Analysis and Perception of Speech Under Physical Task Stress

Keith W. Godin, John H.L. Hansen

Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
godin@ieee.org, john.hansen@utdallas.edu

Abstract

It is known that speech under physical task stress degrades speech system performance. Therefore, an analysis of speech under physical task stress is performed across several parameters to identify acoustic correlates. Formal listener tests are also performed to determine the relationship between acoustic correlates and perception. To verify the statistical significance of all results, student-t statistical tests are applied. It was found that fundamental frequency decreases for many speakers, that utterance duration increases for some speakers and decreases for others, and that the glottal waveform is quantifiably different for many speakers. Perturbation of two speech features, fundamental frequency and the glottal waveform, is applied in listener tests to quantify the degree to which these features convey physical stress content in speech. Finally, the enhanced understanding of physical task stress speech provided here is discussed in the context of speech systems.

Index Terms: physical task stress, stress analysis

1. Introduction

In many application environments of speech technology, speakers experience some form of stress or emotion that ultimately affects the speech signal. Several emotion and stress types have been examined across a wide variety of parameters [1, 2], but speech under physical task stress, which can be found in sports and military applications of speech technology, has not generally been an area of active research. Physical task stress has been shown in one study to degrade speech system performance [3]. In addition, other forms of cognitive, emotional, and situational stress have been shown to vary significantly from neutral speech, as well as impact speech system performance [4]. However, to date no studies have offered an analysis of the parameters of the acoustic waveform of speech under physical task stress, a key starting point for developing speech systems that are robust to speech under physical task stress.

Some aspects of speech under physical task stress are quite different from other types of stress and emotional speech. Generally, speakers attempt to mitigate the effects of task stress on their speech, while balancing the needs of their respiratory system, unlike many instances of emotional speech, where the speaker may be allowing unconscious changes in their speech to take place, or may perhaps even actively modify their speech signal to communicate information about their state. Also, physical task stress is strongly correlated with a readily measurable non-speech aspect of the body, heart rate, while many other stress types are the result of unmeasurable changes in the body or cognitive state. This perhaps leaves open the possibility for systems that can quantify and adapt to the amount of physical task stress the speaker is experiencing.

In this study, we first introduce the physical task stress corpus used in the study. We then examine several aspects of the speech waveform to determine in what ways it changes under physical task stress relative to neutral speech. Next, listener tests are performed to evaluate the perceptual salience of two parameters found to be strongly correlated with physical task stress, by measuring listener performance on a binary classification task of "is this person exercising or sitting?". The test is performed on both unmodified speech, and on speech processed to have a mean fundamental frequency (F_0) or glottal waveform which is the same as the opposite task condition. Finally, we offer a discussion of our results and how they may help in speech system development, and directions for future research.

2. Corpus

The corpus used in this study is UT-Scope[5]. UT-Scope was collected at the University of Texas at Dallas and includes speech produced under four conditions: physical task stress, cognitive stress, Lombard effect, and neutral. To measure the amount of physical task stress induced in the speakers, heart rate (HR) in beats per minute (BPM) was also recorded for the neutral, cognitive stress, and physical task stress portions of the database. Figure 1 shows average heart rate versus task time for all speakers. The range of HR for neutral is 91–95 BPM, while in physical task stress it rises to the range of 110–133 BPM, confirming the increase in stress level from neutral for the physical task stress speech in UT-Scope.

The cognitive stress/physical task stress portion of UT-Scope has 118 total sessions, with 77 unique speakers. Of those who completed the physical task, 9 are male native speakers, and 42 are female native speakers of American English, self-reported. For each task, the subject was prompted to say the same 35 sentences that were played back to the subject. The physical task stress was induced using an elliptical stair stepper. Recordings were done in an ASHA certified double-walled soundbooth at a 44.1 kHz sampling rate with a close-talking mic, a throat mic, and a far field mic (approx. 1 m). Heart rate measurements were made throughout the recording sessions at 15 s intervals with an athletic heart rate monitor. A summary of the subset of UT-Scope used in this study is found in Table 1.

