INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Changes in glottal source parameter values with light to moderate physical load

Heather Weston[1], *Laura L. Koenig*[2], *Susanne Fuchs*[1]

[1] Leibniz-Zentrum Allgemeine Sprachwissenschaft, Germany
[2]Adelphi University, USA
weston@leibniz-zas.de, koenig@haskins.yale.edu, fuchs@leibniz-zas.de

Abstract

Engaging in everyday physical activities, like walking, initiates physiological processes that also affect parts of the body used for speech. However, it is currently unclear to what extent such activities affect phonatory processes, and in turn, the voice. The present exploratory study investigates how selected glottal source parameters are affected by light and moderate physical activity. Recordings of sustained vowel /a/ were obtained from 39 female speakers of German at rest, and during low-intensity and moderate-intensity cycling. Ten glottal source parameters thought to reflect different physiological states were investigated using VoiceSauce. Even during light activity, significant increases were found in f0, strength of excitation and H1, and a decrease in harmonics-to-noise ratio at higher frequencies. During moderate-intensity activity, significant effects were stronger and found for most parameters. However, considerable intra- and interspeaker variability was observed. These findings may be relevant for applications in automatic speaker-state recognition. They also underscore the importance of investigating individual-level responses to better understand stress–voice interactions.

Index Terms: human speech production, automatic stress recognition, physical task stress, voice quality

1. Introduction

People frequently speak while doing things and going places, especially since the advent of cell phone technology. However, it is not fully clear how such physical activity affects the voice. Even everyday activities, such as brisk walking, initiate physiological adaptations that also affect parts of the body used for speech, particularly the respiratory system. These changes may affect the laryngeal mechanism and, in turn, acoustic parameters of the speech signal. The present exploratory study investigates how selected glottal source parameters are affected by different levels of physical activity.

For phonation to occur, the vocal folds must be adducted and a transglottal pressure difference must be present [1]. When either or both of these processes are perturbed, as is thought to be the case during physical activity, certain changes in the behavior of the vocal folds can be predicted to occur. The most widely obtained finding is that fundamental frequency (f0), perceived as pitch, increases when individuals cope with diverse types of stressors, including physical load [reviews: 2, 3]. This is thought to be due to increased muscle activation, which may increase tension in the vocal folds, and increased respiratory drive, which could increase airflow through the glottis. These physiological changes would also affect other aspects of the glottal source, or "voice quality."

In a broad sense, voice quality refers to certain perceptual characteristics of the voice, such as "breathy" or "hoarse." In this investigation, we understand voice quality in the narrow sense, pertaining to the behavior of the vocal folds. Because the vocal folds vibrate across time and space, voice quality is inherently multidimensional, for both speakers and listeners [e.g., 4, 5]. For that reason, it is important to investigate multiple glottal source parameters, as they are thought to reflect different aspects of glottal settings [6, 7].

Few studies have investigated changes in glottal source parameters during light, everyday physical activity, such as brisk walking. The present exploratory study is led by the theoretical hypothesis that physical activity affects laryngeal valving, and that this in turn may affect glottal source parameters. The general hypothesis is that a given glottal source parameter will change when an individual engages in light/moderate physical activity. To investigate this, 10 acoustic parameters were chosen that are thought to reflect different physiological phenomena:

- **Harmonics-to-noise ratio** (HNR) quantifies additive noise in the voice signal, and may reflect variation in vocal fold length and tension as well as the degree of adduction, and turbulent noise in the glottis
- **H1*-H2***, a measure of spectral tilt, is related to the proportion of the glottal cycle in which the glottis is open, which may in turn relate to phonatory effort and vocal fold thickness [8]
- **H1*** has been proposed as an acoustic correlate of glottal constriction with less variance than H1*-H2* [9]
- **H1*-A1*** and **H1*-A3*** capture increases in noise energy in lower and higher frequencies, respectively, that would arise from turbulence generated by a posterior glottal chink [6]
- **Strength of excitation** (SoE) quantifies the abruptness of the closing phase in the glottal cycle [10]
- **Fundamental frequency** (f0) reflects the frequency at which the vocal folds vibrate, which corresponds to vocal pitch and varies with the tension, mass/thickness and/or length of vocal folds [11]

