Annals of Otology, Rhinology & Laryngology
2017, Vol. 126(10) 712–716

DOI: 10.1177/0003489417728088
journals.sagepub.com/home/aor
SAGE

Validation of an Algorithm for Semi-automated Estimation of Voice Relative Fundamental Frequency

Yu-An S. Lien, PhD[1], Elizabeth S. Heller Murray, MS, CCC-SLP[2], Carolyn R. Calabrese, MS, CCC-SLP[2], Carolyn M. Michener, BS[2], Jarrad H. Van Stan, PhD, CCC-SLP[3,4], Daryush D. Mehta, PhD[3,4,5], Robert E. Hillman, PhD, CCC-SLP[3,4,5], J. Pieter Noordzij, MD[6], and Cara E. Stepp, PhD[1,2,6]

Abstract

Objectives: Relative fundamental frequency (RFF) has shown promise as an acoustic measure of voice, but the subjective and time-consuming nature of its manual estimation has made clinical translation infeasible. Here, a faster, more objective algorithm for RFF estimation is evaluated in a large and diverse sample of individuals with and without voice disorders.

Methods: Acoustic recordings were collected from 154 individuals with voice disorders and 36 age- and sex-matched controls with typical voices. These recordings were split into training and 2 testing sets. Using an algorithm tuned to the training set, semi-automated RFF estimates in the testing sets were compared to manual RFF estimates derived from 3 trained technicians.

Results: The semi-automated RFF estimations were highly correlated (r = 0.82-0.91) with the manual RFF estimates.

Conclusions: Fast and more objective estimation of RFF makes large-scale RFF analysis feasible. This algorithm allows for future work to optimize RFF measures and expand their potential for clinical voice assessment.

Keywords

acoustics, voice assessment

Introduction

Vocal hyperfunction (VH), "a hypertonic state of both intrinsic and extrinsic laryngeal musculature,"[1] is associated with the majority of voice disorders. Current clinical assessment often relies on subjective measures based on auditory perception (eg, voice quality), visual perception (eg, endoscopic imaging), and manual palpation of neck musculature,[2] which are prone to reliability issues.[3-5] This can make evaluations completed by different clinicians difficult to interpret. Objective measures may be useful adjuncts to subjective measures. An acoustic measure that has shown promising results for the assessment of VH is relative fundamental frequency (RFF), a measure of the fundamental frequency (f_o) of an onset or offset vocal cycle relative to steady-state (Figure 1). Relative fundamental frequency is lower in individuals with VH compared to those with healthy voices,[6] potentially due to their increased baseline laryngeal tension.[7]

Adoption of RFF for clinical and research applications is currently hampered by its time-consuming manual estimation. At least 6 RFF speech sequences are needed for a reliable estimate,[8] requiring 20 to 40 minutes of analysis. Incorporating this additional time into voice evaluation is clinically infeasible. In addition, the current protocol requires extensive training for operators to be reliable. Technicians must first individually inspect each RFF instance and make a subjective decision about the location of the boundary between voiced and voiceless speech. Then, Praat's f_o detection algorithm[9] is used to identify 10 offset and onset cycles. An objective method is particularly needed to accomplish the first step of the manual process (identifying the boundaries between voiced and voiceless

[1]Department of Biomedical Engineering, Boston University, Boston, Massachusetts, USA
[2]Department of Speech, Language, and Hearing Sciences, Boston University, Boston, Massachusetts, USA
[3]Massachusetts General Hospital Institute of Health Professions, Boston, Massachusetts, USA
[4]Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital, Boston, MA, USA
[5]Department of Surgery, Harvard Medical School, Cambridge, MA, USA
[6]Department of Otolaryngology – Head and Neck Surgery, Boston University School of Medicine, Boston, Massachusetts, USA

Corresponding Author:
Yu-An S. Lien, Department of Biomedical Engineering, Boston University, 635 Commonwealth Avenue, Boston, MA 02215-1300, USA.
Email: slien@bu.edu

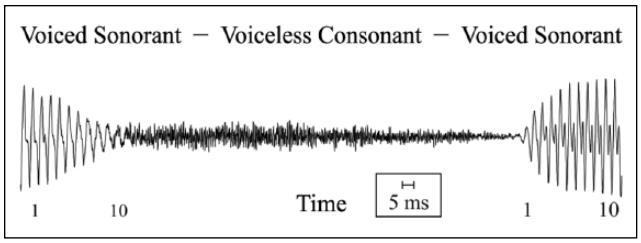

Figure 1. An acoustic waveform of a relative fundamental frequency (RFF) instance, /ufu/. The 1st and 10th vocal cycles are denoted.

speech) because f_o estimation methods are less effective at identifying voicing offsets (voiced to unvoiced transitions) and onsets (unvoiced to voiced transitions)[10] and user decisions about this boundary are subjective: Technician decisions are based on individual interpretations and potentially different criteria. Thus, the development of faster and more objective RFF estimation is warranted. The purpose of this study is to test the results of an algorithm developed to meet this need against manual estimates of RFF.

