Godin and Hansen *EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:29
DOI 10.1186/s13636-015-0072-7

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Physical task stress and speaker variability in voice quality

Keith W. Godin[1] and John H. L. Hansen[1,2*]

Abstract

The presence of physical task stress induces changes in the speech production system which in turn produces changes in speaking behavior. This results in measurable acoustic correlates including changes to formant center frequencies, breath pause placement, and fundamental frequency. Many of these changes are due to the subject's internal competition between speaking and breathing during the performance of the physical task, which has a corresponding impact on muscle control and airflow within the glottal excitation structure as well as vocal tract articulatory structure. This study considers the effect of physical task stress on voice quality. Three signal processing-based values which include (i) the normalized amplitude quotient (NAQ), (ii) the harmonic richness factor (HRF), and (iii) the fundamental frequency are used to measure voice quality. The effects of physical stress on voice quality depend on the speaker as well as the specific task. While some speakers do not exhibit changes in voice quality, a subset exhibits changes in NAQ and HRF measures of similar magnitude to those observed in studies of soft, loud, and pressed speech. For those speakers demonstrating voice quality changes, the observed changes tend toward breathy or soft voicing as observed in other studies. The effect of physical stress on the fundamental frequency is correlated with the effect of physical stress on the HRF ($r = -0.34$) and the NAQ ($r = -0.53$). Also, the inter-speaker variation in baseline NAQ is significantly higher than the variation in NAQ induced by physical task stress. The results illustrate systematic changes in speech production under physical task stress, which in theory will impact subsequent speech technology such as speech recognition, speaker recognition, and voice diarization systems.

Keywords: Physical task stress; Glottal waveform analysis; Speech variability; Speaker variability

1 Introduction

Exercising or otherwise performing a strenuous physical task, referred to here as physical task stress, influences the behavior of the speech production system. The study of physical task stress speech is the search for a link between stress and the resulting behaviors and to the acoustic and perceptual results of those behaviors. Several studies have added to the catalog of behaviors studied in terms of the changes caused by physical task stress. In this study, we examine voice quality. The combination of speech with exercise, or physical task stress, results in physiological and behavioral responses that depend on the fitness of the speaker, the age of the speaker, the specific task, the speaker's fatigue level, the time in the task, and other factors.

An overview representation of speech production is shown in Fig. 1. Here, typical speech production occurs when an excitation source excites the resonant structure of the vocal tract, resulting in an output speech waveform (Deller et al. [1]). If we consider a traditional vowel, the figure shows the resulting glottal cycle (closed phase, open phase) which excites the corresponding configured vocal tract. Articulators in the vocal tract need to be continuously positioned to produce fluent speech over time. If the subject is performing some type of external physical task stress, these factors will influence speech production with respect to airflow from the lungs, glottal excitation source structure, and vocal tract articulator positioning. The notion of stress level can be described using the Yerkes-Dodson human performance and stress curve, as shown in Fig. 2 (motivated by [2]). Here, there

* Correspondence: john.hansen@utdallas.edu
[1]Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas 75080, USA
[2]Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601N. Floyd Road, EC33, Richardson, TX 75080-1407, USA

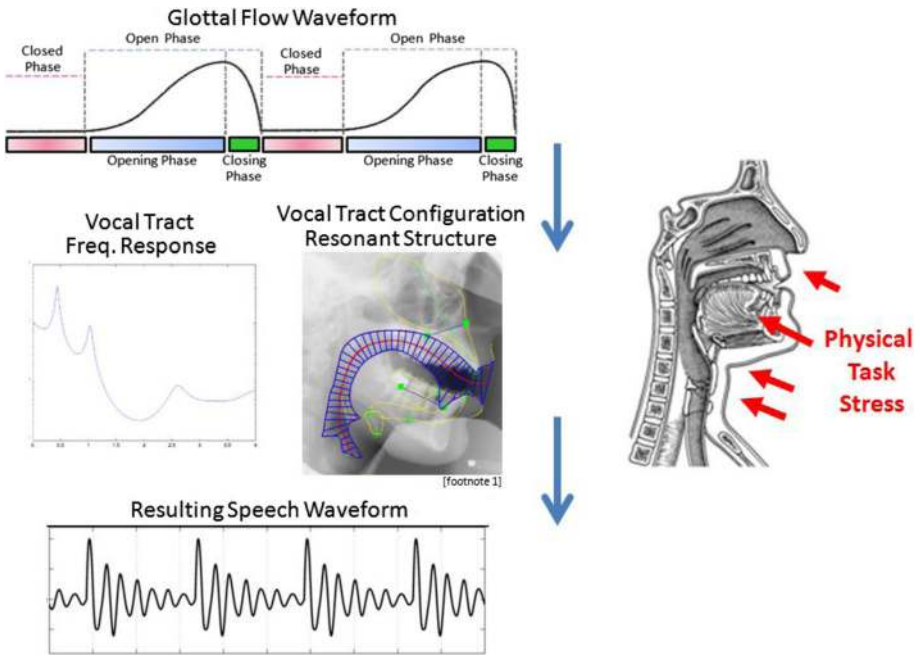

Fig. 1 Illustration of physical speech production: glottal source excitation driving vocal tract frequency response, resulting in output speech waveform; physical task stress impacts excitation source muscles, vocal tract articulator muscles, and airflow structure, all impacting output speech signal (note: the physiological vocal tract configuration shown here was adapted from a figure available at the following location: [1]http://www.loria.fr/~laprie/ACS/tubesOfa.png)

is a continuum of stress, beginning from inactive or calm conditions, to optimum stress, to fatigue, to exhaustion, to excessive levels resulting in panic/breakdown. As such, physical task stress is not simply a binary event which is turned on/off but a continuum. The impact of physical task stress on speech production and how that impacts voice quality is the focus here.

1.1 Background on speaking and exercise

Speaking and exercising compete for some of the same resources, and exercise affects the speech production system. Conversely, speaking during exercise affects exercise performance, influencing heart rate, ventilation, tidal volumes, and perception of dyspnea or air hunger. During exercise, speakers decrease their ventilation while speaking in order to make controlled utterances [3–7], then compensate in the period between utterances by significantly increasing their ventilation past baseline [4, 8]. When speaking segments are so long that recovery periods of sufficient length do not occur with enough frequency, the speaker is forced to place breathing pauses at linguistically inappropriate places [3]. The effect of exercise on speech breathing is significant and consistent enough that it may be used as a feature in the automatic detection of exercise from the speech signal [9]. Studies are inconsistent and conflicting regarding the question of whether speaking increases [3] or does not increase [5] heart rate relative to the exercise-only heart rate at the same VO_2 task level. Finally, speech production during exercise results in reduction of oxygen intake and an increase in blood lactic acid [5], decreasing physical performance and hastening fatigue.

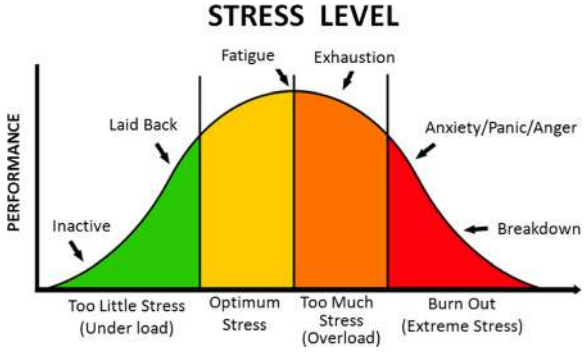

Fig. 2 Illustration of the Yerkes-Dodson human performance and stress curve [2]

Significant inter-speaker variability was observed across these physiological variables including oxygen uptake, heart rate, and blood lactate [5]. Also, while perceived speech production difficulty is strongly correlated with the difficulty of the exercise task [10–12], significant inter-speaker variability has been observed despite these correlations. The strength of the correlation may be increased when the subject pool is more uniformly fit and more generally homogeneous, such as the expert cyclists studied in Rodriguez-Marroyo et al. [11, 13].

