Fundamental Frequency Estimation of Low-quality Electroglottographic Signals

*Christian T. Herbst, and †,‡Jacob C. Dunn, *Vienna, Austria, and †‡Cambridge, UK

Summary: Fundamental frequency (f_o) is often estimated based on electroglottographic (EGG) signals. Because of the nature of the method, the quality of EGG signals may be impaired by certain features like amplitude or baseline drifts, mains hum, or noise. The potential adverse effects of these factors on f_o estimation have to date not been investigated.

Here, the performance of 13 algorithms for estimating f_o was tested, based on 147 synthesized EGG signals with varying degrees of signal quality deterioration. Algorithm performance was assessed through the standard deviation σ_{fo} of the difference between known and estimated f_o data, expressed in octaves.

With very few exceptions, simulated mains hum, and amplitude and baseline drifts did not influence f_o results, even though some algorithms consistently outperformed others. When increasing either cycle-to-cycle f_o variation or the degree of subharmonics, the SIGMA algorithm had the best performance (max. $\sigma_{fo} = 0.04$). That algorithm was, however, more easily disturbed by typical EGG equipment noise, whereas the NDF and *Praat*'s auto-correlation algorithms performed best in this category ($\sigma_{fo} = 0.01$).

These results suggest that the algorithm for f_o estimation of EGG signals needs to be selected specifically for each particular data set. Overall, estimated f_o data should be interpreted with care.

Key Words: Electroglottography−EGG−Fundamental frequency−f_o

INTRODUCTION

Fundamental frequency (f_o) is one of the key parameters used for the quantitative description of voice signals.[1−5] f_o represents the rate of vibration of the laryngeal sound generator, typically consisting of the vocal folds in humans and most mammals. f_o detection is performed under the assumption that the analyzed sound source exhibits periodic vibration.

A time series such as the (acoustic) voice signal is said to be periodic when it precisely repeats itself at certain intervals, mathematically expressed as

$$x(t \pm nT_o) = x(t) \tag{1}$$

where t is time, n is a positive integer, and T_o is the period,[6] that is, the duration of one glottal cycle. The smallest possible value of T_o of a periodic time series that satisfies Equation 1 is called the fundamental period of that time series, and its inverse is the fundamental frequency:

$$f_o = \frac{1}{T_o} \tag{2}$$

Several different ways of denoting fundamental frequency are used in the literature (eg, *F0* or f_0 with a subscript zero). However, a recent consensus paper suggests to use the denotation f_o with a lower letter f and the character $_o$ (for "oscillatory") instead of the zero (for "zero harmonic") as a subscript.[7]

f_o is often confused with "pitch." f_o is a property of the vibration of a physical system, measured in hertz [Hz]. In contrast, pitch is a psychoacoustic quantity, defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high."[8] In quite a few cases the two quantities approximate each other, but not always. Hence, the term "pitch" should only be used if (human) perception is addressed, and be avoided when laryngeal sound generation is described as a physical system.

Voice, as practically any other biosignal, is never purely periodic. Rather, it is nearly periodic at best (some authors use the term "quasi-periodic," which, however, is reserved for describing a signal with two individual fundamental frequencies[6,9]). For one, f_o traces typically contain linear or quadratic terms, introduced by gradual changes of f_o. Additionally, even the most steady vocalizations contain slight cycle-to-cycle alterations—see Ref. 9 for a very good discussion. More severe phenomena are constituted by irregularity/chaos, subharmonics ("period doubling," "period tripling," etc) or multiphonia or biphonation, constituted by two independent sound sources.[11,12] These issues make f_o detection nontrivial, particularly so in pathologic voices, certain singing styles, and in animal bioacoustics, where often the laryngeal sound source exhibits nonlinear phenomena like irregularity, subharmonics, and bifurcations between different vibratory states.[13]

Strictly speaking, f_o can thus not be calculated for voice signals, because f_o is a property of purely periodic signals. Consequently, there is always a certain degree of inherent inaccuracy in any f_o estimation. In the words of Owren and Linker (1995), "All pitch extraction techniques are found to fail under some circumstances, which places a burden on the investigator to consistently monitor the performance of

Accepted for publication January 4, 2018.
Conflict of interest: There are no competing interests declared.
From the *Bioacoustics Laboratory, Department of Cognitive Biology, University of Vienna, Vienna, Austria; †Behavioural Ecology Research Group, Department of Biology, Faculty of Science & Technology, Anglia Ruskin University, Cambridge, UK; and the ‡Division of Biological Anthropology, University of Cambridge, Cambridge, UK.
Address correspondence and reprint requests to Christian T. Herbst, Bioacoustics Laboratory, Department of Cognitive Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria. E-mail: herbst@ccrma.stanford.edu
Journal of Voice, Vol. 33, No. 4, pp. 401−411
0892-1997

https://doi.org/10.1016/j.jvoice.2018.01.003

each routine being used."[5] Regrettably, apart from some informal recommendations,[14] no rigorously established limit or respective error ranges for the acceptable degree of irregularity have been established. This makes comparison of f_o data ranges presented in different studies highly problematic.

One additional complication of f_o detection is sometimes introduced by the degeneration of the analyzed acoustic signal by background noise. Lacking anechoic chambers or other adequately sound-treated rooms, in a medical setting this problem can be circumvented by directly assessing the process of laryngeal sound production, for example, via the glottal area waveform,[15] derived by analysis of endoscopic laryngeal high-speed videos.[16,17] However, the respective equipment is expensive and not always available.

A noninvasive alternative for assessing laryngeal vibration is electroglottography (EGG), pioneered by Fabre in 1957.[18] In EGG, a high-frequency, low-voltage current is passed between two electrodes, which are placed on either side of the thyroid cartilage. Changes in vocal fold contact area during vocal fold vibration result in admittance variations, and the resulting EGG signal is proportional to the relative vocal fold contact area.[19] A number of parameters quantitatively describing the laryngeal sound generation process can be extracted from a properly recorded EGG signal.[19] Among others, the EGG signal is an ideal candidate for assessment of the (time-varying) f_o because it is neither influenced by vocal tract acoustics nor by background noise.

Even under optimal conditions there can be a certain degree of distortion in an acquired EGG signal—see, for example, Ref. 20 for a discussion. Further quality degeneration of the EGG signal can be introduced by inadequately positioned EGG electrodes (eg, caused by excessive larynx or neck movement); reduced conductivity between EGG electrodes and the larynx due to tissue fat, beards, or fur (in animals); radio signals or mains hum interference with the utilized electroglottograph; or noise introduced during (potentially wireless) transmissions of the EGG signals from the electroglottograph to the recording device. Furthermore, EGG signals from pathologic voice production can consist of nonperiodic sequences. All these influences potentially pose challenges for f_o detection, as described earlier.

For these reasons, we decided to formally assess the performance of a number of algorithms for f_o estimation when analyzing a set of synthesized EGG signals with six types of artificially induced distortion of quality. One key aim of this study is to assess how the *Praat* software package, one of the standard tools for f_o assessment in animal bioacoustics, performs in relation to eight other algorithms mainly known from human speech processing.

