Automated Assessment of Glottal Dysfunction Through Unified Acoustic Voice Analysis

Ian Vince Mcloughlin, Olivier Perrotin, Hamid Sharifzadeh, Jacqui Allen, Yan Song

Automated Assessment of Glottal Dysfunction Through Unified Acoustic Voice Analysis

[*]Ian Vince McLoughlin, [†]Olivier Perrotin, [‡]Hamid Sharifzadeh, [§]Jacqui Allen, and [‖]Yan Song, *Singapore, †CNRS, Grenoble INP, France, ‡§Auckland, New Zealand, and ‖Hefei, China

Abstract: This paper uses the recent glottal flow model for iterative adaptive inverse filtering to analyze recordings from dysfunctional speakers, namely those with larynx-related impairment such as laryngectomy. The analytical model allows extraction of the voice source spectrum, described by a compact set of parameters. This single model is used to visualize and better understand speech production characteristics across impaired and nonimpaired voices. The analysis reveals some discriminative aspects of the source model which map to a physiological class description of those impairments. Furthermore, being based on analysis of source parameters only, it is complementary to any existing techniques of vocal-tract or phonetic analysis. The results indicate the potential for future automated speech reconstruction systems that adapt to the method of reconstruction required, as well as being useful for mainstream speech systems, such as ASR, in which front-end analysis can direct back-end models to suit characteristics of impaired speech.

Keywords: Laryngectomy−Whispers−Glottal flow model−Distorted speech.

INTRODUCTION

Public health statistics reveal that voice impairment is a significant issue, affecting around 20% of the UK population at some point in their lives,[1] and around 7.5% in the US annually, with 4% reporting an impairment lasting a week or longer.[2,3] Meanwhile, the growing prevalence of speech-enabled automatic interaction systems for public and commercial services exacerbates the issues faced by those with impaired speech. Commercial speech systems do not always operate effectively with impaired speech input,[4,5] even where the speech is intelligible to human listeners. Technological solutions to this issue involve either preprocessing to reconstruct the impaired speakers voice,[6] or making the underlying systems more robust to impairments. In either case, a first step is likely to be automatic analysis of the input signal to determine the effect of any impairment on the speech.

A broad categorization of speech impairments could be those that affect (a) articulation, including distortion or deletion of speech sounds, (b) fluency including speaking speed, pauses, and formation of sounds, and (c) voice quality, including defects in pitch, loudness, or timbre. The origins of these three categories of speech impairments can be broadly attributed to either neurologic (or motor speech disorders) and non-neurologic disorders,[7] although in practice some individuals experience a complex mix of impairments. Examples of neurologic speech disorder include dysarthria and apraxia.[8,9] Dysarthria collectively describes a group of neurologic speech disorders that cause distortion in the accuracy of the muscular movements involved in breathing, phonation, resonation, and articulation of speech production. Apraxia, by contrast, is the inability to move the lips or tongue to form the correct sounds due to impaired ability to plan and program the sensorimotor orders needed to direct movements resulting in phonation, despite having the desire as well as the physical ability to do so.[7,10] Other cognitive and linguistic disorders that have neurologic origin include mutism, stuttering, aprosodia, echolalia, and foreign accent syndrome.[11]

Speech disorders which are not located in the nervous system are identified as being non-neurologic in origin. These are either psychogenic disorders, most commonly aphonia (loss of voice) and hoarseness,[12] or psychogenic disorders resulting from musculoskeletal defects.[7] Psychogenic disorders, including schizophrenia or depression, can also manifest through voice loss or vocal hoarseness. Musculoskeletal defects result from causes such as trauma, abnormality of bone or cartilage from birth, or following surgery such as laryngectomy.[7,12] Face or neck surgery which changes the shape of the vocal chambers by partial removal of muscles, cartilage, or bone can significantly affect speech production. Laryngectomy, the total or partial surgical withdrawal of the larynx, is often a treatment for throat cancer. Voice loss is a common side effect.[13]

Although speech impairment is not usually life threatening, it can have a profound effect on daily life and well-being.[4] There has thus been significant research effort on understanding the characteristics of distorted speech in recent years.[14−17] This has incited efforts on speech reconstruction, speech recognition, and speech enhancement systems aiming to improve the quality of life for affected individuals.[6,18−21] Research is progressing rapidly, such that computational voice reconstruction systems for some specific impairments are becoming

Accepted for publication August 25, 2020.

From the *Singapore Institute of Technology, Singapore; †University Grenoble Alpes, CNRS, Grenoble INP, France; ‡School of Computing, Unitec Institute of Technology, Auckland, New Zealand; §Department of Otolaryngology, North Shore Hospital, Auckland, New Zealand; and the ‖National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.

Address correspondence and reprint requests to Ian Vince McLoughlin, Singapore Institute of Technology, Singapore. E-mail: ian.mcloughlin@singaporetech.edu.sg

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2020.08.032

mature technology, yet the input parameters necessary to control the type and degree of reconstruction are currently either fixed or hand-tuned for individual speakers.[6,18–21]

This paper aims to advance research knowledge in terms of automatically deriving input parameters, using a recently defined voice decomposition method; glottal flow model (GFM) based iterative adaptive inverse filtering (IAIF), which extracts a meaningful and well-defined compact set of parameters to describe an analyzed voice signal.[22] GFM-IAIF separates vocal tract and lip contributions from the voice source signal. In experiments on normal speech involving a glottal source, GFM-IAIF has been shown able to accurately analyze different voice qualities for natural and synthesized speech.[22] It has recently been used to reconstruct a speech glottal source from analyses of whisper source, using a single decomposition model.[21] This successful extrapolation to non-speech voice sources has led – in this paper – to its application to a set of voice recordings obtained from patients who have undergone glottal or larynx treatment and surgery (including excision), alongside baseline normal speech. The analysis will demonstrate the ability of GFM-IAIF to derive parameters from voice source signals that differentiate the category of impairment present. Furthermore, it can represent this in a consistent set of analytical parameters which are suitable for control of future reconstruction or impaired ASR technology.

In the remainder of the paper, Section 2 will first examine the underlying speech production model which we parameterize through GFM-IAIF. Section 3 will outline the data used while Section 4 will examine the analysis results to determine whether the technique is effective at assessing glottal dysfunction. Finally, Section 5 will conclude the paper.

SPEECH MODEL AND ANALYSIS

General Speech Production

In normal speech, voiced phonemes are generated through periodic vibration of the vocal folds, producing glottal harmonic air flow into the upper vocal tract, and which exits through the oral and nasal chambers.[23]

The vocal fold vibration is periodic, beginning with a glottis opening phase in which the folds are pulled apart under the influence of subglottal pressure. As the pressure releases, then the natural elasticity of the vocal folds draws them together in the closing phase, blocking the tracheal airflow. Then, the effect of the constriction plus sustained diaphragm contraction increases subglottal pressure until it is sufficient to trigger the next opening phase. In normal speech, the opening and closing phases of the glottis tend to influence distinct regions of the frequency spectrum,[24] the former providing a major resonance near the fundamental frequency often called the "glottal formant," and the latter through the contribution of high frequencies. Figure 1 (top left panel) depicts one glottal flow period, and the frequency response of its derivative (right), from the widely used LF-model.[25] The opening phase contribution can be modeled using a second-order all-pole resonant filter with a ± 20 dB/decade slope. The position F_{GF} and bandwidth B_{GF} of the glottal formant are influenced by the relative duration of the open phase over a period as well as the glottal pulse asymmetry.[26,27] The more abrupt the closing phase, the more high frequencies are present. Conversely, the smoother the closure, the more the high frequencies are attenuated. This is modeled by a –20 dB/decade first order low-pass filter of varying cutoff frequency F_{ST}. A smoother closure reduces F_{ST}, and vice versa. The overall frequency attenuation that results from the glottal formant shape and position of F_{ST} is commonly called spectral tilt ST, expressed in dB/decade. These parameters are illustrated in the right panel of Figure 1. This spectral description can be summarized by

