Alku, Paavo; Murtola, Tiina; Malinen, Jarmo; Kuortti, Juha; Story, Brad; Airaksinen, Manu; Salmi, Mika; Vilkman, Erkki; Geneid, Ahmed

OPENGLOT – An open environment for the evaluation of glottal inverse filtering

Published in:
Speech Communication

DOI:
10.1016/j.specom.2019.01.005

Published: 01/02/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

OPENGLOT – An open environment for the evaluation of glottal inverse filtering

Paavo Alku[a)], Tiina Murtola[a)], Jarmo Malinen[b)], Juha Kuortti[b)], Brad Story[c)], Manu Airaksinen[a)], Mika Salmi[d)], Erkki Vilkman[e)], Ahmed Geneid[e)]

[a)]Department of Signal Processing and Acoustics, Aalto University, School of Electrical Engineering, P.O. Box 12200, FI-00076 Aalto, Finland
[b)] Department of Mathematics and Systems Analysis, Aalto University, School of Science, P.O. Box 11100, FI-00076 Aalto, Finland
[c)] Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721, USA
[d)] Department of Mechanical Engineering, Aalto University, School of Engineering, P.O. Box 14300, FI-00076 Aalto, Finland
[e)] Department of Otorhinolaryngology and Phoniatrics - Head and Neck Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Abstract

Glottal inverse filtering (GIF) refers to technology to estimate the source of voiced speech, the glottal flow, from speech signals. When a new GIF algorithm is proposed, its accuracy needs to be evaluated. However, the evaluation of GIF is problematic because the ground truth, the real glottal volume velocity signal generated by the vocal folds, cannot be recorded non-invasively from natural speech. This absence of the ground truth has been circumvented in most previous GIF studies by using simple linear source-filter synthesis techniques with known artificial glottal flow models and all-pole vocal tract filters. Moreover, in a few previous studies, physical modeling of speech production has been utilized in synthesis of the test data for GIF evaluation. The evaluation strategy in previous GIF studies is, however, scattered between individual investigations and there is currently a lack of a coherent, common platform to be used in GIF evaluation. In order to address this shortcoming, the current study introduces a new environment, called OPENGLOT, for GIF evaluation. The key ideas of OPENGLOT are twofold: the environment is versatile (i.e., it provides different types of test signals for GIF evaluation) and open (i.e., the system can be used by anyone who wants to evaluate her or his new GIF method and compare it objectively to previously developed benchmark techniques). OPENGLOT consists of four main parts, Repositories I-IV, that contain data and sound synthesis software. Repository I contains a large set of synthetic glottal flow waveforms, and speech signals generated by using the Liljencrants-Fant (LF) waveform as an artificial excitation, and a digital all-pole filter to model the vocal tract. Repository II contains glottal flow and speech pressure signals generated using physical modeling of human speech production. Repository III contains pairs of glottal excitation and speech pressure signal generated by exciting 3D printed plastic vocal tract replica with LF excitations via a loudspeaker. Finally, Repository IV contains multichannel recordings (speech pressure signal, electroglottogram, high-speed video of the vocal folds) from natural production of speech. After presenting these four core parts of OPENGLOT, the article demonstrates the platform by presenting a typical use case.

Keywords: speech production, glottal flow, glottal inverse filtering, evaluation tool

1 Introduction

Glottal inverse filtering (GIF) refers to a technology to estimate the source of voiced speech, the glottal volume velocity waveform, either from the acoustic speech pressure waveform recorded outside the lips by microphone (Strube, 1974; Wong et al., 1979) or from the oral flow captured with a specially-constructed pneumotachograph mask (Rothenberg, 1973). GIF techniques first estimate the vocal tract filtering effect computationally, presenting it typically in the form of a digital filter. The effect of the vocal tract is then canceled from the recorded speech signal by filtering the signal through the inverse model of the vocal tract. In case the input to GIF is the pressure signal recorded outside the lips, the lip radiation effect (i.e., conversion of the flow at the lips into a pressure signal in the free field) must be taken into account. For low frequencies, the lip radiation effect can be estimated as the time-derivative of the flow at the lips (Flanagan, 1972) and it is digitally modeled typically by a first-order differentiator.

GIF methods have been developed by many researchers over the past five decades. Well-known examples of GIF techniques are, for example, the closed phase (CP) covariance analysis (Wong et al., 1979), iterative adaptive inverse filtering (IAIF) (Alku, 1992) and the zeros of the z-transform (ZZT) technique (Bozkurt et al., 2005). Readers interested in the history of GIF are referred to the review articles by Alku (2011) and Drugman et al. (2014). Despite its long history, GIF is still today a topic of active research as evidenced by several recently proposed techniques such as the quasi-closed phase (QCP) analysis (Airaksinen et al., 2014), inverse filtering based on extended Kalman filtering (Sahoo and Routray, 2016), and the iterative optimal pre-emphasis technique (Mokhtari and Ando, 2017). GIF is a reasonable method to study human speech production particularly because the methodology enables non-invasive analysis of the glottal flow. Therefore, GIF has been used in recent decades in several areas of speech and voice science such as in the analysis of voice quality (Childers and Lee, 1991; Gobl and Chasaide, 1992) and vocal emotions (Cummings and Clements, 1995; Gobl and Ní Chasaide, 2003; Airas and Alku, 2006), in the research of intensity regulation of speech (Gauffin and Sundberg, 1989; Titze and Sundberg, 1992; Alku et al., 2006a), as well as in the study of occupational voice (Vilkman et al., 1997; Vilkman, 2004; Lehto et al., 2008) and singing (Sundberg et al., 2005; Arroabarren and Carlosena, 2006; Björkner et al., 2006). In addition to these fundamental research-oriented applications, GIF has also been used in more technologically-oriented studies, particularly in the area of speech synthesis (Raitio et al., 2011; Airaksinen et al., 2016; Sorin et al., 2017; Hwang et al., 2018; Cui et al., 2018) but also in automatic speaker recognition (Plumpe et al., 1999; Kinnunen and Alku, 2009). Despite the fact that GIF analysis is relatively easy to conduct, one has to keep in mind that most GIF methods suffer from drawbacks such as the linearity assumption of the speech production process and poor estimation accuracy in the analysis of high-pitched voices (Drugman et al., 2014).

The ultimate goal in all GIF methods is to estimate the time-domain waveform of the true glottal flow generated by the vocal folds with maximum accuracy. However, absolute flow values, including the DC flow, cannot be measured unless the GIF method under investigation is based on the use of a calibrated pneumotachograph mask. Since the majority of the developed GIF methods do not take advantage of the oral flow recorded by the pneumotachograph mask, this article will treat from now on only those GIF methods that use as input the speech pressure signal recorded outside the lips. Ideally, evaluation of the accuracy of this kind of GIF method calls for comparing the following two time-domain signals: (1) the ground truth, i.e., the true glottal volume velocity waveform generated by the vocal folds, and (2) the estimate, i.e., the waveform computed with a GIF method using the speech pressure signal recorded outside the lips. Unfortunately, the former signal is extremely difficult, if not impossible, to be acquired with any flow sensor, at least if natural production of speech is to be preserved. This absence of the ground truth constitutes a serious principal obstacle to the accuracy evaluation of any GIF method. This problem has been circumvented in previous GIF studies by taking advantage of a few main evaluation methodologies that will be described in the following. It is worth emphasizing that the selection of the evaluation strategy is scattered between investigations, and in the study area of glottal inverse filtering there is currently both a lack of coherent evaluation strategy and a lack of a common platform to cope with the problem caused by the missing ground truth.

Since direct measurement of the glottal flow during natural production of speech is complicated, only a few GIF studies have used such measurements. Cranen and Boves, however, published a series of articles in the 1980s using four types of simultaneous measurements from the production of natural speech (Cranen and Boves, 1982, 1985b,a, 1988). Their studies involved the following four information signals: photo-electric glottogram (PEG), electroglottogram (EGG), acoustic speech signal recorded in the free field, and, most importantly, direct pressure measurements from two positions below the glottis in the trachea and from two positions above the glottis in the pharynx. The pressure measurements were conducted using a miniaturized transducer mounted on a plastic catheter. The catheter was inserted via the nasal passage and through the glottis in such a manner that two of the measurement points were below the glottis and two above the glottis. The authors reported that the catheter did not interfere with phonation at low and medium levels of vocal effort and that the test subjects, which were five adult males (Cranen and Boves, 1982), were able to produce different kinds of speech signals varying from sustained vowels to spontaneous speech. By assuming plane wave propagation, the pressure measurements were used to estimate both the tracheal and pharyngeal flow signals. In principle, the glottal flows computed using the direct pressure measurement approach by Cranen and Boves could be used as reference signals in GIF evaluation, since the data collected also included the corresponding default input of GIF, the acoustic speech signal. Observations reported, however, suggest that the method might not yield consistent flow estimates for all vowels (Cranen and Boves, 1985a) and that the pharyngeal flow estimates might be less convincing than the tracheal flow estimates (Cranen and Boves, 1985b). To the best of our knowledge, the data collected by Cranen and Boves has not been made publicly available and has not been used in GIF evaluation except for the preliminary study reported by Cranen and Boves (1982).