Physical stress causes behavioral changes in the speech production system, resulting in acoustical differences compared to speech produced in neutral conditions. The most commonly studied acoustic parameter is the fundamental frequency (F0), which typically increases in physical task stress. In Godin and Hansen [14], mean F0 increased by 60 % of speakers, similarly for 7 of 10 subjects in Koblick [15], while Johannes et al. [16] observed increases for all speakers, with a more uniformly fit subject pool. Furthermore, Johannes et al. [16] designed their study to include a task of increasing difficulty and measurements of F0 throughout and proposed a nonlinear plateau model for the change in F0 due to stress. They noted that the anchor frequencies and the height of each plateau in their model were speaker-dependent. In contrast, Mohler [10] observed a linear increase in F0 with increases in VO_2. While most studies considering speech during physical tasks use aerobic exercises as stimuli, Orliko [17] measured speech production characteristics before and during a weightlifting task. Mean F0 was not affected, nor was phonatory airflow nor pitch perturbation coefficient, but the F0 coefficient of variation increased.

Studies have also considered vocal intensity, noise-to-harmonics ratio, and jitter, which all may increase in physical task stress [15]. One study suggests that these increases are correlated with the underlying increase in heart rate (Orlikoff and Baken, [18]). Godin and Hansen [14] found that the standard deviation of F0 increased by 2 % of speakers and decreased by 24 % of speakers, suggesting a reduced prosodic range in physical task stress. They found that utterance duration increased by 30 % of the speakers, as well as decreased by 43 % of the remaining speakers. Changes in duration may be related to the breathing strategies discussed above, and the inter-speaker differences here suggest that different speakers employ different strategies. The glottal open quotient and the first two formants are also affected by physical task stress [19]. A qualitative comparison of low and high vowels to plosives and fricatives suggested that the vowels were more affected by physical task stress than the plosives and fricatives [20] and further that nasal phones are more affected by physical task stress than plosives and fricatives [20]. This may be caused by the decline in nasal resistance during physical stress, which might in turn affect the acoustic properties of the upper vocal tract [21]. Variability across speakers in response to physical task stress is a theme across these studies, where, as cited above, Koblick [15], Godin and Hansen [14], Baker et al. [3], and Godin et al. [22] observed parameter shifts for a majority but not all speakers. Godin and Hansen [19] observed changes for all speakers but found statistically significant differences in shift of these parameters across speakers, and Johannes et al. [16] observed shifts in F0 for all speakers but noted that the parameters of their model were speaker-dependent. The significant inter-speaker variability in the physiological and behavioral effects of stress as observed in, e.g., [3, 5], should result in significant inter-speaker variability in the acoustic correlates of stress. Significant speaker variability in acoustic correlates has also been noted for other types of stress [23, 24].

A recent study, Godin et al. [22], studied the effects of physical task stress on voice quality. That study measured six parameters, the harmonic richness factor (HRF), normalized amplitude quotient (NAQ), H1−H2 ratio (H1H2), F1F3syn [25], harmonics-to-noise ratio (HNR), and spectral slope (SS). Each of these six parameters is sensitive primarily to changes in the vocal fold behavior or related acoustical properties, rather than to the upper vocal tract. In plotting the distribution of each parameter in neutral and stress across all speakers, they found very little change in the overall distribution of the parameter sample values. However, when focusing on measurements from individual speakers, they observed effects of physical stress on these parameters for a subset of speakers. As with any examination of the effects of an outside influence on the behavior of the speech production system, we must approach our analysis from a speaker-dependent perspective. This study expands on Godin et al. [22] to look more closely at a subset of these voice quality measurements.

2 Acoustic measures of stress and voice quality

Voice quality is the acoustic result of phonatory behavior [26]. Voice qualities include modal (neutral), creaky, breathy, whispery, tense, and lax [26] and depend on the tension and compression of the vocal folds, among other factors. Voice quality varies naturally throughout speech, carries paralinguistic information, and may depend on social context, mood, and intent [27, 28]. Variations in vocal fold health, tension, temperature, configuration, and other aspects result in significant acoustic differences as well as different voice qualities. These changes may be made consciously, as in the case of loud or soft vocal effort [29, 30], may be the result of emotions or stressors [29, 31, 32], or may be the result of unconscious communication habits [28]. Thus, acoustic measures over the speech signal may be strongly associated with particular classes of vocal fold behavior and physiology.

Estimation of the glottal flow waveshape through inverse filtering of the speech waveform, and parameterization of the waveshape estimate, is the primary method by which to derive acoustic parameters that measure voice quality. Care is needed to ensure an effective vocal tact model from traditional linear prediction (LP), since the error residual from LP analysis is not guaranteed to represent the true glottal flow waveform, since it also encodes any error

residual from poor vocal tract spectral modeling. The study by Gavidia-Ceballos and Hansen [33] explored this issue for subjects with various forms of vocal fold cancer and successfully employed estimates of vocal tract structure from parallel speakers to more accurately suppress vocal tract structure for glottal flow waveform analysis. The study by Cummings and Clements [34] considered inverse filtering with a parametric model of the resulting glottal waveform shape for speech under stress and emotion. Earlier analysis of the glottal source structure suggested this would be possible (Hansen and Clements [35]). With respect to quality, glottal pulse width, glottal pulse skewness, abruptness of glottal closure, and turbulent noise component may be indicative of voice type variation [36]. Lower open quotient and closing quotient are related to breathy voice, and higher closing quotient is related to pressed voice [37]. Higher AC flow and increased subglottal pressure is associated with loud voice, while lower AC flow and lower subglottal pressure is associated with soft voice [38]. Harmonics-to-noise ratio has been extensively studied and is strongly correlated with breathy and rough voice quality [39]. Aspiration noise is also a significant factor in voice quality and may be estimated using the F1F3syn parameter [25].

More recently, the normalized amplitude quotient (NAQ) has been demonstrated to be strongly correlated with voice quality variations and robust to noise and estimation errors [28, 37, 40]. Drugman et al. [40] showed that NAQ, H1H2 ratio, and harmonic richness factor (HRF) measured for soft, modal, and loud speech resulted in significantly different distributions in these parameters for a corpus of a single speaker. On their corpus of a single speaker, NAQ had a higher distribution mean for soft speech, middle distribution mean for the modal speech, and a sharper, lower-mean distribution for the loud speech. While there was less separation across speech types for the H1−H2 ratio, the harmonic richness factor was lower for soft speech.

Speech, even in the absence of stressors or significant emotions, has variations in voice quality that carry paralinguistic cues such as affirm, deny, or backchannel [28, 41], and many other external influences can affect voice quality, such as depression [42], circadian rhythm and fatigue [43, 44], cognitive load [29, 45, 46], emotions [29, 31, 45, 47, 48], and aging [49]. Also, baseline (modal) values of voice quality measurements such as NAQ vary significantly across speakers [28]. Spontaneous, continuous speech, typical of conversations, has voice quality characteristics that differ significantly enough across speakers that they may be used as features for automatic speaker identification systems [50, 51]. For these reasons, in order to measure voice quality of a given speech segment, the measure must be normalized for the underlying speaker variation regarding age, mood, conversational context, fatigue, and other factors.

Like depression, emotions, circadian rhythm, and conversational context, physical task stress can be expected to induce changes in voice quality, driven by the physical demands of exercise and the competition between the speaking and breathing tasks. As physical task stress is an external factor that drives behavior and physiology rather than a specific phonatory behavior itself, we may not expect a direct link between the parameters of the physical task or the fitness of the speaker and the resulting acoustic measures.

