Prosodic Correlates of Individual Physiological Response to Stress

Serguei Pakhomov¹, Michael Kotlyar²

¹Center for Clinical and Cognitive Neuropharmacology
²Department of Experimental and Clinical Pharmacology
University of Minnesota, Minneapolis, USA
pakh0002@umn.edu, kotly001@umn.edu

Abstract

Response to stress is an important health risk factor. We compared several methods based on automatic speech analysis for extracting prosodic information from spontaneous speech of 19 subjects participating in a study of the effects of bupropion (a medication that increases smoking cessation rates and treats depression) on physiological measures of stress: plasma concentrations of epinephrine (adrenaline), heart rate and blood pressure. These findings indicate that automated speech analysis may be used for non-invasive stress response measurement.

Index Terms: spontaneous speech analysis, voice pitch, mental stress

1. Introduction

Stress induces measurable physiological changes in the human body (e.g., hormone secretion, heart rate, blood pressure, electrical skin conductivity) that have a profound effect on health[1-4]. In particular, the degree of physiologic reaction to stress is associated with the risk of developing cardiovascular disease, sleep disorders, overeating and obesity, immunodeficiency, and the ability to discontinue addictive behavior such as smoking and alcohol consumption[1].

The sympathetic nervous system responds rapidly to stress with the release of epinephrine (a.k.a. adrenaline) leading to increases in blood pressure and heart rate among other physiological effects that reflect the intensity of the response. In addition to physiological changes, stress elicits behavioral changes evident in speech and language use that are arguably more subtle and difficult to measure. While speech and language characteristics may be manually extracted from speech and analyzed with relatively high precision, such approaches are labor-intensive, expensive, and are not scalable to large numbers of individuals in a timely fashion. In healthcare in particular, ease and scalability of such measurements are critical. Furthermore, direct physiological methods of assessing stress reactivity via hormone blood levels are not only expensive but also invasive, potentially harmful and may not be well tolerated by some patients. Thus, it is important to develop non-invasive and easy-to-use instruments for assessing stress reactivity both in the clinical and research settings.

We introduce and compare three computational methods for assessing behavioral reactivity to stress by automatically analyzing spontaneous speech produced under experimental conditions. Our preliminary results suggest that automated methods can be used to extract characteristics that correlate with a physiologically important marker of epinephrine concentrations in the blood.

2. Related Work

Many of the previous studies of acoustic characteristics of speech under stress have relied on examination of existing corpora of speech including the Speech Under Simulated and Actual Stress (SUSAS) corpus[5, 6], NATO SUSC-0 military fighter and air-traffic-controller communications[7], and the Columbia-SRI-Colorado (CSC) corpus of deceptive and non-deceptive speech[8-10]. These corpora have been constructed by manually transcribing the speech at the phoneme-level or word level with subsequent automatic alignment with the audio signal.

A number of acoustic characteristics of speech were examined in these studies including the fundamental frequency (F0), intensity, duration, formant locations, and spectral slope features. F0 and its variability have been consistently identified as sensitive to the expression of emotions in speech since one of the earliest studies of this phenomenon[11]. These features were also found to be the most reliable in classifying stressed and unstressed speech[5, 12].

In addition to the acoustic speech characteristics, a number of researchers have examined distributional and prosodic characteristics of silent and filled pauses. In the context of deception detection truthful speakers were found to use fewer pauses (silent or vocal) than deceptive speakers[13]. The context of deceptive speech is somewhat different from the context of speech under stress; however, there is some overlap between these areas, as the act of deception may be viewed as a stressful condition, at least for some speakers. Thus, examination of the silent pause characteristics such as their distribution and duration in the context of speech under stress is warranted.

Most of the computational linguistic studies of speech under stress evaluate the performance of their approaches on subjectively defined reference standards (e.g., manual classification of voice recordings by human judges[12]). In contrast, the evaluation in our study relies on one of the most objective markers of stress – increased epinephrine secretion, in addition to less direct measures such as heart rate and blood pressure.