Due to difficulties in acquiring the ground truth from real speech, some of the previous GIF studies have used non-acoustic information signals recorded during the natural production of speech in assessment of glottal flows estimated by GIF. The most widely used of such non-acoustic information signals are EGG (Colton and Conture, 1990), laryngeal film analysis (Krishnamurthy and Childers, 1981), and high-speed video (HSV) of the vocal folds (Eysholdt et al., 1996). Performance evaluation of GIF with the help of non-acoustic information signals has been conducted, for example, by comparing time-based parameters, such as the open quotient, computed from the simultaneously recorded EGG signals to those extracted from the estimated glottal flows (Fröhlich et al., 2001; Fu and Murphy, 2006; Bone et al., 2010; Ghosh and Narayanan, 2011). Even though simultaneous recordings of different information signals undoubtedly help in obtaining a more versatile insight into the speech production process, the procedure suffers from two drawbacks when used in GIF evaluation. First, the glottal area function, which can be estimated, for example, using high-speed digital imaging of the vocal folds is not directly proportional to the corresponding glottal flow due to vocal tract inertance (Titze, 2006a). Second, some of the non-acoustic signals, particularly EGG and HSV, are unable to give reliable information for all types of phonation and for all parts of the flow pulse during the glottal cycle. As an example, EGG is typically capable of indicating accurately only the abrupt instants of glottal closures while its performance deteriorates in the estimation of more gradual glottal opening instants (Baer et al., 1983). Similarly, glottal area functions estimated from HSV might suffer from low image quality and poor time resolution, especially when processing high-pitched speech (Hertegård et al., 2003).

When a new GIF algorithm is proposed and when its performance in the processing of natural speech is demonstrated, the existence of a flat, horizontal closed phase is typically used as the criterion for the goodness of the glottal flow estimate. This evaluation criterion was used already in early analog inverse filtering studies (e.g., Lindqvist-Gauffin, 1964). Since this kind of criterion is vague and might not be correct in soft phonation, researchers interested in GIF evaluation have replaced natural utterances with synthetic speech generated using a known, artificial glottal flow waveform. This evaluation strategy has been used in various studies (e.g., Strik, 1998; Fu and Murphy, 2006; Airaksinen et al., 2014; Sahoo and Routray, 2016) by utilizing, for example, the Liljencrants-Fant (LF) (Fant et al., 1985) pulse as the synthetic glottal source. Synthetic test data has almost exclusively been computed by a (linear) convolution between synthetic glottal source signals and digital all-pole filters modeling the vocal tract of vowel sounds. Even though this approach is easy to conduct and makes generating a large set of test sounds possible, the methodology suffers from a chicken and egg

problem: the GIF method under evaluation is typically based on the same general assumptions about voice production (e.g., linear, time-invariant relationship between the source and tract) which the test sound synthesis also relies on. Therefore, the use of these kinds of straightforward synthetic test vowels might bias the evaluation results.

As an alternative to the above-mentioned strategy, some previous studies have taken advantage of physical modeling of human voice production in the evaluation of GIF algorithms (Alku et al., 2006b, 2009; Chien et al., 2017). This approach is in principle different from the use of synthetic sounds referred to above because the generation of the test data is based on simulation of physical laws in sound production and transmission rather than on a convolution of a pre-selected artificial glottal pulseform and a digital filter modeling the vocal tract. As an example of the physical modeling approach, the GIF evaluation reported by Alku et al. (2009) used the following procedures in the generation of test vowels with a known, physically-produced glottal excitation. First, self-sustained vocal fold vibration was simulated with three masses coupled to one another through stiffness and damping elements according to Story and Titze (1995). The input parameters consisted of physical and physiological issues such as, for example, lung pressure, vocal fold length and thickness, which were transformed to mechanical model parameters such as mass, stiffness, and damping. The vocal fold model was coupled to the pressures and flows in the trachea and vocal tract according to Titze (2002), thus allowing for self-sustained oscillation. Second, acoustic wave propagation in the trachea and vocal tract was simulated with the wave-reflection technique (Strube, 1982) by discretizing the area functions of the trachea and vocal tract into short cylindrical sections. The physically-generated reference used in the GIF evaluation, the glottal flow, was determined by the interaction of the glottal area with the time-varying pressures present inferior and superior to the glottis according to Titze (2002). The vocal tract was represented by an area function, specific to the underlying vowel to be produced (the vowel [a] was used in Alku et al., 2009). The vocal tract area function was measured from vowel productions by natural talkers using magnetic resonance imaging (MRI) (Story et al., 1996).

A new type of experimental evaluation environment for GIF was proposed by Chu et al. (2013). They constructed a special apparatus consisting of a 3-microphone impedance head that is connected to a physical vocal tract model, which is made of stacked plexiglas dics. The impedance head is first used to measure the input impedance of the vocal tract seen at the glottis. Once the tract has been measured, a known periodic glottal flow is injected at the glottis through an audio interface and amplifier. The radiated pressure signal can then be measured at the mouth opening of the apparatus. Chu et al. (2013) demonstrated their evaluation environment by synthesizing several glottal flow pulseforms with varying values of the open quotient. Two simple inverse filtering methods, IAIF (Alku, 1992) and direct inverse filtering (Airas, 2008), were compared by estimating the glottal flow from the radiated pressure signal and by using the injected glottal flow as the reference.

The previous studies referred to above indicate that many candidate methods have been proposed to assess GIF accuracy in the absence of the ground truth. Individual studies typically take advantage of just one or two of the evaluation procedures described above. Moreover, the test procedure is typically designed for an individual GIF study and the test data is not necessarily made publicly available. Given this regrettable situation, the current study was launched aiming at an environment that could be used jointly by all researchers interested in the evaluation of inverse filtering algorithms. The key idea is to provide an environment that is both versatile, i.e., the tool provides different types of test signals for GIF evaluation, and open, i.e., the system can be used by any developer of GIF algorithms who wants to evaluate her or his new inverse filtering method and compare it objectively to previously developed benchmark techniques. The environment, entitled the open environment for evaluation of glottal inverse filtering (OPENGLOT), consists of the following four data repositories: (1) A large set of synthetic glottal flow waveforms and speech pressure signals, including different vowels and phonation types generated by using the LF waveform as an artificial excitation and a digital all-pole filter to model the vocal tract. (2) A set of glottal flow and speech pressure signals generated using physical modeling of human speech production. (3) A set of glottal excitation and sound pressure signals that have been obtained by feeding an LF-modelled artificial glottal excitation via a loudspeaker into a plastic vocal tract model and by recording the output sound at the mouth-

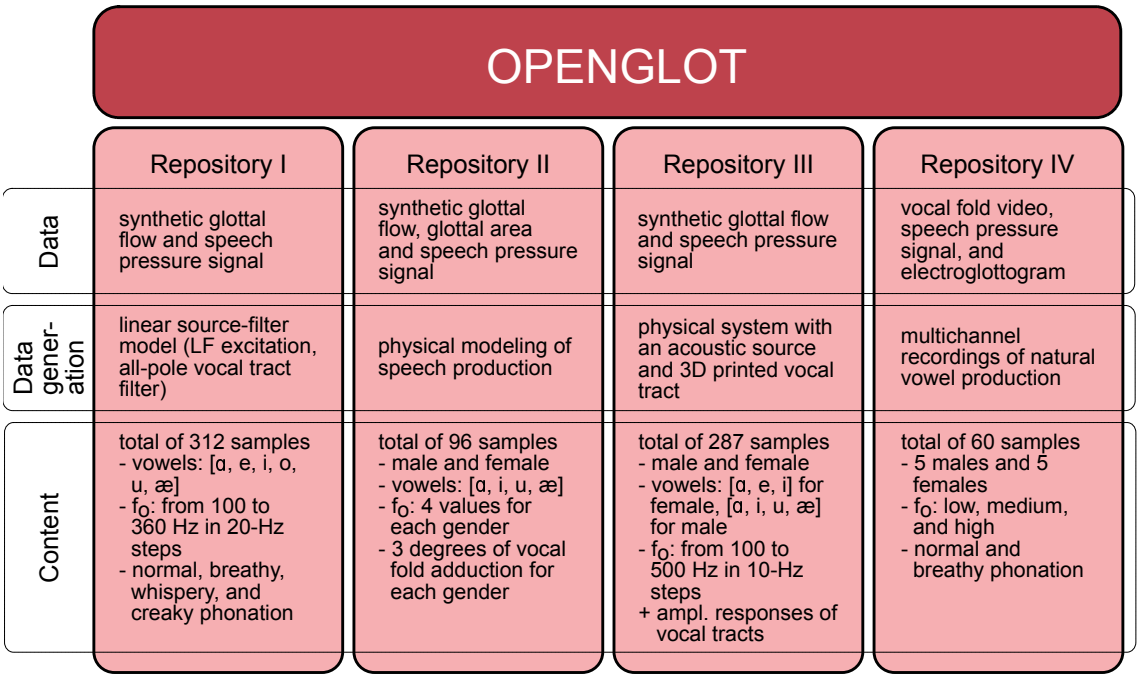

Figure 1: General structure of OPENGLOT.