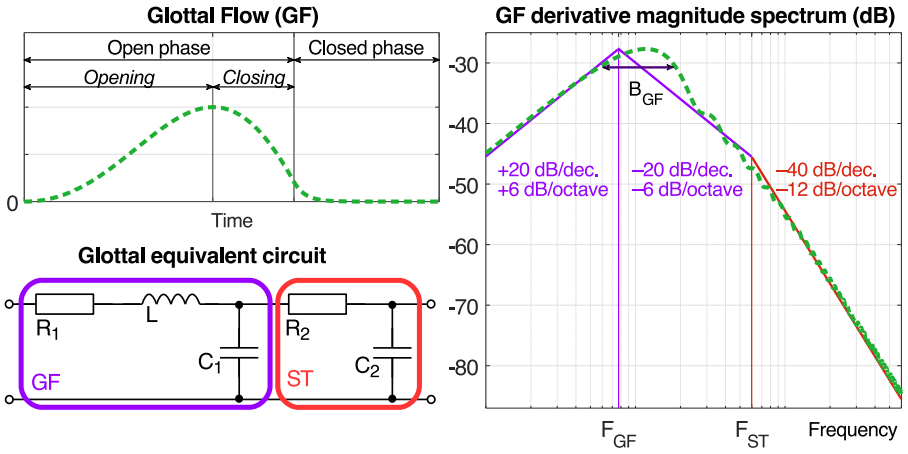

FIGURE. 1. Illustration of the LF-model showing the glottal flow physical process (top left), the glottal flow equivalent circuit (bottom left), and the glottal flow derivative spectrum (right panel) with glottal formant frequency and bandwidth (F_{GF}, B_{GF}), and spectral tilt cutoff frequency (F_{ST}).

modeling the glottal flow derivative with a third order filter having an anticausal complex conjugate pole pair $\{a, a^*\}$ (to model glottal formant) and one causal real pole b (to provide spectral tilt)[28]:

$$G(z) = \frac{1}{(1-az^{-1})(1-a^*z^{-1})(1-bz^{-1})} \quad (1)$$

This relates well to voice quality, eg, a tensed voice has higher F_{GF} and wider B_{GF} and weaker spectral tilt than a breathy voice.[29] Close correlation between vocal effort and spectral tilt[30,31] was a major motivation for using a three-pole GFM. This has been validated for voice quality analysis and modification,[32] expressive singing and speech synthesis,[33,34] whisper-to-speech conversion,[21] and in this paper is now considered for analyzing the effects of dysfunction on glottis parameters. For convenience, we will use the term "glottal formant" even for speakers who lack a glottis. In those cases, it describes the spectral shape of the fundamental frequency excitation that serves as their functional glottis replacement.

Meanwhile the vocal tract is modeled through the resonances (poles) that it introduces in the glottal flow spectrum. Additional pairs of poles are included to model the effect of nasal attenuation (zeros),[23] yielding an auto-regressive model comprising N_v pairs of complex conjugate poles $\{c_i, c_i^*\}$:

$$V(z) = \left\{ \prod_{i=1}^{N_v} (1-c_i z^{-1})(1-c_i^* z^{-1}) \right\}^{-1} \quad (2)$$

Finally, lip radiation is modeled as a derivative filter with coefficient d close to 1,[23] $L(z) = 1-dz^{-1}$, and the speech signal can be modeled as $S(z) = E(z)G(z)V(z)L(z)$, where $S(z)$ is the z-transform of the speech signal, and $E(z)$ the z-transform of a mixture of flat-spectrum periodic and aperiodic excitations. Having discussed the spectral model of voice production, we now consider how dysfunction affects production before turning our attention to obtaining model parameters from speech using glottal inverse filtering.

Dysfunctional Speech Production

While there are probably a wider variety of failure modes for speech production than there are intelligible operating modes, we can nevertheless discuss some categorical aspects of speech dysfunction.

In relation to the spectral model, disorders may affect single or multiple parts of the system, from exhalation discontinuities through glottal function impairment, unusual arrangement of vocal tract, mouth or nasal cavities, teeth or lips, and impaired dynamics of motion for any and all articulators. Laryngectomy and glossectomy – surgical removal of part of the larynx or tongue respectively – are typical examples of medical interventions that result in changed glottal and vocal tract functionality. Meanwhile nerve lesions or impaired motor signals in the brain can affect the dynamics of speech production.

In a total laryngectomy, the entire larynx including vocal folds, thyroid, cartilage, hyoid bone, and some associated tissue are typically surgically removed.[13] The result, postsurgery, is that harmonic air flow is absent,[35] and in many cases airflow from the vocal tract is diverted via a stoma rather than through the mouth. In general, the absence of a larynx also means that speakers require alternative forms of glottal source to drive speech production, even if the lung exhalation, vocal tract, mouth and nasal cavity $V(z)$, and lip radiation $L(z)$ are present and working without impairment.

Some laryngectomees (those who are postlaryngectomy) learn to produce an alaryngeal voice by modifying intraoral pressure, and hence airflow, with a surgical prosthesis. A few learn to control constrictions in their airway to provide the required pressure modulation.[4] However, the resulting voice is usually feeble or drained, and sounds very hoarse.[36] Furthermore the majority of laryngectomees are unable to learn how to articulate intelligible speech in this way.[37]

Many therefore tend to use an external electrolarynx or use pseudo-whispers[6] for communication – which is possible only for those without a diverted airflow. In the former case, the electrolarynx mostly provides a new excitation $E(z)$ without introducing the timbral characteristics of vocal fold vibrations that are normally encoded in $G(z)$. In the latter case, speech without a pitch source $G(z)$, means the vocal tract is excited only with turbulent airflow, which is effectively a whisper.[38] Without periodic excitation, prosodic information including intonation, stress, tone, and rhythm are largely absent,[39,40] and it has been found that the resulting vowel space (namely the boundaries in F1-F2 frequency space formed by the different vowel regions) is translated compared to neutral speech.[41]

A partial laryngectomy involves surgical removal of only part of the larynx, usually one side. This typically allows for normal airflow and full function of other articulators within the voice production apparatus.[18] Individuals with disease or damage to just one side of the focal folds, or damage to the nerves affecting only one side of the glottis (eg, after heart surgery complications), effectively have similar partial functionality. Some degree of glottis control is often possible, and those individuals therefore have a pitch source, albeit one characterized by reduced dynamic control and severe asymmetry, resulting in voice quality degradation.

Others who retain a full glottis may have different impairments such as those caused by ingested substance damage, or from laryngeal papillomatosis (eg, HPV), ulcers, tissue buildup and so on. These again affect dynamic and static characteristics, and are manifested through voice quality degradation and reduced articulatory control. Anecdotally, it appears that many individuals with vocal fold damage (including some partial laryngectomees) are able to learn to compensate for that damage and regain nearly-full speech capabilities over time.

TABLE 1.
Details of Participating Subjects

ID	Bio	G	R	B	A	S	C	I	Notes
Total laryngectomy									
5	72 M	3	2	3	3	1	2	2	Very whispery
6	– M	3	2	3	3	3	3	3	Quiet whispers
11	81 M	3	3	2	3	3	3	3	Oesophageal, 8yrs
14	60 M	3	2	3	3	2	3	3	Quiet whispers
16	64 M	3	2	3	3	2	3	3	Whispery voice
Total laryngectomy with TEP									
7	57 M	3	2	1	1	3	1	1	11yrs
12	76 M	3	3	2	2	2	2	1	
13	– M	3	3	1	2	3	2	2	
15	60s M	3	3	2	1	3	2	1	2yrs
17	79 M	2	3	1	1	2	1	1	
18	80 F	1	2	3	3	1	0	1	Clear whisper 20yrs
19	– M	3	3	2	2	3	2	1	9 and 5yrs
20	73 M	2	3	1	2	2	1	1	9yrs
Partial laryngectomy/function									
1	60s M	2	3	1	1	2	0	0	3 yrs
2	40s F	2	2	0	1	2	0	0	
3	60s M	1	1	1	1	0	0	0	4 wks
4	52 F	2	2	1	1	2	0	0	8 wks, tracheotomy
8	40 M	1	1	1	0	2	0	0	2yrs, botox
9	78 M	2	2	1	2	2	0	1	"many years"
10	69 M	3	3	0	1	2	0	0	7yrs, radiotherapy
Reference									
21	38 M	0	0	0	0	0	0	0	Clear voice
22	52 F	0	0	0	0	0	0	0	Clear voice

Notes: GRBAS, C (cadence) and I (intelligibility) impairment scales are rated 0 = normal, 1 = mild, 2 = moderate, 3 = severe.