Sustained vowels were chosen as speech material. Commonly investigated in clinical settings to assess voice quality, sustained vowels are free from prosodic influences and coarticulation effects and thus provide a foundation to better understand the general mechanism of glottal changes during speech during physical activity. The results may be relevant for the voice-based detection of speaker states. While researchers are developing automatic means of detecting, e.g., stress at work [e.g., 12] or vocal pathologies [review: 13], it is still necessary to understand the vocal effects of other types of stressors, like physical load, and their potential differences.

http://dx.doi.org/10.21437/Interspeech.2021-1881

2. Methods

2.1. Participants

A total of 48 female native speakers of standard German participated in an experiment with three speech tasks. Speakers were recruited via a study database and paid for participation. All gave informed consent; none reported any hearing, speech or breathing pathologies; none were smokers.

The present study analyzes 39 participants (19–34 years of age; $\bar{x} = 23.5$). Nine participants were excluded from analysis because they did not produce sustained vowels at a consistent pitch after repeated instruction and demonstration prior to and during the experiment.

2.2. Stimuli and equipment

The point vowels /a, i, u/ were presented in randomized order three times each per condition. Each vowel was embedded in the carrier phrase *sie sagt* ('she says VOWEL'). Vowels were sustained for 2–3s.

The exercise task was performed on a low-noise ergometer (daum electronic, Germany) in a lab setting (Fig. 1) at room temperature (20–22°C). Speech was recorded at a sampling rate of 22050 Hz using a head-mounted microphone (beyerdynamic, Germany) placed 4cm from the corner of the mouth at a 90° angle.

Figure 1: *Laboratory set-up: presentation screen at eye level (A); low-noise ergometer (B); head-mounted microphone with battery pack on mic stand (C).*

2.3. Design and procedure

2.3.1. Experimental design

A within-participant design was used with three conditions: CONTROL (sitting still), LIGHT intensity activity and MODERATE intensity activity. The order of the conditions was fixed (CONTROL > LIGHT > MODERATE) because the physiological response to exercise persists for some time following activity; thus, randomizing condition order would make it difficult to ensure that CONTROL truly measured baseline values.

2.3.2. Calculating exercise intensity as target heart rate

Exercise intensity refers to the physical effort required to perform a given activity. In line with standard practice, we defined intensity as a percentage of maximal heart rate (HR_{max}): LIGHT = 35% and MODERATE = 65%. These values reflect guidance provided by health organizations [e.g., 14]. Target heart rates were calculated for each participant using a standard method in sports science, the Karvonen formula, given in (1). This formula uses age and resting pulse (an indication of physical fitness) to calculate an individual's target heart rate for a given exercise intensity. This step was necessary to make the physical effort required in the experimental conditions comparable across participants, regardless of their age or level of fitness.

$$\text{target HR} = [(HR_{max} - HR_{rest}) \times \%\ \text{intensity}] + HR_{rest} \quad (1)$$

Following [15], HR_{max} was predicted as: $208 - (0.7 \times \text{age})$. HR_{rest} was estimated by having each participant lie down for 10 minutes and taking the average heart rate measured over minute 11 using a wrist-worn heart rate monitor (Scosche Rhythm), which was worn throughout the experiment. To validate that the calculated heart rates reflected the intended exercise intensities, participants were asked to rate their level of perceived exertion using the Borg scale [16]. The average rating for the LIGHT condition was 9.7 ("very light") and for the MODERATE condition 13.8 ("somewhat hard").

2.3.3. Procedure

Vowels were presented one at a time on a monitor at eye level. Participants were instructed to read the carrier phrase and sustain the vowel until the screen went blank. The CONTROL condition was followed by a 4-minute cycling warm-up to reach target HR for the LIGHT condition. The procedure was repeated for the MODERATE condition. Heart rate was monitored using a tablet connected to the wrist-worn monitor and modified, if needed, by adjusting resistance on the bike between trials.

2.4. Parameter extraction and statistics

Only vowel /a/ is analyzed here. A total of 349 tokens were analyzed (39 participants × 3 conditions × 3 trials = 351; two mispronounced tokens were excluded).