end of the tract. The plastic vocal tract is a 3D printed replica of a natural speaker's vocal cavity constructed using MRI data. (4) A set of multichannel data including three simultaneously recorded information signals (speech pressure signal, EGG, high-speed imaging of the vocal folds) from natural productions of vowels by five female and five male speakers. In the following section, the four parts of OPENGLOT are first described, after which Section 3 demonstrates a typical use case in which two known GIF algorithms are compared using the proposed tool.

2 The OPENGLOT system

2.1 General

OPENGLOT is an open, web-based system that allows speech researchers free access to download signals (both audio and high-speed video of the vocal folds) and software of vowel productions (both natural and synthetic utterances). The system has been primarily designed for the evaluation of glottal inverse filtering algorithms. Therefore, OPENGLOT provides a multitude of time-domain signal pairs of speech pressure, $p(n)$, and glottal excitation, $g(n)$, so that the user can feed $p(n)$ as an input to the GIF algorithm under evaluation and test the method's accuracy by comparing the obtained glottal flow estimate against $g(n)$. In order to enable versatile evaluation of GIF methods, the OPENGLOT system provides signal pairs of $p(n)$ and $g(n)$ that have been produced with three principally different synthesis approaches. In addition, the system provides multichannel data from natural production of vowels. Altogether, OPENGLOT consists of four main components (Fig. 1), data repositories I-IV, which will be described in detail in the following sub-sections. It should be pointed out that the four repositories are independent in the sense that the user can either take advantage of all of them in a sequential manner or just use some of them.

2.2 Data Repository I: synthetic sounds generated with a linear source-filter model

Repository I includes a multitude of synthetic signal pairs, produced with an 8 kHz sampling rate. Each pair consists of an LF excitation and the corresponding speech pressure signal that has been obtained by filtering the LF excitation with a digital all-pole filter model of the vocal tract. In order to synthesize sounds with greatly different glottal source characteristics, both the phonation type and fundamental frequency (f_o) of the LF excitation were varied. The phonation type was varied by expressing the excitation pulse in the form of dimensionless LF parameters (E_e, R_a, R_g, and R_k) and using parameter value combinations corresponding to normal, breathy, whispery, and creaky phonation according to data published by Gobl (1989). f_o was varied between 100 and 360 Hz with a step of 20 Hz. Vocal tract was modeled by simulating resonance structures of vowels [ɑ, e, i, o, u, æ] using a 9^{th} order all-pole filter. Table 1 shows the first three resonances (f_{Rn}, $n = 1, 2, 3$, using the notation of Titze et al. (2015)) of each vowel. The fourth resonance is set at $f_{R4} = 3500$ Hz for all vowels. In total, the procedure yielded 52 different LF excitation waveforms (four phonation types x 13 f_o values) and 312 different speech pressure signals (6 vowels x 52 LF excitation waveforms). The repository also includes the MATLAB code that was used to generate the data so that the user can either download the signals or the synthesis software.

Table 1: Resonance frequencies (in Hz) for each vocal tract configuration in Repository I.

Vowel	f_{R1}	f_{R2}	f_{R3}
ɑ	730	1090	2440
e	530	1840	2480
i	390	1990	2550
o	570	840	2410
u	440	1020	2240
æ	660	1720	2410

2.3 Data repository II: Synthetic sounds generated with physical modeling of speech production

Repository II contains a collection of signals that were generated with an airway modulation model of speech production (cf. Story, 2013) configured to represent both an adult male and adult female. The model includes a tracheal airspace, a kinematic representation of the vibrating portion of the vocal fold medial surfaces that modulate the glottal airspace (Titze, 1984, 2006b), and a vocal tract that can be shaped to produce acoustic resonances representative of a wide variety of vowels. Fig. 2 shows a schematic representation of the model with settings for an adult male speaker. Vocal tract configurations of the vowels [ɑ, i, u, æ] and the trachea are plotted as area functions (cross-sectional area as a function of distance from the glottis), and the vocal folds are shown as sheet-like structures representing the medial surfaces of the vocal folds. For the adult male model, the vocal tract has an overall length of 17.5 cm and the vocal folds are set to a rest length of 1.5 cm. Although not shown in the figure, there is a side branch resonator, intended to represent the piriform sinuses, coupled to the main vocal tract at a location +2.4 cm from the glottis.

Airway modulations produced at the level of the vocal folds operate on two time scales, one that is representative of their vibrational frequency (approximately 80–400 Hz) and another for the much slower adductory and abductory movements of the medial surfaces that take place during the unvoiced parts of speech. The output of the kinematic vocal fold model is the glottal area $A_g(t)$, which, when aerodynamically and acoustically coupled to the trachea and vocal tract (Liljencrants, 1985; Story, 1995; Titze, 2002), generates a glottal flow signal, $U_g(t)$, the primary sound source for vowels. Examples of these two signals are shown in Fig. 2 above the vocal fold model. The sound pressure, $P_{out}(t)$ radiated at the lip termination, is generated by a wave propagation algorithm that includes losses such as yielding walls, heat conduction, and viscosity, as well as radiation impedance. This

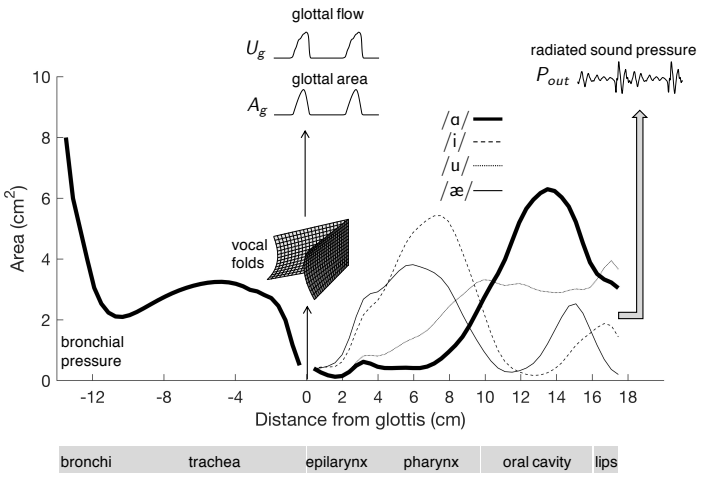

Figure 2: Schematic representation of the airway modulation model configured to represent an adult male speaker. Vocal tract configurations of the vowels [ɑ, i, u, æ] and the trachea are plotted as area functions, and the vocal folds are shown as sheet-like structures representing the medial surfaces of the vocal folds. The origin of the plot is located at the glottis. When configured as an adult male, the vocal tract has an overall length of 17.5 cm and the vocal folds are set to a rest length of 1.5 cm. Although not shown in the figure, there is a side branch resonator, intended to represent the piriform sinuses, coupled to the main vocal tract at a location +2.4 cm from the glottis.

signal, shown for the vowel [ɑ] in Fig. 2, is analogous to a microphone signal in a recording of natural speech.