MATERIALS AND METHODS

Synthetic test signals

A set of synthesized EGG signals at various stages of corruption were generated at a sampling frequency of 48,000 Hz. Each synthesized signal had a duration of 2 seconds. The f_o information for each glottal cycle within a synthesized signal was derived randomly from a Gaussian distribution centered around 1000 Hz with a standard deviation of 500 Hz. Only f_o data between 100 Hz and 2000 Hz were considered. This extended range was chosen to encompass the singing voice range of humans and vocalizations of some nonhuman mammals.

The f_o values were sorted in ascending order, and the resulting information was used to drive a kinematic vocal fold vibration model.[21] The model's default parameters were used ($Q_a = 0.3$; $Q_s = 3.0$; $Q_b = 1.0$; $Q_p = 0.2$). This process resulted in synthetic EGG signals with nonlinearly increasing f_o, as illustrated in Figure 1. The time offset and the period of the resulting glottal cycles within each synthesized signal were stored for later comparison with the analysis results from the tested algorithms.

As mentioned in the introduction, there are a number of factors that can introduce distortions into the recorded EGG signal and make f_o estimation problematic. To test the potential effect of these factors, the following features were introduced into the synthesized EGG signals at various degrees:

1. **Random f_o variation**: When generating the individual EGG cycles, their respective period was allowed to vary randomly within a certain range. This processing step was introduced after sorting the f_o values retrieved from the Gaussian distribution (see previous discussion). The final f_o of consecutive cycles within each synthesized signal was determined by
$$f_o'(t) = f_o[1 + \alpha(RND[0..1] - 0.5)] \qquad (3)$$
where α is the f_o random factor, which was varied between 0 (no f_o variation) and 0.3. A comparison between α and the relative average perturbation (RAP), a voice quality parameter to assess pathologic human voice production,[22] suggests a relationship of $RAP = 0.2118\,\alpha + 0.0029$, $R_2 = 0.9996$. (The y-intercept of 0.0029 was introduced by the nonlinear increase of f_o in the synthesized signals.) As a reference, for healthy humans, RAP values of 0.0021−0.0089 were reported,[19] which would be the equivalent of $\alpha = [-0.0038...0.0283]$. Pathologic voices were measured to have RAP values of 0.0068−0.0452, corresponding to $\alpha = [0.0187...0.1997]$.
2. **Subharmonics**: The presence of subharmonics, a relatively common feature in mammalian vocalization, was simulated by scaling the amplitude of every other synthesized EGG glottal cycle by $(1 - \beta)$, where the factor β was varied between 0 and 0.3. Nonzero values of β resulted in the appearance of period-2 subharmonics (period doubling). The parameter value range follows Bergan and Titze,[23] who found that the perceptual pitch-drop of an octave occurred at amplitude modulation rates of 10%−30%.
3. **Amplitude drift**: The temporal variation of the EGG signal amplitude was simulated by introducing a sinusoidally varying amplitude modulation at an arbitrarily

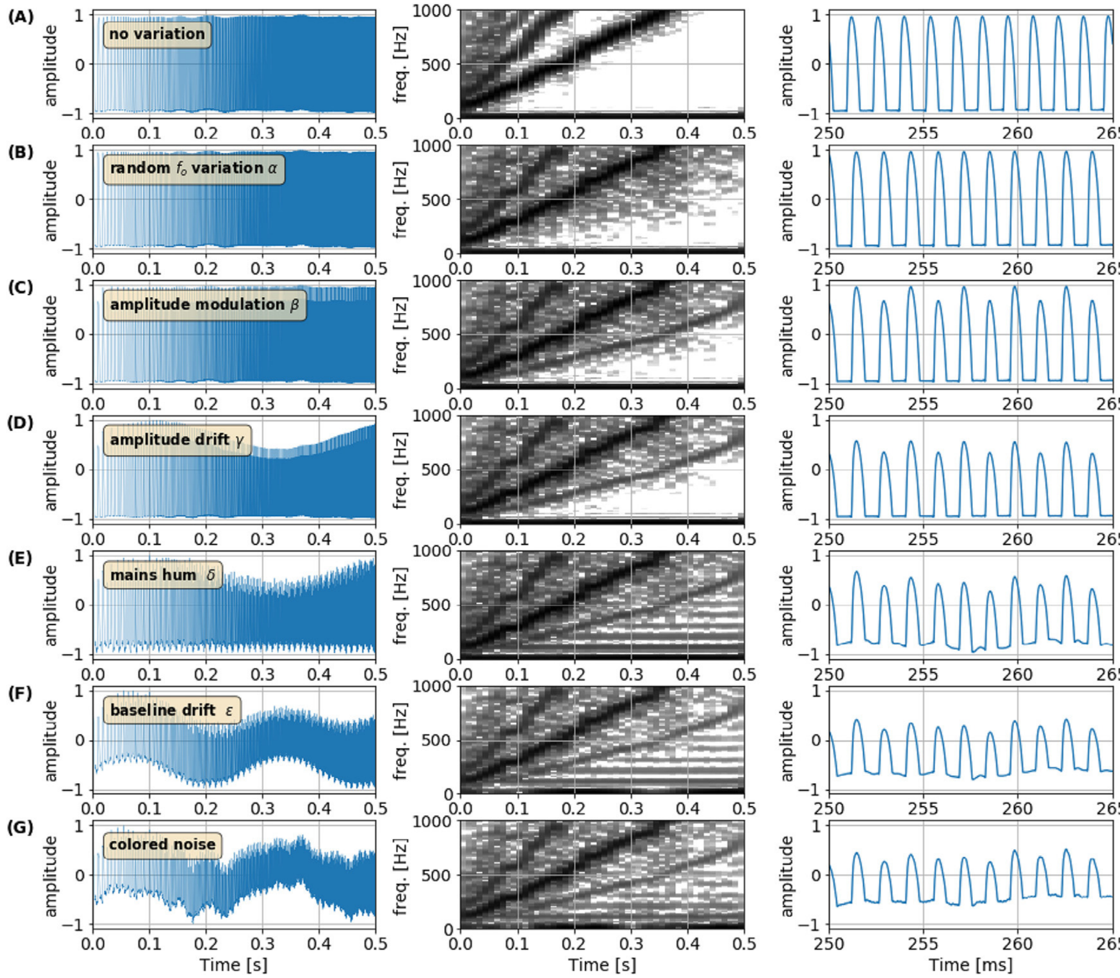

FIGURE 1. Illustration of EGG signal synthesis, cumulatively introducing the features which degenerate the EGG signal quality at various stages (see text). The left panels show the EGG signal (reduced to 0.5 seconds to increase the clarity of the illustration). The middle panels contain a narrow-band spectrogram of the EGG signal. In the right panels the first ten glottal cycles of each signal are displayed. (**A**) Undistorted synthesized signal; (**B**) random f_o variation, a = 0.15; (**C**) introduction of subharmonics, b = 0.15; (**D**) amplitude drift added, g = 0.3; (**E**) mains hum added, d = 0.15; (**F**) baseline drift added, e = 0.4; and (**G**) typical EGG equipment noise added, SNR = 15 dB.

fixed rate of $f_{AM} = 2.27$ Hz. In particular, the EGG signal was multiplied by

$$(1-\gamma) + \frac{\gamma[1+\sin(2\pi f_{AM} t)]}{2} \qquad (4)$$

where the amplitude modulation factor γ was varied between 0 and 0.6 across synthesized signals.