Method

Participants and Recording Procedure

A control group (C) of 36 adults (27 females) aged 18 to 85 years (M = 41, SD = 19) all reported no prior history of speech, language, or hearing disorders. A group of 154 adults (116 females) aged 18 to 83 years (M = 41, SD = 17) with voice disorders (VD) had been diagnosed with a voice disorder by a board-certified laryngologist. A speech-language pathologist judged participants' overall severity of dysphonia (0-100) using the Consensus Auditory-Perceptual Evaluation of Voice.[11] The 0th, 25th, 50th, 75th, and 100th percentiles of overall severity in the VD group were 0.0, 8.7, 22.9, 54.4, and 99.1, respectively. In the VD group, 54 individuals had nonphonotraumatic VH (muscle tension dysphonia, defined as VH without vocal fold damage), 81 had "secondary" VH (symptoms of VH accompanied by vocal fold lesions, inflammation, edema, and/or glottal insufficiency), and 19 had voice disorders not primarily associated with VH (eg, gastroesophageal reflux disease, spasmodic dysphonia, Parkinson's disease). Participants completed written consent in compliance with either the Boston University Institutional Review Board or the Massachusetts General Hospital (MGH) Institutional Review Board.

Roughly half the speakers were recorded in a waiting area or quiet room in Boston Medical Center (BMC) using a Shure WH20XLR microphone (Shure, Niles, Illinois, USA), sampled at 44.1 kHz with 16-bit resolution. The remaining speakers were recorded in a sound-treated room at MGH using a Sennheiser MKE104 microphone (Sennheiser, Wedemark, Germany), sampled at 20 kHz with 16-bit resolution. The VD participants in the 2 settings varied in severity, etiology, and occupation.

Participants produced a set of 3 /afa/ utterances, took a breath, produced a set of 3 /ifi/ utterances, took a breath, and produced a set of 3 /ufu/ utterances; they were instructed to use their typical pitch and loudness. These stimuli were chosen as they yield low intraspeaker variability compared to running speech stimuli and other voiceless phonemes.[12] These tokens also shorten the recording protocol and facilitate algorithmic processing.

Manual RFF Analysis

Manual RFF analysis was independently performed on each audio sample by 3 trained technicians using Praat software to estimate pulse timings before and after each voiceless consonant.[9] Technicians manually altered Praat settings on a per sample basis. Technicians then decided if the sample should be rejected due to aperiodicity, glottalization, or lack of steady-state voicing. Offset f_o values were normalized (in semitones; STs) to the f_o of the first offset cycle and onset f_o values to the f_o of the 10th onset cycle (cycles furthest from the consonant).

Each technician reestimated 15% of their samples in a different sitting. The intrarater reliability, calculated using the Pearson's product-moment correlation coefficients and the resultant RFF values, ranged from 0.90 to 0.95. The interrater reliability, calculated using the intraclass correlation coefficients, type (2, k, absolute), was 0.97.

Semi-automated Algorithm Design

The RFF estimation algorithm was implemented in MATLAB (MathWorks, Natick, Massachusetts, USA) in 4 steps. First, fricatives and vowels of the 3 RFF instances in the acoustic waveform were identified using ratios of high- and low-frequency energy. Second, the speaker's f_o range was estimated via autocorrelation from the vowels. The acoustic signal was band-pass filtered ±3 ST around the f_o range. After this pre-processing, starting from the center of the fricative in the band-pass filtered signal, a sliding window set to the reciprocal of the f_o estimate was shifted backward and forward to find the peaks and troughs of potential vocal cycles in the offset and onset sonorants, respectively. Third, the boundary between voiced and voiceless waveforms was determined on a cycle-by-cycle basis by analyzing 3 parameters: the number of zero-crossings, shape dissimilarity between adjacent cycles, and peak-to-peak amplitude. A threshold for each of these parameters was chosen by maximizing the effect size of the difference between potential voiced and unvoiced segments. This threshold was then further tuned to manual estimates by adjusting the threshold in step sizes of 5% from 80% to