3 Speech parameters for physical task stress analysis

In the analysis of speech under stress, a range of speech parameters are possible. In the area of speech under stress analysis, Hansen [29, 52] considered 200 speech parameters spanning the domains of glottal spectrum, pitch/fundamental frequency, duration, intensity, vocal tract spectral structure. Further analysis was considered for military communication applications of speech under stress by NATO RSG.10 [53], USAF [54]. These feature analysis studies lead to advancements in robust speech recognition under stress [29, 55–57] and a tutorial overview of a number of stress compensation techniques based on voiced-transition-unvoiced speech tagging as well as neural network and source generator compensation of stress [58]. An additional application domain included advancements in automatic detection of speech under stress using signal processing advancements derived from the Teager energy operator (TEO) [59], TEO-CB-AutoEnv ([60]). More recently, nonlinear TEO-based advancements have been considered for stress detection using sub-band filterbank weighting for various actual speech under stress scenarios [61, 62]. While these have explored a range of stress conditions, specific speech under physical task stress was not addressed. As such, it is believed that alternative features could also be explored for the present study. As such, it is believed that alternative features could also be explored for the present study. “In this study, the UTSCOPE-Physical Task Stress corpus (see Table 1) is employed for analysis. The Corpus consists of 78 subjects collected in both neutral and physical task stress conditions, as well as being balanced across gender (male/female), native/non-native, read/spontaneous speech conditions.”

We have selected three parameters to study the voice quality effects of physical task stress. Fundamental frequency is widely studied and serves as a comparison with prior work. Harmonic richness factor (HRF) and normalized amplitude quotient (NAQ) have been selected because past studies have quantified the relationship of these parameters to specific speaking behaviors

Table 1 UT-SCOPE-physical speech corpus: details of corpus including speakers, audio, sessions, and transcription effort. The corpus is used for speaker ID and stress classification for two stress conditions: cognitive stress and physical stress. The cognitive load is simulated by having subjects play a driving game on an interactive game console. The physical stress is induced by requiring subjects to maintain a 10-mph pace based on visual speed display on an elliptical stair-stepping machine

UT-scope-physical corpus

Total speaker	78		
Native speaker	51	Nonnative speaker	27
Male speaker	9 (native)	Female speaker	42 (native)
Total audio size	28 h	Audio size/speaker	15 min
Recording Format	16 bit raw	Sampling rate	44.1 kHz
Microphone types	1.) Sound level meter 2.) Shure Beta 54 Close talk mic 3.) Two Shure MX391/S far-field mics		
Session types	1.) Spontaneous 2.) Read		
Session contents	1.) 3 min spontaneous parts for cognitive and physical 2.) 35 TIMIT sentences spoken in neutral, cognitive, and physical stress, prompted through headphones		
Available transcription	1.) 35 TIMIT sentences		
Transcribed data	The prompted parts of the first sessions of all native speakers have sentence and word transcripts. The word and phone transcripts were created with forced alignment. All word transcripts were corrected by hand. Phone transcripts for 10 of the female native speakers were corrected by hand. In neutral and in stair-stepper mode – there is heart-rate information for subjects in metadata directory. 51 native speakers – sentence and word transcripts. 38 of female speakers also have phone level transcripts.		

including pressed speech and soft speech. This facilitates our investigation into whether the effects of physical stress can be described in terms of these speaking types.

NAQ and HRF can be reliably estimated if an inverse filtered glottal waveform is available. Past studies have shown that care should be exercised in applying vocal tract inverse filtering for glottal source waveform estimation when voice characteristics are under pathology [33], since determining the exact glottal closure instant (GCI) is not always possible. In general, glottal inverse filtering is a significant area of research interest. Here, the GLOAT toolkit is used for GCI detection [63], fundamental frequency estimation [64], and glottal inverse filtering [40]. Kane and Gobl [65] demonstrated that voice quality variation has a significant effect on the accuracy of GCI detection, which is critical for correct glottal inverse filtering, but their data suggested that the SEDREAMS method used here is reliable enough for speech analysis, despite voice quality variation.

3.1 Fundamental frequency

The fundamental frequency (F0) has been the primary object of study of speech under physical task stress. Most studies have concluded that stress results in an increase in F0. However, there is significant speaker variability in the effects of physical stress on F0, as Godin and Hansen [14] noted an increase in the F0 by just 61 % of speakers and a decrease by 14 % of speakers. To reduce F0 estimation errors such as doubling or halving, we have set the allowable range to 120–400 Hz.

3.2 Normalized amplitude quotient

The normalized amplitude quotient (NAQ) is the ratio of the maximum amplitude of the glottal flow to the

minimum of the glottal flow derivative, normalized by the fundamental period and the sampling frequency [66]. NAQ is sensitive to variations caused by breathy and pressed phonation [66] and to soft and loud speech [64]. It is known that NAQ increases for breathy phonation, and decreases for pressed phonation, relative to neutral speech.

3.3 Harmonic richness factor

The harmonic richness factor (HRF) is the ratio of the sum of the amplitudes at the harmonics in the glottal waveform to the amplitude of the component at the fundamental frequency [36]. In Childers and Lee [36], the HRF of modal voicing was higher than that for breathy voicing by 6.8 db. In Drugman and Alwan [64], there were clear shifts in the distribution of HRF between loud, modal, and soft voicing. In our implementation of HRF, we have used only the fundamental and the first nine harmonics. This ensures that, unless the F0 exceeds 800 Hz, all measurements of HRF sum over the same number of harmonics, eliminating the dependence of HRF on F0.

4 UT-SCOPE-Phy-II corpus: speech under physical task stress corpus

This study uses physical task stress data from the UT-SCOPE corpus [67] and the UT-SCOPE-Phy-II corpus [19]. The protocol for the corpus development is based on the following steps: (i) the subject was first asked to produce neutral examples of speech based on prompts provided through Sennheiser HD 650 open air headphones from an audio stream originating from a computer; (ii) the sentences were presented sequentially with a pause inserted between each to allow for an effective but relaxed speech pace; (iv) once neutral speech was captured/completed, the subject was positioned on a Conversion II Elliptical/Stair Stepper machine (see Fig. 3) and continued voice data collection again with prompts presented through headphones. Each subject was allowed to take breaks, as per IRB approved protocol. Pure tone testing was done with the 650 open air headphones to ensure no attenuation existed between inside and outside the ear units (from Fig. 3)—this ensured that each subject experienced no occlusion effect with the headphones and could hear themselves. Sound level testing was also performed on the Elliptical/Stair Stepper machine, with no appreciable device noise levels measured at 1-ft distance from the floor foot pedal area (i.e., no machine noise from the stair stepper was captured in any of the audio recordings).