The signals included in Repository II consist of glottal area, glottal flow, and radiated sound pressure generated for all combinations of three variable settings: four vocal tract configurations representative of the vowels [ɑ, i, u, æ], four fundamental frequencies of 82, 110, 156, and 220 Hz, and three degrees of vocal fold adduction (ξ_{02}) consisting of 0.03, 0.06, and 0.09 cm. The latter variable ξ_{02} is the distance of the vocal processes from the glottal midline; thus, large values of this setting tend to produce breathy voice qualities, whereas small values tend toward a pressed voice quality. For each simulation (i.e., each set of signals), the bronchial pressure was set to 8000 dyn/cm^2 and the duration was 1.0 s. The f_o was constant during the intervals 0–0.3 s and 0.7–1.0 s, but between 0.3 and 0.7 s, the f_o was increased and decreased to 1.15 times and 0.85 times the baseline f_o value, respectively. In all, there are 48 sets of $A_g(t)$, $U_g(t)$, and $P_{out}(t)$ signals generated by the model when configured as an adult male speaker. The sampling frequency for all signals is $f_s = 44100$ Hz.

Fig. 3 shows an example set of signals generated with the airway modulation model for the case where the baseline $f_o = 110$ Hz, the vocal tract was configured as an [ɑ] vowel, and the vocal fold adduction was set to $\xi_{02} = 0.06$ cm. The time-dependent f_o contour is plotted in the lower panel, and the A_g, U_g, and P_{out} signals are shown in the upper three panels, respectively. The vertical lines indicate five segments (labelled A to E) where there are parametric differences that may be useful for testing algorithms. The segments are defined as:

A: This segment extends from 0 to 0.15 s (sample interval $[1 : 6615]$). The f_o is constant at the baseline value, but there is ramp in bronchial pressure from 0 to 8000 dyn/cm^2, and a change in adduction starting from 0.05 cm beyond the baseline value of ξ_{02} to the baseline value itself. The latter two parameter variations generate a soft onset of voicing.

B: This segment extends from 0.15–0.30 s (sample interval $[6616 : 13230]$). The f_o is constant at the baseline value throughout the segment, as are all other parameters.

C: Extending from 0.30–0.70 s (sample interval $[13231 : 30870]$), this segment contains a variation

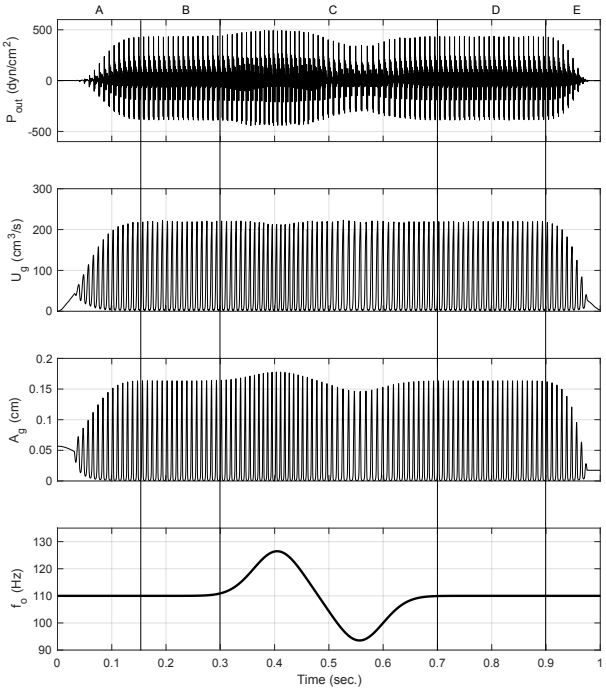

Figure 3: Example set of signals in Repository II obtained with model settings typical for an adult male. From top to bottom, the panels show radiated sound pressure P_{out}, glottal flow U_g, glottal area A_g, and fundamental frequency f_o. Note that the f_o contour was prescribed whereas the three signals were the output of the simulation. The segments A-E are detailed in the text.

in f_o in which the baseline value is first increased by a factor of 1.15 and then decreased by a factor of 0.85. At 0.70 seconds, the f_o returns to the baseline value.

D: This segment extends from 0.70–0.90 s (sample interval [30871 : 39690]). Just as in segment B, the f_o is constant at the baseline value throughout the segment, as are all other parameters.

E: During this final segment that extends from 0.90–1.0 s (sample interval [39691 : 44100]), the f_o and an adduction both remain constant at their baseline values, but the bronchial pressure is ramped from 8000 to 0 dyn/cm^2.

A second set of signals was similarly generated based on the model configured to represent an adult female speaker. To transform the model to be typical for a female speaker, the vocal tract length was set to 15.5 cm, and the four vocal tract area functions shown previously in Fig. 2 were modified based on the method reported in Story et al. (2018). In addition, the resting vocal fold length was set to 1.1 cm, the four fundamental frequencies were set to 175, 196, 220, and 294 Hz, and the degree of adduction was set to 0.05, 0.07, and 0.09 cm. All other parameter values were identical to the values used in the signals generated with model settings typical for an adult male.

The calculated vocal tract resonance frequencies (i.e., formants) are provided in Table 2. These values were derived from calculations of the frequency response of each area function based on a transmission line technique (Sondhi and Schroeter, 1987; Story et al., 2000) that includes losses due to yielding walls, viscosity, heat conduction, and radiation at the lips. The effect of the piriform sinus was also included in these calculations. More details about the arrangements of data structures in Repository II are available on the OPENGLOT web page, see http://research.spa.aalto.fi/projects/openglot/.

Table 2: Calculated resonance frequencies (in Hz) for each vocal tract configuration in Repository II.

	Male				Female			
Vowel	f_{R1}	f_{R2}	f_{R3}	f_{R4}	f_{R1}	f_{R2}	f_{R3}	f_{R4}
ɑ	752	1095	2616	3169	848	1210	2923	3637
i	340	2237	2439	3668	379	2634	4256	5395
u	367	1180	2395	3945	420	1264	2714	4532
æ	693	1521	2435	3252	795	1700	2692	3740

2.4 Data Repository III: sounds generated with a physical system consisting of an acoustic glottal source and 3D printed vocal tract replicas

Repository III consists of sound pressure signals that have been measured from a physical system that includes a 3D printed plastic replica of the human vocal tract, excited by a custom sound source and software detailed in Hannukainen et al. (2017). The anatomic data for the vocal tract replicas were acquired by 3D Magnetic Resonance Imaging (MRI) from prolonged productions of five Finnish vowels ([ɑ, e, i, u, æ]). This MRI data was collected from one female (age 26 years, vowels [ɑ, e, i]) and one male (age 26 years, vowels [ɑ, i, u, æ]) test subject using arrangements described in Aalto et al. (2014). The MRI geometries were visually inspected for imaging errors, and their numerically computed resonance structures were compared with the formants of the speech samples that were recorded during the MR imaging.

From each of the seven MR images, the air/tissue interface was extracted by the algorithm detailed in Ojalammi and Malinen (2017). The surface models were further processed to volume models of 2 mm wall thickness so that the interior surface of the volume model corresponded to the air/tissue interface in the MR images. The files were printed by material extrusion process (Fused Deposition Modeling, FDM) from ABSplus thermoplastic with support structures consisting of SR-30 Soluble Support Material (Stratasys). Stratasys uPrint SE was used for printing with a layer thickness of 0.254 mm and a nozzle diameter of 0.35 mm. Much of the support material could be removed using WaveWash Support Cleaning System with Ecoworks cleaning agent. A custom arrangement was used to efficiently pump cleaning agent through the prints in order to completely dissolve the remaining

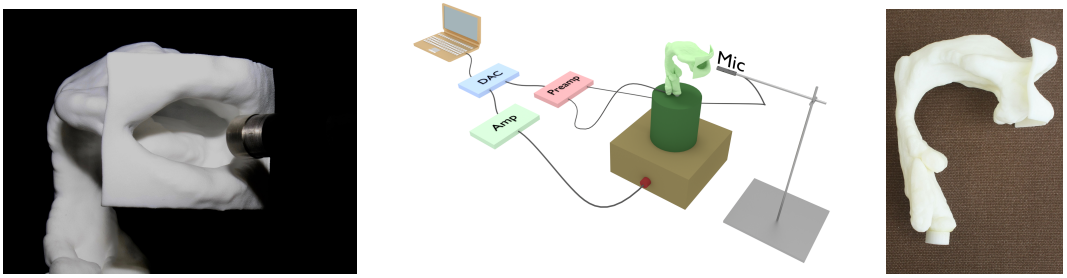

Figure 4: Left panel: Close-up of the measurement arrangement of [ɑ] from a 3D printed vocal tract model. Middle panel: An illustration of the full measurement system (without soundproofing material) used in generating synthetic vowels in Repository III. Right panel: 3D printed vocal tract of the vowel [ɑ].

support material inside the model. The final quality of these prints was inspected visually and by resonance measurements.