Sentence labels are available for the male and female native speakers of American English, with word and phone segmentation labels available for the females. Sentence labels were applied by human labelers, with word and phone labels applied using forced alignment.

This work was supported by a grant from TSWG and the Univ. of Texas at Dallas under Project Emmitt.

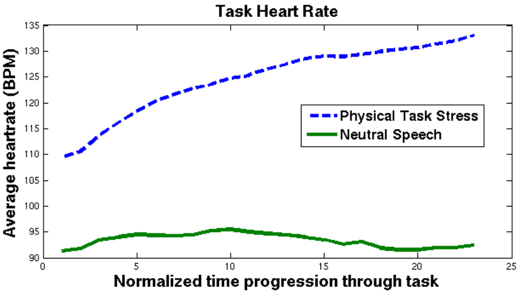

Figure 1: Average heart rate of all speakers versus task time (approx. 5min) for speakers under neutral and physical exertion.

Parameter	Male speakers	Female speakers
# of speakers	9	42
Average age (yrs)	22.3	23.6
Age range	19-33	18-45
Sentences/task	35	
Tasks	Neutral, Physical exertion	
Native language	American English	
Microphone	Close-talking	
Speech style	Prompted	

Table 1: Aspects of the subset of UT-Scope used in this study.

3. Analysis

In this section we analyze the physical task stress speech across several speech parameters to understand the ways the acoustic signal changes (or does not change) in physical task stress speech. We consider fundamental frequency (F_0) first, as it is a widely studied parameter found to vary under many conditions.

3.1. F_0 Distribution

The F_0 for each 10ms frame was computed with WaveSurfer [6] using the ESPS algorithm with an analysis window of 75ms, with the F_0 minimum set to 120 Hz for females and 80 Hz for males, and the maximum to 400 Hz. A distribution was formed for each condition (physical and neutral) for each speaker, selecting only F_0 values lying within a prompted utterance as labeled by human labelers. Two 1-sided t-tests, with a 99% confidence level, were used to compare the distribution means for each speaker.

Table 2 shows the analysis results. Here, 60.8 % of the speakers had a statistically significant increase in their mean F_0 under physical task stress, 13.6 % of the speakers had a statistically significant decrease in their mean F_0 under physical task stress, and 25.5 % of the speakers had no statistically significant change.

Statistical test result	% of speakers
F_0 greater in physical task	60.8
F_0 lower in physical task	13.6
F_0 same in physical task	25.5

Table 2: Comparing each speaker's mean F_0 within each condition at a 99% confidence level.

Statistical test result	% of speakers
F_0 σ greater under physical task	1.96
F_0 σ lower under physical task	23.5
F_0 σ same under physical task	74.5

Table 3: Comparing each speaker's distribution of utterance F_0 σ within each condition using two 1-sided t-tests at a 99% confidence level.

Statistical test result	% of speakers
Duration shorter under physical task	43.1
Duration longer under physical task	31.4
Duration same under physical task	25.5

Table 4: Comparing utterance duration of same utterance in physical vs. neutral conditions using two 1-sided t-tests at a 99% confidence level.

3.2. Standard Deviation of F_0 in an Utterance

Second, we consider the standard deviation, σ, of the F_0 within an utterance, to determine whether F_0 varies more or less under physical task stress. The standard deviation of the F_0 within each utterance was computed, yielding 35 measurements of the distribution of σ per condition per speaker. Two 1-sided student-t tests were used to compare the means of these distributions for each speaker to determine in which direction, if any, the utterance F_0 σ changed.

The analysis results are shown in Table 3. Only 25.4 % of the speakers showed a statistically significant difference in the mean of utterance F_0 σ, with 23.5 % of the speakers having a lower utterance F_0 σ. For most speakers (74.5 %), no statistically significant change in utterance F_0 σ was found. We conclude that physical task stress has a negligible effect on the short term variability most speakers impart in their F_0.