In summary, when looking at glottal dysfunction, we could divide this into categories of complete absence of glottal function, partial glottal function,[1] or full function with quality impairment, where "function" implies the ability to dynamically control the glottal component of speech production. Vocal tract dysfunction classes might be viewed similarly as being complete absence of dynamic vocal tract control, partial control or full control with degraded tract characteristics. The question is then whether it is possible to infer the degree of degradation through an analysis of the dynamic attributes of $G(z)$ and $V(z)$ from recorded speech.

Source-Filter Decomposition Methods

For over half a century,[42,43] glottal inverse filtering has been studied as a way to decompose speech into source and filter components. The simplest method is direct linear prediction after pre-emphasis to extract VT tube model coefficients[44] with single tap long-term prediction to extract the excitation periodicity.[23] Iterative Adaptive Inverse Filtering (IAIF)[45] yielded a significant improvement on this by using first order LPC analysis to define the pre-emphasis filter, then iteratively estimating the glottis and vocal tract filters. IAIF models spectral tilt into the VT filter, and hence IOP-IAIF[46] was proposed as a further improvement to separately model it, although this involved unconstrained glottal complexity. Finally, two of the current authors proposed GFM-IAIF,[22] where the IAIF first order glottal model is replaced by a third order filter, following Equation 1. Note that the GFM-IAIF glottis filter is fully causal compared to Equation 1, yet it does not affect the magnitude spectrum, from which are extracted the spectral parameters (eg, F_{GF}, B_{GF}, and F_{ST}) that we now use for analysis of dysphonic speech. GFM-IAIF has also recently been demonstrated in the conversion of postlaryngectomy speech to phonated speech,[21] lending empirical support to its effectiveness at modeling dysfunctional speech.

DYSFUNCTIONAL SPEECH DATA AND ANALYSIS

Data Collection

Data was obtained from 22 volunteers (M = 18, F = 4) spanning various degrees of glottis-related impairment, plus two healthy reference speakers as listed in Table 1. Three subjects (4, 8, 10) had undergone nonsurgical larynx treatment. Four had partial laryngectomy or related surgery (eg, thoracotomy), and 13 had a total laryngectomy. Among the latter, eight spoke with a tracheoesophageal puncture (TEP) to redirect air through the oesophagus (serving as an

[1]Bearing in mind that "glottal function" may be provided by something (eg, prosthesis or alternative articulatory organ) that is not actually a glottis.

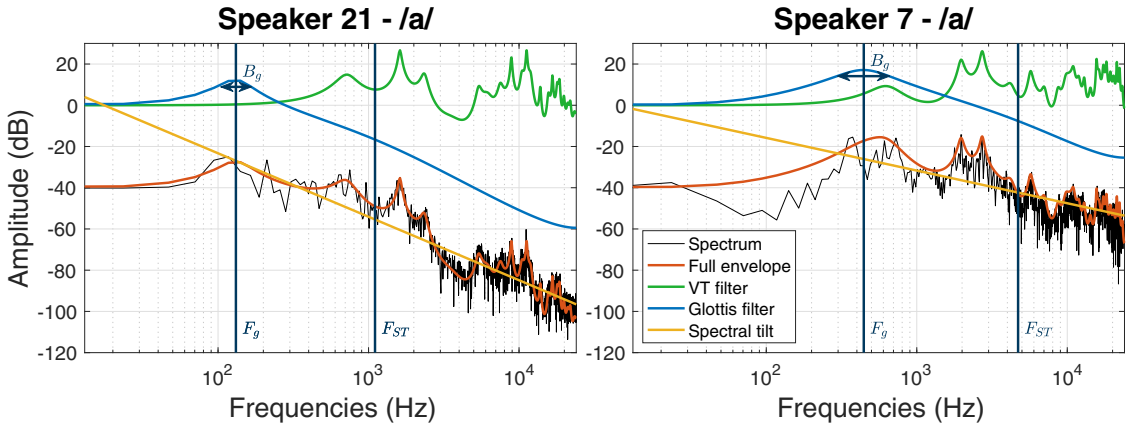

FIGURE. 2. Examples of glottis + lips (blue) and VT (green) filters extraction with GFM-IAIF. Orange: combined contribution, overlapping with the speech spectrum of a given frame (black). Vertical lines: glottal formant and spectral tilt frequencies. Yellow: glottis + lip spectral tilt. Left: healthy speaker; Right: speaker with TEP.

alryngeal oscillation source). Subjects varied in age from 38 to 81, and were recorded in an audiology room at North Shore hospital, Auckland, New Zealand. Subjects were seated with a ZOOM H4n recorder placed approximately 30 cm away from their mouth. Recordings were made in stereo with 16-bit resolution at a sample rate of 96 kHz, downsampled to 16 kHz mono for the purpose of analysis. During recorded sessions, subjects were first asked to describe their condition, then asked to recite digits from 0 to 9. Next, they read a sequence of 11 vowels in carrier words, /hxd/, where "x" is replaced in turn by different vowels (as in[41]), and finally a set of 15 TIMIT sentences. Subjects who were comfortably able to speak as well as whisper (namely subjects 1−4, 8−10, 15, 21, 22) were asked to repeat each recitation list firstly with their usual speaking voice and then using whispers. Stronger speakers were asked to repeat each list three or four times, while weaker speakers repeated everything at least twice. Misspoken words or sentences were repeated immediately, with erroneous sections removed prior to postprocessing, leaving around 8.8 hours of recorded speech frames. Table 1 includes a subjective categorization of voices using the GRBAS scale[47,48] agreed by the first, third, and fourth authors. GRBAS is an auditory-perceptual evaluation method for evaluating hoarseness in terms of overall grade (G) of dysphonia, (R) raspiness, (B) breathiness, (A) asthenicity, and (S) vocal strain. We added a subjective assessment of (C) cadence (speech rate) and intelligibility, along with duration since last treatment (in months or years), and other information, where known.

Data Analysis

The detailed analysis presented in this paper was performed only on manually segmented frames. All audio segments were listened to and quality checked by expert listeners prior to analysis. Segmented recordings were divided into 80% overlapping 512 sample speech frames (32 ms at $f_s = 16$ kHz). GFM-IAIF[2] was performed on each frame to yield third order glottal component $G(z)$ and $f_s/1000 + 2 = 18$th order $V(z)$. The frequency responses of $G(z)$ and $V(z)$ were computed along with the glottis signal spectrum after de-convolving the VT component in accordance with.[22] To provide an intuitive illustration of this, Figure 2 plots normal and dysfunctional signals for extracted glottis + lip $G(z)L(z)$ (blue) and VT $V(z)$ (green) filters from one analysis frame (black), respectively. The combined contribution of $G(z)V(z)L(z)$ gives the full spectral envelope of the speech signal (orange) in each case.