Vowels were delimited manually in the PRAAT program [17]. Vowel onset and offset were defined using the second formant [18]. F2 was determined to be present at "the time of onset of the first vertical striation extending upward through the frequency regions of the first and second formants without interruption" [19].

A PRAAT script was used to extract a 200ms section of the vowel for analysis. The section was extracted 500ms after vowel onset to obtain an early-production sample of the vowel. A fixed point for extraction was chosen as an alternative to a percentage of total duration to allow for repeatability and consistency across speakers [20].

VoiceSauce [21] was used to extract the glottal source parameters. The audio files were downsampled to 16 kHz. The STRAIGHT algorithm [22] was used to estimate f0 (range: 100–400Hz) and the Snack Sound Toolkit [23] was used to estimate formants (maximum value: 5500 Hz; 5 formants). VoiceSauce records measurements at 1ms intervals for all parameters except for strength of excitation (5ms intervals), but because each series of observations is indexed in time, adjacent observations may be correlated and thus cannot be considered independent. Consequently, the mean of all observations was calculated for each vowel interval.

3. Results

3.1. Preliminary data exploration

Assumptions for linear regression models were checked for each parameter. Normality was assessed visually using histograms, quantile-quantile plots and box plots, and objectively with a Shapiro-Wilk test. The data generally showed a normal distribution. Homoscedasticity was checked with Levene's test and was found for all parameters.

3.2. Linear mixed model: group results

Statistical analyses were run in R (Version 4.0.4) to assess differences in baseline and response to physical load. Using the packages *lmerTest* (Version 3.1-3) and *lme4* (Version 1.1-26), a linear mixed model was fitted for each parameter, with condition as the independent variable (three levels: CONTROL, LIGHT, MODERATE; control as reference level), the outcome for the given parameter as the dependent variable, and random slopes and intercepts for speakers (*lmer* syntax: parameter ~ condition + (1 + condition | speaker), data). The residuals were plotted and checked for normality using the methods in 3.1.

The linear mixed models tested the general null hypothesis that there is no change in the parameter during light/moderate physical activity. For f0, an increase was predicted during light and moderate physical activity. Table 1 summarizes the results for each parameter. Data are plotted in Figure 2.

Table 1: *Linear mixed effects for each parameter; bold face indicates statistical significance.*

parameter	CONTROL–LIGHT		CONTROL–MODERATE	
f0	t = 4.00	***p* < .001**	t = 10.61	***p* < .001**
SoE	t = 2.50	***p* = .0167**	t = 7.65	***p* < .001**
H1*	t = 2.89	***p* = .006**	t = 4.71	***p* < .001**
H1*-H2*	t = -.17	*p* = .86	t = -.57	*p* = .57
H1*-A1*	t = -.08	*p* = .93	t = .44	*p* = .66
H1*-A3*	t = -1.46	*p* = .15	t = -2.37	***p* = .023**
HNR05	t = -2.69	*p* = .10	t = -2.76	***p* = .009**
HNR15	t = -2.38	***p* = .0226**	t = -4.51	***p* < .001**
HNR25	t = -2.84	***p* = .0073**	t = -5.36	***p* < .001**
HNR35	t = -4.01	***p* < .001**	t = -6.54	***p* < .001**

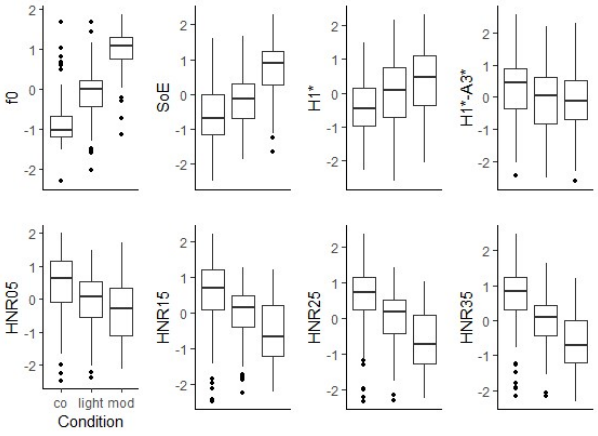

Figure 2: *Parameters showing significant changes during physical activity (z-scores).*

In the LIGHT and MODERATE activity conditions, there were strong significant effects ($p < 0.001$) for f0 and HNR35, showing an increase and decrease, respectively. In the MODERATE condition, there were also strong significant effects for SoE and H1*, which increased, and HNR15 and HNR25, which decreased. The HNR parameters generally showed stronger effects with larger frequency windows. No significant changes were found for H1*-H2* and H1*-A1*.