The UT-SCOPE and UT-SCOPE-Phy-II corpora were collected under the same protocols and are used together. Expanding on the protocol from above, both corpora include a segment of 35 prompted TIMIT sentences spoken in both neutral and physical task stress (presented through headphones). A spontaneous speech portion is also available but was not used in this study. These sentences comprise the analysis data set used in this study. Having the same sentence spoken in both tasks reduces the phonetic variability for analysis of the effects of physical task stress. Sessions from 66 female native speakers of American English are used in this study. We choose to consider the female speakers because we had a larger sample size and did not want to introduce gender as

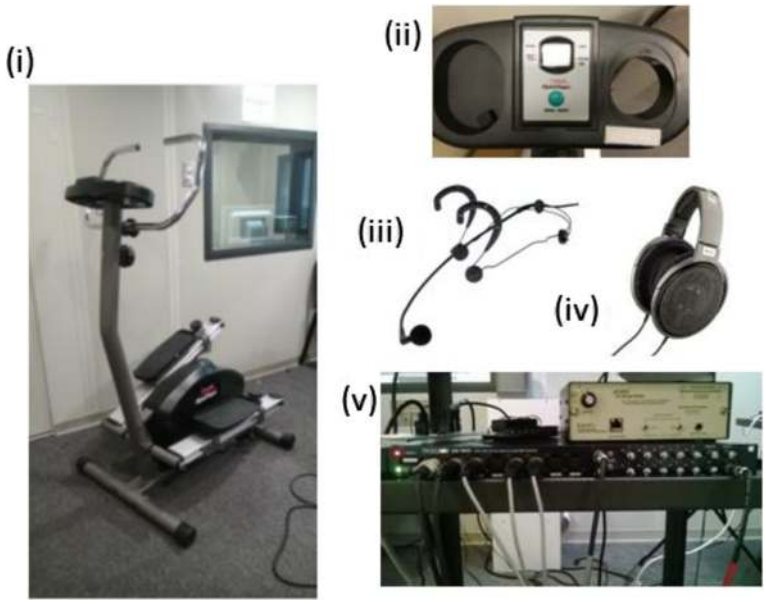

Fig. 3 Equipment used for physical task stress speech data collection: (i) Conversion II Elliptical/Stepper, (ii) digital screen display for Stepper, (iii) Shure Beta 53 close-talking microphone, (iv) Sennheiser HD 650 open air headphones, (v) Tascam US-1641 multi-channel digital recorder

another variable in the study. All participants were at least 18 years of age at the time of the study. A Conversion II Elliptical/Stair Stepper machine (Fig. 3, along with other equipment) was used to induce the physical task stress. Each speaker was asked to maintain an approximately 10-mph pace on the machine (there is a digital readout which indicates speed and allows the subject to maintain the requested pace). Having the same task for each subject resulted in different levels of exertion for each speaker, and therefore, there is a diversity of exertion levels in the corpus. The data was collected inside 13-by-13 ft ASHA-certified single-walled sound booth, with the subject wearing a Shure Beta 53 close-talking microphone.

Ground truth exertion level was determined by the subject's percentage of maximum oxygen uptake (VO_2 max). Heart rate data was also measured and recorded and included with the UT-SCOPE corpora, which is correlated with exertion level. Both UT-SCOPE collection protocols include the use of a chest worn heart rate monitor that samples the speaker heart rate every 15 s. Figure 4 shows heart rate (HR) (in beats per minute (BPM)) from the start of data collection until completion for (i) neutral speech entry only, (ii) cognitive task stress with speech entry using race car simulator, and (iii) physical task stress with speech entry using stair stepper. Each reading is taken every 15 s, so the 65 readings for the physical stress task plot represent 16 min 15 s elapsed time. A comparable plot is obtained for each subject in the corpus. The physical task stress HR is significantly higher than the neutral HR. A cognitive task was also performed during the collection which involved using an interactive video race-car system. The results here show that the cognitive task did not affect the HR significantly. The HR increase in physical stress demonstrates that the speakers are under significant stress in performing the task. It should be noted that during the collection process, no speaker actually had difficulty speaking, suggesting they did not exceed the ventilatory threshold [68]. The HR may be used with the Karvonen method to estimate the exertion level (i.e., l below) for a speaker [69] as follows:

$$l = \frac{\text{HR}-\text{RHR}}{\text{MHR}-\text{RHR}}, \tag{1}$$

where HR is the current heart rate, RHR is the resting heart rate, and MHR is the maximum heart rate. As considered in Godin and Hansen [20], the maximum heart rate is estimated as [70], where A is the age of the subject in years:

$$\text{MHR} = 208.9 - 0.7A \tag{2}$$

To highlight the subject's response characteristics during speech production, Fig. 5 shows a scatter plot of the 67 female UT-SCOPE-Physical Stress speakers in two dimensional space representing age versus exertion level. The plot shows that most subjects are clustered in the 18–28 years age range, with some ranging between 30 and 60 years. Exertion levels generally fall in the range of 0.2–0.65. A correlation coefficient between age and

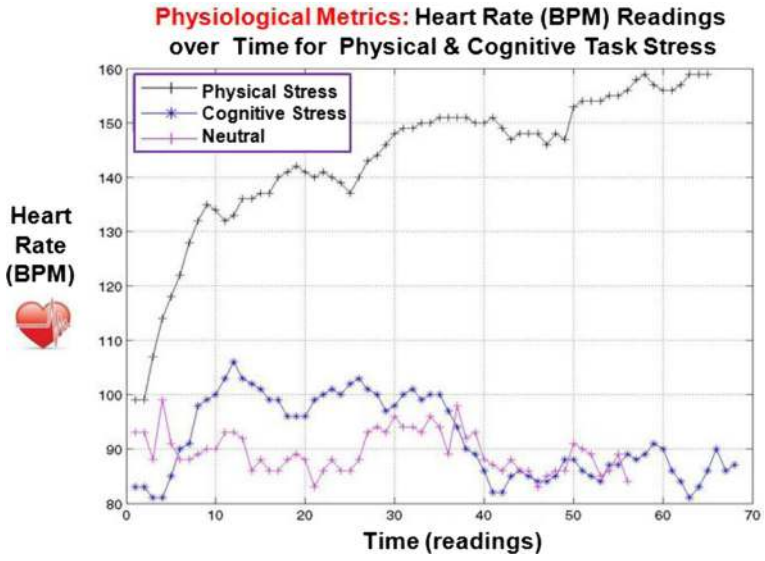

Fig. 4 Heart rate during tasks from a sample speaker from UT-SCOPE-physical corpus. Shown are heart rate (in beats per minute) from the start of data collection until completion for (i) neutral speech entry only, (ii) cognitive task stress with speech entry using race car simulator, and (iii) physical task stress with speech entry using stair stepper. Each reading is taken every 15 s, so 65 readings for the physical stress task plot represents 16-min–15-s elapsed time

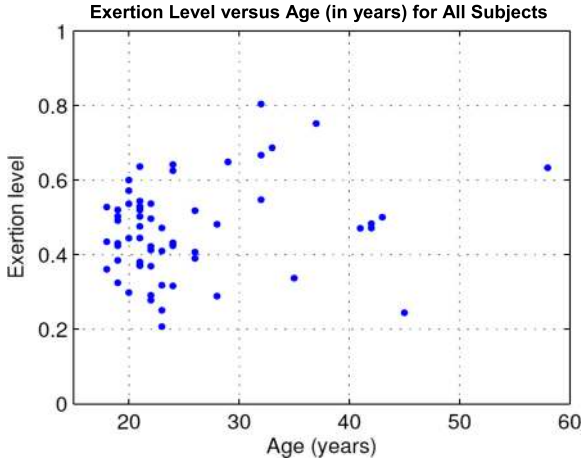

Fig. 5 Age versus exertion level for all speakers in the combined corpus

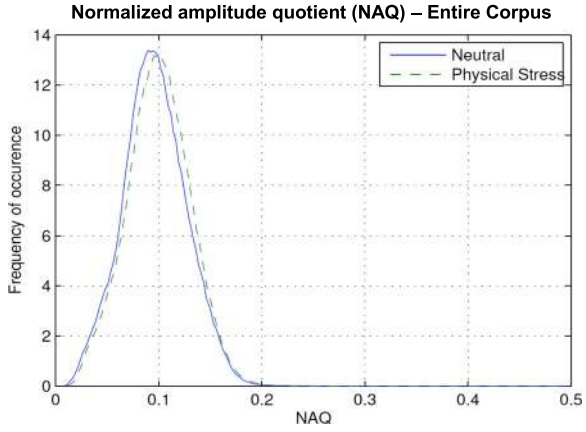

Fig. 7 The normalized amplitude quotient (NAQ) across entire corpus

exertion level of −0.046 suggests that speaker age was not an explanatory factor for exertion level. It also appears that other factors such as overall speaker fitness, or potentially general fatigue level at the time of the recording, played a greater role in determining the exertion level than the subject's age.