120% to maximize the performance in a training set (see the following). Finally, a peak or trough was accepted as part of the offset vocal cycle if all cycles before it satisfied at least 2 out of the 3 parameter thresholds. Lastly, RFF instances that did not meet certain criteria (eg, those without at least 10 vocal cycles) were rejected. Otherwise, RFF was calculated based on the identified vocal cycles, similar to the manual process. Full details about algorithm design and implementation are found in chapter 6 of Lien.[13]

Semi-automated Algorithm Performance Evaluation

RFF Instance Identification. The accuracy of identifying fricative locations was calculated using a graphical interface in MATLAB with visualizations of the waveform. For those that were incorrectly identified, the interface was then used by an operator to correct the fricative location before evaluating the remainder of the algorithm. This allowed the user to briefly visualize each individual instance, allowing for quality control. This procedure is not fully automated but likely to provide more reliable results by end-users than a "blind" approach. In addition to allowing for correction of potential algorithm errors, it provides information about signal quality (eg, recording issues, competing noise, or inappropriate stimuli).

RFF Accuracy. The semi-automated RFF estimates were compared to the manual RFF estimates using Pearson's product-moment correlation coefficients, root mean square error (RMSE), and average difference of offset 10 and onset 1 estimates (automated estimates subtracted from manual estimates). A training set of 126 speakers (two-thirds of total data set) was used to tune the boundary between voiced and voiceless waveforms to maximize the correlation coefficient between manual and semi-automated RFF for offset cycle 10 and onset cycle 1 (values used in previous studies[6,7]). The training set was chosen to have similar makeup to the overall sample with respect to overall severity, voice disorder type, and recording condition (BMC vs MGH). The training set included 20 speakers from the C group (4 recorded at BMC and 16 recorded at MGH) and 106 speakers from the VD group (59 recorded at BMC and 47 recorded at MGH; 0th, 25th, 50th, 75th, and 100th percentiles of overall severity of 0.0, 8.5, 22.5, 54.7, and 99.1, respectively). Two testing sets were composed of the remaining 64 participants: (1) a group of 8 controls and 27 speakers with VD recorded at BMC and (2) a group of 8 controls and 21 speakers with VD recorded at MGH. The overall severity of the VD speakers in each group varied; the 0th, 25th, 50th, 75th, and 100th percentiles of overall severity was 2.3, 16.4, 35.6, 73.6, and 91.4 for the BMC test group and 4.3, 6.5, 13.5, 28.1, and 63.3 for the MGH group, respectively. These 2 test groups were used to evaluate the algorithm, using the boundary settings determined in the training set. As in the first step of evaluation, the algorithm allows the user to visualize each instance and the location of the cycles that were used for analysis. For the purposes of this study, the cycles used were not adjusted by the user; however, this feature allows future users to quickly gain more information about the basis of the resulting RFF values. The semi-automated RFF estimates were compared to manual RFF estimates for all speakers with at least 1 usable RFF instance in the testing set.

Results

Usable RFF Tokens

The average number of usable offset and onset semi-automated RFF instances per speaker in the BMC testing set were 4.5 (SD = 2.3) and 4.5 (SD = 2.2), respectively; these were lower than the average usable instances via manual estimation (7.8, SD = 1.7; 6.5, SD = 2.1). Out of the 35 speakers, 1 did not have any usable offset RFF instances, and 2 did not have any usable onset RFF instances.

The average number of usable offset and onset semi-automated RFF instances per speaker in the MGH testing set were 7.0 (SD = 2.1) and 8.2 (SD = 1.4), respectively; again, these were lower than the average usable instances via manual estimation (8.9, SD = .3; 8.6, SD = .8). Out of the 29 speakers, all had at least 1 usable offset and onset RFF instance.

Examining the algorithm's "usable" samples data from both test groups, there was a strong effect of order. There was a reduction in usable samples within each set of utterances as a function of order as well as a reduction of usable samples as a function of the set order. The percentage of usable samples for offset and onset RFF for the /afa/ stimulus were 100% and 97% for the first production, 94% and 94% for the second production, and 83% and 91% for the third production. The percentage of usable samples for offset and onset RFF for the /ifi/ stimulus were 77% and 83% for the first production, 67% and 72% for the second production, and 61% and 59% for the third production. The percentage of usable samples for offset and onset RFF for the /ufu/ stimulus were 42% and 51% for the first production, 27% and 42% for the second production, and 16% and 28% for the third production.