3.3. Descriptive remarks: speaker differences

In addition to the group patterns, we saw some differences between speakers. Figure 3 shows data from six speakers (anonymized with three-letter strings) for the parameter f0. Speakers were chosen to illustrate variation phenomena and are not representative of the entire dataset. The violin plots show all observations across the 200ms intervals; the mean for each repetition is shown with a black dot. Three main types of variation were observed:

1. Between-speaker variation at baseline (e.g., **ite** vs. **jwi**)
2. Within-speaker variation in the f0 of tokens produced within a single condition (**brx**, MODERATE; **rvh**, **vnu**, LIGHT)
3. Between-speaker variation in response to physical activity, whereby the parameter value increased (**brx**, **hox**, **vnu**), did not change (**ite**, **jwi**) or decreased (**rvh**)

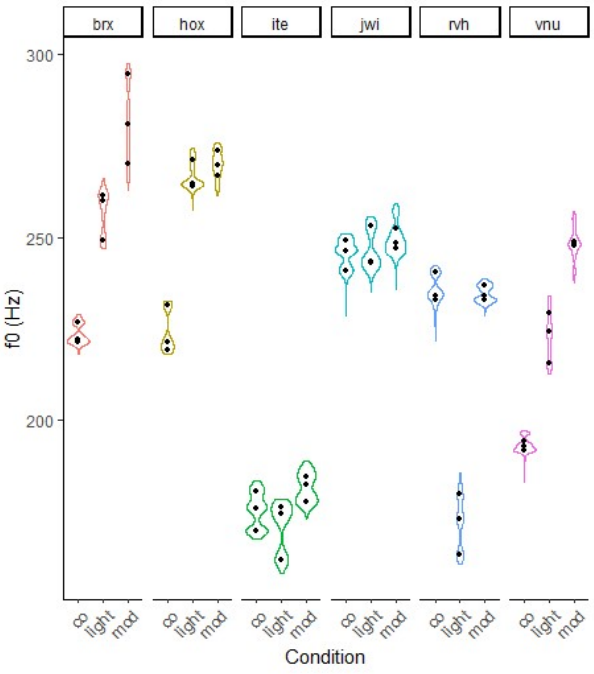

Figure 3: *Observations for six speakers showing between- and within-speaker variability.*

Some within-speaker variation was observed in eight speakers (20%), who produced tokens of sustained /a/ that differed in f0 by more than 20 Hz in one condition only (e.g., **brx**). Four speakers did this in the MODERATE condition, two in the LIGHT condition and two in the CONTROL condition.

Between-speaker variation was observed in how speakers responded to physical load, particularly in the LIGHT condition. Figure 4 shows changes in f0 for all speakers. While the majority (67%) showed an increase in f0 during light activity,

10 speakers (25%) showed no change (<5 Hz) and 3 showed a sharp decrease. This contrasts with the moderate condition, in which almost all speakers showed an increase compared to baseline.

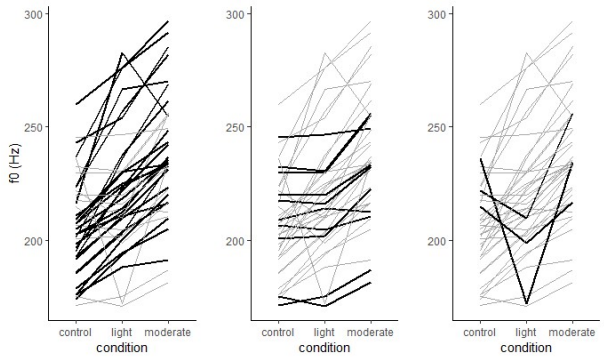

Figure 4: *Between-speaker variation in change in f0 during light physical activity;* from left*: increase, no change (<5Hz), decrease.*