3.3. Utterance Duration

Next, we consider the duration of an utterance under physical task stress compared to the same utterance as spoken by the same speaker under neutral conditions. Two one-sided student-t tests were used to find if the difference in duration had a distribution with mean statistically greater or less than zero at the 99% confidence level.

Table 4 summarizes the analysis results, which show that 25.5 % of the speakers have no statistically significant difference in the duration of their sentences, while 31.4 % of the speakers spoke longer under physical task and 43.1 % spoke shorter. We conclude that duration of prompted sentences is often affected by physical task, but that the manner of the effect is speaker dependent.

3.4. Percentage of Voiced Frames in an Utterance

Next we consider the percentage of voiced frames in an utterance. As done for F_0 above, WaveSurfer was used to identify which 10ms frames of the recordings were voiced and which were unvoiced. Using the sentence labels obtained from human labelers, the percentage of voiced frames in each sentence was computed, yielding 35 measurements per speaker per condition. For each speaker, distributions for neutral and physical task stress speech were formed and the means of these distributions compared using two 1-sided student-t tests.

Table 5 summarizes the results. Approximately 88 % of the

Statistical test result	% of speakers
Greater % voiced in physical task	1.96
Lower % voiced in physical task	88.2
Same % voiced in physical task	9.80

Table 5: Testing whether speakers phonate fewer or greater frames per utterance in physical task, at a 99% confidence level.

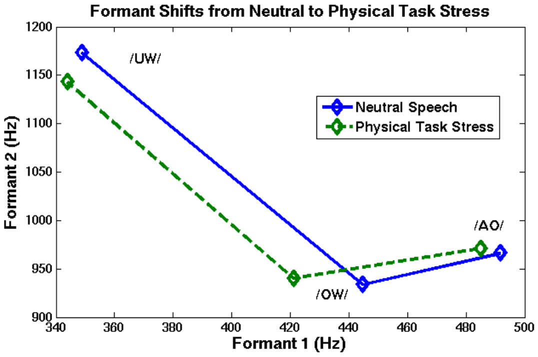

Figure 2: Three formants moving from neutral to physical task stress within the formant space.

speakers showed a statistically significant decrease in the percentage of frames in an utterance considered voiced, and just 1.96 % showed a statistically significant increase. We conclude that a reduction in the amount of voiced speech is a primary indicator of physical task stress in speech.

3.5. Shifts of Formants F1 and F2

We next consider the first two formants in an effort to determine whether their locations consistently shift under physical task stress. Phone alignments were used to extract the center 50 % of three vowels under both neutral and physical task stress conditions. WaveSurfer [6] was used to extract the formant locations, using a 12^{th} order LP analysis. The following results apply to sessions for which phone alignments are available, which comprise 37 female speakers.

A plot of the vowels in the two dimensional space of the first two formants is shown in Figure 2, where it can be seen that the formant space shifts inward. To test whether these shifts are statistically significant, a distribution was formed for each condition for each formant across all speakers, and the two distributions compared across conditions. Table 6 shows that just the shift of the first formant of /OW/ is statistically significant. We conclude that formant location shifts are not a primary indicator of physical task stress.

3.6. Glottal Volume Velocity Waveform

Finally, we consider the glottal volume velocity waveform. The glottal volume velocity waveform has been found to vary under a number of stressed speech styles, such as angry, loud, and Lombard [7]. The inverse filter method described in [8] is used to extract glottal waveforms from speech.

Phoneme	Shift in F1 significant?	Shift in F2 significant?
UW	-	-
OW	✓	-
AO	-	-

Table 6: Testing whether formant shifts are statistically significant (marked with X) for each vowel, at a 99% confidence level.

