Vowel Context and Speaker Interactions Influencing Glottal Open Quotient and Formant Frequency Shifts in Physical Task Stress

Keith W. Godin, John H.L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas, Richardson, TX, USA

godin@ieee.org, john.hansen@utdallas.edu

Abstract

Physical task stress is known to affect the fundamental frequency of speech. This study of two American English vowels /IY/ and /AH/ investigates whether physical task stress affects the center frequencies of formants F1 and F2, and whether it affects the glottal open quotient, and whether these effects are different for different speakers, the different vowels, and two different vowel contexts. Formant center frequencies are measured from the acoustic waveform, and the glottal open quotient is measured from the electroglottograph signal. The study finds in general that the production of vowels is affected by physical task stress. In particular, the study finds that F1, F2, and the glottal open quotient are affected by physical task stress. It also finds that the effects of stress on F1 vary for different speakers, and that the effects of stress on the glottal open quotient vary for different combinations of speaker and vowel.

Index Terms: physical task stress, open quotient, electroglottograph

1. Introduction

Stressors, tasks, emotions, environmental noise, and fatigue affect the production of speech and the resulting acoustic wave, and research is active in these areas. Physical exercise, known in the literature as physical task stress, occurs in a variety of real life situations, and the literature notes the problems caused by such stress on speech and speaker recognition systems [1, 2]. It is of interest that physical task stress, unlike most types of speech variability, such as emotions, is strongly correlated with several physiological variables [3]. The study of the effects of physical task stress on the acoustic speech wave and the speech production process are not only of interest to more fully understand the breadth of human speech variability, but also to stimulate development of new approaches to the design of robust speech systems.

Previous studies have confirmed that the effects of physical task stress on speech production are more than durational and fundamental frequency effects, and more than added breaths in non-speech regions. The extent of the effects of physical task stress on speech has not been established, but three studies [4, 5, 6] have examined these effects. Several acoustic parameters closely related to vocal fold behavior have been studied; for example Johannes et. al. [4] observed fundamental frequency (F0) increases were observed for all speakers, and Godin and Hansen [5] observed such changes for most speakers. Johannes

et. al. [4] found in particular that F0 increases were associated with increases in stress level, but found the relationship non-linear, characterized by wide, disjoint plateaus. Further, Godin and Hansen [5] observed an increased F0 for 60% of speakers, a decrease in the proportion of frames voiced in an utterance for 88% of speakers, and an increase in glottal open quotient, as measured by an algorithm employing inverse filtering, for 47% of speakers and a decrease for 27% of speakers.

In examining physical task stress speech with a broader, spectral measure of speech dissimilarity, Godin and Hansen [6] suggested that different phone classes are affected differently by physical task stress. It was observed in that study that the average short-time spectrum of nasal phones was more affected than the average spectra of plosives and fricatives. Their results also suggest the possibility that physical task stress has a greater impact on the production of high vowels than on low vowels.

The present study is concerned with examining the effects of physical task stress on the production of vowels. This study characterizes the vowel production process by the first and second formants and the glottal open quotient. The research questions for this study are first, whether physical task stress affects these parameters, and second, whether physical task stress interacts with speaker identity, vowel height, and vowel context to influence the vowel production process. Rotstein et. al. [7] suggested that changes in speech production processes may not be associated with any particular level of exertion across speakers, thus, speakers respond to stress differently. An interaction between speaker and task in influencing formants and open quotient would support this finding and provide a complementary perspective on the way that speakers might differ in their responses to stress, in addition to the evidence already provided by [5]. Motivated by the findings of [6], this study includes one high vowel and one low vowel for comparison. Significant interaction between task and vowel identity would suggest that different vowels are affected differently by physical task stress.

Average formant frequencies for a given vowel are primarily determined by the fixed factors of an individual vocal tract, and by the speaker's habits of articulator placement. Formant frequencies vary around these averages due to variations in articulator placement, and to a lesser extent due to factors such as acoustic coupling with the sub-glottal system [8] which may be related to variations in the glottal open quotient [9]. Thus, the glottal open quotient is of interest not only as a way to characterize vocal fold behavior, but as a possible causal factor in formant shifts in physical task stress. In regards to the glottal open quotient, this study is primarily concerned with establishing whether the glottal open quotient is affected by physical task stress.