4. Discussion

We found significant changes in several glottal source parameters during light- and moderate-intensity physical activity. Fundamental frequency was found to increase, even during light activity. A higher frequency of vibration of the vocal folds suggests a change in the mass/thickness, length and/or tension of the vocal folds. It may also reflect higher driving pressure, arising from the greater respiratory effort under physical load. A significant increase in the strength of excitation (SoE) measure was also found. This is thought to reflect higher vocal fold contact and may also suggest that the vocal folds are closing more forcefully on each cycle. At the same time, an increase in H1* was found, which is thought to reflect greater glottal opening. While this may seem to contradict greater SoE, it is possible that there is a greater chink under effort while at the same time there is greater vocal-fold contact along midline. Though further investigation is needed, these results demonstrate the utility of using multiple measures to capture the complexity of how vocal fold characteristics may change under physical load.

A significant decrease was found for the harmonics-to-noise ratio, though the added noise did not appear to cluster in any particular frequency regions. A decrease in HNR reflects a greater degree of aperiodicity in the speech waveform and suggests phonatory perturbation, or possibly less stable control over vocal fold length and tension. However, perceptual assessments of several speakers with the greatest decreases under physical load did not yield the impression that they were breathier during physical activity. Further investigation is needed. The decrease in the harmonics-to-noise ratio is difficult to interpret, because multiple laryngeal configurations correspond with intraharmonic noise.

Taken together, these results may be interpreted in the following ways: during light physical activity, the vocal folds may be generally stiffened and/or subglottal pressure heightened, which increases glottal excitation and fundamental frequency. Initially, the folds may simply oscillate faster, but the overall glottal cycle with its opening and closing phases does not change. However, as physical load intensifies, there may be further changes to the shape of motion (e.g., opening quotient). These changes may be a reaction to physical load or may somehow facilitate phonation in the body's altered physiological state, for example, when there is increased competition for the work of the lungs.

Apart from these general results, we found between-speaker variability in almost all parameters, though of course, differences in f0 were also reflected in the harmonics-related measures. The differences between speakers suggest that individuals react differently to physical load. Speaking while engaging in activity requires finely tuned coordination of the motor, respiratory, articulatory and laryngeal systems. Differences in physiology, fitness, or even personality traits (e.g., ability to endure discomfort, such as dyspnea) may make this coordination more or less economical in individuals.

These findings may be relevant for voice-based applications that automatically recognize speaker states, especially via mobile communications systems: the acoustic parameters used to detect, e.g., stress levels at work may also be affected by physical load. Thus, investigating the acoustic consequences of everyday physical load can contribute to a better understanding of the potential differences between diverse load types that affect the voice.

One limitation of the study is the focus on the low vowel /a/, which does not allow us to generalize to other vowels or contexts. In a further analysis, we will thus investigate the other point vowels as well as vowels extracted from connected speech in the same corpus. While sustained vowels are widely used in linguistic and clinical studies, speakers typically do not produce them in daily life. It has been observed that in healthy speakers f0 measurements taken from connected speech may be more consistent than those from sustained vowels [24]. Comparison of data from different speech contexts is thus also of methodological interest.

5. Conclusion

Significant differences in f0, SoE, H1* and HNR were found with light physical load, at an intensity equivalent to brisk walking. This suggests that acoustic differences could be present during everyday activities. Moderate physical load had a larger effect on all parameters, suggesting greater changes to laryngeal and respiratory behavior.

Some variation was observed, both between and within speakers. Speakers differed in their baseline values but also in their response to physical load, with some participants showing opposite patterns. Variation between speakers was greatest in response to light load. The response to moderate activity was generally consistent across participants. Within-speaker variation was observed in about 20% of speakers, with f0 varying by 20–45 Hz in their tokens of sustained /a/ in a single condition, mostly under physical load. It is not clear whether this variation reflects a speaker-specific response to physical load or to the task of producing a sustained vowel. In any case, investigating individual differences may prove useful in better understanding the ways in which stress–voice interactions can be manifested across the population.

6. Acknowledgements

This research was supported by the French National Research Agency and the German Research Foundation as part of the SALAMMBO project. The authors thank Jörg Dreyer for help with data collection, Hélène Serré for discussions on statistics, and three anonymous reviewers for their helpful comments.