The average heart rate across all speakers for both tasks (neutral and physical task stress) is shown in Fig. 6. The dashed lines mark one standard deviation in the data. The wide distributions suggest a greater level of subject variability in heart rate. The change in mean from neutral to physical task stress shows the average increase in heart rate ranges from 38.8 to 54.9 beats/min for the subjects in the corpus.

5 Effects of physical stress on voice quality

The NAQ and HRF are measured for every detected glottal cycle in the speech data for all speakers. Figure 7 shows the distribution of the NAQ measurements across the entire corpus. The blue, neutral condition has a mean of 0.096, and the green which represents the physical task stress condition has a mean of 0.10. Likewise, Fig. 8 shows the distribution of the HRF measurements across the entire corpus. The blue, neutral condition has a mean of 10.9, while the green representing the physical task stress condition has a mean of 10.6. In both cases, the shifts in these parameters from neutral are in the direction of the "soft" voicing style, as discussed in Drugman et al. [40], or in the direction of "breathy" voicing as discussed by Alku et al. [66]. However, the shifts observed here are much smaller than those observed for those discrete phonation types (i.e., breathy or soft), suggesting that physical task stress does not affect voice quality. However, these plots do in fact mask the underlying speaker differences. In a speaker-dependent analysis, we compare the shifts in the mean NAQ and HRF for one speaker. Some speakers have large shifts in these parameters. Speaker FJF3 is an example, with a shift in

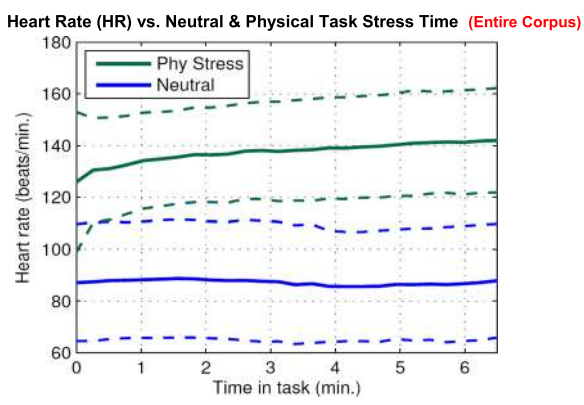

Fig. 6 Mean HR across speakers during (i) neutral with speech only entry and (ii) physical task stress combined with speech data entry. *Dashed lines* mark one standard deviation range over time

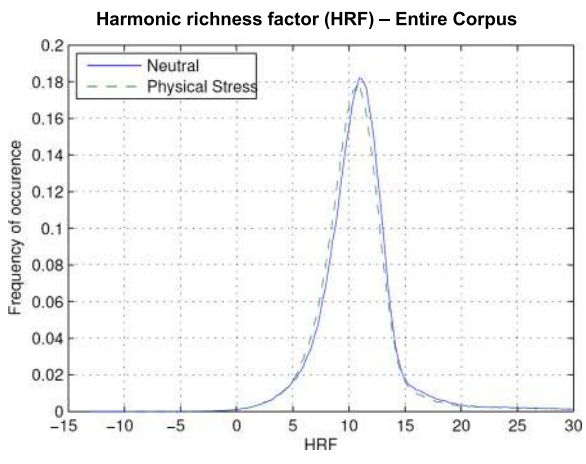

Fig. 8 The harmonic richness factor (HRF) across entire corpus

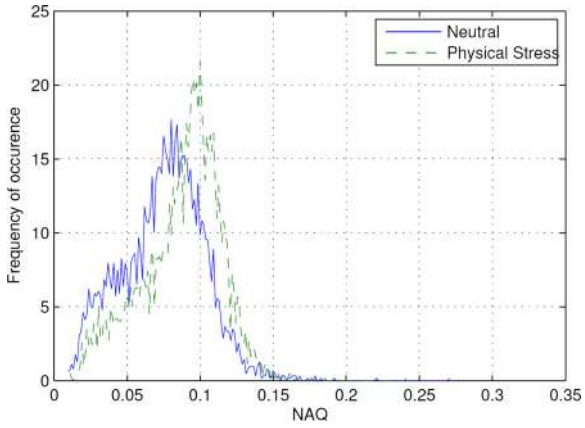

Fig. 9 The normalized amplitude quotient (NAQ) for the single female speaker (FJF3)

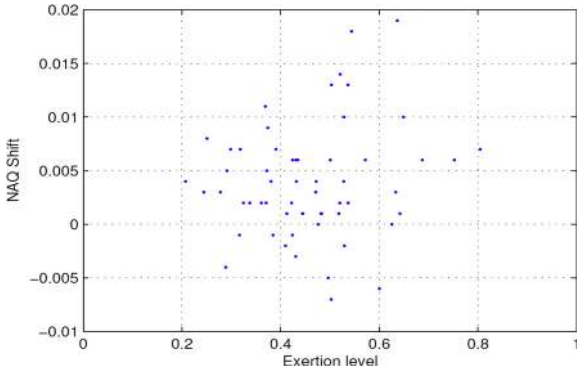

Fig. 11 The normalized amplitude quotient (NAQ) shift versus exertion level: each point represents one speaker; there is significant variability in mean NAQ shift across speakers; however, this is not correlated with exertion level, with a correlation value of $r = 0.108$

HRF among the largest at −1.33. Figure 9 shows the change in the distribution of NAQ from neutral to physical task stress, showing a trend in the same direction from neutral as for "breathy" or "soft" voice qualities [40, 66]. Figure 10 shows that the HRF is also affected by physical task stress for the same example subject FJF3 (similar trends were seen in other subjects as well).

Continuing with this specific speaker-dependent analysis, we consider the variation across speakers in the shift of mean NAQ and HRF and investigate the correlation of this shift with speaker exertion level. Figure 11 shows a scatter plot of exertion level versus mean shift in NAQ, where each point is one speaker. These results show there is a wide speaker variation in mean NAQ shift, ranging from −0.00627 to 0.0187. It was observed in Alku et al. [66] and Drugman et al. [40] that known behavioral changes such as soft voicing or pressed voicing resulted in changes for NAQ which were greater than about 0.01. Here, 10 speakers had changes in NAQ greater than 0.01. The results from this figure suggests that changes in exertion level explains little of the shift in mean NAQ, where the correlation coefficient is $r = 0.108$. While the exertion level change is positive for almost all of the speakers, the voice quality changes are significantly different across speakers, with some trends toward pressed voicing, with others toward soft voicing.

Figure 12 shows the shift in the exertion level and the shift in the HRF, where each point is one speaker. As with NAQ, we see significant speaker variability in the shift in mean HRF, with shifts ranging from −1.911 to 0.682. In Alku et al. [66] and Drugman et al. [40], shifts greater than about 1 were associated with known changes in voice quality, such as soft or pressed voicing. Here, seven speakers had an HRF shift greater than 1. The shift in mean HRF is not related to shifts in the exertion level, with a correlation coefficient of $r = 0.012$.