The manually segmented vowel-only analysis from the /hxd/ carrier words amounted to 36,800 frames (over the 14 vowels and 22 speakers). The final experiment reported in this paper was performed on all recorded material, without manual segmentation or checking. The frame size, sample rate and GFM-IAIF analysis procedures were identical to that for the /hxd/ vowels, but was conducted over 1,060,479 frames. During analysis, three parameters were computed from the glottis filter coefficients: the glottal formant center frequency F_{GF}; the glottal formant quality factor $Q_{GF} = F_{GF}/B_{GF}$, where B_{GF} is the glottal formant bandwidth (full width half maximum); and the spectral tilt cutoff frequency F_{ST}. These were derived from the equivalent analogue model filter of the glottis filter, shown in the bottom left of Figure 1, where:

$$G(s) = \frac{1}{(s^2LC_1 + sC_1R_1 + 1)} \cdot \frac{1}{(1 + sR_2C_2)}$$

Specifically, the resonance of the first clause described by L, C_1, and R_1 models the glottal formant centre frequency while the resonance of the second clause described by R_2 and C_2 models the spectral tilt cutoff frequency. We find both frequencies by simply differentiating the denominators, $2\pi F_{GF} = (LC_1)^{-0.5}$ and $2\pi F_{ST} = (R_2C_2)^{-1}$. The quality factor can be derived by setting the denominator of the first clause to zero. If $s^2LC_1 + sC_1R_1 + 1 = 0$ then the quadratic rule gives us roots at $\{-C_1R_1 \pm (C_1{}^2R_1^2 - 4LC_1)^{0.5}\}/\{2LC_1\}$.

[2]Using code available from https://github.com/operrotin/GFM-IAIF.

For these to be real, it is obviously necessary for $C_1{}^2R_1^2-4LC_1>0$. Setting to the balance point of 0 gives $C_1R_1{}^2>4L$. If $Q_{GF}{}^2=L/(CR^2)$, then Q_{GF} must be 0.5 or greater for the 2 roots to be real (and they are equal when $Q_{GF}=0.5$). Note that the glottal damping factor ζ_{GF} is $1/(2Q_{GF})$ (since $\zeta_{GF}=CR/(2\sqrt{LC})$ and as in any second order resonant system, it is critically damped when this is unity, is underdamped when less than one, and overdamped when greater than one. In practice the parameters F_{GF}, Q_{GF}, and F_{ST}, are extracted from the analysis with the aid of the bilinear transform.[49,50]

The left and right vertical lines on Figure 2 show the glottal formant and spectral tilt cutoff frequencies, respectively. The left panel shows an /a/ vowel uttered by a control (nondysfunctional) speaker, while the right panel displays the same vowel uttered by a laryngectomee using a tracheoesophageal puncture valve (TEP). The higher values of F_{GF} and F_{ST} for the second speaker lead to a flatter spectrum, and show how these parameters can indicate various glottal functions. Additionally, to quantify the spectral "flatness" in the high frequency region, the glottal spectral tilt slope ST, in dB/decade, is computed as the linear regression of the glottis spectrum on a log-log scale. For the sake of plotting, we added the lip filter slope (+20 dB/decade) to the spectral tilt shown in yellow on Figure 1 to match the glottis + lip $G(z)L(z)$ spectral envelope. Note that this +20 dB/decade is not added in the subsequent analysis.

ANALYSIS RESULTS

GFM-IAIF was used to extract the glottal $G(z)$ and vocal tract $V(z)$ components from each frame, then converted on a frame-wise basis to F_{GF}, Q_{GF}, F_{ST}, and ST as noted in Section 3.2. These parameters are now explored in different ways in this section. Note that F_{GF}, Q_{GF} and F_{ST} are expressed on a logarithmic scale. As noted above, analysis is conducted over two subsets of recorded data; firstly the manually segmented vowel-only analysis on 14 vowels in carrier words, namely *had, hard, head, heard, heed, heyd, hid, hide, hoard, hod, hood, hoyd, had*, and *who'd*. Secondly, analysis of the full recordings from each subject including the phonetically balanced TIMIT sentences, the conversations and the carrier words. While the former subset is carefully separated into spoken and whispered recordings, the latter subset includes the full range of speech comprising normally voiced and normally unvoiced phonetic components. It should be noted that the actual degree of phonation exhibited in the former case is subject-dependent, and hence it could be best described as being "as phonated as possible" (whether that uses natural larynges or not). The following subsections now examine the GFM-IAIF analysis results from each subjects' speaking characteristics and the speech contents.

TABLE 2.
Significance of the Four Factors on F_{GF}, Q_{GF}, F_{ST} and ST Assessed by a Kruskal-Wallis Rank Sum Test Using a χ^2 Distribution

		F_{GF}	Q_{GF}	F_{ST}	ST
Factor	*df*		χ^2		
Vowel	13	181	287	1111	769
Phonation	1	4223	9850	7.63*	1870
Group	3	3443	5943	1265	15446
Speaker	21	10415	9588	6737	22977

All Have $P<10^{-15}$ except *($P<10^{-2}$).

Carrier Sentence Analysis

We first explore the vowel-only subset by selecting only the vowel from within the /hxd/ carrier sentences. We assess the effect of four different factors: vowel; phonation mode (normal or whisper); subject group (control, partial, TEP, total); and speaker; on the four GFM-IAIF output parameters. Table 2 gives the statistical significance of these factor on the GFM-IAIF parameters, assessed by a Kruskal-Wallis rank sum test using a χ^2 distribution. All factors have significant effects ($P<10^{-2}$ for the effect of phonation mode on F_{ST}; $P<10^{-15}$ for the others) with various strengths, detailed in the following sections.

Effect of Vowels

The first row of Figure 3 plots the range of glottal formant center frequencies and quality factors, and glottal spectral tilt cutoff frequencies and slopes from vowel carriers, collated for all speakers. Each of these can be seen in the LP model description of Figure 1, including the quality factor which is related to the bandwidth of the glottal formant as $Q_{GF}=F_{GF}/B_{GF}$. The box plots span the 25th−75th percentiles and indicate medians with a notch. Whiskers extend across the entire range, apart from outliers which are marked with small black circles when present. Although all parameters reject the null hypothesis of the Kruskal-Wallis rank-sum test (first line of Table 2), the effect size is small. Figure 3 displays homogeneous distribution across all uttered vowels for each parameter. This observation proves the efficiency of GFM-IAIF as a source-filter separation method: there is little vowel variability that is associated to the vocal tract in the glottis-related parameters.

Effect of Phonation Mode

The second row of Figure 3 shows the range of the four GFM-IAIF parameters depending on the phonation mode (whisper vs normal). Remember that normal speech recordings are "as phonated as possible," depending on the patients' abilities to produce phonation; whisper speech recordings, when produced, are all unphonated. Clearly there is a noticeable difference between whisper and normal mode in most of the plots, but primarily so in terms of the glottal formant quality factor and spectral tilt. The former

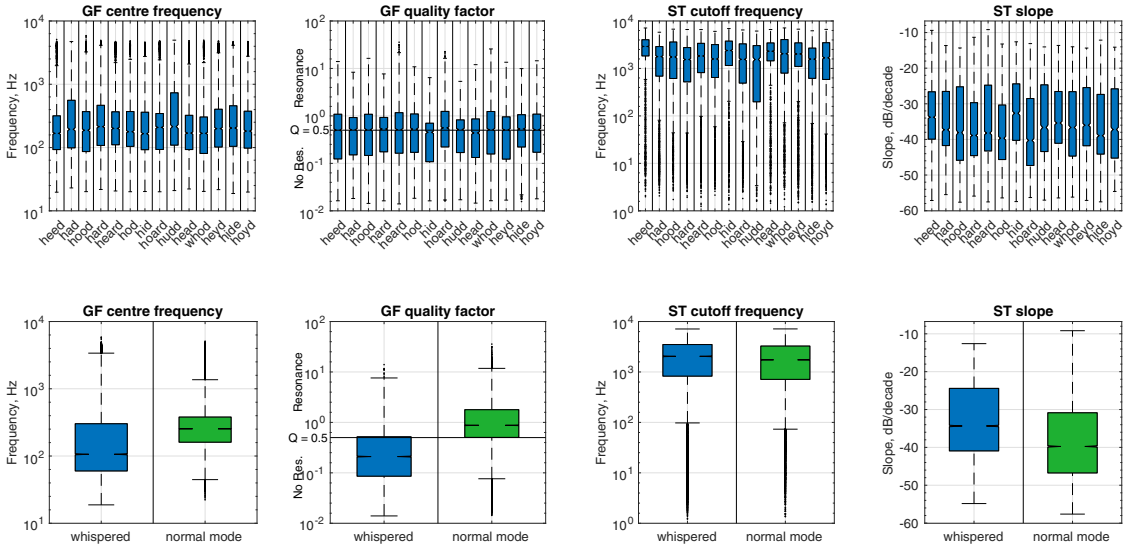

FIGURE. 3. Box plots of glottis spectral parameters per vowel (top row) and per phonation mode (bottom row). From left, each row plots the log glottal formant frequency, the log glottal formant quality factor, the log spectral tilt cutoff frequency, and the spectral tilt slope.