Previous studies of physical task stress speech have relied

28 – 31 August 2011, Florence, Italy

solely on the study of the acoustic speech wave and on physiological measures, such as heart rate, unrelated to the speech production process. This study uses an electroglottograph (EGG) as the source for the measure of the glottal open quotient (OQ), providing a non-acoustic perspective on the effects of physical task stress on speech production. The nature of errors in acoustic algorithms that occur due to physical task stress speech has not been established for any known algorithm, and so such an alternative perspective can help to play a confirmatory role in a study, in that the character of errors of measurements drawn from the EGG signal is probably affected by physical task stress in a different way than measurements from the acoustic speech signal.

A new corpus of physical task stress speech was collected for this study. The remainder of this paper presents the data collection process, the measurement and analysis process, measurement results, and finally the conclusions and future work.

2. Speech data collection

The data for this study is collected in a similar manner to UT-Scope [10] which was used in [5] and [6]. This study adds an electroglottograph to the data collection process, as well as vowel-consonant-vowel (VCV) and consonant-vowel (CV) utterances.

Data collection was performed in an ASHA-certified single walled soundbooth. Speech data was recorded from a Shure Beta 53 head worn close-talking microphone. The electroglottograph used was model EG2-PCX from Glottal Enterprises. A Polar S520 heart rate watch with chest-worn sensor was used to record heart rate during both the neutral and physical task stress segments. Acoustic and EGG signals were recorded to a Fostex D824 digital recorder at 44.1kHz, 16 bits/sample, and downsampled to 16kHz for processing. The physical task stress is induced by a Stamina Conversion II Elliptical/Stepper machine in the elliptical mode. Subjects were instructed to hold an approximately 10mph pace. This constant work load resulted in different levels of exertion for each speaker, because each speaker had a different level of physical fitness. 2 male participants and 2 female participants were recorded.

35 TIMIT sentences, 8 vowel-consonant-vowel (VCV), and 8 consonant-vowel (CV) utterances were prompted through headphones. The consonants were two stops /T/ and /D/, and two nasals /M/ and /N/. The two vowels were a high vowel /IY/ and a low vowel /AH/. Each consonant was paired with each vowel to create the CV utterances. Each VCV utterance then had the same vowel in both places, i.e. /IYDIY/ was one of the VCV utterances but /IYDAH/ was not. The 8 VCV and CV utterances were placed in a random order that was fixed for all speakers, and this order was repeated 5 times by each speaker both while seated in the neutral task, and during the physical task stress task.

Initial phone segmentations for the recordings were found using forced alignment, and the resulting alignments were hand corrected by an experienced transcriptionist. The remainder of this study used only the middle 80% of frames of vowels from the VCV and CV utterances formed using the stop consonants, as the drawn out final vowel prompted in the collection will provide a more stable region from which to make the formant and open quotient measurements than the connected speech. Only the final vowel of each VCV utterance was used for analysis. Therefore because there were 5 instances of /TIY/ in neutral by a given speaker and 5 instances of /IYTIY/, there were 10 instances of /IY/ in /T/ context used in the following experiment,

Table 1: MANOVA with speaker, vowel, preceding consonant, and task as factors, and F1, F2, and OQ as response variables. $\alpha = 0.05$. The four-way interaction was omitted when an initial MANOVA indicated insignificance with $p = 0.9981$.

Factor	Approx. F	p	Sig.
Speaker	254.0	$< 2.200 * 10^{-16}$	✓
Vowel	9928.9	$< 2.200 * 10^{-16}$	✓
Consonant	11.5	$4.266 * 10^{-07}$	✓
Task	7.2	0.0001258	✓
Spkr:Vowel	84.3	$< 2.200 * 10^{-16}$	✓
Spkr:Cons.	2.1	0.0284929	✓
Vowel:Cons.	10.2	$2.465 * 10^{-06}$	✓
Spkr:Task	2.8	0.0034863	✓
Vowel:Task	2.1	0.1062017	
Cons:Task	1.2	0.3285535	
Spkr:Vowel:Cons.	2.3	0.0171800	✓
Spkr:Vowel:Task	1.9	0.0435645	✓
Spkr:Cons.:Task	1.6	0.1068049	
Vowel:Cons.:Task	0.6	0.5993058	

per speaker, per task, as well as 10 instances of /IY/ in /D/ context, 10 instances of /AH/ in /T/ context, and 10 instances of /AH/ in /D/ context. With 2 tasks and 4 speakers, this results in 320 recorded vowel utterances used in this study.