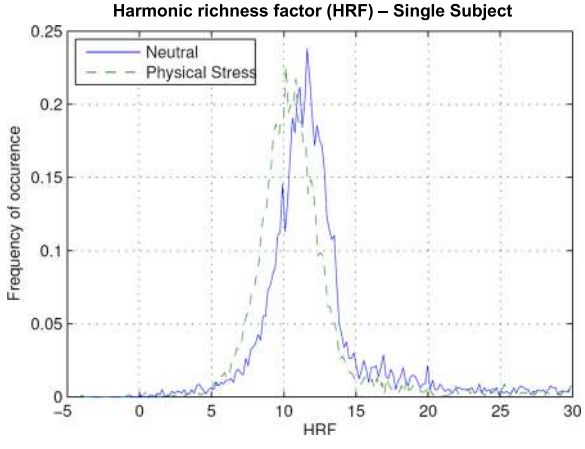

Fig. 10 The harmonic richness factor (HRF) for the single female speaker (FJF3)

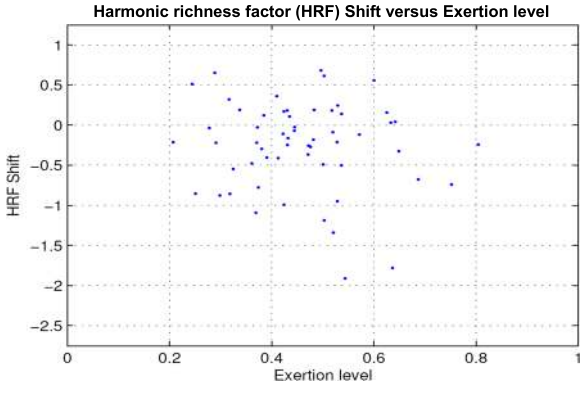

Fig. 12 The harmonic richness factor (HRF) shift versus exertion level: each point represents one speaker; there is significant variability in mean HRF shift across speakers; however, this is not correlated with exertion level, with a correlation value of $r = -0.012$

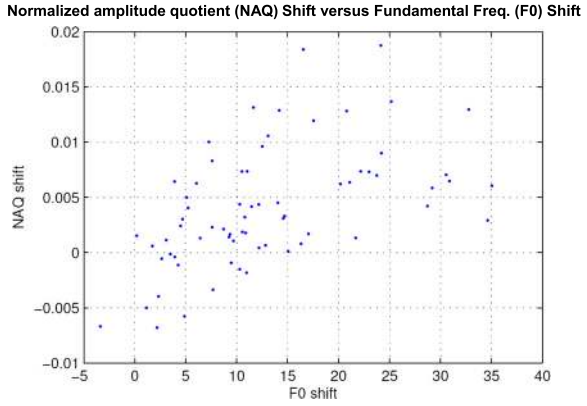

Fig. 13 The normalized amplitude quotient (NAQ) shift versus fundamental frequency (F0) shift for each speaker across the entire corpus. These two are correlated, with a correlation value of $r = -0.53$

On a subset of this data set, Godin and Hansen [14] found that F0 increased for the physical task stress condition by 60 % of speakers. Figure 13 shows the relationship between the shift in mean F0 and the shift in mean NAQ for each speaker. There is a strong correlation between ($r = 0.53$) shift in mean F0 and shift in mean NAQ, despite the fact that specific F0 dependence is normalized out of the NAQ. Next, Fig. 14 shows the relationship between the shift for each speaker in F0 and the shift for each speaker in HRF. The correlation between shift in F0 and shift in HRF is $r = -0.34$. The figures provide a more specific individualized view of the speaker variation in response to physical task stress compared with the much earlier study results from Godin and Hansen [14]. Here, we see that the shift in F0 is actually along a continuum and does not display a discrete break between speakers who shifted their F0 versus those who did not.

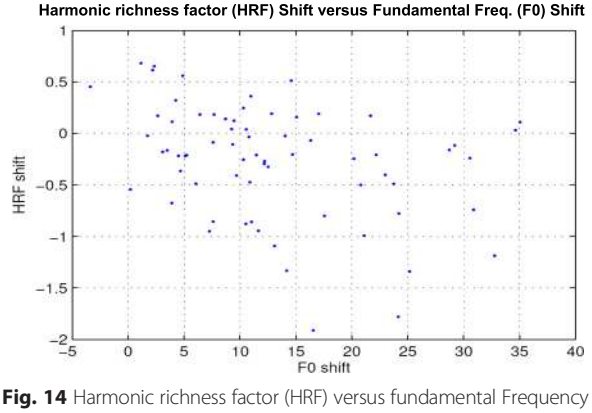

Fig. 14 Harmonic richness factor (HRF) versus fundamental Frequency (F0) for each speaker across the entire corpus. These two have some degree of correlation, with a correlation value of $r = -0.34$

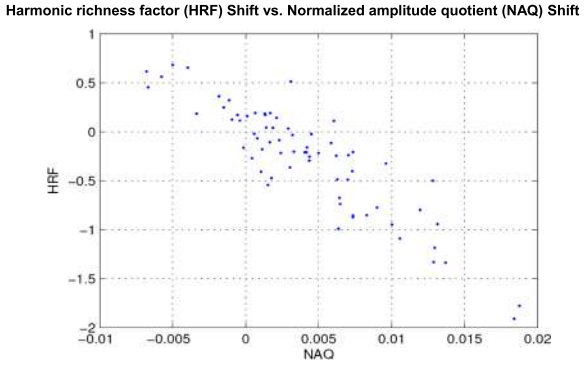

Fig. 15 Harmonic richness factor (HRF) shift versus normalized amplitude quotient (NAQ) shift for each speaker across the entire corpus. These two have a very strong degree of correlation, with a correlation value of $r = -0.89$

Next, a direct comparison of the effects of physical task stress on the HRF and NAQ reveal they are strongly correlated, with a correlation value of $r = -0.89$, as is demonstrated by the scatter plot in Fig. 15.

Finally, we consider speaker variation with respect to the mean NAQ and HRF measurements for both neutral and physical task stress. The motivation for this is that in Figs. 7 and 8 which represent NAQ and HRF across the entire corpus, they in fact hide significant individual speaker differences in the baseline NAQ and HRF measurements, which are now shown in Fig. 16. Here, Fig. 16 shows a scatter plot of the per speaker mean NAQ for neutral versus speaker mean NAQ for physical task stress. As was noted in Campbell and Mohktari [28], this figure confirms the wide variation in neutral voice quality across speakers. The strong correlation between neutral mean NAQ and stress mean NAQ suggests that inter-speaker variation in baseline voice quality is significantly greater than the voice quality changes induced by the physical task stress itself.

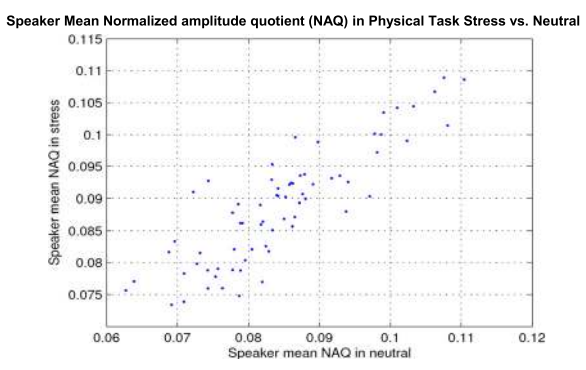

Fig. 16 Individual speaker differences in mean NAQ for both neutral and physical task stress

6 Conclusions

Physical task stress is an external factor imposed on the speech production system that competes for limited physical resources when subjects are performing simultaneous tasks (i.e., speaking and physical task). Speaking while exercising increases the actual and perceived difficulty of the task [3], and exercising while speaking results in significant changes to the fundamental frequency (F0), formant structure (location, etc.), pause placement, open quotient, and other measureable speech parameters. Based on this evidence, this study undertook an investigation of the voice quality changes induced by physical task stress, with particular attention paid to the speaker differences in the measured response. Compared with previous studies in the area of physical task stress [19, 20, 22] and voice quality [28, 36, 39], this study uses a larger corpus of speech, with speakers of varying fitness levels, with significantly more inter-subject variability, while retaining low phonetic variability.