is to be expected since whispered speech does not include glottis resonance, and hence all whispers are located in the "no resonance" region while the median phonated samples are in the "resonance" region. Not all are strongly resonant, and as we will see later this is because the quality of phonation varies across subjects. The boundary point $Q_{GF} = 0.5$ is a convenient demarcation between the resonance region (a clear glottal formant exists) and the no-resonance region (the glottal formant is either extremely flat or is not detectable). Spectral tilt is another differentiating factor. The slope is far shallower for whispered speech due to the absence of glottal formants. The Kruskal-Wallis rank-sum test (second row of Table 2) showed that the phonation factor has a significant effect on Q_{GF} ($\chi^2 = 9850$, $P < 10^{-15}$) and ST ($\chi^2 = 1870$, $P < 10^{-15}$). Both the glottal formant center frequency F_{GF} and the spectral tilt cutoff frequency F_{ST} are less discriminative parameters for this test. In particular, the effect of phonation mode on F_{ST} is extremely small ($\chi^2 = 7.63$, $P < 10^{-2}$). We can nevertheless observe that the variability of F_{GF} is smaller for normal speech than whispered speech, deriving from the higher stability of voiced sounds compared to unvoiced.

Effect of Subject Group

The Kruskal-Wallis rank-sum test showed a significant effect of the subject group on all four GFM-IAIF parameters (third row of Table 2). For each parameter, a post hoc Dunn test was used for a pairwise comparison between the distributions associated to the four factor levels (control, partial, TEP, total), and all pairs were assessed significantly different ($P < 10^{-15}$). This indicates that all four glottis parameters are able to discriminate well across various degrees of glottis-related impairment. Since we showed in the previous section that the phonation mode has a strong effect on the distribution of some glottis parameters, we continue our analysis by separately grouping whisper (unphonated) and normal (as phonated as possible) speech for each subject. The statistics over all speakers within the four dysfunction groups are now plotted in Figure 4. Utterances have been separated into those which are phonated as much as possible (top row) and those which are whispered or unphonated (bottom row). We again subjected the GFM-IAIF parameters from the speech and whisper vowel analysis plotted in Figure 4 to a Kruskal-Wallis ranked-sum test with the null hypothesis that results from each group belong to the same distribution. All parameters from the speech segments strongly reject the null hypothesis (for F_{GF}, $\chi^2 = 4583$, for Q_{GF}, $\chi^2 = 8705$, for F_{ST}, $\chi^2 = 906$ and for ST, $\chi^2 = 10,346$ all with $df = 3$ and $P < 10^{-15}$). Meanwhile the whisper segments also reject the null hypothesis, but less definitively (for F_{GF}, $\chi^2 = 1873$, for Q_{GF}, $\chi^2 = 751$, for F_{ST}, $\chi^2 = 526$ and for ST, $\chi^2 = 6437$, again with $df = 3$ and $P < 10^{-15}$). For each parameter, a pairwise comparison between all distributions with a post hoc Dunn test shows that all pairs are significantly different ($P < 0.02$ for F_{ST} in whisper mode between total and TEP and between partial and control; $P < 10^{-15}$ for all the others), except for Q_{GF} in whisper mode for the pair partial vs control ($P = 0.3$).

Overall, we see quite high levels of differentiation between groups in all parameters. In particular, following an increasing degree of impairment (control, partial, TEP, total), we observe an increase of the glottal formant frequency F_{GF}, a decrease of the glottal formant quality factor Q_{GF}, an increase of the spectral tilt cutoff frequency F_{ST}, and an increase of the spectral tilt slope ST. In the case of fully phonated speech, F_{GF} is directly linked to the frequency of vibration of the vocal folds. In the presence of noise in the glottis signal, we showed that F_{GF} captures the position of the dominant frequency region of this noise.[21] Therefore, in the control group we observe an F_{GF} distribution around a few hundred Hertz, the order of magnitude of vocal fold

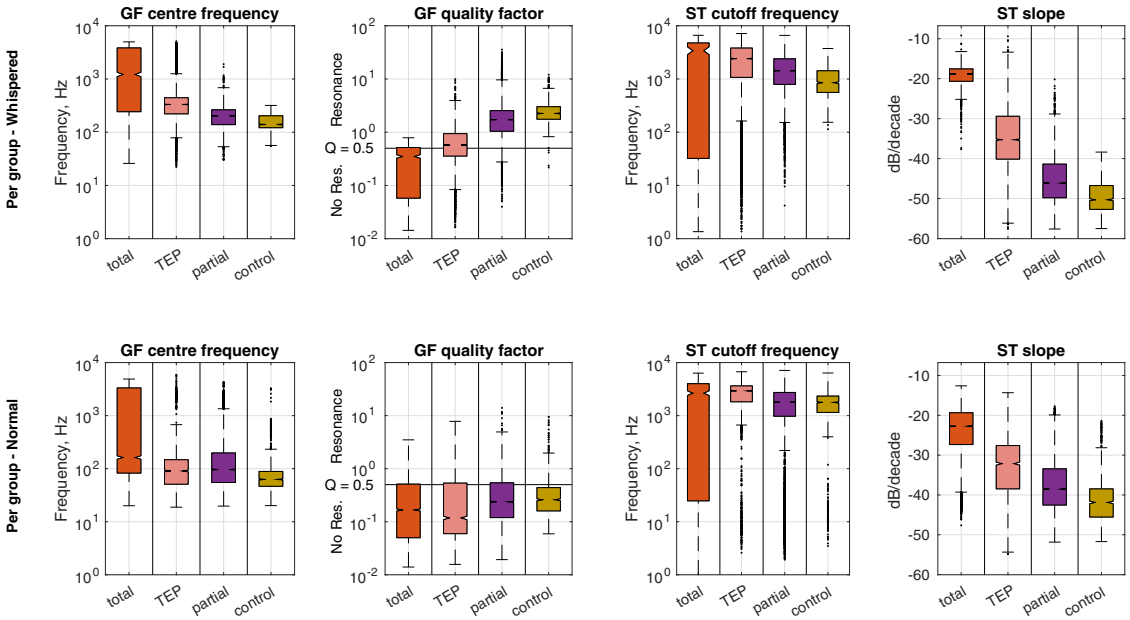

FIGURE. 4. Box plots of glottis spectral parameters per group for normal speech (top row) and whispered speech (bottom row). From left, each row plots the log glottal formant frequency, the log glottal formant quality factor, the log spectral tilt cutoff frequency, and the spectral tilt slope. The dysfunctional group of speakers on each plot from left to right is total laryngectomy (orange), TEP (salmon), partial laryngectomy (dark purple) and control speakers (gold).

vibration. The increase of F_{GF} with the degree of impairment follows the increasing amount of noise in the glottal signal, linked to the progressive loss of phonation. Glottal formant quality factor, Q_{GF}, is perhaps the most interesting parameter with clear differentiation between: "control" and "partial" groups which involve clear resonances; the "TEP" group that tends to have lower resonances; the "total" group firmly located in the "no resonance" region. This resonance is the signature of the strong harmonics that correspond to the frequency of vocal folds aperture, and is therefore salient for healthy phonated speech. In unphonated speech, the absence of vocal fold vibration leads to the absence of the corresponding resonance. Again, the gradual decrease of Q_{GF} between the "control" and "total" group follows the gradual loss of phonation caused by the various impairments. The increase of spectral tilt cutoff frequency with impairment has a clear effect on the spectral tilt slope parameter, where the slope for the total laryngectomy group is much shallower than control speech. In fact, if we add the effect of lip radiation (+20 dB/decade), we can note that total laryngectomy speech has an almost flat spectrum when measured at the lips. Like Q_{GF}, the variation across the four groups is one of the most discriminative for glottis impairment. Looking now at the distribution variances, the "control" group has lower variance than other groups for all parameters except the ST slope, where the total laryngectomy patients have the lowest. While speech impairment can generally be expected to increase variance by reducing the stability of speech production, the "total" group has *reduced* variance. We believe this is due to the loss of controlled articulation in the produced speech.