RFF Accuracy

Semi-automated RFF estimates for each speaker are plotted as a function of their manual RFF estimates in Figure 2. The semi-automated RFF estimates tended to have a smaller range compared to the manual RFF estimates. In the BMC testing set, the correlation coefficient and RMSE were .82

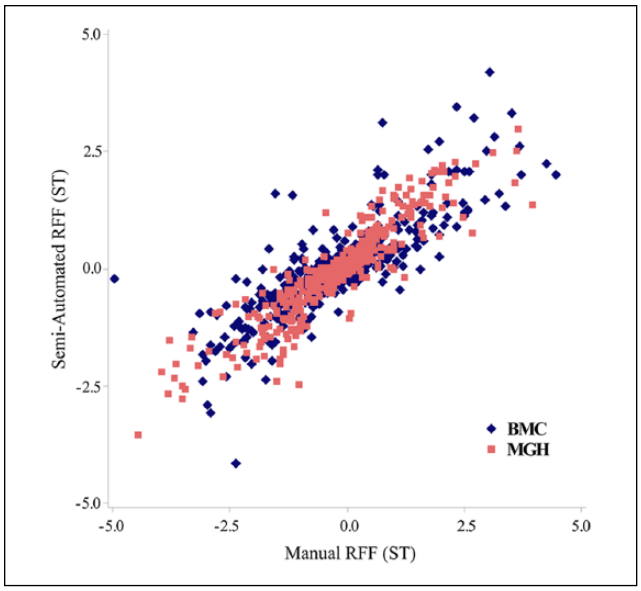

Figure 2. Testing sets: semi-automated relative fundamental frequency estimates plotted as a function of manual relative fundamental frequency estimates.

and .37 ST, respectively. The average difference in the BMC testing set between the manual estimates and semi-automated estimates was –.41 ST for offset 10 values and .10 ST for onset 1 values. In the MGH testing set, the correlation coefficient and RMSE were .91 and .28 ST, respectively. The average difference in the MGH testing set between the manual estimates and semi-automated estimates was –.22 ST for offset 10 values and .11 ST for onset 1 values.

Discussion

In both testing groups, semi-automated RFF estimates were highly correlated ($r \geq 0.82$) with manual RFF estimates (comparable to correlations between manual RFF estimates completed by different technicians).[6-8,12,14] The RMSE between the estimates was .28 to .27 ST, a difference that is partially explained by the fact that the semi-automated RFF estimates were derived from a band-pass filtered waveform with a pass-band around the speaker's f_o, whereas the manual RFF estimates were derived directly from the raw waveform. Band-pass filtering reduces the high-frequency content of the signal, which has a smoothing effect, reducing the cycle-to-cycle variation in f_o[15] and thus the RFF. Due to this bias, in future, semi-automated RFF estimates should not be directly compared to manual estimates.

Two different samples were included in this study to test the robustness of the algorithm in diverse populations. In general, the automated RFF values were more similar to the manual RFF values in the MGH group than in the BMC group, with a higher correlation and lower RMSE. The BMC VD speakers had higher average overall severity than the MGH VD speakers (40.4 vs 21.0), and the testing samples differed in terms of signal quality. The MGH speakers were recorded in a sound-treated room, whereas the BMC speakers were not. While all our samples were qualitatively deemed of sufficient quality for inclusion, they were recorded under varying conditions and with varying instrumentation, representing the diversity present in current clinical practice. Thus, differences in the algorithm performance between the 2 testing groups could be due to differences in recording condition, differences in overall severity, or both. Regardless, these results give best-case (low severity, ideal recording conditions) and worst-case (high severity, nonoptimal recording conditions) parameters for future clinical and research applications: On average, users of these algorithms can expect differences of –.41 ST to –.22 ST for RFF offset 10 values and .10 to .11 ST for onset 1 values. These differences are systematic and consistent with the algorithmic processing: Automated values will tend to be higher for offset 10 and lower for onset 1.

An unexpected finding was that there was a strong effect of order on the percentage of the algorithm's samples that were usable. The study was not designed to study this, so the order of production was not randomized: Participants produced a set of 3 /afa/ utterances, took a breath, produced a set of 3 /ifi/ utterances, took a breath, and produced a set of 3 /ufu/ utterances. Since the study was not designed to examine this effect, it is not clear to what extent this is an actual function of order or rather a result of stimuli type or breath group.

Although the RFF estimates computed with these algorithms were found to have systematic differences when compared to manual estimates, it is not clear that these are indeed errors. Manual RFF estimation, although the current gold standard, is subjective. The algorithmic estimates are objective and may provide more clinically appropriate information; thus far, they have been successfully applied to study the relationship between RFF and laryngeal tension,[16] whether RFF is sensitive to hydration and vocal loading,[17] and whether RFF can discriminate between hyperfunctional voice disorders with and without vocal lesions.[18]

Finally, like many acoustic measures, RFF estimation requires periodicity. In the VD group examined, only a small percentage did not have any sufficiently periodic samples to use for RFF estimates, even though one-fourth of the samples had overall severity ratings of 54.2 or greater. Nevertheless, the requirement for a periodic signal limits applicability to mild to moderate dysphonia. Future algorithms may benefit from using RFF in conjunction with overall measures of periodicity, expanding its usefulness across the spectrum of severity.