We expected that physical stress would induce a greater variety of phonation behaviors. This might have resulted in a flattening of the NAQ and HRF histograms (i.e., becoming more uniform in distribution). Instead, in the global distributions of voice quality parameters for HRF and NAQ, rather small overall changes were observed, suggesting a corresponding small overall change in phonation behavior. However, for a subset of speakers, shifts in the mean values of NAQ and HRF were consistent with significant changes in voice quality, with trends toward either breathy or soft voice dimensions. These changes were not correlated with an elevated exertion level but were instead correlated with an increased fundamental frequency (F0). Research on speech and exercise has suggested that exercise results in both an increased vocal fold tension and increased subglottal pressure, relative to neutral speech production. This would suggest that, for those whom the voice quality is affected, the voice quality should move toward the pressed or tense voice, rather than the breathy or soft voice observed here in the current study. Further investigation of the relationship between physical changes caused by physical task stress and the voice quality changes is required in order to explain these results. It is in fact a major challenge to exactly measure physical airflow and actual excitation structure during speech production while subjects are performing physical tasks (i.e., without the measurement devices/instruments themselves introducing new variables into the problem).

It has been shown that listeners can perceive physical stress in speech [14], and therefore, there must be perceptual artifacts that consistently identify stressed speech across speakers. If voice quality is an inconsistent indicator of physical tasks stress, it is likely that inappropriate pause placement, formant shifts, and increased F0 play a more significant role in the perception of physical task stress than voice quality.

Finally, significant variation was observed in the baseline neutral measurements for NAQ and HRF across speakers. The *inter-speaker variation in baseline was significantly greater than the variation induced by physical task stress*. As observed in Godin et al. [22], this inter-speaker variation makes it difficult to consistently employ voice quality parameters individually for stress detection, and therefore, probabilistic classifiers may rely more on the correlation between these parameters than on the raw values of individual parameters themselves for detection of voice quality changes.

Competing interests

The authors are with CRSS - Center for Robust Speech Systems (UTDallas) and have competing interests.

Authors' contributions

KG and JH worked collaboratively on this study. KG performed the analysis of the features for speech under physical task stress and on the initial manuscript draft. He also contributed to updates in the review process. JH contributed to the technical logistics and manuscript writing and updates, as well as response to reviewers. Both authors read and approved the final manuscript.

Acknowledgements

This project was funded by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

Received: 23 December 2013 Accepted: 23 December 2013
Published online: 08 October 2015

References

1. J Deller, JHL Hansen, J Proakis, *Discrete-Time Processing of Speech Signals*, 2nd edn. (IEEE Press, New York, 2000)
2. AT Welford, *Stress and Performance*. Ergonomics, 2007, pp. 567–580
3. E Baker, J Hipp, H Alessio, Ventilation and speech characteristics during submaximal aerobic exercise. J. Speech. Lang. Hear. Res **51**, 1203–1214 (2008)
4. JH Doust, JM Patrick, The limitation of exercise ventilation during speech. Respir. Physiol. **46**, 137–147 (1981)
5. Y Meckel, A Rotstein, O Inbar, The effects of speech production on physologic responses during submaximal exercise. Med. Sci. Sports Exerc. **34**(8), 1337–43 (2002)
6. EF Bailey, JD Hoit, Speaking and breathing in high respiratory drive. J. Speech Lang. Hear. Res. **45**, 89–99 (2002)
7. JD Hoit, RW Lansing, KE Perona, Speaking-related dyspnea in healthy adults. J. Speech Lang. Hear. Res. **50**, 361–374 (2007)
8. JE Luketic, *The Effect of Inspiratory Muscle Strength Training on Ventilation and Dyspnea During Simultaneous Exercise and Speech* (Master's thesis, Miami University, Oxford, 2007)
9. SA Patil, *Alternate Sensor Based Speech Systems for Speaker Assessment and Robust Human Communication*. PhD thesis, CRSS: Center for Robust Speech Systems (The University of Texas at Dallas, Richardson, 2009)
10. JG Mohler, Quantification of dyspnea confirmed by voice pitch analysis. Bull. Eur. Physiopathol. Respir. **18**, 837–50 (1982)
11. JA Rodriguez-Marroyo, G Villa, J Garcia-Lopez, C Foster, Relationship between the talk test and ventilatory thresholds in well trained cyclists. J. Strength Cond. Res. **27**(7), 1942–1949 (2013)
12. A Rotstein, Y Meckel, O Inbar, Perceived speech difficulty during exercise and its relation to exercise intensity and physiological responses. Eur. J. Appl. Physiol. **92**, 431–436 (2004)