For comparison, additional prints were produced using powder bed fusion from polyamide (PA 2200) material using EOS Formiga P100. Because this kind of 3D printing technology does not require support structures, challenges related to dissolving blockages of support material inside the prints were avoided. The polyamide prints appear somewhat more elastic than the ABS prints used for the measurements. More importantly, it was observed in comparisons that the acoustic resonances, measured from the ABS prints, are very sensitive to residual support structures if some of it remains undissolved. The material of the print was observed to not affect the measured resonance frequencies very much, even though the polyamide prints produce slightly wider resonance peaks.

For each of the seven vocal tract models, two kinds of acoustic data are provided in Repository III:

1. Amplitude response for a frequency range of between 80 Hz and 7350 Hz, obtained by producing a constant pressure amplitude, logarithmic, sinusoidal sweep at the vocal folds position of the model.

2. Sound pressure vowel signal, produced by reconstructing the LF pulse excitation waveform at fundamental frequencies between 100 Hz and 500 Hz at the vocal folds position of the model.

In both kinds of data, the output signal was recorded ca. 10 mm in front of the lips of the vocal tract model, see Fig. 4 (left panel). The Brüel & Kjæll 4188 condenser microphone was used as a free-field microphone, coupled to an RME Babyface digitizer. An inexpensive electret capsule of generic type was embedded near the vocal folds position inside the sound source to produce the reference signal. Custom software written on MATLAB (R2017a) was used for signal processing; Playrec (a MATLAB utility, Humphrey (2011)) and Audacity v.1.3.14 were used as interfaces to the RME Babyface. All experiments were carried out in an anechoic chamber.

The specification, design, signal processing algorithms, and construction of the instrumentation is described in detail by Hannukainen et al. (2017). Similar physical sound generation systems have also been used in a few other studies (e.g., Epps et al., 1997; Wolfe et al., 2001; Chu et al., 2013). It is not possible to design a suitable sound source to meet ideal acoustic requirements. Fortunately, most of the physical non-idealities can be compensated numerically as reported in Hannukainen et al. (2017).

For all seven vocal tract models, a large number of measurements were carried out. There were 26 amplitude sweeps for each model. To obtain vowel signals, LF waveforms were generated by varying f_o in steps of 10 Hz, resulting in 40 frames 10 seconds in length for each model. From these samples, a segment of 200 ms was extracted near the middle of the frame for Repository III.

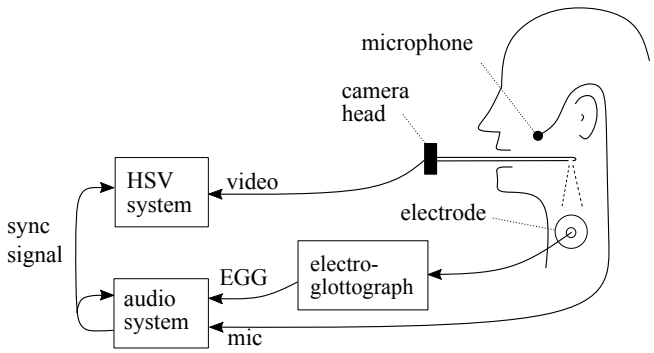

Figure 5: Schematic diagram describing collection of data on natural production of speech for Repository IV. Vocal fold movements are imaged using a rigid endoscope connected to the HSV system. Simultaneously, the audio system records EGG and speech signal from a microphone headset. A custom synchronization signal is recorded with the video, EGG, and microphone signals.

2.5 Data Repository IV: multichannel recordings of natural vowel production

Repository IV differs from the other three repositories in that it includes recordings from natural speakers. The repository includes three simultaneously recorded information signals: pressure signal recorded with a free-field microphone, electroglottogram (EGG), and high-speed video (HSV) of the vocal folds. Five male and five female speakers were each instructed to produce a vowel sound using two types of phonation (normal and breathy) with three pitch levels (low, medium, and high). The speakers were asked to produce the vowel [i], with their tongues as far forward as possible, in order to obtain the clearest possible view of the glottis. Due to the HSV endoscope, however, the resulting sounds ranged between [æ] and [œ].

Fixed targets for pitch and phonation type were not used; instead, the speaker was asked to change his or her voice production so that the six utterances generated (2 phonation types x 3 pitches) were perceptually different. The production was monitored by an experienced experimenter who asked the speaker to repeat the task if a sufficiently large difference in the phonation type and pitch was not observed. Full listing of the the 60 recorded sounds (200 ms each), including the gender and age information of the speakers, is given in Table 3. Although the recorded utterances have a large variability, the use of fixed targets would have made the already challenging vocalization tasks more difficult for the speaker and would have lead to an increased number of repetitions and, hence, increased discomfort for the test subject. The duration of the measurements depended on the speaker's skills, tolerance for the endoscope and the heat emitted by its light source, as well as anatomical and technical factors. Typically, 2–3 hours were required per speaker to record the six utterances.

Fig. 5 illustrates the process of data collection schematically. The measurements were performed at Helsinki University Central Hospital using the KayPentax Color High-Speed Video System (model 9710) with a spatial resolution of 512 x 512 pixels and a temporal resolution of 2000 frames/s. EGG was acquired with a Glottal Enterprises electroglottograph (EG2-PCX2). A DPA omnidirectional headset microphone (model 4065-BL) was set 6.5 cm from the centre of the speaker's mouth. The microphone signal and EGG were recorded using a MOTU UltraLite-mk3 Hybrid audio interface connected to a MacBook Pro running OS X (v. 10.9.5) and AudioDesk 4. To enable synchronization of the audio signals with the video, a synchronization signal comprising binary frequency-shift keyed code at the beginning of each second was adopted. This signal was played in AudioDesk simultaneously with the recording and directed from the audio interface to the high-speed unit's audio capture module, and was looped back as input to the audio interface.

The microphone and EGG signals were high-pass filtered (cut-off frequency 60 Hz, linear phase) and synchronized to the high-speed video by aligning the synchronization signals and shifting them to

Table 3: Multichannel data of natural vowel production collected in Repository IV from five male (M01–M05) and five female (F01–F05) speakers with fundamental frequency (f_o) computed for each recoding by task (phonation type, pitch).

			f_o (Hz)		
Speaker	Age	Phonation type	low	medium	high
M01	36	normal	120	122	195
		breathy	107	98	122
M02	61	normal	94	118	297
		breathy	99	113	227
M03	28	normal	125	130	173
		breathy	107	147	204
M04	26	normal	114	152	178
		breathy	89	104	168
M05	29	normal	114	282	273
		breathy	94	109	219
F01	26	normal	168	231	282
		breathy	163	154	376
F02	25	normal	232	256	526
		breathy	167	295	516
F03	28	normal	179	229	353
		breathy	150	188	242
F04	40	normal	186	293	321
		breathy	183	255	383
F05	25	normal	185	281	405
		breathy	213	243	446

account for various delays. The delays, including propagation delays and internal delays within and between the measurements systems, were estimated to be approximately 1.6 ms for males and 1.5 ms for females. After the completion of this alignment, the maximum error in the synchronization of the EGG signal to the video is ±0.5 ms (one frame). In addition, the alignment of the microphone signal to the EGG can have an error of ±0.08 ms at most, due to the estimation of the propagation delays.

3 A case of typical use

In order to demonstrate OPENGLOT, an example of typical use of the platform is described next. In this example, the (hypothetical) user compares two existing GIF algorithms, IAIF (Alku, 1992) and QCP (Airaksinen et al., 2014), using the data provided by the OPENGLOT environment. IAIF and QCP were selected as GIF methods for this example because they have been implemented in the same tool, Aalto Aparat (Alku et al., 2017), and this tool was easy to use to carry out all the GIF analyses of the example. For this demonstration, all four repositories are utilised, but the order used is I, III, II, and IV. In this order, the two repositories (I and III), where the GIF-computed glottal flows are compared against the corresponding LF excitation waveforms, will be demonstrated one after the other. Likewise, the use of the two repositories (II and IV) containing glottal area information are demonstrated consecutively.