4. **Mains hum**: A mains hum signal with a duration of 2 seconds was synthesized as a harmonic series at 100 Hz f_o with a steadily decaying spectral envelope, using a spectral slope of -6 dB per octave. A total of 20 harmonics were included. This mains hum signal was scaled by the factor δ and then added to the normalized synthesized EGG. The amplitude scaling factor δ was varied between 0 (no mains hum added) and 0.3.
5. **Baseline drift**: The baseline offset of the signal was allowed to vary sinusoidally at an arbitrarily defined fixed rate. In particular, the following baseline drift was added to the synthesized signals: $0.5\,\varepsilon\,[1 + sin(2\pi f_{BD} t)]$, where $f_{BD} = 1.71$ Hz. The factor ε was varied between 0 (no baseline drift) and 0.8 across the synthesized signals.
6. **Noise**: Finally, colored noise was added to the synthesized EGG signal to simulate various signal-to-noise ratios (SNR). SNR was varied in a range of -5 dB to 35 dB. These values were taken from a recent study which reported these surprisingly low SNR ranges for EGG signals recorded from humans in laboratory conditions.[24] The noise was generated by scaling the frequency components of white noise in the frequency domain during a forward-backward Fourier transform. The amplitudes for the frequency-dependent scaling were derived from averaged noise contained in previously recorded EGG signals.[24]

Deviating from the "best case" of the synthesized EGG signal, six data sets were generated, where each of the aforementioned six parameters was varied in isolated fashion at 21 equidistantly spaced intervals (see Figure 1 for an example). Additionally, one data set was generated where all six parameters were varied at once (termed "compound" scenario in the remainder of this text). In this fashion, a total of 147 (21 × 7) synthesized EGG signals were produced.

Evaluated algorithms

Owing to frequently occurring linear/quadratic trends and cycle-to-cycle aberrations in a voice signal, quantitative analysis focuses on the time-varying rather than the mean f_o. This is either achieved by short-term windowed approaches,[25] where short portions of the voice signal are evaluated at consecutive time instants, or by estimation of the so-called glottal closure instants (GCI).[26] GCI detection operates on the assumption that the major sound generation event occurs at the instant of glottal closure, that is, (partial) collision of the laryngeal or syringeal tissue, and that each glottal cycle has a period (sometimes called "epoch"[27]) that is determined by two consecutive GCIs. Recalling Equation 2, the time-varying f_o is then found by taking the inverse of the period.

f_o estimators and GCI detectors may operate on different computational principles. The majority of them are rooted in either the time domain (looking at similarities or recurrent features in the voice signal) or in the frequency domain (by further analyzing the time-varying spectrum of the voice signal, as produced by a Fourier transform). Alternative approaches include, for example, wavelet analysis[28−30] or phase space analysis.[31] Description of the concepts involved in the various algorithms is beyond the scope of this paper. The reader is referred to the landmark textbook by Hess.[25] Some key approaches are described in Owren and Linker's summary.[5] Further, good overviews are given by Talkin[32] and Drugman et al,[33] the latter of which discusses some more recent developments.

Overall, a surprisingly large number of f_o estimators have been described in the literature. In 1983, Hess already stated that "literally hundreds of pitch-determination methods and algorithms have been developed."[25] In Appendix S1 of this paper, we include a nonexhaustive list of 75 f_o and GCI estimators, addressing some past and recent developments, and providing web links to free source code or software applications where applicable.

Given the number of options, a practical solution had to be found for selecting the algorithms tested in this paper. Besides focusing on the algorithms included in the *Praat* software package,[34] our main selection criterion was (1) free availability of the algorithm source code, and (2) ability to operate the algorithm within a free software environment, that is, the *Linux* operating system, and, where applicable, *GNU Octave*,[35] the free equivalent to *MATLAB*. A total of 13 such algorithms were included in this study:

- five algorithms from the *Praat* software, version 5.4.06. The following methods were tested in this study: "to Pitch (ac)," "to Pitch (SHS)," "to PointProcess (periodic, cc)," "to PointProcess (periodic, peaks)," and "to PointProcess (zeroes)." These algorithms are referred to as **Praat (AC)**, **Praat (SHS)**, **Praat (periodic cc)**, **Praat (periodic peaks)**, and **Praat (zeros)**, respectively, for the remainder of this text. Preliminary analysis suggested that *Praat*'s methods "to Pitch (SPINET)" and "to PointProcess (extrema)" produced greatly inferior results. These two algorithms were thus excluded from this report;
- the **DECOM** algorithm[36] presented in[37];
- the **DYPSA** GCI algorithm, introduced by Kounoudes et al[38] and described in further detail by Naylor et al[39];
- The **NDF** (nearly defect-free) f_o detector,[40] implemented as MulticueF0v14.m, version 2016-06-30;
- The **RAPT** algorithm,[32] implemented as fxrapt.m in the voicebox package;
- David Talkin's **REAPER** algorithm (unpublished work; https://github.com/google/REAPER);
- the **SIGMA** GCI detector, developed by Thomas and Naylor[30];
- the **SWIPE'** algorithm, developed by Camacho and Harris[41];
- the **YAGA** GCI detector, developed by Thomas et al[42];

Web links for downloading the software of the algorithms utilized in this comparison can be found in the supplementary materials.

All algorithms were controlled through a set of custom scripts written in *Python* by author C.T.H., operated on Ubuntu *Linux* 16.04 LTS. *Praat* and the compiled C-code of REAPER were accessed through command-line pipes. All other algorithms were available as *MATLAB* (MathWorks, Natick, Massachusetts) code. They were thoroughly tested in *GNU Octave* 4.0 and were then embedded into the custom *Python* code through *Python*'s oct2py wrapper module for *MATLAB*/*Octave* code.[43] For all algorithms, the respective standard parameters were used, except for the upper and lower limits, which (where possible) were specified as 100 Hz and 2000 Hz, respectively. The upper frequency limits of REAPER and the voicebox-based DYPSA, RAPT, and SIGMA algorithms had to be changed from 500 Hz to 2000 Hz in the respective source code. All f_o detection algorithms (*Praat* (AC), *Praat* (SHS), NDF, RAPT, REAPER, SWIPE') were operated at a time step of 1 ms.