13. JA Rodriguez-Marroyo, J Garcia-Lopez, C-E Juneau, JG Villa, Workload demands in professional multi-stage cycling races of varying duration. Br. J. Sports Med. **43**, 180–185 (2007)
14. KW Godin, JHL Hansen, *Analysis and Perception of Speech Under Physical Task Stress*. ISCA INTERSPEECH-2008, 2008, pp. 1674–1677. Brisbane, Australia
15. HM Koblick, *Effects of Simultaneous Exercise and Speech Tasks on the Perception of Effort and Vocal Measures in Aerobic Instructors* (Master's thesis, Univ. of Central Florida, Orlando, 2004)
16. B Johannes, P Wittels, R Enne, G Eisinger, CA Castro, JL Thomas, AB Adler, R Gerzer, Non-linear function model of voice pitch dependency on physical and mental load. Eur. J. Appl. Physiol. **101**, 267–276 (2007)
17. RF Orliko, Voice production during a weightlifting and support task. Folia Phoniatr. Logop. **60**, 188–194 (2008)
18. RF Orliko, RJ Baken, The effect of the heartbeat on vocal fundamental frequency perturbation. J. Speech Hear. Res. **32**, 576–582 (1989)
19. KW Godin, JHL Hansen, *Vowel context and speaker interactions influencing glottal open quotient and formant frequency shifts in physical task stress*. ISCA INTERSPEECH-2011, 2011, pp. 2945–2948
20. KW Godin, JHL Hansen, Analysis of the effects of physical task stress on the speech signal. J. Acoust. Soc. Am. **130**, 3992–3998 (2011)
21. LG Olson, KP Strohl, The response of the nasal airway to exercise. Am. Rev. Respir. Dis. **135**(2), 356–359 (1987)
22. KW Godin, T Hasan, JHL Hansen, *Glottal Waveform Analysis of Physical Task Stress Speech*. ISCA INTERSPEECH-2012, Wed-SS6-15, 2012, pp. 1–4. Portland, OR
23. MHL Hecker, KN Stevens, G von Bismark, CE Williams, Manifestations of task-induced stress in the acoustic speech signal. J. Acoust. Soc. Am. **44**(4), 993–1001 (1968)
24. JHL Hansen, S Patil, Speech Under Stress: Analysis, Modeling and Recognition. Speaker Classification I: Fundamentals, Features, and Methods, (Springer Publishing, 2007), pp. 108–137
25. CT Ishi, *A New Acoustic Measure for Aspiration Noise Detection*. ISCA INTERSPEECH-2004, 2004. Jeju Island, Korea
26. C Gobl, AN Chasaide, Acoustic characteristics of voice quality. Speech Comm. **11**, 481–490 (1992)
27. N Campbell, *Changes in Voice Quality Due to Social Conditions*. Proc. Inter. Congress on Phonetic Science, 2007, pp. 2093–2096
28. N Campbell, P Mohktari, *Voice Quality: The 4th Prosodic Dimension*. Proc. Inter. Congress on Phonetic Science, 2003, pp. 2417–2430
29. JHL Hansen, *Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition* (PhD thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, 1988)
30. C Zhang, JHL Hansen, *Analysis and Classification of Speech Mode: Whispered Through Shouted*. ISCA Interspeech-2007, 2007, pp. 2289–2292
31. C Gobl, AN Chasaide, The role of voice quality in communicating emotion, mood, and attitude. Speech Comm. **40**, 182–212 (2003)
32. CE Williams, KN Stevens, Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. **52**(4B), 1238–1250 (1972)
33. L Gavidia-Ceballos, JHL Hansen, Direct speech feature estimation using an iterative EM algorithm for vocal cancer detection. IEEE Trans. Biomed. Eng. **43**(4), 373–383 (1996)
34. KE Cummings, MA Clements, Analysis of glottal excitation of emotionally styled and stressed speech. J. Acoust. Soc. Am. **98**, 88–98 (1995)
35. JHL Hansen, MA Clemments, Evaluation of speech under stress and emotional conditions. J. Acoust. Soc. Am. **82**, S17 (1987)
36. DG Childers, CK Lee, Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. **90**(5), 2394–2410 (1991)
37. P Alku, E Vilkman, A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. Folia Phoniatr. Logop. **48**, 250–254 (1994)
38. EB Holmberg, RE Hillman, JS Perkell, Glottal airflow and transglottal air pressure measurements for male and female speaker in soft, normal, and loud voice. J. Acoust. Soc. Am. **84**, 511–529 (1988)
39. G de Krom, Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. J. Speech Hear. Res. **38**, 794–811 (1995)
40. T Drugman, B Bozkurtb, T Dutoit, Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. Speech Comm. **53**, 855–866 (2011)
41. CT Ishi, K-I Sakakibara, H Ishiguro, N Hagita, A method for automatic detection of vocal fry. IEEE Trans. Audio Speech Lang. Process. **16**(1), 47–56 (2008)
42. E Moore, J Torres, A performance assessment of objective measures for evaluating the quality of glottal waveform estimates. Speech Comm. **50**, 56–66 (2008)
43. M Artkoski, J Tommila, A-M Laukkanen, Changes in voice during a day in normal voices without vocal loading. Logoped. Phoniatr. Vocol. **27**, 118–123 (2002)
44. AL Bouhuys, HK Schutte, DGM Beersma, GLJ Nieboer, Relations between depressed mood and vocal parameters before, during and after sleep deprivation: a circadian rhythm study. J. Affect. Disord. **19**, 249–258 (1990)
45. KE Cummings, MA Clements, *Analysis of Glottal Waveforms Across Stress Styles*. IEEE ICASSP-90: Inter. Conf. Acoustics, Speech, and Signal Processing, 1990
46. TF Yap, J Epps, EHC Choi, E Ambikairajah, TX Dallas, *Glottal Features for Speech-Based Cognitive Load Classification*. IEEE ICASSP-2010: Inter. Conf. Acoustics, Speech, and Signal Processing, 2010, pp. 5234–5237
47. M Lugger, B Yang, *Cascaded Emotion Classification via Psychological Emotion Dimensions Using a Large Set of Voice Quality Parameters*. IEEE ICASSP-2008: Inter. Conf. Acoustics, Speech, and Signal Processing, 2008
48. R Sun, E Moore, Affective Computing and Intelligent Interaction, vol. 6975 of Lecture Notes in Computer Science, chapter Investigating Glottal Parameters and Teager Energy Operators in Emotion Recognition, (Springer, 2011), pp. 425–434
49. SE Linville, J Rens, Vocal tract resonance analysis of aging voice using long-term average spectra. J. Voice **15**, 323–330 (2001)
50. J Gudnason, M Brookes, *Voice Source Cepstrum Coefficients for Speaker Identification*. IEEE ICASSP-2008: Inter. Conf. Acoustics, Speech, and Signal Processing, 2008
51. MD Plumpe, TF Quatieri, DA Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE. Trans. Speech. Audio. Process. **7**(5), 569–86 (1999)
52. JHL Hansen, *Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition*, 1989, pp. 31–32. Boston, Mass
53. JHL Hansen, C Swail, AJ South, RK Moore, H Steeneken, EJ Cupples, T Anderson, CRA Vloeberghs, I Trancoso, P Verlinde, *The Impact of Speech Under 'Stress' on Military Speech Technology, published by NATO Research & Technology Organization RTO-TR-10, AC/323(IST)TP/5 IST/TG-01*, 2000
54. JHL Hansen, SE Bou-Ghazale, G Zhou, R Sarikaya, *Speech Processing in Noise, Stress, and Lombard Effect, Research Monograph published by DoD, AFRL-IF-RS-TR-1999-208*, 1999
55. SE Bou-Ghazale, JHL Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE. Trans. Speech. Audio. Process. **8**(4), 429–442 (2000)
56. JHL Hansen, D Cairns, ICARUS: a source generator based real-time system for speech recognition in noise, stress, and Lombard effect. Speech Comm. **16**(4), 391–422 (1995)
57. JHL Hansen, M Clements, Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress. IEEE. Trans. Speech. Audio. Process. **3**(5), 407–415 (1995)
58. JHL Hansen, Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Comm. Special Issue Speech Under Stress. **20**(2), 151–170 (1996)
59. D Cairns, JHL Hansen, Nonlinear snalysis and detection of speech under stressed conditions. J. Acoust. Soc. Am. **96**(6), 3392–3400 (1994)
60. G Zhou, JHL Hansen, JF Kaiser, Nonlinear feature based classification of speech under stress. IEEE. Trans. Speech. Audio. Process. **9**(2), 201–216 (2001)
61. JHL Hansen, W Kim, M Rahurkar, E Ruzanski, J Meyerhoff, Robust emotional stressed speech detection using weighted frequency subbands. EURASIP J. Adv. Signal Process. Article ID 906789, 10 (2011)
62. JHL Hansen, E Ruzanski, H Boril, J Meyerhoff, TEO-based speaker stress assessment using hybrid classification and tracking schemes. Int. J. Speech Technol. **15**(3), 295–311 (2012)
63. T Drugman, M Thomas, J Gudnason, P Naylor, T Dutoit, Detection of glottal closure instants from speech signals: a quantitative review. IEEE Trans. Audio Speech Lang. Process. **20**, 994–1006 (2012)
64. T Drugman, A Alwan, *Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics*. ISCA INTERSPEECH-2011, 2011, pp. 1973–1976
65. J Kane, C Gobl, Evaluation of glottal closure instant detection in a range of voice qualities. Speech Comm. **55**, 295–314 (2013)
66. P Alku, T Backstrom, E Vilkman, Normalized amplitude quotient for parametrization of the glottal flow. J. Acoust. Soc. America. **112**, 701–710 (2002)

67. A Ikeno, V Varadarajan, S Patil, JHL Hansen, *UT-Scope: Speech Under Lombard Effect and Cognitive Stress*. IEEE Aerospace Conf.-2007, 2007, pp. 1–7. Big Sky, Montana
68. AL Webster, S Aznar-Lain, Intensity of physical activity and the "talk test". ACSM's Health. Fitness J. **12**, 13–17 (2008)
69. JA Davis, VA Convertino, A comparison of heart rate methods for predicting endurance training intensity. Med. Sci. Sports. 7, 295–298 (1975)
70. H Tanaka, KD Monahan, DR Seals, Age-predicted maximal heart rate revisited. J. Am. Coll. Cardiol. **37**, 153–156 (2001)