3. Parameter measurement methods

For the analysis, each vowel instance is broken into 20ms frames with 10ms overlap, and the middle 80% of frames are selected for inclusion in the analysis. The first and second formants and the glottal open quotient are measured from each frame, and averaged across the frames to form one summary statistic of each measurement for each vowel instance. WaveSurfer [11] is used to measure the formants. The default settings are sufficient to take the measurements.

The DECOM algorithm [12] is used to measure glottal open quotient from the EGG signal. The parameter values of the DECOM algorithm are set differently than those in original algorithm description, in order to cope with the particulars of this data set. It was informally observed that double closing peaks in the DEGG signal occurred more often in physical task stress, and that when these occurred in the vicinity of changing F0, which also occurred more often in physical task stress, the algorithm chose peaks from the autocorrelation of very low lag, resulting in a very high estimated fundamental frequency. This was solved by setting the D parameter to 1, rather than 0.5 as suggested in the algorithm description. Also, the DECOM description specifies a pitch synchronous analysis, but instead are used fixed frames of 20ms length and 10ms skip, corresponding to the frames from the acoustic analysis.

The DECOM algorithm specification stipulates that a glottal open quotient measurement cannot be reliably made when double DEGG closing or opening peaks are detected [12], because the doubled peaks introduce ambiguity as to the instant of glottal closure or opening. This strength of the algorithm — specifying when unreliable measurements are likely — can turn into a weakness when doubled closing peaks occur during an entire utterance and an OQ measurement cannot be made. For these cases the vowel was simply dropped from the analysis. This resulted in a total of 274 vowels for analysis.

Table 2: ANOVA with F1 as response variable, and speaker, vowel, preceding consonant, and task as factors. A Bonferroni correction to a nominal $\alpha = 0.017$ is applied to achieve a true $\alpha = 0.05$ across the three ANOVAs in this study. The four-way interaction was omitted when an initial ANOVA indicated insignificance with $p = 0.782561$.

Factor	F	p	Sig.
Speaker	425.417	$< 2.200 * 10^{-16}$	✓
Vowel	$> 10^3$	$< 2.200 * 10^{-16}$	✓
Consonant	33.3789	$2.293 * 10^{-08}$	✓
Task	3.8617	0.050531	
Spkr:Vowel	90.7492	$< 2.200 * 10^{-16}$	✓
Spkr:Cons.	3.7960	0.010922	✓
Vowel:Cons.	29.6113	$1.279 * 10^{-07}$	✓
Spkr:Task	4.6925	0.003319	✓
Vowel:Task	0.5461	0.460616	
Cons.:Task	0.5142	0.474025	
Spkr:Vowel:Cons	4.6288	0.003613	✓
Spkr:Vowel:Task	1.8393	0.140575	
Spkr:Cons.:Task	0.1499	0.929684	
Vowel:Cons.:Task	1.2016	0.274086	

Table 3: ANOVA with F2 as response variable, and with speaker, vowel, preceding consonant, and task as factors. A Bonferroni correction to a nominal $\alpha = 0.017$ is applied to achieve a true $\alpha = 0.05$ across the three ANOVAs in this study. The four-way interaction was omitted when an initial ANOVA indicated insignificance with $p = 0.9651$.

Factor	F	p	Sig.
Speaker	865.7	$< 2.2 * 10^{-16}$	✓
Vowel	$> 10^3$	$< 2.2 * 10^{-16}$	✓
Consonant	0.2826	0.5955103	
Task	12.1290	0.0005878	✓
Spkr:Vowel	239.284	$< 2.2 * 10^{-16}$	✓
Spkr:Cons.	2.4818	0.0615530	
Vowel:Cons.	0.9731	0.3248803	
Spkr:Task	0.6947	0.5560811	
Vowel:Task	1.5859	0.2091125	
Cons.:Task	2.6008	0.1080939	
Spkr:Vowel:Cons.	1.8300	0.1422366	
Spkr:Vowel:Task	0.8182	0.4848991	
Spkr:Cons:Task	2.3381	0.0741706	
Vowel:Cons:Task	0.0304	0.8616865	

4. Results

The data set of 274 vowels is analyzed with a 4 way MANOVA with F1, F2, and OQ as response variables and speaker, task, vowel, and preceding consonant as factors. The results of the MANOVA are shown in Table 1. The interaction between speaker, vowel, and consonant is significant. This expected result that speakers produce different vowels differently depending on the vowel and on the preceding consonant, suggests that the data collection and measurement process has correctly characterized the vowel production process. The interaction between speaker, vowel, and task is also significant. This suggests that physical task stress affects the production of vowels differently depending on the speaker and the vowel. On the other hand, these results do not suggest that physical task stress affects the production of vowels differently depending on the preceding consonant.