Combining algorithm outputs

Preliminary assessment of the performance of the algorithms suggested that there was no single algorithm that performed best under all conditions. Rather, the SIGMA GCI detector and the *Praat* autocorrelation (AC) f_o estimator showed the most robust performance in different subsets of the synthesized data (see Results). In an attempt to consolidate the benefits of these two algorithms, a custom analysis approach (denoted as CUSTOM for the remainder of this paper) was implemented as follows: SIGMA GCI data were converted to f_o information at a time-step of 1 ms. For each data point (totaling 2000 for 2 seconds of synthesized sound), the difference between f_o data from *Praat* AC and

SIGMA was computed, expressed in octaves. If that difference was below a certain threshold, an f_o data point was generated by the CUSTOM algorithm (NaN otherwise). The threshold was arbitrarily defined as 5% of an octave. Preliminary tests with a more rigorous threshold of 1/120 octave (ie, 10 musical cents), which approximates the just-noticeable difference for pitch perception in humans,[44] considerably decreased the usefulness of the CUSTOM algorithm, due to the great number of rejected data points even at slight levels of EGG signal quality degeneration.

Testing procedure

Including the CUSTOM algorithm, 14 algorithms were tested on the 147 EGG signals described earlier, resulting in a total number of 2058 observations. Prior to f_o calculation and GCI detection, the EGG signals were band-pass filtered twice using a third-order Butterworth filter with cutoff frequencies at 20 Hz and 4800 Hz. The second consecutive application of the filter was performed on the time-inverted input signal to negate phase distortion effects. The application of the band-pass filter was deemed appropriate, because comparable preprocessing steps would be performed in "real" data analysis situations. The cutoff frequencies were chosen carefully so as not to distort the analyzed signals.

Evaluation of performance

The output of f_o estimators and GCI detectors is fundamentally different in nature. Although the f_o estimators produce equidistantly spaced data points (every 1 ms in the case of this study) representing the time-varying (quasi-instantaneous) f_o information, the GCI detectors provide estimates of the time offsets of presumed glottal closure instants. To prevent adding any bias to the analysis (neither in favor of either f_o estimation nor GCI detection methods), we initially decided to compare the performance of all tested algorithms in both domains.

When comparing two frequencies, their difference in hertz is meaningless as an absolute value. A relative measure needs to be established instead. For the purpose of this study, the frequency differences between known and estimated f_o values were expressed in octaves[45]:

$$\Delta oct = \log_2\left(\frac{f_{SYNTH}}{f_{EST}}\right) \tag{5}$$

For performance evaluation in the f_o domain, the glottal cycle information from the systhesized signal was converted to a time-series of f_o data at intervals of 1 ms. Based on this information, the following three parameters were calculated:

- A success metric, expressing the number of produced f_o data points in percent:

$$\rho_{f_o} = 100\frac{m}{n} \tag{6}$$

where n is the total number of possible data points (2000 for 2 seconds of synthesized sound) and m is the number of actually detected data points.

- Applying Equation 5, the average of the absolute differences between known f_o information from the synthesized signals and estimated f_o data was computed as follows:

$$\mu_{f_o} = \frac{1}{n}\sum_{0}^{n-1}|\Delta oct[i]| \tag{7}$$

- Similarly, the standard deviation of f_o estimation was computed as:

$$\sigma_{f_o} = \sqrt{\frac{1}{n}\sum_{0}^{n-1}(\Delta oct[i])^2} \tag{8}$$

The performance metrics parameters ρ_{GCI}, μ_{GCI}, and σ_{GCI} for GCI detection were calculated in analogy to those for f_o estimation, with the difference that n was defined as the total number of glottal cycles in the respective synthesized signal.

Preliminary inspection of the algorithm performance data revealed no remarkable differences between the f_o-based (ρ_{fo}, μ_{fo}, and σ_{fo}) and the respective GCI-based values (ρ_{GCI}, μ_{GCI}, and σ_{GCI}), suggesting that conversion between f_o and GCI information did not introduce noteworthy artifacts into the data. Furthermore, there were no substantial differences of trends between the μ_{fo} and σ_{fo} parameters. For these reasons, the remainder of this text focuses on the f_o-related parameters ρ_{fo} and σ_{fo} alone.

RESULTS

Detailed results of f_o detection from one representative signal are shown in Figure 2. An overview of the parameters ρ_{fo} and σ_{fo} for all analyzed scenarios is given in Figures 3 and 4. Detailed μ_{fo} success rates and σ_{fo} scores for all analysis scenarios are provided in supplementary Tables S1 and S2.

With a few exceptions of extreme EGG signal modifications in the "compound" scenario and for extreme SNR values, most algorithms produced data for more than 90% of the possible 2000 data points per synthesized signal (Figure 3). Exceptions to this trend were found in the RAPT and DECOM algorithms, which typically had ρ_{fo} values of about 90% and 80%, respectively. The CUSTOM algorithm deviated from its typical 95% f_o detection success rate when the random f_o variation α was increased above 0.1 and when the amplitude modulation β was greater than 0.12, suggesting that above these critical values, the f_o readings from the two algorithms upon which the CUSTOM algorithm is based (ie, *Praat*'s "to Pitch (AC)" and SIGMA—see Methods) deviated by more than 5% of an octave.

Three of the analyzed algorithms (DYPSA, REAPER, and YAGA), all designed with the purpose of analyzing human speech, had problems recognizing f_o above

