp. 1117–1138, 2014.

[20] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Commun.*, vol. 11, pp. 109–118, 1992.

[21] D. Y. Wong, J. D. Markel, and A. H. Gray, “Least squares glottal inverse filtering from the acoustic speech waveform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, Aug. 1979.

[22] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, “Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering,” *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3289–3305, 2009.

[23] M. Airaksinen, T. Raitio, B. Story, and P. Alku, “Quasi closed phase glottal inverse filtering analysis with weighted linear prediction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[24] M. Airaksinen, T. Backstrom, and P. Alku, “Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 929–939, May 2017.

[25] A. Oppenheim, R. Schafer, and T. Stockham, “Nonlinear filtering of multiplied and convolved signals,” *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 3, pp. 437–466, Sep. 1968.

[26] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, “Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1929–1939, Sep. 2013.

[27] W. Ding, H. Kasuya, and S. Adachi, “Simultaneous estimation of vocal tract and voice source parameters based on an ARX model,” *IEICE Trans. Inf. Syst.*, vol. E78-D, no. 6, pp. 738–743, 1995.

[28] H.-L. Lu and J. O. Smith III, “Joint estimation of vocal tract filter and glottal source waveform via convex optimization,” in *Proc. 1999 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 1999, pp. 79–82.

[29] K. Funaki, Y. Miyanaga, and K. Tochinai, “Recursive ARMAX speech analysis based on a glottal source model with phase compensation,” *Signal Process.*, vol. 74, pp. 279–295, 1999.

[30] M. Fröhlich, D. Michaelis, and H. W. Strube, “SIM—Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals,” *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 479–488, 2001.

[31] D. Vincent, O. Rosec, and T. Chonavel, “Estimation of LF glottal source parameters based on an ARX model,” in *Proc. INTERSPEECH*, 2005, pp. 333–336.

[32] G. Degottex, A. Roebel, and X. Rodet, “Phase minimization for glottal model estimation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1080–1090, Jul. 2011.

[33] P. Milenkovic, “Glottal inverse filtering by joint estimation of an AR system with a linear input model,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 1, pp. 28–42, Feb. 1986.

[34] 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 Phoniatr. Logop.*, vol. 58, no. 2, pp. 102–113, 2006.

[35] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, “A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 316–328, Feb. 2016.

[36] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, “Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis,” in *Proc. INTERSPEECH*, 2011, pp. 2681–2684.

[37] A. Yamauchi *et al.*, “Age- and gender-related difference of vocal fold vibration and glottal configuration in normal speakers: Analysis with glottal area waveform,” *J. Voice*, vol. 28, no. 5, pp. 525–531, 2014.

[38] T. V. Ananthapadmanabha and G. Fant, “Calculation of true glottal flow and its components,” *Speech Commun.*, vol. 1, no. 3–4, pp. 167–184, 1982.

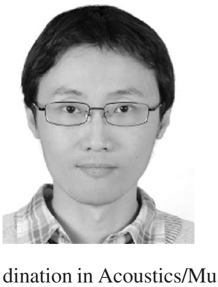
[39] P. Birkholz, D. Jackel, and B. J. Kroger, “Simulation of losses due to turbulence in the time-varying vocal system,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1218–1226, May 2007.

[40] P. Alku, T. Bäckström, and E. Vilkman, “Normalized amplitude quotient for parametrization of the glottal flow,” *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, 2002.

[41] I. R. Titze and J. Sundberg, “Vocal intensity in speakers and singers,” *J. Acoust. Soc. Amer.*, vol. 91, no. 5, pp. 2936–2946, 1992.

[42] D. G. Childers and C. K. Lee, “Vocal quality factors: Analysis, synthesis, and perception,” *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[43] V. M. Espinoza, M. Zañartu, J. H. Van Stan, D. D. Mehta, and R. E. Hillman, “Glottal aerodynamic measures in adult females with phonotraumatic and non-phonotraumatic vocal hyperfunction,” *J. Speech, Lang. Hearing Res.*, to be published.



Yu-Ren Chien received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the National Taiwan University, Taipei City, Taiwan, in 2000, 2002, and 2016, respectively. Since 2016, he has been a Postdoctoral Researcher at Reykjavik University, Reykjavik, Iceland. He was a Research Assistant in the Institute of Information Science, Academia Sinica, Taipei City, Taiwan. From 2007 to 2008, he was a Senior Engineer in the Realtek Semiconductor Corp., Hsinchu, Taiwan. In 2013, he was a Visiting Ph.D. Student in the Institute for Research and Coordination in Acoustics/Music, Paris, France. His research interests include music signal processing and speech acoustics.

Daryush D. Mehta (S’01–M’11) received the B.S. degree (*summa cum laude*) in electrical engineering from the University of Florida, Gainesville, FL, USA, in 2003, the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2006, and the Ph.D. degree in speech and hearing bioscience and technology from Harvard-MIT Division of Health Sciences and Technology, MIT, in 2010. He is currently the Director of the Voice Science and Technology Laboratory, Massachusetts General Hospital Voice Center, Boston, MA, USA, an Assistant Professor of Surgery in Harvard Medical School, Boston, MA, USA, and an Adjunct Assistant Professor in the MGH Institute of Health Professions, Boston, MA, USA. He is also an Honorary Senior Fellow in the Department of Otolaryngology, University of Melbourne, Melbourne, Vic, Australia.

Jón Guðnason (M’96) received the M.Sc. degree from the University of Iceland, Reykjavik, Iceland, and the Ph.D. degree from the Imperial College London, London, UK. He is currently a Lecturer of electrical engineering at Reykjavik University Iceland, Reykjavik, Iceland, and the Chairman of the Center for Analysis and Design of Intelligent Agents, Reykjavik, Iceland. He held research positions in the Imperial College London, and Columbia University, New York, NY, USA. He is a member on the board of IEEE Iceland Section and is a member of ISCA. His research interests include speech processing and Icelandic language technology.

Matías Zañartu (S’08–M’11) received the B.S. degree in acoustical engineering from the Universidad Tecnológica Vicente Pérez Rosales, Santiago, Chile, and the M.S. and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, USA. He is an Assistant Professor in the Department of Electronic Engineering and the Head of the Biomedical Engineering Research Track in the Advanced Center for Electrical and Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile. His research interests include the development of digital signal processing, system modeling, and biomedical engineering tools that involve speech, audio, and acoustics. His recent research efforts have revolved around developing quantitative models that describe nonlinear effects in human speech production, and applying these physiological descriptions for the development of communication and clinical technologies. He is a member of the Acoustical Society of America, the Institute of Electrical and Electronics Engineers, and the American Speech-Language-Hearing Association.

Thomas F. Quatieri (S’78–M’79–SM’87–F’99) received the B.S. degree (*summa cum laude*) from Tufts University, Medford, MA, USA, in 1973, and the S.M., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1975, 1977, and 1979, respectively. He holds a faculty appointment in the Harvard Speech and Hearing Bioscience and Technology Program under the Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA, USA. He is a Senior Member of the Technical Staff with MIT Lincoln Laboratory (MIT LL), Lexington, MA, USA, involved in applying speech, auditory, and neuromotor science to detection and monitoring of neurological disorders and cognitive stress conditions. He is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, ICSA, and the Acoustics Society of America. He received four IEEE best paper awards in speech and signal processing and the 2010 MIT LL Best Paper Award for an IEEE TASLP article. He led the MIT LL team that took first place in the 2013 and 2014 AVEC Depression Challenges, as well the 2014 MIT LL Team Award for vocal and facial biomarkers.