Statistical test result	% of speakers
Glottal duty cycle greater	47.1
Glottal duty cycle lower	27.5
Glottal rise time greater	19.6
Glottal rise time lower	72.5
Glottal fall time greater	43.1
Glottal fall time lower	41.2
Glottal spectral slope greater	47.1
Glottal spectral slope lower	19.6

Table 7: Testing whether speakers vary the shape and frequency content of their glottal waveform, at a 99% confidence level.

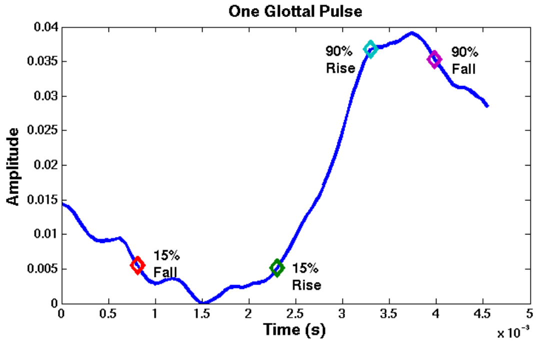

Figure 3: An example of an automatically extracted and labeled glottal pulse waveform from neutral speech.

To analyze the spectral content of the glottal waveforms, the glottal spectral slope is computed. To analyze the time-domain shape of the extracted glottal waveforms, we extract three parameters, the 15-90% rise time, the 90-15% fall time, and the 15-15% duty cycle. To extract these, first the rising and falling segments of the waveform are marked with the 15% and 90% points. The 15% rise point is found by moving earlier in time from the maximum to the first point with amplitude less than 15% of the maximum. Next the 90% rise point is found by moving later in time from the 15% rise point to the first point greater than 90% of the maximum. The falling points are computed similarly. An example extracted waveform is shown in Figure 3.

Once the rise and fall points are computed, the rise time, fall time, and duty cycle are found by subtraction. Student-t statistical tests were then used to compare the means of the distributions of the four parameters, including spectral slope. The results of the tests are shown in Table 7. The table shows several speaker dependent variations in the glottal waveform shape, though there is a tendency (72.5 %) towards decreased rise times.

4. Perceptual Tests

The analyses described in Sec. 3 revealed that changes in F_0 and the glottal waveform are strong correlates with physical exertion. Listener tests were therefore performed to determine the strength of these acoustic correlates as perceptual cues, and also to establish listener performance on a stress classification task. Here, 10 subjects were asked to classify 84 utterances, describing the speaker as either “exercising”, or “seated, resting”.

4.1. Listener Test Procedure

The physical task speech had heavy breathing surrounding the utterances, so as a first step all of the utterances were closely

Category	Perf.	Sig.
Unprocessed neutral	84.4 %	N/A
Unprocessed physical	68.9 %	N/A
Neut. $< F_0 >$ shifted to phy.	82.8 %	-
Phy. $< F_0 >$ shifted to neut.	44.4 %	✓
Replace neut. glottal waveform w/ neut.	48.9 %	✓
Replace phy. glottal waveform w/ phy.	66.7 %	-
Replace neut. glottal waveform w/ phy.	61.6 %	-
Replace phy. glottal waveform w/ neut.	71.7 %	-

Table 8: Results of listener tests and statistical comparisons.

cropped in time to remove their context. This helped to ensure that listeners made their classification decision based on the speech itself, and not on surrounding breath sounds.

To test the strength of F_0 and the glottal waveform as perceptual cues, some of the utterances were modified so that they were in some way shifted towards the counterpart utterance in the opposite condition. To shift the F_0 of some utterances, a PSOLA technique was applied so that the given utterance had a mean F_0 equal to the mean F_0 of the same speaker's utterance in the opposite condition.

For the glottal waveform tests, the glottal inverse filtering method described in [8] was used to extract glottal waveforms from the voiced portions of the speech, which were then replaced with waveforms extracted from the opposite condition. The utterances were then reconstructed by inverting the process and concatenating them with the unmodified unvoiced portions of the utterances.