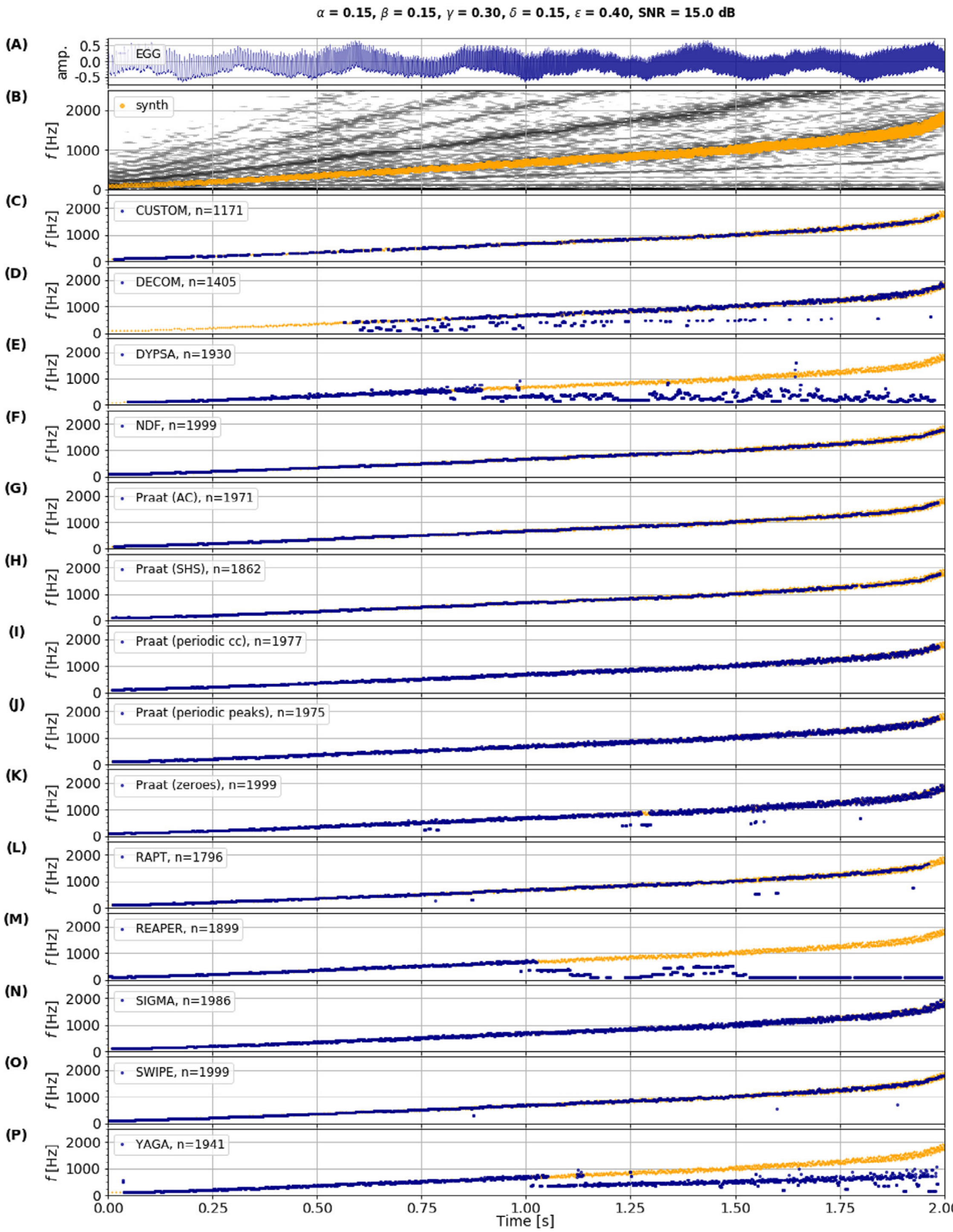

FIGURE 2. Detailed results of f_o detection from the synthesized signal depicted in Figure 1. (**A**) Synthesized EGG signal; (**B**) narrow-band spectrogram of synthesized EGG signal, known f_o data superimposed; (**C**)−(**P**) f_o detection results for all evaluated algorithms (dark dots), superimposed upon known f_o data (light dots).Data from GCI detectors were converted to equidistantly spaced f_o values (see Methods). The illustrated synthesized EGG signal represents the “compound” case 10 in Figures 3G and 4G.

ca. 1000 Hz. Consequently, they were the worst-performing algorithms analyzed. The error benchmark σ_{fo} for the DECOM algorithm was typically around 10% of an octave, rising considerably with increased random f_o variation α. All other algorithms started out with acceptable σ_{fo} ratings for EGG signals at lesser degrees of EGG signal quality distortion. However, increased random f_o variation had a tendency to gradually increase σ_{fo} in all algorithms except DYPSA, REAPER, and YAGA. Overall, the CUSTOM and SIGMA algorithms had the best performance when testing for random f_o variation (Figure 4A).

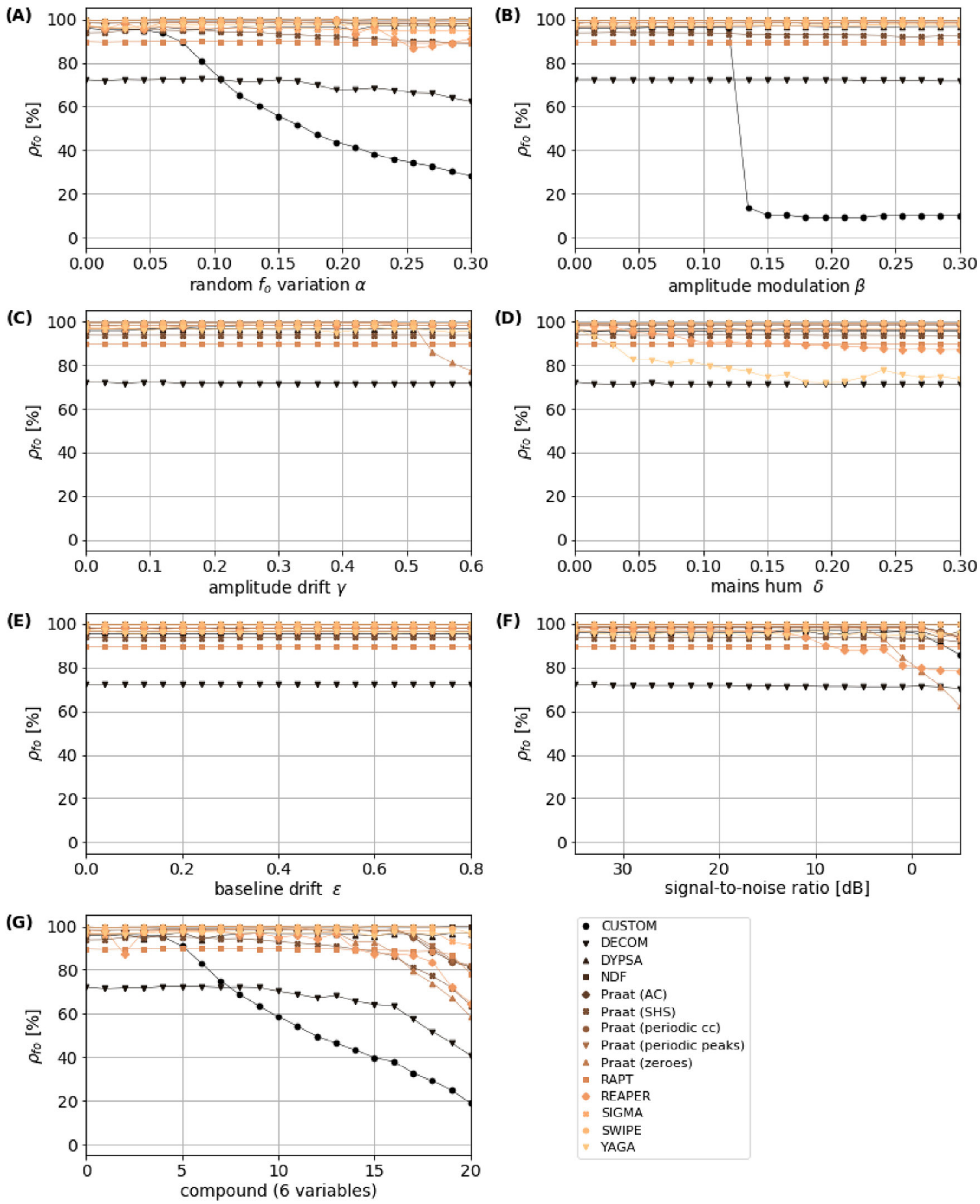

FIGURE 3. f_o Data point resolution metric ρ_{fo} for all analyzed algorithms and all synthesized EGG signals. (**A**)−(**F**) ρ_{fo} as a function of the six simulated influence factors on EGG signal quality. (**G**) Effect of simultaneous change of all six influence factors.