Briefly comparing between modes, ordering between groups (ie, highest to lowest) is preserved for all parameters (except F_{GF} in the "partial" group), but we see far closer alignment and similarity between measures across the four groups in whisper mode. Also, variance on all parameters tends to be higher. This closer similarity between distributions derives from the fact that both healthy and unimpaired whispers are produced with similar unphonated airflow. However, these airflows are different in nature: from the lungs for healthy speakers and partial laryngectomy; from a valve for those with TEP, and from the oesophagus for total laryngectomy patients, which helps to explain the remaining differences between distributions.

Effect of Speaker

The Kruskal-Wallis rank-sum test showed a significant and large effect of speakers on all four GFM-IAIF parameters (last row of Table 2). This suggests a high variability between groups. Looking now at individual speakers, Figure 5 plots the GFM-IAIF parameter statistics for vowel-only analysis conducted on each speakers' normal speaking mode, ie, this excludes whispers for speakers who can produce fully or partially phonated speech. Data is then collated and presented color-coded over different subject groups. From left to right on each subplot, these are total laryngectomy ("total"), partial laryngectomy or partial glottal function ("partial"), tracheoesophageal puncture ("TEP") and unimpaired speakers ("control"), all as identified in Table 1. Considering first the glottal formant center frequency, we see again that variance is low in the "control"

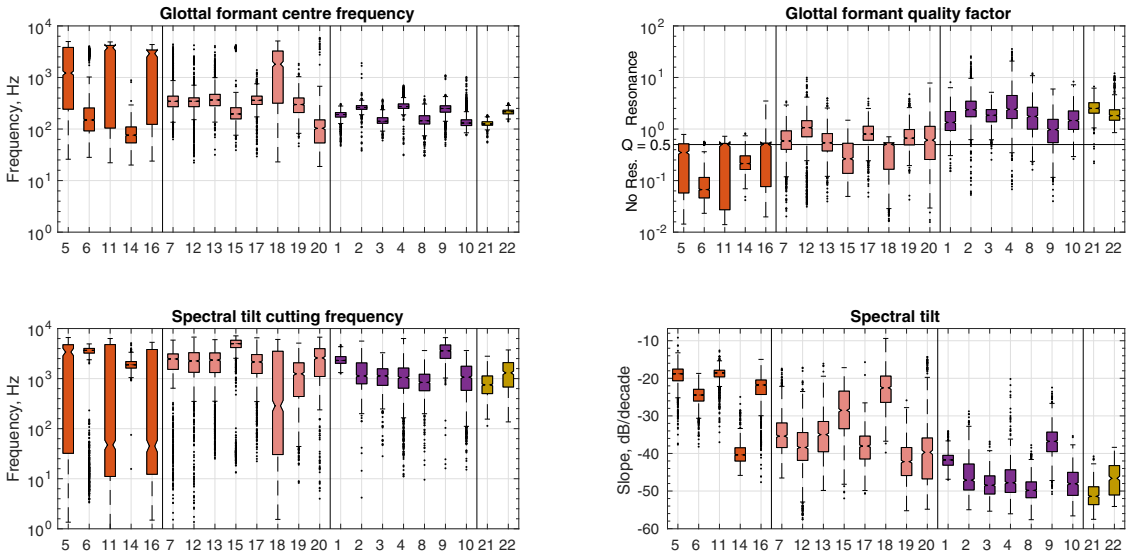

FIGURE. 5. Box plots of glottis spectral parameters per speaker for normal speaking mode. Clockwise from top left, the log glottal formant frequency, its log quality factor, the glottis spectral tilt slope and its log cutoff frequency. Colors identify the types of speaker, from left to right on each plot, these are total laryngectomy (orange), TEP (salmon), partial laryngectomy (dark purple) and control speakers (gold).

and "partial" groups, but much higher for most of the "total" and "TEP" speakers. The female speakers (2, 4, 18, and 22) tend to have higher glottal formant frequency, obviously given that female pitch tends to be higher than male pitch, since F_{GF} is usually proportional to pitch.[24] However the difference between male and female pitch appears to be smaller than the difference between groups. Glottal formant quality factor, Q_{GF}, also reveals substantial variation between speakers within the "TEP" and "total" groups, which tend to have weaker resonances. Table 1 included subjective information regarding intelligibility, cadence and grade of impairment. It is notable that less impaired subjects tend to have higher Q_{GF}. This may be due to the anecdotal evidence that some speakers in the "partial" group (including those mentioned) have found ways of speaking that are able to compensate for, or partially mitigate, the effects of their glottis damage. The spectral tilt cutoff frequency reveals high intragroup variations, that are also observed for glottal formant center frequency. Namely, subjects 5, 11, and 16 from the "total" group and 18 from the "TEP" group present similar distributions. This is also true for glottal formant quality factor and spectral tilt. Listening to those subjects reveals that almost all articulation that is present is as a result of turbulent airflow − there is little or no pitch evident in each case. Finally, we can see that *ST* also exhibits high intra-group variations for the "total" and "TEP" groups, suggesting different effects of impairment, even though intrasubject variation can be low.

Spectral Envelope for Vowel Production Per Group

GFM-IAIF involves separate LPC analyses of the vocal tract and for the glottal components of voice, yielding separate sets of parameters describing $V(z)$ and $G(z)$. The resulting median spectral envelopes are plotted in Figure 6 with a

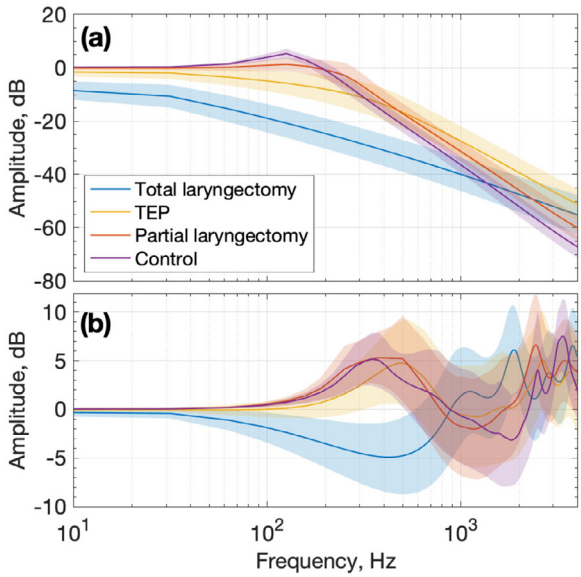

FIGURE. 6. Spectral envelope response for the (a) glottis and (b) vocal tract from the speech of all patients, grouped according to their dysfunction class.

log-frequency, log-amplitude scale, for the four speaker groups. Each median is shaded to indicate a span of one standard deviation. Unlike the previous box plots, which show statistics of the GFM-IAIF glottis parameters, these plots show the full spectral response, much as in Figure 2.

The glottis envelope at the top of Figure 6 shows strong differentiation between groups, with the control group exhibiting the expected clear glottal formant bump, a steep slope towards higher frequencies, along with a change in slope. The partial laryngectomy group resembles this closely, but with a lower bandwidth bump at slightly higher frequency (ie, Q_{GF} is lower). The TEP is similar in shape but

lacks any distinct glottal formant, whereas the "total" group appears to completely lack glottal influence.