7. References

[1] J. van den Berg, "Myoelastic-aerodynamic theory of voice production," *Journal of Speech and Hearing Research*, vol. 1, no. 3, pp. 227–244, Sep. 1958.

[2] C. Kirchhübel, D. M. Howard, and A. W. Stedmon, "Acoustic correlates of speech when under stress: research, methods and future directions," *International Journal of Speech, Language and the Law*, vol. 18, no. 1, pp. 75–98, 2011.

[3] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice stress analysis: a new framework for voice and effort in human performance," *Frontiers in Psychology*, vol. 9, 1994, Nov. 2018.

[4] A. Lovato, M. R. Barillari, L. Giacomelli, L. Gamberini, and C. de Filippis, "Predicting the outcome of unilateral vocal fold paralysis: a multivariate discriminating model including grade of dysphonia, jitter, shimmer, and Voice Handicap Index-10," *Annals of Otology, Rhinology & Laryngology*, vol. 128, no. 5, pp. 447–452, May 2019.

[5] J. Kreiman, B. R. Gerratt, and G. S. Berke, "The multidimensional nature of pathological voice quality," *The Journal of the Acoustical Society of America*, vol. 96, no. 3, pp. 1291–1302, Sep. 1994.

[6] H. Hanson, "Glottal characteristics of female speakers: acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, Jan. 1997.

[7] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among male and female talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–856, Feb. 1990.

[8] Z. Zhang, "Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1493–1507, Apr. 2016.

[9] Y. Chai and M. Garellek, "Using H1 instead of H1–H2 as an acoustic correlate of glottal constriction," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, p. 3008, Oct. 2019.

[10] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp.1602–1613, Nov. 2008.

[11] I. R. Titze, "Physiologic and acoustic differences between male and female voices," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1699–1707, Apr. 1989.

[12] A. König, K. Riviere, N. Linz, H. Lindsay, J. Elbaum, R. Fabre, A. Derreumaux, and P. Robert, "Measuring stress in health professionals over the phone using automatic speech analysis during the COVID-19 pandemic: observational pilot study," *Journal of Medical Internet Research*, vol. 23, no. 4, e24191, Apr. 2021.

[13] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control*, vol. 51, pp. 181–199, May 2019.

[14] Physical Activity Guidelines Advisory Committee, *Physical Activity Guidelines Advisory Committee Report, 2008*. Washington, DC: U.S. Department of Health and Human Services, p. D-3, 2008, retrieved from https://health.gov/sites/default/files/2019-10/ CommitteeReport_7.pdf, 10 Aug. 2020.

[15] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *Journal of the American College of Cardiology*, vol. 37, no. 1, pp. 153–156, Jan. 2001.

[16] G. A. Borg, "Psychophysical bases of perceived exertion," *Medicine and Science in Sports and Exercise*, vol. 14, no. 5, pp. 377–381, 1982.

[17] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.1.39, retrieved from http://www.praat.org/, 2 Feb 2021.

[18] E. Fischer-Jørgensen and B. Hutters, "Aspirated stop consonants before low vowels, a problem of delimitation – its causes and consequences," *Annual Report of the Institute of Phonetics at the University of Copenhagen*, vol. 15, pp. 77–102, 1981.

[19] A. L. Francis, V. Ciocca, and J. M. C. Yu, "Accuracy and variability of acoustic measures of voicing onset," *The Journal of the Acoustical Society of America,* vol. 113, no. 2, pp. 1025–1032, Feb. 2003.

[20] M. Blomgren and M. Robb, "How steady are vowel steady-states?," *Clinical Linguistics & Phonetics*, vol. 12, no. 5, pp. 405–415, 1998.

[21] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences,* Hong Kong, China, Aug. 2011, pp. 1846–1849.

[22] H. Kawahara, A. de Cheveigné, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite," *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, Nov./Dec. 1998, paper 0659.

[23] K. Sjölander, Snack sound toolkit, KTH Stockholm, Sweden, online at http://www.speech.kth.se/snack, 2004.

[24] J. Fitch, "Consistency of fundamental frequency and perturbation in repeated phonations of sustained vowels, reading, and connected speech," *Journal of Speech and Hearing Disorders*, vol. 55, pp. 360–363, May 1990.