For most of the algorithms, the occurrence of subharmonics appeared to be a crucial factor that led to abrupt increases in σ_{fo} over an amplitude modulation range of $0.1 > \beta > 0.24$ (Figure 4B). In each of these cases, the respective algorithm started to latch onto the subharmonic energy components in the signal. The respective threshold values were found at NDF: $\beta = 0.21$; *Praat* (AC): $\beta = 0.14$; *Praat* (SHS): $\beta = 0.24$; *Praat* (periodic cc): $\beta = 0.14$; *Praat* (periodic peaks): $\beta = 0.14$; RAPT: $\beta = 0.15$; and SWIPE: $\beta = 0.12$. As with random f_o variation, the CUSTOM and SIGMA algorithms had the best performance with increased amplitude modulation.

No noteworthy trends were found with variation in amplitude drift, mains hum, or baseline drift—Figures 4C-E. (Preliminary experiments conducted without band-pass filtering the signals before analysis revealed the same trends, even for baseline drifts.) The only exception was the NDF algorithm, which suffered an abrupt decrease in performance for amplitude drifts $\gamma < 0.42$, and the DECOM algorithm, which achieved reduced σ_{fo} values for $\gamma < 0.45$ and $\delta < 0.04$.

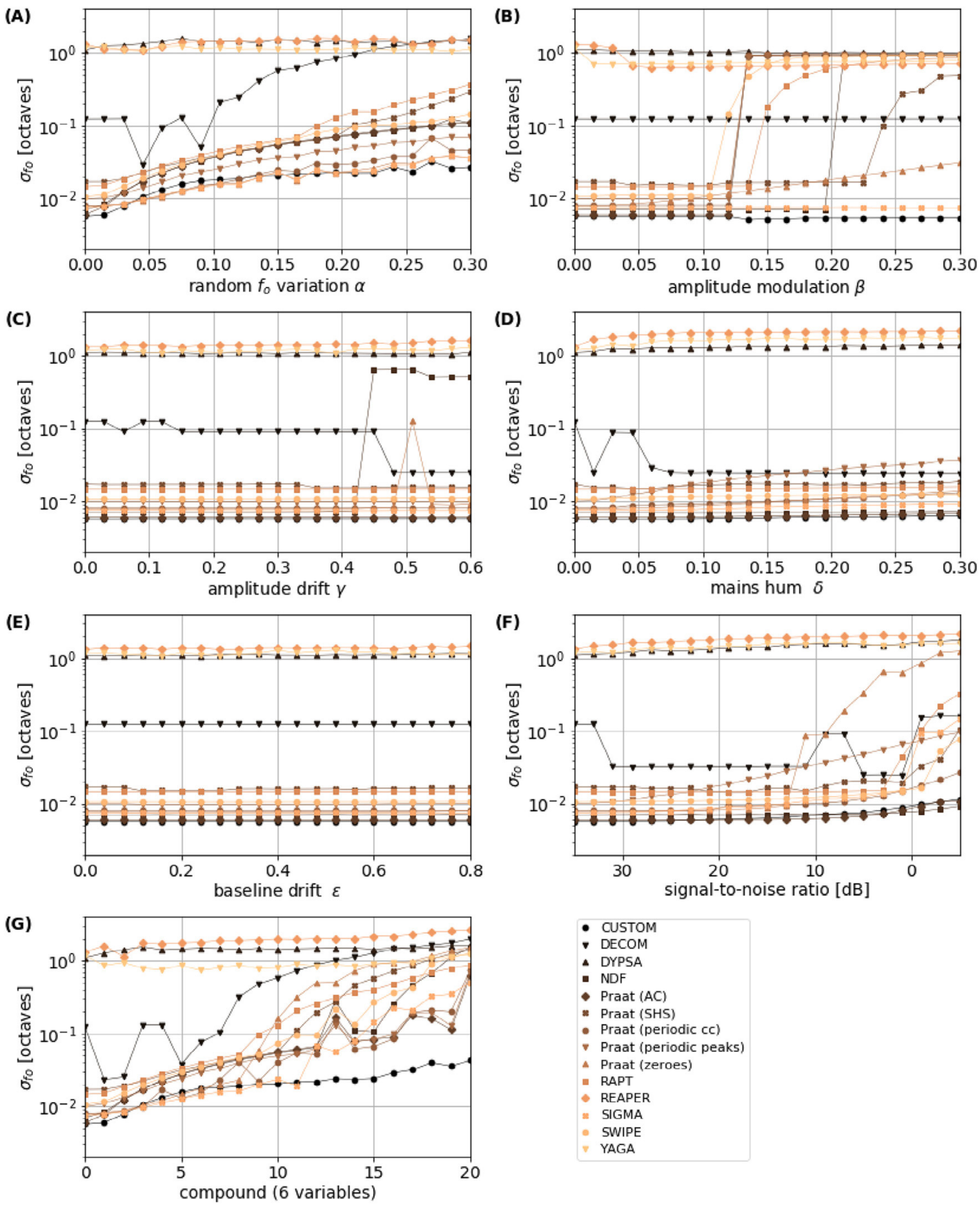

FIGURE 4. f_o Detection performance metric σ_{fo} for all analyzed algorithms and all synthesized EGG signals. (**A**)−(**F**) σ_{fo} as a function of the six simulated influence factors on EGG signal quality. (**G**) Effect of simultaneous change of all six influence factors.

Finally, typical EGG equipment noise seemed to be an important factor, influencing a number of algorithms (Figure 4F): There was an almost linear correlation between SNR of noise and σ_{fo} in the "*Praat* (periodic peaks)" algorithm. More abrupt degenerations of performance (measured by increasing σ_{fo}) were found for the following algorithms at respective thresholds: *Praat* (SHS): SNR = −5 dB; Praat (zeroes): SNR = 11 dB; RAPT: SNR = 1 dB; SIGMA: SNR = −1 dB; SWIPE: SNR = −3 dB. The CUSTOM, NDF, and Praat (AC) algorithms appeared to perform particularly well under the influence of noise, with terminal values of $\sigma_{fo} = 0.01$ at an SNR of −5 dB.

Algorithm performance for linear combinations of the six influence factors described above are shown in the "compound" scenario illustrated in Figure 4G. The CUSTOM algorithm had a notably better performance (ie, lower σ_{fo} values) than all other algorithms, particularly at higher degrees of EGG signal deterioration. This performance success was, however, counterbalanced by the algorithm's lowered success rates μ_{fo} (Figure 3G).