Now consider the vocal tract envelope at the bottom of Figure 6. Looking in the lower frequency half of the plot, we again see large differences between "total" and "control" groups, while "TEP" and then "partial" are more similar to "control." All have formants at higher frequency (the right-hand side of the plot). Formants are resonances of the vocal tract, which all groups still possess, and are able to form – providing enough energy is coupled into the VT in order to form the resonances. While the "control" and "partial" groups have similar first and even second formants, the "TEP" formants are shifted higher in frequency and the first formant of "total" speakers is very much higher (as also noted in McLoughlin et al. and Sharifzadeh et al.[6,41]) An issue commonly reported by "total" laryngectomy speakers[4] is difficulty in being heard in noise, or low perceived volume, which is clearly evident in Figure 6 where the envelope energy between 100 and 1000 Hz is significantly lower than for all other groups. In summary, the GFM-IAIF derived envelopes in Figure 6 effectively highlight the differences between groups in terms of not only the main glottal features, but also the main vocal tract features.

Results overall show that as expected, laryngectomy patients without TEP achieve almost no glottal resonance, whereas the use of TEP provides an improvement in terms of glottal function, although the degree of improvement varies among individuals. Meanwhile, those with partial glottal function tended to produce a glottal formant that is as strong and almost as sharp in bandwidth as for the control speakers. Interestingly, the degree of variation intrasubject was higher in the groups with greater dysfunction, indicating that the unimpaired group, and those who communicate effectively with partial glottal damage, may have a tighter degree of vocal control. Intersubject variation may also be higher in the groups with greater dysfunction, although testing with more subjects would be required to confirm this.

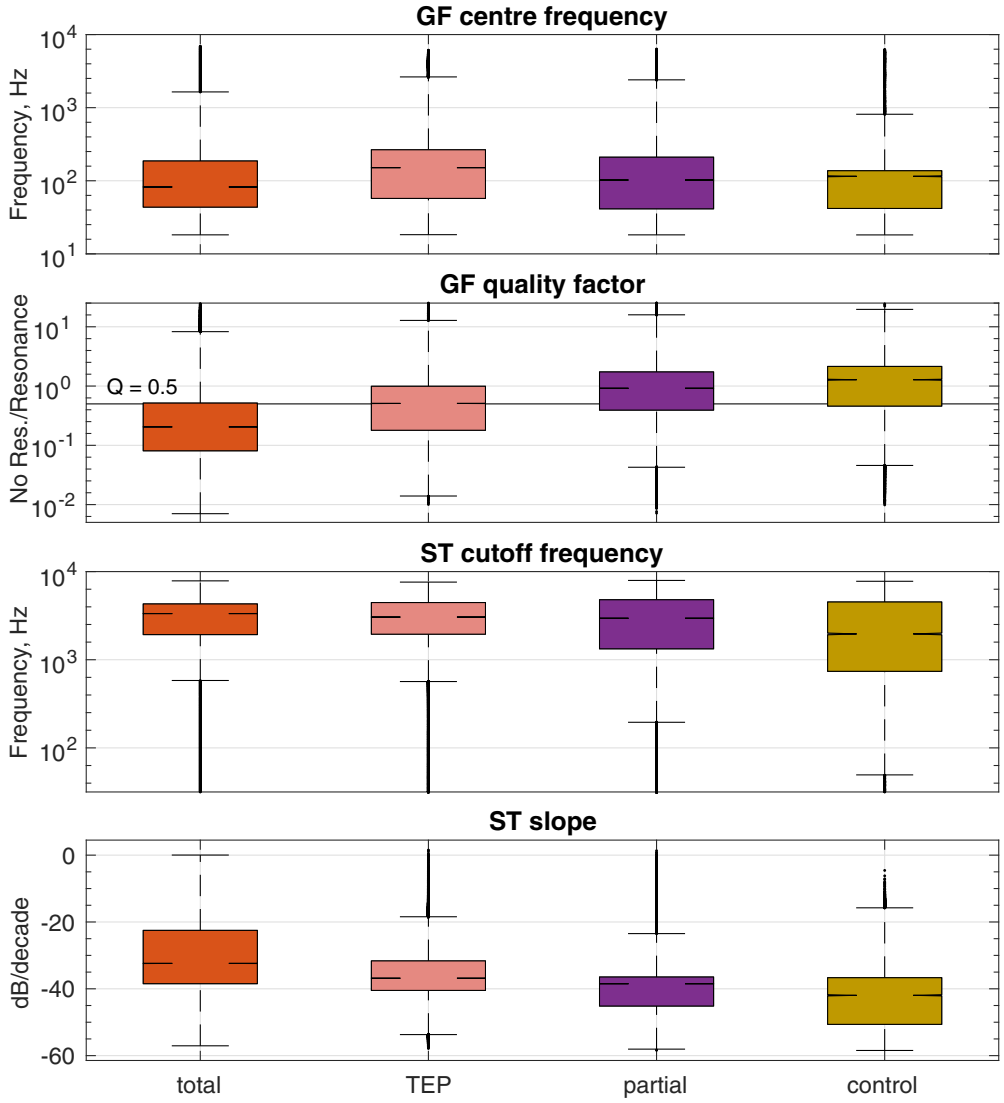

FIGURE. 7. Plots of the four GFM-IAIF derived parameters over different groups for all recordings of their normal mode speech.

Comparing the characteristics of individual GFM-IAIF parameters (particularly with reference to Figure 4), we conclude that both glottal formant quality factor and spectral tilt slope are most discriminative between groups, and furthermore that the former may be more discriminate between speaking modes (as phonated as possible vs unphonated).

Analysis of General Speech

The parameters for all speakers (excluding only deliberate whispers for those who are able to produce phonated speech) are now analyzed over all recorded material, and plotted in Figure 7. It should be stressed that this analysis includes all frames from all recordings of sentences, phrases and discrete words. It not only encompasses vowels and consonants but also includes silence between words, and thus represents a very much worse case analysis compared to the highly controlled vowel-only analysis above. Immediately, it is clear that the quality of the data is reduced − with much wider spans of parameters in each group. For the GF center frequency, the ordering from highest to lowest has now changed slightly from that of Figure 4. Specifically, TEP now has a slightly higher average frequency than for the total group. ST cutoff frequency is similarly affected, and both changes are due to the fact much of the analyzed data is now unvoiced consonants, whereas previously it included only voiced (where able) vowels. Meanwhile, however, both Q_{GF} and particularly ST remain discriminative ($\chi^2 = 1.55 \times 10^5$ and 1.06×10^5 respectively, with $df = 3$ and $P < 10^{-16}$ for both).

CONCLUSION

This paper applied the recent GFM for iterative adaptive inverse filtering (IAIF) to analyze recordings from a set of dysfunctional speakers. Subjects included those with larynx-related impairment such as laryngectomy, those using a tracheoesophageal puncture (TEP), and those with partial or unilateral glottal damage, as well as reference speech from unimpaired speakers. The GFM-IAIF decomposition into three parameters describing glottal formant center frequency, quality factor, and spectral tilt cutoff frequency, plus the additional measure of spectral tilt, show an ability to provide discriminative information for the four classes of total laryngectomy, partial glottal function, laryngectomy with TEP and control speakers. Both the visualizations of the data and the subsequent statistical analysis reveal the ability of GFM-IAIF to extract this useful information.

Interestingly, the analysis of unphonated and phonated speaking modes (eg, Figure 4 bottom row) corroborates the common assumption that post-laryngectomy speech − at least in terms of glottal function − can be usefully modeled by unimpaired speakers who are whispering, although a significant (20 dB/decade) spectral tilt and a formant shift should be applied to unimpaired whispers for them to better resemble total laryngectomy speech.

By employing a single algorithm to analyze voice recordings, this research leads towards future speech reconstruction for impaired speakers which can automatically and dynamically adjust reconstruction method based upon the type and degree of reconstruction required. Results could also be useful in adaptive front-end analysis for speech systems such as ASR, to adjust back-end models to suit the class of impaired speech being input.

Finally, we reiterate that GFM-IAIF analysis is complementary to techniques which use vocal-tract or phonetic information − which may also be discriminative between classes. This allows the possibility in future of developing combined methods which fuse the result of glottal, vocal and phonetic analysis to enhance the ability to understand, process or reconstruct impaired input speech in a way which is adaptive to its characteristics.