Three factorial ANOVAs are run with the same four factors, one for each of the response variables F1, F2, and OQ. To ensure an overall alpha of $\alpha = 0.05$ for the three ANOVAs in this study, a Bonferroni correction is applied, resulting in a nominal alpha of $\alpha = 0.017$. Table 2 shows the ANOVA for the first formant (F1). The interaction between speaker, vowel and consonant is significant. This implies that the change in frequency of the first formant from speaker to speaker depends on the vowel and on the preceding consonant. As with the MANOVA over the three response variables, this interaction is common knowledge and verifies to some extent the success of the formant center frequency measurements made in this study. The interactions between task and vowel and between task and preceding consonant are not significant. Thus we cannot reject the null hypothesis that stress affects the first formant for each vowel the same way, and cannot reject the null hypothesis that stress affects the first formant the same way for each preceding consonant. However, the interaction between speaker and task is significant. This suggests that physical task stress affects the first formant differently for each speaker.

The interaction between speaker and task is probably dominated by significant changes having different sign for different speakers. The first formant increased in physical task stress by between 7-15Hz for the two female speakers, and by 20Hz for one male speaker. The first formant of the other male speaker instead decreased by 20Hz.

The results of the ANOVA for the second formant are shown in Table 3. Again, a Bonferroni correction was applied to the α. There were no significant three way interactions, though the p value for the interaction between speaker, consonant, and task is near significance, suggesting further study is needed to investigate this finding. The interaction between speaker and vowel is significant, an expected result, as different speakers are known to have different second formants for different vowels. Task is a significant main effect. Thus, while we cannot reject our null hypothesis that physical task stress affects the second formant the same way across all speakers and all vowels, the data suggests that physical task stress affects the second formant.

The results of the ANOVA for the open quotient (OQ) are shown in Table 4. The interaction between speaker, vowel, and task is significant. This suggests that different speakers respond to physical task stress in a different way by changing their average open quotient differently for each vowel. For example, informally looking at the means, without performing post-hoc pairwise significance testing, the average OQ for speaker F001 appeared to decrease in physical task stress for both vowels /AH/ and /IY/, while the average OQ for speaker M003 appeared to decrease in physical task stress for vowel /AH/ and increase in physical task stress for vowel /IY/.

5. Conclusions

The results support the conclusion that the production of vowels is affected by physical task stress. Three pieces of evidence — the interaction between speaker and task influencing the first formant, task as a significant main effect influencing the second formant, and the interaction between speaker, vowel, and task influencing the open quotient — support the conclusion that physical task stress affects the first formant, second formant, and the open quotient in the production of vowels.

The interaction between speaker and task influencing the first formant, and between speaker, vowel, and task influencing

Table 4: ANOVA with open quotient (OQ) as response variable, and with speaker, vowel, preceding consonant, and task as factors. A Bonferroni correction to a nominal $\alpha = 0.017$ is applied to achieve a true $\alpha = 0.05$ across the three ANOVAs in this study. The four-way interaction was omitted when an initial ANOVA indicated insignificance with $p = 0.9904$.

Factor	F	p	Sig.
Speaker	220.426	$< 2.200 * 10^{-16}$	✓
Vowel	49.1723	$2.271 * 10^{-11}$	✓
Consonant	0.0048	0.94463	
Task	3.0049	0.08427	
Spkr:Vowel	4.0066	0.00826	✓
Spkr:Cons.	0.6685	0.57210	
Vowel:Cons.	1.1754	0.27936	
Spkr:Task	2.6399	0.05010	
Vowel:Task	4.4747	0.03541	
Cons:Task	0.1149	0.73497	
Spkr:Vowel:Cons.	1.0301	0.37988	
Spkr:Vowel:Task	3.5102	0.01595	✓
Spkr:Cons.:Task	2.7085	0.04581	
Vowel:Cons.:Task	0.6588	0.41776	

the open quotient, suggests that different speakers respond differently to physical task stress. This is new, detailed evidence of this phenomenon and is in line with the evidence from [7] and [5] that speakers differ in their responses to physical task stress in a variety of ways.

Finally, the interaction between speaker, vowel, and task influencing the open quotient implies that the open quotient of different vowels are affected differently by physical task stress. While the results do not support rejecting the null hypothesis that the formants of different vowels are affected differently by physical task stress, low p values in this study suggest that there may be an effect present that is small, or the effect could be related more strongly to factors not considered in this study. If the effect exists, it could be uncovered by additional speaker data affording additional statistical precision.