DISCUSSION

This study examines the performance of a number of f_o and GCI detection algorithms when analyzing a special class of signals, that is, EGG signals with increasing complexity and at various stages of signal quality degradation. A total of six influence factors were assessed in this study: two inherent to the voice signal itself (random f_o variation and subharmonics), and four types of signal degradations (amplitude and baseline drifts, mains hum, and typical EGG equipment noise). Mains hum, amplitude drift, and baseline drift all appeared to have a lesser influence on algorithm performance. The opposite was true for the other three factors—alteration in cycle-to-cycle variation (introduced by random f_o variation), subharmonics (introduced by amplitude modulation of odd cycles), and typical EGG equipment noise all had a clear impact on algorithm performance.

The somewhat disquieting main finding of this study is that there does not seem to exist one single "best" algorithm for analyzing EGG signals at various stages of complexity and degradation. For high-quality, low-noise EGG signals (eg, those typically acquired in excised larynx settings) the SIGMA algorithm seems to be the best choice. In signals with low SNRs, such as those collected *in vivo* from humans with a certain degree of fat tissue or phonating with incomplete glottal closure,[24] or signals with suboptimal EGG electrode placement, the SIGMA algorithm does not appear to be the best choice. In those cases, NDF or *Praat*'s AC algorithm would appear to be better suited. However, the performance of both NDF and *Praat*'s AC algorithm is negatively affected by the occurrence of subharmonics.

In an attempt to consolidate these trends, a CUSTOM approach was introduced in this paper, combining the virtues of both SIGMA and *Praat*'s AC algorithm. This CUSTOM algorithm showed the best performance overall (particularly in the "compound" scenario), but the improved performance came at the expense of discarding a large proportion of the analyzed data in situations where the outputs of the SIGMA and AC algorithms did not converge. This obvious trade-off between data quality and quantity can, to a certain degree, be controlled via the CUSTOM algorithm's threshold setting (see Methods).

Some of the analyzed algorithms are intended to operate on certain types of signals.[36] This may partially explain why the DYPSA, REAPER, and YAGA algorithms failed to produce meaningful data outside the typical f_o ranges of human speech. Therefore, inferior performance of an algorithm in this study does not constitute a reason to conclude that the respective algorithm is inferior *per se*.

When studying the literature, the following pattern emerged: Most of the proposed f_o and GCI detection algorithms were introduced by comparing their results with those from some other algorithms (differing across the various studies, and typically basing the tests on different input signals across different studies). Interestingly, in all of these cases, the respectively proposed algorithm had comparable or better performance than all other algorithms. Five non-mutually exclusive explanations can be found for this phenomenon:

(a) Owing to progress in the field of engineering the newly introduced algorithms become increasingly better over the years;
(b) Some algorithms work better for a certain type of data (eg, noisy data[46] or special voice production types[47,48]) than others;
(c) Different methods of estimating algorithm performance result in different outcomes[49];
(d) The authors of studies might have had a certain (potentially unconscious) *a priori* bias toward their "own" algorithm, which may have influenced them in choosing test data and competing algorithms for their performance tests; or finally,
(e) The authors may have made the mistake to train their algorithm on the chosen test data, leading to an over-specialized algorithm performance which cannot be generalized to other data sets.

Surprisingly, even studies that are only concerned with comparing algorithm performance (without introducing a new algorithm) do not converge to identical recommendations,[48−54] suggesting that estimating algorithm performance might be as complex a task as f_o or GCI detection itself. One way to address this issue is by consensually establishing databases of test signals with known properties. Advancing this notion, we have made all synthesized EGG signals utilized in this study available as supplementary materials.

Some of the considerations concerning standardizing algorithm performance evaluation also apply to this study. The f_o range for synthesized signals was somewhat arbitrarily chosen to be in the range of 100−2000 Hz, in consideration of the human singing voice and the vocalizations of some nonhuman mammals. Furthermore, whereas three of the parameters for determining the synthesized EGG signals were chosen in relation to known value ranges (random cycle-to-cycle variation α, amplitude modulation β, and SNR), the value ranges of the other three parameters had to be defined in an arbitrary fashion, based on the first author's long-term experience with EGG signals. Different values may naturally lead to different performance evaluation results. This is particularly true for the "compound" case, where all six parameters were varied in unison. In fact, preliminary tests with different, more extreme value ranges produced slightly different trends. For this reason, we have refrained from computing an overall metric of success across all synthesized signals. Such a metric would only apply to the given test data set and could not be generalized.

CONCLUSION

This study corroborates the insight that f_o detection is highly nontrivial.[5,25,32] No single "best algorithm" was found for the special class of signals analyzed in this study. Thus, no recommendation for one single all-purpose f_o detection algorithm can be given. Rather, the nature of EGG data needs to be studied carefully before choosing an appropriate algorithm, and the insights from this study can help with that choice. Such an informed approach is recommended, rather than defaulting to a commonly used algorithm.

In summary, some main insights from this study are as follows: The researcher should never blindly trust a chosen f_o detection algorithm. *Ex post facto*, computed f_o data should always be assessed "by eye," for example, via f_o traces superimposed upon narrow-band spectrograms. Furthermore, f_o data reported in the literature should not be taken at face value, particularly if the authors did not disclose (1) which f_o detection algorithm was chosen; (2) how the utilized f_o detection algorithm was chosen; and (3) whether (and how) the computed data were double-checked manually. There is an inherent degree of uncertainty and error in such data, due to the difficulties in automated f_o detection described in this paper.

Acknowledgments

This work is supported by an APART grant from the Austrian Academy of Sciences (to C.T.H.).

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at doi:10.1016/j.jvoice.2018.01.003.