Conclusion

An algorithm that provides faster, more objective RFF estimation was examined. The semi-automated RFF estimations

were highly correlated with the manual RFF estimates, although the degree of correlation was impacted by sample characteristics. Future users of these algorithms can expect differences of –.41 ST to –.22 ST for RFF offset 10 values and .10 to .11 ST for onset 1 values when compared to manual estimates. These systematic differences are small, and due to the objective nature of the algorithms, these estimates are more reliable than current manual methods. Future research to improve RFF algorithms should include estimates of voice overall severity and acoustic signal quality (eg, room acoustics) in their design.

Acknowledgments

Thanks to Defne Abur, Christina Stevens, Alexandria Martinson, Laura Enflo, Theodore Kahn, Melissa Cooke, Amanda Fryd, and Molly Bresnahan.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants DC012651 (CES) and DC011588 (REH) from the National Institute on Deafness and Other Communication Disorders.

References

1. Holmberg EB, Doyle P, Perkell JS, Hammarberg B, Hillman RE. Aerodynamic and acoustic voice measurements of patients with vocal nodules: variation in baseline and changes across voice therapy. *J Voice*. 2003;17(3):269-282.
2. Morrison MD, Nichol H, Rammage LA. Diagnostic criteria in functional dysphonia. *Laryngoscope*. 1986;96(1):1-8.
3. Stepp CE, Heaton JT, Braden MN, Jetté ME, Stadelman-Cohen TK, Hillman RE. Comparison of neck tension palpation rating systems with surface electromyographic and acoustic measures in vocal hyperfunction. *J Voice*. 2011;25(1):67-75.
4. Yiu EM, Lau VC, Ma EP, Chan KM, Barrett E. Reliability of laryngostroboscopic evaluation on lesion size and glottal configuration: a revisit. *Laryngoscope*. 2014;124(7):1638-1644.
5. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20(1):14-22.
6. Stepp CE, Hillman RE, Heaton JT. The impact of vocal hyperfunction on relative fundamental frequency during voicing offset and onset. *J Speech Lang Hear Res*. 2010;53(5):1220-1226.
7. Stepp CE, Merchant GR, Heaton JT, Hillman RE. Effects of voice therapy on relative fundamental frequency during voicing offset and onset in patients with vocal hyperfunction. *J Speech Lang Hear Res*. 2011;54(5):1260-1266.
8. Eadie TL, Stepp CE. Acoustic correlate of vocal effort in spasmodic dysphonia. *Ann Otol Rhinol Laryngol*. 2013;122(3):169-176.
9. Praat: doing phonetics by computer [Computer program] Version 5.3.042012.
10. Quatieri TF. *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall; 2001.
11. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Amn J Speech Lang Pathol*. 2009;18:124-132.
12. Lien Y-AS, Gattuccio CI, Stepp CE. Effects of phonetic context on relative fundamental frequency. *J Speech Lang Hear Res*. 2014;57(4):1259-1267.
13. Lien Y-AS. Optimization and automation of relative fundamental frequency for objective assessment of vocal hyperfunction. Doctoral dissertation, Boston University; 2015.
14. Smith AB, Robb MP. Factors underlying short-term fundamental frequency variation during vocal onset and offset. *Speech Lang Hear*. 2013;16(4):208-214.
15. Deliyski DD, Shaw HS, Evans MK. Influence of sampling rate on accuracy and reliability of acoustic voice analysis. *Logoped Phoniatr Vocol*. 2005;30:55-62.
16. McKenna VS, Heller Murray ES, Lien YAS, Stepp CE. The relationship between relative fundamental frequency and a kinematic estimate of laryngeal stiffness in healthy adults. *J Speech Lang Hear Res*. 2016;59(6):1283-1294.
17. Fujiki RB, Chapleau A, Sundarrajan A, McKenna V, Sivasankar MP. The interaction of surface hydration and vocal loading on voice measures. *J Voice*. 2017;31(2):211-217.
18. Heller Murray EH, Lien YAS, Michener CM, et al. Relative fundamental frequency distinguishes between non-phonotraumatic and phonotraumatic vocal hyperfunction. *J Speech Lang Hear Res*. 2017;60:1507-1515.