6. Future work

Based on informal observations made during this study, future work could explore the possibility that double closing or opening peaks in the DEGG signal occur more often in physical task stress, and evaluate possible causes for such a finding. Table 5 shows the p values of two ANOVAs for opening peak count and closing peak count as response variables. It appears here that the closing peaks count is unaffected by physical task stress, as the low p value for the interaction between vowel and task is due to changing signs in tiny differences of means and accepting this as significant may very well constitute a Type I error. The results for the opening peaks are more clearly related to task and other factors, including vowel, consonant, and speaker. Examining the means reveals a significant decrease in mean opening peak count in stress from neutral. As doubled opening and closing peaks in the DEGG signal can be related to special types of phonation [12], determining whether this observation is just a measurement artifact, a finding particular only to this data set, or something more, is of interest for future work.

This study has considered four speakers. Future work would do well to expand the field of participants. For larger study groups of speakers, gender might be a suitable replacement for speaker identity as a study factor.

Table 5: p values from two separate ANOVAs with closing peak count (CPC) and opening peak count, respectively, as response variable, and with speaker, vowel, preceding consonant, and task as factors.

Factor	Closing peaks p	Opening peaks p
Speaker	$5.405 * 10^{-05}$	$< 2.200 * 10^{-16}$
Vowel	0.1024	$2.509 * 10^{-07}$
Consonant	0.1678	0.0154
Task	0.3981	0.0002
Spkr:Vowel	0.0668	0.0246
Spkr:Cons.	0.0048	0.8338
Vowel:Cons.	0.2678	0.3068
Spkr:Task	0.3288	$4.329 * 10^{-12}$
Vowel:Task	0.0053	0.5666
Cons:Task	0.6825	0.6547
Spkr:Vowel:Cons.	0.1095	0.9467
Spkr:Vowel:Task	0.2595	$4.237 * 10^{-05}$
Spkr:Cons:Task	0.3630	0.8855
Vowel:Cons:Task	0.8821	0.1522

7. References

[1] M. S. Entwistle, "The performance of automated speech recognition systems under adverse conditions of human exertion," *Intl. J. of Human Computer Interaction*, vol. 16, no. 2, pp. 127–140, 2003.

[2] S. A. Patil and J. H. L. Hansen, "The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification," *Speech Comm.*, vol. 52, pp. 327–340, 2010.

[3] Y. Meckel, A. Rotstein, and O. Inbar, "The effects of speech production on physologic responses during submaximal exercise," *Medicine and Sci. in Sports and Exercise*, vol. 34, pp. 1337–43, Aug. 2002.

[4] B. Johannes, P. Wittels, R. Enne, G. Eisinger, C. A. Castro, J. L. Thomas, A. B. Adler, and R. Gerzer, "Non-linear function model of voice pitch dependency on physical and mental load," *Eur. J. Appl. Physiology*, vol. 101, pp. 267–276, 2007.

[5] K. W. Godin and J. H. L. Hansen, "Analysis and perception of speech under physical task stress," in *INTERSPEECH 2008*, (Brisbane, Australia), pp. 1674–1677, Sep. 2008.

[6] K. W. Godin, "Classification based analysis of speech under physical task stress," Master's thesis, University of Texas at Dallas, Richardson, TX, Richardson, TX, USA, Dec. 2009.

[7] A. Rotstein, Y. Meckel, and O. Inbar, "Perceived speech difficulty during exercise and its relation to exercise intensity and physiological responses," *Eur. J. Appl. Physiol*, vol. 92, pp. 431–436, 2004.

[8] K. N. Stevens, *Acoustic Phonetics*, p. 73. Cambridge, MA: MIT Press, 1998.

[9] A. Barney, A. D. Stefano, and N. Henrich, "The effects of glottal opening on the acoustic response of the vocal tract," *Acta Acustica United with Acustica*, vol. 93, pp. 566–577, 2007.

[10] A. Ikeno, V. Varadarajan, S. Patil, and J. H. L. Hansen, "UT-Scope: Speech under lombard effect and cognitive stress," in *IEEE Aerospace Conf. 2007*, (Big Sky, Montana), pp. 1–7, 2007.

[11] K. Sjolander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. ICSLP*, 2000.

[12] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. of the Acoustical Soc. of Am.*, vol. 115, pp. 1321–1332, 2004.