REFERENCES

1. Baken RJ, Orlikoff RF. *Clinical Measurement of Speech and Voice*. Vol. 2. 2nd ed. San Diego, CA: Singular Publishing, Thompson Learning; 2000.
2. Fischer J, Noser R, Hammerschmidt K. Bioacoustic field research: a primer to acoustic analyses and playback experiments with primates. *Am J Primatol*. 2013;75:643–663.
3. Fitch WT, Hauser MD. Vocal production in nonhuman primates: acoustics, physiology, and functional constraints on 'honest' advertisement. *Am J Primatol*. 1995;37:191–219.
4. Fletcher NH. Acoustic systems in biology: from insects to elephants. *Acoust Aust*. 2005;3:83–88.
5. Owren MJ, Linker CD. Some analysis methods that may be useful to acoustic primatologists. In: Zimmermann E, Newman JD, Jürgens U, eds. *Current Topics in Primate Vocal Communication*. New York: Springer; 1995:286.
6. Titze IR. Workshop on acoustic voice analysis. Summary statement. National Center for Voice and Speech. 1995.
7. Titze IR. Some consensus has been reached on the labeling of harmonics, formants, and resonances. *J Voice*. 2016;30:129.
8. ANSI. USA standard acoustical terminology (including mechanical shock and vibration). *Tech Rep*. 1960 S1.1-1960.
9. Bergé P, Pomeau Y, Vidal C. *Order Within Chaos: Towards a Deterministic Approach to Turbulence*. Paris: Hermann and John Wiley & Sons; 1984.
10. Roark RM. Frequency and voice: perspectives in the time domain. *J Voice*. 2006;20:325–354.
11. Herzel H, Berry D, Titze IR, et al. Analysis of vocal disorders with methods from nonlinear dynamics. *J Speech Hear Res*. 1994;37:1008–1019.
12. Titze IR, Baken RJ, Herzel H. Evidence of chaos in vocal fold vibration. In: Titze IR, ed. *Vocal Fold Physiology: Frontiers Basic Science*. San Diego, CA: Singular Publishing Group; 1993:143–188.
13. Fitch WT, Neubauer J, Herzel H. Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. *Anim Behav*. 2002;63:407–418.
14. Friedrich G, Dejonckere PH. Das Stimmdiagnostik−Protokoll der European Laryngological Society (ELS)—erste Erfahrungen im Rahmen einer Multizenterstudie. *Laryngo Rhino Otol*. 2005;84:744–752.
15. Childers DG, Naik JM, Larar JN, et al. Electroglottography, speech, and ultra-high speed cinematography. In: Titze IR, Scherer R, eds. *Vocal Fold Physiology and Biophysics of Voice*. Denver, CO: Denver Center of Performing Arts; 1983:202–220.
16. Deliyski DD, Hillman RE. State of the art laryngeal imaging: research and clinical implications. *Curr Opin Otolaryngol Head Neck Surg*. 2010;18:147–152.
17. Hertegard S. What have we learned about laryngeal physiology from high-speed digital videoendoscopy? *Curr Opin Otolaryngol Head Neck Surg*. 2005;13:152–156.
18. Fabre P. Un procédé électrique percuntané d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute fréquence; premiers résultats (A non-invasive electric method for measuring glottal closure during phonation: High frequency glottography; first results). *Bull Acad Nat Med*. 1957;141:66–69.
19. Hampala V, Garcia M, Svec JG, et al. Relationship between the electroglottographic signal and vocal fold contact area. *J Voice*. 2016;30: 161–171.
20. Baken RJ. Electroglottography. *J Voice*. 1992;6:98–110.
21. Titze I. A four-parameter model of the glottis and vocal fold contact area. *Speech Commun*. 1989;8:191–201.
22. Koike Y. Application of some acoustic measures for the evaluation of laryngeal dysfunction. *Stud Phonol*. 1973;VII:17–23.
23. Bergan C, Titze IR. Perception of pitch and roughness in voice signals with subharmonics. *J Voice*. 2001;15:165–175.
24. Herbst CT, Schutte HK, Bowling DL, et al. Comparing chalk with cheese—The EGG contact quotient is only a limited surrogate of the closed quotient. *J Voice*. 2017;31:401–409.
25. Hess W. *Pitch Determination of Speech Signals: Algorithms and Devices*. Heidelberg, Germany: Springer-Verlag; 1983.
26. Tuan VN, D'Alessandro C. Robust glottal closure detection using the wavelet transform. *Proc Eur Conf Speech Technol*. 1999;2808:2805.
27. Ananthapadmanabha T, Yegnanarayana B. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans Acoust*. 1979;27:309–319.
28. Kadambe S, Boudreaux-Bartels GF. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans Inf Theory*. 1992;38:917–924.
29. Manfredi C, Aniello MD, Bruscaglioni P, et al. A comparative analysis of fundamental frequency estimation methods with application to pathological voices. *Med Eng Phys*. 2000;22:135–147.
30. Thomas MRP, Naylor PA. The SIGMA algorithm: a glottal activity detector for electroglottographic signals. *IEEE Trans Audio Speech Lang Process*. 2009;17:1557–1566.
31. Hagmüller M, Kubin G. Poincaré sections for pitch mark determination. In: *ITRW on Non-Linear Speech Processing (NOLISP 05)*. 2005:1–7.
32. Talkin D. A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn WB, Paliwal KK, eds. *Speech Coding and Synthesis*. New York: Elsevier; 1995:495–518.
33. Drugman T, Alku P, Alwan A, et al. Glottal source processing: from analysis to applications. *Comput Speech Lang*. 2014;28:1117–1138.
34. Boersma P, Weenink D. *Praat: Doing Phonetics by Computer*. Amsterdam, The Netherlands: Institute of Phonetic Sciences, University of Amsterdam; 2017.
35. Eaton JW, Bateman D, Hauberg S, et al. GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations. 2015.

36. Henrich N. *DECOM*. 2006.
37. Henrich N, d'Alessandro C, Doval B, et al. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *J Acoust Soc Am*. 2004;115:1321–1332.
38. Kounoudes A, Naylor PA, Brookes M. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. *ICASSP*. 2002:349–352.
39. Naylor PA, Kounoudes A, Gudnason J, et al. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Trans Audio Speech Lang Process*. 2007;15:34–43.
40. Kawahara H, De Cheveigné A, Banno H, et al. Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. *Interspeech*. 2005537–540.
41. Camacho A, Harris JG. A sawtooth waveform inspired pitch estimator for speech and music. *J Acoust Soc Am*. 2008;124:1638–1652.
42. Thomas MRP, Gudnason J, Naylor PA. Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. *IEEE Trans Audio Speech Lang Process*. 2012;20:82–91.
43. Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. 2001.
44. Rossing T. *The Science of Sound*. London, UK: Addison-Wesley Publishing Company; 1990.
45. Young RW. Terminology for logarithmic frequency units. *J Acoust Soc Am*. 1939;11:134–139.
46. Drugman T, Drugman T, Alwan A. Joint robust voicing detection and pitch estimation based on residual harmonics. *Interspeech*. 2011.
47. Kane J, Gobl C. Evaluation of glottal closure instant detection in a range of voice qualities. *Speech Commun*. 2013;55:295–314.
48. Babacan O, Drugman T, Henrich N, et al. A Quantitative Comparison of Glottal Closure Instant Estimation Algorithms on a Large Variety of Singing Sounds. *Interspeech*. 2013:1–5.
49. Rabiner L, Cheng M, Rosenberg A, et al. A comparative performance study of several pitch detection algorithms. *IEEE Trans Acoust*. 1976;24:399–418.
50. Jang S-J, Choi S-H, Kim H-M, et al. Evaluation of performance of several established pitch detection algorithms in pathological voices. In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* Vol. 2007. 2007:620–623.
51. Cheng MJ, Rabiner LR, Rosenberg AE, et al. Comparative performance study of several pitch detection algorithms. *J Acoust Soc Am*. 1975;58(S1):S61–S62.
52. Parsa V, Jamieson DG. A comparison of high precision F0 extraction algorithms for sustained vowels. *J Speech Lang Hear Res*. 1999; 42:112–126.
53. Titze IR, Liang H. Comparison of F0 extraction methods for high precision voice perturbation measurements. *J Speech Hear Res*. 1993; 36:1120–1133.
54. Tsanas A, Zañartu M, Little MA, et al. Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering. *J Acoust Soc Am*. 2014;135:2885–2901.