Repository I contains pairs of glottal flow and speech pressure signal to be used in assessing the accuracy of a GIF algorithm. This data has been generated by assuming that voice production is a linear process between the excitation and vocal tract filter and that the latter can be modelled as an all-pole filter. Fig. 6 (left panel) shows the glottal flow estimates computed by IAIF and QCP as well as the original LF excitation waveform in the case when the user analyzed the vowel [ɑ] with f_o=100 Hz and normal phonation from Repository I. Three time quotients (open quotient OQ, closing quotient ClQ, and normalized amplitude quotient NAQ) commonly used to parameterize glottal flow pulses are

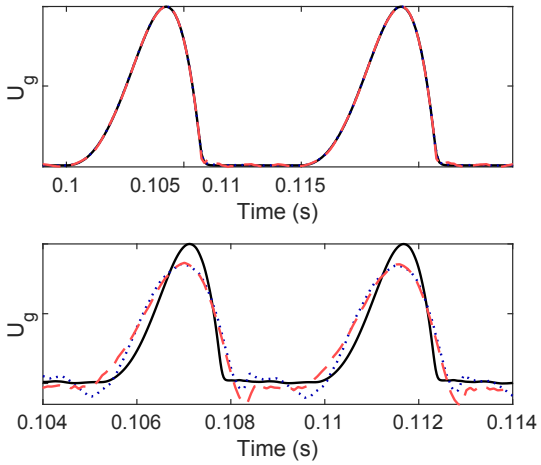

Figure 6: Examples of the use of Repositories I and III in GIF evaluation. Reference glottal flow (solid), as well the glottal flows estimated using the IAIF method (dashed) and the QCP method (dotted) are shown for two data pairs. Top panel: [ɑ] with f_o=100 Hz and normal phonation from Repository I. Bottom panel: female vocal tract geometry of the vowel [e] with $f_o = 220$ Hz from Repository III.

shown in Table 4 for the reference flow and the flow estimates of the example. It can be seen that the two GIF methods produce virtually identical parameter values for this example, and the differences between them and the LF ground truth are small.

Table 4: Parameters describing reference glottal flow and glottal area pulses (where applicable) and the same parameters for the estimated glottal flows in the examples of typical use detailed in Section 3.

	Parameter	Reference U_g	Reference A_g	IAIF	QCP
Repository I	OQ	0.690	-	0.721	0.728
	ClQ	0.236	-	0.267	0.270
	NAQ	0.089	-	0.088	0.091
Repository II	OQ	0.422	0.421	0.483	0.512
	ClQ	0.103	0.155	0.095	0.126
	NAQ	0.025	0.084	0.029	0.040
Repository III	OQ	0.609	-	0.768	0.705
	ClQ	0.187	-	0.288	0.276
	NAQ	0.083	-	0.165	0.144
Repository IV	OQ	-	0.835	0.778	0.819
	ClQ	-	0.423	0.245	0.271
	NAQ	-	0.249	0.123	0.132

Good GIF accuracy demonstrated by the previous example using data of Repository I is desirable but it does not necessarily indicate that the algorithms to be evaluated perform well when the data starts to deviate from the basic GIF assumptions or when f_o increases. The data in Repository III can be used in GIF evaluation in a similar manner to Repository I: the source signal remains LF-based but as physical measurements of using the 3D printed vocal tract models do not fully fit the all-pole filter assumption with linear source–filter coupling, the test data becomes more challenging for GIF algorithms. Fig. 6 (right panel) shows the reference glottal flow for female vocal tract geometry of the [e] vowel with $f_o = 220$ Hz as well as the glottal flow estimates obtained using the IAIF and QCP methods. Both algorithms produce glottal excitation estimates that differ from the reference signal both visually and in terms of pulse parameters (Table 4), although the parameters of the QCP-

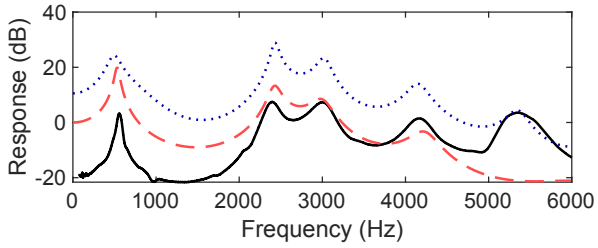

Figure 7: Amplitude responses of the vocal tract for the Repository III example with female vocal tract geometry of the [e] vowel: average response from sweeps of the 3D printed vocal tract (solid), vocal tract response estimated using the IAIF method (dashed), and the QCP method (dotted).

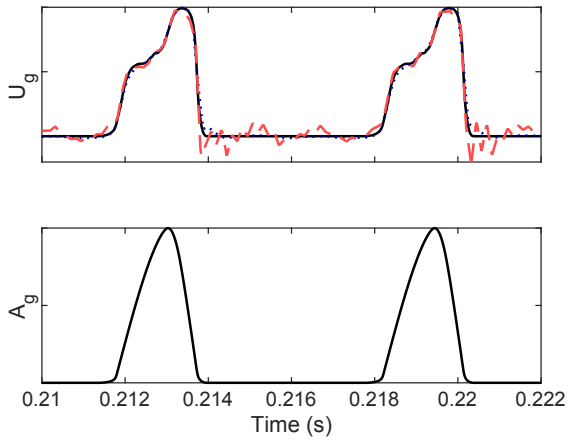

Figure 8: An example of the use of Repository II in GIF evaluation. Top panel: Reference glottal flow (solid), as well the glottal flows estimated using the IAIF method (dashed) and the QCP method (dotted). Bottom panel: Area of the glottal opening. Data corresponding to male vocal tract geometry of the [æ] vowel with $f_o = 156$ Hz and $\xi_{02} = 0.03$ cm is used.

based pulses are slightly closer to the reference values. Repository III also enables the comparison of estimated vocal tract responses to those measured from the 3D prints. Fig. 7 shows that, for this example with female vocal tract geometry of the [e] vowel, both GIF methods match the resonance structure well, although the matching of f_{R1} is slightly better with IAIF.

Repository II provides speech pressure–glottal flow pairs, as well as additional information regarding the utterances, such as the area of the glottal opening. The physical model used to generate the data contains non-linear source–filter interaction, and hence poses a challenge for GIF algorithms. Fig. 8 shows glottal flow reference and estimates for male vocal tract geometry of the [æ] vowel with $f_o = 156$ Hz and $\xi_{02} = 0.03$ cm. The sample is taken from the first stable phonation segment (segment B in Fig. 3). The delay between the pressure and glottal signals due to the 17.5 cm vocal tract has been removed. Both IAIF and QCP are capable of producing pulse shapes with similar formant ripple to the reference pulse, but they struggle to reproduce the long closed phase. The parameters of the flow estimate obtained using IAIF match the reference values better (Table 4), even though visual evaluation favors the smoother QCP flow estimate.

Compared to Repositories I-III, Repository IV cannot be used in direct evaluation of GIF algorithms because this repository does not include the ground truth glottal excitation. However, the HSV and EGG data of Repository IV provide complementary information of the glottal function, to which glottal

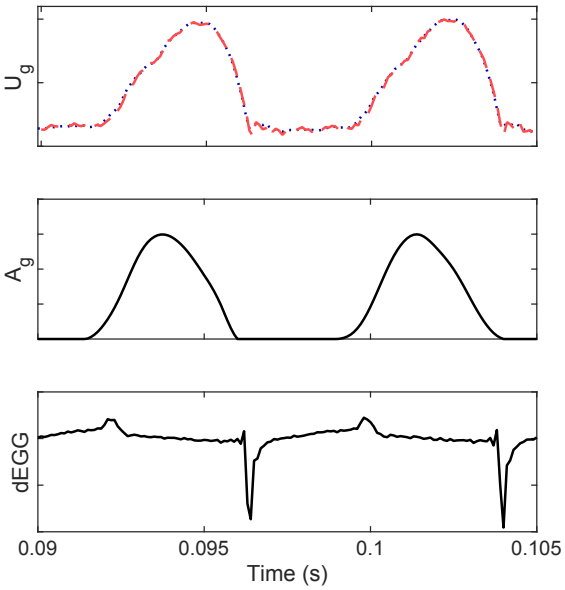

Figure 9: An example of the use of Repository IV to support GIF evaluation. Top panel: Glottal flow estimated using the IAIF method (dashed) and the QCP method (dotted). Middle panel: Area of the glottal opening extracted from the HSV data. Bottom panel: Time derivative of the EGG. The utterance with normal phonation and medium pitch from speaker M03 is used.

flow estimated using GIF can be compared. In addition, the data in Repository IV has the benefit of representing natural speech production.