The 84 utterances were partitioned into 8 groups. Two groups of 10 utterances, one from each condition, were left unprocessed. Two groups of 11 utterances were processed as described above to shift their pitch. Two groups, each comprising 10 utterances, served as control groups for the glottal processing technique. These were processed using glottal waveforms from the same condition. Two groups served as experimental groups, each of 11 utterances. These utterances were synthesized using glottal waveforms extracted from that speaker's opposite condition utterance. The order of all 84 utterances was randomized, and then presented to listeners in a formal test.

4.2. Listener Test Results

The listener test results are summarized in Table 8. The table shows that the listeners correctly classified 84.4 % of the unprocessed neutral utterances, and 68.9 % of the physical task stress utterances. Student-t tests were also used to make comparisons between the test results. The results for pitch shifted utterances were compared with those from unprocessed utterances of the same condition. Shifting the pitch of the physical stress utterances caused a statistically significant decrease in listener performance of more than 20 %. Shifting the pitch of the neutral speech did not have an effect on performance.

The results for the utterances which underwent glottal waveform replacement from the same condition were compared with unprocessed utterances to determine if the processing method had an effect on listener performance. The processing method did not have a statistically significant effect on the listeners' ability to mark utterances as physical task stress, though the processing decreased performance on neutral utterances to chance levels. In comparing the utterances with glottal waveforms swapped from opposite conditions to those with waveforms from the same conditions, no statistically significant shift was found.

5. Discussion

Several fundamental results regarding speech under physical task stress have been considered. A change in mean F_0 and a decrease in the percent of voicing of an utterance have been shown to be speaker independent acoustic correlates of physical task stress. Utterance duration and four parameters of the glottal waveform have been shown to be speaker dependent acoustic correlates of physical task stress. Formants, as well as the amount that a speaker varies his or her F_0 have been shown not to change due to physical task stress.

Furthermore, two results have been presented regarding the perception of physical task stress. It has been shown that listeners can achieve a strong, but not perfect, performance on classifying speech as being either physical task stress speech or neutral speech. It has also been shown that F_0 is a perceptual correlate of physical task stress speech.

The results presented in the analysis and perception of physical task stress in speech reflect meaningful changes in speech production which will help improve models for speech and speaker systems. The following observations can be made:

- *Speech coding*: Both F_0 and the glottal waveform structure will vary and therefore coding schemes must reflect these variations for the listener to perceive physical stress in the voice.
- *Speech recognition*: Though ASR systems deemphasize the effects of the excitation waveform, changes in the structure of the glottal waveform, which affect spectral slope, are expected to impact the word error rate (WER) of ASR systems. Also, a reduction in the percentage of voiced frames and changes in utterance duration will likely impact WER.
- *Speaker recognition*: Changes in F_0 and utterance duration will generally not impact speaker ID systems. However, changes in the overall glottal waveform structure and a reduction in voiced frames is expected to negatively impact speaker ID equal error rates (EER).

Future research in speech under physical task stress could consider compensation methods to address these variations for improved system performance in coding, ASR, and speaker ID.

6. References

[1] J. H. L. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*. PhD thesis, Georgia Inst. of Tech., July 1988.

[2] C. E. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates," *JASA*, vol. 52, pp. 1238–1250, Oct. 1972.

[3] M. S. Entwistle, *Training methods and enrollment techniques to improve the performance of automated speech recognition systems under conditions of human exertion*. PhD thesis, Department of Psychology, University of South Dakota, 2005.

[4] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, pp. 151–173, Nov. 1996.

[5] V. Varadarajan, J. H. L. Hansen, and A. Ikeno, "UT-Scope - a corpus for speech under cognitive/physical task stress and emotion," in *LREC Workshop on Speech Under Emotion*, May 2006.

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

[7] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *JASA*, vol. 98, pp. 88–98, July 1995.

[8] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *JASA*, vol. 90, pp. 2394–2410, Nov. 1991.