REFERENCES

1. Royal College of Speech and Language Therapists. Key statistics about speech and language therapy. 2017. [Online]. Available: https://www.rcslt.org/influencing/key_stats.
2. Bhattacharyya N. The prevalence of voice problems among adults in the United States. *Laryngoscope*. 2014;124:2359–2362.[Online]. Available: https://doi.org/10.1002/lary.24740
3. Morris MA, Meier SK, Griffin JM. Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. *Disabi Health J*. 2016;9:140–144. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1936657415001016
4. Norgate S, Oswald N. Laryngectomy is not a tragedy (2nd ed.). *Cancer Laryngectomee Trust,*1984.
5. Mustafa MB, Rosdi F, Salim SS, et al. Exploring the influence of general and specific factors on the recognition accuracy of an {ASR} system for dysarthric speaker. *Expert Syst Appl*. 2015;42(8):3924–3932. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417415000482
6. McLoughlin IV, Sharifzadeh HR, Tan SL, et al. Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation. *ACM Trans Accessible Comput (TACCESS)*. 2015;6:12.
7. Duffy JR. *Motor speech disorders: substrates, differential diagnosis, and management*. United States: Elsevier - Health Sciences Division, St Louis; 2012.
8. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. *J Speech Lang Hearing Res*. 1969;12:246–269.
9. Ackermann WZH. Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *J Neurol Psychiatry*. 1991;54:1093–1098.
10. Keller E, Vigneux P, Laframboise M. Acoustic analysis of neurologically impaired speech. *Brit J Disord Commun*. 1991;26:75–94.
11. Lawrenc D, Shriberg A, Fourakis M, et al. Extensions to the speech disorders classification system. *J Clin Linguist Phonet*. 2010:795–824.
12. Ross ED, Rush AJ. Diagnosis and neuroanatomical correlates of depression in brain-damaged patients. *J Arch Gen Psychiatry*. 1981;38:1338–1344.
13. Pietruch R, Michalska M, Konopka W. Methods for formant extraction in speech of patients after total laryngectomy. *Biomed Signal Proc and Control*. 2005;1:107–112.
14. Dehqan A, Yadegari F. Correlation of vhi-30 to acoustic measurements across three common voice disorder. *J Voice*. 2017;31.
15. Brockmann M, Drinnan MJ, Stock C. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel,gender,vocal intensity, and fundamental frequency effects in a typical clinical task. *J Voice*. 2011;25:44–53.
16. Fischer E, Goberman AM. Voice onset in parkinson disease. *J Commun Disord*. 2010;43:21–34.

17. Allen J, Sharifzadeh H, McLoughlin I. Acoustic analysis and computerised reconstruction of speech in laryngectomised individuals. *137th Annual Meeting of American Laryngological Association (ALA), Chicago*. ALA; 2016.
18. Sharifzadeh HR, McLoughlin I, Ahmadi F. Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. *IEEE Trans Biomed Eng*. 2010;57:2448–2458.
19. Toda T, Nakagiri M, Shikano K. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Trans Audio Speech Lang Process*. 2012;20:2505–2517.
20. Sharifzadeh H, Rassouliha AH, McLoughlin I. A training-based speech regeneration approach with cascading mapping models. *Elsevier Comput Electr Eng*. 2017;62:601–611.
21. Perrotin O, McLoughlin IV. Glottal flow synthesis for whisper-to-speech conversion. *IEEE/ACM Trans Audio SpeechLang Process*. 2020;28:889–900.
22. Perrotin O, McLoughlin I. A spectral glottal flow model for source-filter separation of speech. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK*. IEEE; 2019:7160–7164.
23. McLoughlin IV. *Speech and Audio Processing: a MATLAB-based approach*. Cambridge University Press; 2016.
24. Doval B, d'Alessandro C, Henrich N. The spectrum of glottal flow models. *Acta Acustica*. 2006;92:1026–1046.
25. Fant G, Liljencrants J, Lin Q. A four-parameter model of glottal flow. *R Instit Technol - Dept SpeechMusic Hearing Q Progr Status Rep*. 1985;4.
26. Fant G. The LF-model revisited. transformations and frequency domain analysis. *R Instit Technol - Dept SpeechMusic Hearing Q Progr Status Rep*. 1995;2−3.
27. Henrich N, d'Alessandro C, Doval B. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. In: *Proc. of Eurospeech*, Aalborg, Denmark, September 3-7, 2001, pp. 47−50.
28. Doval B, d'Alessandro C, Henrich N. The voice source as a causal/anticausal linear filter. *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, ser. VOQUAL '03. Geneva, Switzerland: ISCA, August 27-29, 2003, pp. 15−20.
29. Childers D. Vocal quality factors: Analysis, synthesis and perception. *J Acoust Soc Am*. 1991;90:2394–2410.
30. Harwardt C. Comparing the impact of raised vocal effort on various spectral parameters. In: *Proc. of Interspeech, Florence, Italy, August 28−31*, 2011, pp. 2941–2944.
31. Duvvuru S, Erickson M. The effect of change in spectral slope and formant frequencies on the perception of loudness. *J of Voice*. 2013;27:691–697.
32. d'Alessandro C, Doval B. Experiments in voice quality modification of natural speech signals: the spectral approach. *ISCA Speech Synthesis Workshop*. Jenolan Caves House, Blue Mountains, Australia; November 26−29, 1998, pp. 277–282.
33. Feugère L, d'Alessandro C, Doval B. Cantor digitalis: chironomic parametric synthesis of singing. *EURASIP J Audio Speech Music Process*. 2017.
34. Gobl C, Chasaide AN. The role of voice quality in communicating emotion, mood and attitude. *Speech Comm*. 2003;40:189–212.
35. Jensen MK. Recognition of word tones in whispered speech. *WORD*. 1958;14:187–196.
36. Sharifzadeh H, McLoughlin I. *Speech Rehabilitation Methods for Laryngectomised Patients*. Dordrecht: Springer Netherlands; 2010.597−607
37. Olthoff A, Mrugalla S, Laskawi R, Frohlich M, et al. Assessment of irregular voices after total and laser surgical partial laryngectomy. *Arch Otolaryngol-Head Neck Surg*. 2003;129:994–999.09, [Online]. Available: https://doi.org/10.1001/archotol.129.9.994
38. Tartter VC. What's in a whisper? *J Acoust Soc Am*. 1989;86:1678–1683. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/86/5/10.1121/1.398598
39. Eklund I, Traunmller H. Comparative study of male and female whispered and phonated versions of the long vowels of swedish. *Phonetica*. 1997;54:1–21.01
40. Thomas IB. Perceived pitch of whispered vowels. *J Acoust Soc Am*. 1969;46:468–470.
41. Sharifzadeh H, McLoughlin IV, Russell MJ. A comprehensive vowel space for whispered speech. *J Voice*. 2012;26:49–56.
42. Degottex G. *Glottal source and vocal-tract separation: Estimation of glottal parameters, voice transformation and synthesis using a glottal model*. Ph.D. dissertation, Univ. Pierre et Marie Curie (UPMC), Nov. 2010.
43. Alku P. Glottal inverse filtering analysis of human voice production − a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*. 2011;36:623–650.
44. Makhoul J. Linear prediction: a tutorial review. *Proc IEEE*. 1975;63:561–580.
45. Alku P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Comm*. 1992;11:109–118.
46. Mokhtari P, Story B, Alku P, et al. Estimation of the glottal flow from speech pressure signals: evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production. *Speech Comm*. 2018;104:24–38.
47. Wood J, Athanasiadis T, Allen J. Laryngitis. *BMJ*. 2014.
48. Nemr K, Simões Zenari M, Cordeiro GF, et al. GRBAS and Cape-V scales: High reliability and consensus when applied at different times. *J Voice*. 2012;26.812.e17−812.e22, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0892199712000392
49. Perrotin O, Loughlin IM. GFM-voc: a real-time voice quality modification system. In: *Proc. Interspeech 2019, Graz, Austria*, 2019, pp. 3685–3686.
50. Bristow-Johnson R. 2001. March Audio-eq-cookbook. [Online]. Available: https://music.columbia.edu/pipermail/music-dsp/2001-March/041752.html.