Fig. 9 shows the glottal flow estimated using the IAIF and QCP methods, the area of the glottal opening extracted from the HSV data using the algorithm by Lohscheller et al. (2007) with small modifications specified in Murtola et al. (2018), and the time derivative of the EGG (dEGG). The figure demonstrates several phenomena in natural speech production that can be analyzed when the GIF-based glottal flow signal, estimated from speech pressure waveforms of Repository IV, are studied in parallel with the other two information channels (HSV, EGG) of the respository. First, the instants of glottal opening and closure extracted from the glottal flow estimated by GIF coincide, respectively, with the instants of the positive and negative peaks of dEGG. Second, the multichannel data describes how individual pulses of the glottal area are generally more symmetric than the pulses of the glottal flow, a phenomenon that is known as glottal flow skewing (c.f. e.g., Childers et al. (1985); Hertegård and Gauffin (1995)). This phenomenon is visible both in the waveforms (Figs. 8 and 9) and in the ClQ and NAQ values (Table 4). The relationship between glottal flow and area in these examples is shown in more detail in Fig. 10. These Lissajous plots can be used to illustrate and investigate phenomena observed in natural speech, such as formant ripples (illustrated by the non-convex shape of the curve in the left panel) and skewing of the flow (illustrated by the closing phase above the $U_g = A_g$ line in both panels).

4 Summary

Glottal inverse filtering (GIF) is a technique to estimate the glottal flow by cancelling vocal tract resonances from speech. Evaluating the accuracy of GIF algorithms is problematic because the ground truth, the real glottal airflow pulseform generated by the vocal folds, cannot be measured from production of natural speech. The absence of the ground truth has lead to a situation where different

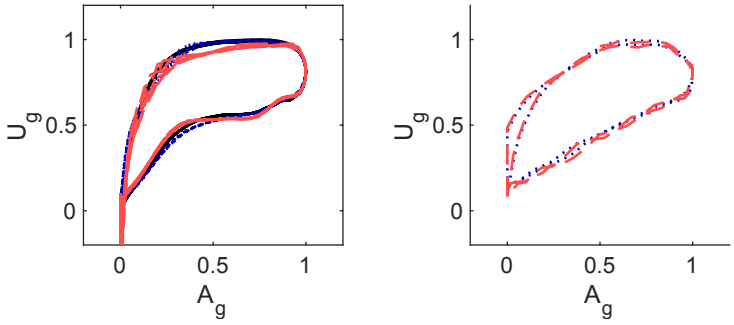

Figure 10: Lissajous plots for the glottal area-glottal flow pairs for the example uses of Repository II and IV. Left panel: Simulated flow (solid black), and glottal flows estimated using the IAIF method (dashed red) and the QCP method (dotted blue) plotted against a simulated glottal area for a sample from Repository II. Right panel: Glottal flows estimated using the IAIF method (dashed red) and the QCP method (dotted blue) plotted against a glottal area extracted from HSV data for a sample from Repository IV.

evaluation methods are used in GIF research. The evaluation methods are typically scattered between individual investigations and there is currently a lack of a common platform to be used in GIF evaluation. In order to address this shortcoming of the study area, a new open GIF evaluation platform, called OPENGLOT, is proposed in this article. OPENGLOT is a publicly available platform that includes data, both natural and synthetic, and software for GIF evaluation. The data and software are organized into four independent parts, Repositories I-IV. The OPENGLOT system is freely available for all speech researchers at http://research.spa.aalto.fi/projects/openglot/.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (projects 284671, 312490). The authors would like to thank the reviewers for their useful comments.

References

D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, T. Soukka, and M. Vainio. Large scale data acquisition of simultaneous MRI and speech. *Appl. Acoust.*, 83:64–75, 2014. doi: 10.1016/j.apacoust.2014.03.003.

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.*, 22(3):596–607, 2014. doi: 10.1109/TASLP.2013.2294585.

M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku. GlottDNN - a full-band glottal vocoder for statistical parametric speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2473–2477, San Francisco, USA, 2016. doi: 10.21437/Interspeech.2016-342.

M. Airas. TKK aparat: An environment for voice inverse filtering and parameterization. *Logoped. Phoniatr. Vocol.*, 33(1):49–64, 2008. doi: 10.1080/14015430701855333.

M. Airas and P. Alku. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, 33(1):26–46, 2006. doi: 10.1159/000091405.

P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun.*, 11(2–3):109–118, 1992. doi: 10.1016/0167-6393(92)90005-R.

P. Alku. Glottal inverse filtering analysis of human voice production - a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011. doi: 10.1007/s12046-011-0041-5.

P. Alku, M. Airas, E. Björkner, and J. Sundberg. An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. *J. Acoust. Soc. Am.*, 120(2):1052–1062, 2006a. doi: 10.1121/1.2211589.

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.*, 58(2):102–113, 2006b. doi: 10.1159/000089611.

P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. H. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *J. Acoust. Soc. Am.*, 125(5):3289–3305, 2009. doi: 10.1121/1.3095801.

P. Alku, H. Pohjalainen, and M. Airaksinen. Aalto Aparat — a freely available tool for glottal inverse filtering and voice source parameterization. In *Subsidia: Tools and Resources for Speech Sciences*, Malaga, Spain, June 2017.

I. Arroabarren and A. Carlosena. Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices. *J. Acoust. Soc. Am.*, 119(4):2483–2497, 2006. doi: 10.1121/1.2177584.

T. Baer, A. Löfqvist, and N. S. McGarr. Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques. *J. Acoust. Soc. Am.*, 73(4):1304–1308, 1983. doi: 10.1121/1.389279.

E. Björkner, J. Sundberg, T. Cleveland, and E. Stone. Voice source differences between registers in female musical theater singers. *J. Voice*, 20(2):187–197, 2006. doi: 10.1016/j.jvoice.2005.01.008.

D. Bone, S. Kim, S. Lee, and S. S. Narayanan. A study of intra-speaker and inter-speaker affective variability using electroglottograph and inverse filtered glottal waveforms. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 913–916, Makuhari, Japan, 2010.

B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal. Proc. Let.*, 12(4):344–347, 2005. doi: 10.1109/LSP.2005.843770.

Y. R. Chien, D. D. Mehta, J. Gudnason, M. Zanartu, and T. F. Quatieri. Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 25(8):1718–1730, 2017. doi: 10.1109/TASLP.2017.2714839.

D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Am.*, 90(5):2394–2410, 1991. doi: 10.1121/1.402044.

D. G. Childers, J. M. Naik, J. N. Larar, A. K. Krishnamurthy, and G. P. Moore. Electroglottography, speech, and ultra-high speed cinematography. In I. R. Titze and R. C. Scherer, editors, *Vocal fold physiology*, pages 202–220. The Dencer Center For The Performing Arts, Denver, 1985.

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. Am.*, 133 (5):EL358–EL362, 2013. doi: 10.1121/1.4798619.

R. H. Colton and E. G. Conture. Problems and pitfalls of electroglottography. *J. Voice*, 4(1):10–24, 1990. doi: 10.1016/S0892-1997(05)80077-3.

B. Cranen and L. Boves. Evaluation of glottal inverse filtering by means of physiological registrations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, page 1988–1991, Paris, France, 1982.

B. Cranen and L. Boves. Aerodynamic aspects of voicing: Glottal pulse skewing revisited. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, page 1085–1088, Tampa, FL, 1985a.

B. Cranen and L. Boves. Pressure measurements during speech production using semiconductor, miniature pressure transducers: Impact on models for speech production. *J. Acoust. Soc. Am.*, 77 (4):1543–1551, 1985b. doi: 10.1121/1.391997.

B. Cranen and L. Boves. On the measurement of glottal flow. *J. Acoust. Soc. Am.*, 84(3):888–900, 1988. doi: 10.1121/1.396658.

Y. Cui, X. Wang, L. He, and F. K. Soong. A new glottal neural vocoder for speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2017–2021, Hyderabad, India, 2018.

K. E. Cummings and M. A. Clements. Analysis of the glottal excitation of emotionally styled and stressed speech. *J. Acoust. Soc. Am.*, 98(1):88–98, 1995. doi: 10.1121/1.413664.

T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech and Language*, 28(5):1117–1138, 2014. doi: 10.1016/j.csl.2014.03.003.

J. Epps, J. R. Smith, and J. Wolfe. A novel instrument to measure acoustic resonances of the vocal tract during phonation. *Meas. Sci. Technol.*, 8(10):1112–1121, 1997. doi: 10.1088/0957-0233/8/10/012.

U. Eysholdt, M. Tigges, T. Wittenberg, and U. Pröschel. Direct evaluation of high-speed recordings of vocal fold vibrations. *Folia Phoniatr. Logop.*, 48(4):163–170, 1996. doi: 10.1159/000266404.

G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):1–13, 1985. doi: 10.1016/0167-6393(89)90001-0.

J. Flanagan. *Speech Analysis, Synthesis and Perception*. Springer, New York, 1972.

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. Am.*, 110(1):479–488, 2001. doi: 10.1121/1.1379076.

Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Trans. Audio Speech Lang. Process.*, 14(2):492–501, 2006. doi: 10.1109/TSA.2005.857807.

J. Gauffin and J. Sundberg. Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, 32(3):556–565, 1989. doi: 10.1044/jshr.3203.556.

P. K. Ghosh and S. S. Narayanan. Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. *Speech Commun.*, 53(1):98–109, 2011. doi: 10.1016/j.specom.2010.07.004.

C. Gobl. A preliminary study of acoustic voice quality correlates. *STL-QPSR*, 4:9–21, 1989.

C. Gobl and A. N. Chasaide. Acoustic characteristics of voice quality. *Speech Commun.*, 11(4–5): 481–490, 1992. doi: 10.1016/0167-6393(92)90055-C.

C. Gobl and A. Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun.*, 40(1–2):189–212, 2003. doi: 10.1016/S0167-6393(02)00082-1.

A. Hannukainen, J. Kuortti, J. Malinen, and A. Ojalammi. An acoustic glottal source for vocal tract physical models. *Meas. Sci. Technol.*, 28(11), 2017. doi: 10.1088/1361-6501/aa85a6.

S. Hertegård and J. Gauffin. Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography. *Journal of Speech and Hearing Research*, 38 (1):85–100, 1995. ISSN 00224685.

S. Hertegård, H. Larsson, and T. Wittenberg. High-speed imaging: applications and development. *Logoped. Phoniatr. Vocol.*, 28(3):133–139, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/14596332`.

R. Humphrey. Playrec, 2011. URL `http://www.playrec.co.uk/`.

M.-J. Hwang, E. Song, K. Byun, and H.-G. Kang. Modeling-by-generation-structured noise compensation algorithm for glottal vocoding speech synthesis system. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Calgary, Alberta, Canada, 2018.

T. Kinnunen and P. Alku. On separating glottal source and vocal tract information in telephony speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 4545–4548, Taipei, Taiwan, 2009. ISBN 9781424423545. doi: 10.1109/ICASSP.2009.4960641.

A. Krishnamurthy and D. Childers. Vocal fold vibratory patterns: Comparison of film and inverse filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '81*, pages 133–136, Atlanta, GA, 1981. Institute of Electrical and Electronics Engineers. doi: 10.1109/ICASSP.1981.1171353. URL `http://ieeexplore.ieee.org/document/1171353/`.

L. Lehto, L. Laaksonen, E. Vilkman, and P. Alku. Changes in objective acoustic measurements and subjective voice complaints in call center customer-service advisors during one working day. *J. Voice*, 22(2):164–177, 2008. doi: 10.1016/j.jvoice.2006.08.010.

J. Liljencrants. *Speech synthesis with a reflection-type line analog*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1985.

J. Lindqvist-Gauffin. Inverse filtering. instrumentation and techniques. *STL-QPSR*, 5:1–4, 1964.

J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, 11(4):400–413, 2007. ISSN 1361-8415. doi: 10.1016/j.media.2007.04.005. URL `http://www.sciencedirect.com/science/article/pii/S1361841507000369`.

P. Mokhtari and H. Ando. Iterative optimal preemphasis for improved glottal-flow estimation by iterative adaptive inverse filtering. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1044–1048, Stockholm, 2017. doi: 10.21437/Interspeech.2017-79.

T. Murtola, P. Alku, J. Malinen, and A. Geneid. Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy. *Speech Commun.*, 96, 2018. ISSN 01676393. doi: 10.1016/j.specom.2017.11.007.

A. Ojalammi and J. Malinen. Automated segmentation of upper airways from MRI vocal tract geometry extraction. In *Proceedings of BIOIMAGING 2017*, pages 77–84, 2017.

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.*, 7(5):569–586, 1999. doi: 10.1109/89.784109.

T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Trans. Audio Speech Lang. Process.*, 19(1): 153–165, 2011. doi: 10.1109/TASL.2010.2045239.

M. Rothenberg. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *J. Acoust. Soc. Am.*, 53(6):1632–1645, 1973. doi: 10.1121/1.1975066.

S. Sahoo and A. Routray. A novel method of glottal inverse filtering. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 24(7):1230–1241, 2016. doi: 10.1109/TASLP.2016.2551864.

M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):955–967, 1987. doi: 10.1109/TASSP.1987.1165240.

A. Sorin, S. Shechtman, and A. Rendel. Semi parametric concatenative tts with instant voice modification capabilities. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1373–1377, Stockholm, 2017. doi: 10.21437/Interspeech.2017-1202.

B. H. Story. *Physiologically-Based Speech Simulation Using AN Enhanced Wave-Reflection Model of the Vocal Tract*. PhD thesis, University of Iowa, 1995.

B. H. Story. Phrase-level speech simulation with an airway modulation model of speech production. *Computer Speech & Language*, 27(4):989–1010, 2013. doi: 10.1016/j.csl.2012.10.005.

B. H. Story and I. R. Titze. Voice simulation with a body-cover model of the vocal folds. *J. Acoust. Soc. Am.*, 97(2):1249–1260, 1995. doi: 10.1121/1.412234.

B. H. Story, I. R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.*, 100(1):537–554, 1996. doi: Doi 10.1121/1.415960.

B. H. Story, A.-M. Laukkanen, and I. R. Titze. Acoustic impedance of an artificially lengthened and constricted vocal tract. *Journal of Voice*, 14(4):455–469, 2000. doi: 10.1016/S0892-1997(00)80003-X.

B. H. Story, H. K. Vorperian, K. Bunton, and R. B. Durtschi. An age-dependent vocal tract model for males and females based on anatomic measurements. *The Journal of the Acoustical Society of America*, 143(5):3079–3102, 2018.

H. Strik. Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *J. Acoust. Soc. Am.*, 103(5):2659–2669, 1998. doi: 10.1121/1.422786.

H. Strube. Time-varying wave digital filters for modeling analog systems. *IEEE Trans. Acoust. Speech Signal Process.*, 30(6):864–868, 1982. doi: 10.1109/TASSP.1982.1163976.

H. W. Strube. Determination of the instant of glottal closure from the speech wave. *J. Acoust. Soc. Am.*, 56(5):1625–1629, 11 1974. ISSN 0001-4966. doi: 10.1121/1.1903487. URL `http://asa.scitation.org/doi/10.1121/1.1903487`.

J. Sundberg, E. Fahlstedt, and A. Morell. Effects on the glottal voice source of vocal loudness variation in untrained female and male voices. *J. Acoust. Soc. Am.*, 117(2):876–885, 2005. doi: 10.1121/1.1841612.

I. R. Titze. Parameterization of the glottal area, glottal flow, and vocal fold contact area. *The Journal of the Acoustical Society of America*, 75(2):570–580, 1984. doi: 10.1121/1.390530.

I. R. Titze. Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *J. Acoust. Soc. Am.*, 111(1):367–376, 2002. doi: 10.1121/1.1417526.

I. R. Titze. Theoretical analysis of maximum flow declination rate versus maximum area declination rate in phonation. *Journal of Speech Language and Hearing Research*, 49(2):439–447, 2006a. doi: 10.1044/1092-4388(2006/034).

I. R. Titze. *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech, 2006b.

I. R. Titze and J. Sundberg. Vocal intensity in speakers and singers. *J. Acoust. Soc. Am.*, 91(5): 2936–2946, 1992. doi: 10.1121/1.402929.

I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D. G. Miller, H. Noé, R. C. Scherer, J. R. Smith, B. H. Story, J. G. Švec, S. Ternström, and J. Wolfe. Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, 137(5):3005–3007, 2015. doi: 10.1121/1.4919349. URL `http://asa.scitation.org/doi/abs/10.1121/1.4919349`.

E. Vilkman. Occupational safety and health aspects of voice and speech professions. *Folia Phoniatr. Logop.*, 56(4):220–253, 2004. doi: 10.1159/000078344.

E. Vilkman, E.-R. Lauri, P. Alku, E. Sala, and M. Sihvo. Loading changes in time-based parameters of glottal flow waveforms in different ergonomic conditions. *Folia Phoniatr. et Logop.*, 49(5):247–263, 1997. doi: 10.1159/000266463.

J. Wolfe, J. Smith, J. Tann, and N. H. Fletcher. Acoustic impedance spectra of classical and modern flutes. *J. Sound Vib.*, 243(1):127–144, 2001. doi: http://dx.doi.org/10.1006/jsvi.2000.3346.

D. Wong, J. Markel, and A. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust. Speech Signal Process.*, 27(4):350–355, 1979. doi: 10.1109/TASSP.1979.1163260. URL `http://ieeexplore.ieee.org/document/1163260/`.