3. Methods

Our study relies on data collected during a pharmacologic study of the effects of bupropion (a medication that increases smoking cessation rates and treats depression, brand name Zyban) on physiological measures of stress response[1].

3.1. Subjects

Participants were recruited from the University of Minnesota and surrounding communities. Only subjects between 18 and 65 years of age that have smoked at least 15 cigarettes per day and were generally healthy (e.g. no current...
psychiatric illness, no serious unstable medical conditions) were included. Subjects with contraindications to bupropion administration or on medication likely to interact with bupropion or interfere with study measures were excluded.

Stress response for each subject was tested during two laboratory sessions (L1 and L2), the first of which occurred prior to smoking cessation and prior to medication administration and second of which occurred on the third day of smoking abstinence (which corresponded to approximately 17 days of medication). Subjects were randomly assigned to receive either bupropion 150 mg twice daily (N=14), matching placebo (N=15), or bupropion with behavioral counseling (N=14). Forty-three subjects completed the entire study of whom 19 (8 men) had speech recording of adequate quality to complete the speech analysis testing described in this report.

3.2. Stress response elicitation

During each laboratory assessment (separated by approximately 4 weeks) subjects performed a mental stress task that consisted of giving a public speech on a predefined topic. All visits occurred during the same time of day – between 8 and 10 a.m. For both visits, a modified Trier Social Stress Test was utilized[14] in which the subjects were presented with a hypothetical stressful scenario involving an interpersonal conflict. For example, on the first visit, the subjects were asked to speak on handling a conflict with a roommate. They were given 5 minutes to formulate their thoughts without taking notes and then were asked to give a 3-min speech in front of 2 people and a videocamera.

Systolic and diastolic blood pressure and heart rate were measured with an automated sphygmomanometer throughout the laboratory session; at 3-min intervals during 30 min relaxation periods and at 1-min intervals during the speech. These measurements were then averaged during the relaxation and speech periods separately to obtain two data points (“at rest” and “during speech”) for each variable.

Epinephrine and norepinephrine levels were measured by blood draw with an indwelling vein catheter that was inserted at the very beginning of the visit to allow the subjects to adjust to it and to allow for drawing blood with minimal interference with the speech session. One blood draw was taken at the conclusion of the first relaxation period and another one 1.5 minutes into the speech. A standard and validated analytical method was used to determine the concentration of epinephrine and norepinephrine in the blood stream. Further details can be found in [1].

3.3. Speech Processing

For the semi-automated approach, the speech collected from subjects was manually transcribed verbatim and subsequently forced-aligned with a Hidden Markov Model speech recognizer (HTK-3.4) [15] using a Mel-frequency Cepstral Coefficient (MFCC) acoustic model (5 HMM states, 11 MFCC parameters with delta and acceleration coefficients) trained on 6.5 hours of spontaneous dialogues contained in the TRAINS corpus[16]. For the fully automated voice-to-phoneme conversion approach, we created a simple phoneme-loop language model based on a vocabulary of 39 phonemes.

The speech recordings in our current study were not originally intended for subsequent acoustic analysis. Therefore, the recordings were made on analog tapes using a standard cassette tape recorder with an omnidirectional condenser lavaliere microphone (RadioShack 33-3013) and subsequently digitized as 16 bit 16 kHz PCM waveform.

We used Praat, a system designed for acoustic analysis[17], for manual examination of the speech signal and to automatically obtain fundamental frequency information. Based on the speech recognizer’s output from these two approaches, we calculated the following variables to represent the acoustic and prosodic characteristics that, based on prior work, are likely to be sensitive to the mild stress induced in the current experiment.

Mean Fundamental Frequency (m-F0) was calculated using Praat. The autocorrelation based algorithm [18] with default parameters of 0.01 seconds as the time step and 75 Hz as the pitch floor was used to estimate the mean pitch of the speech sample.

F0 variability (v-F0) was calculated as the mean variability in F0 estimates computed as specified in the description of m-F0 variable above.

Mean silent pause duration (m-SPD) was calculated by first log transforming the duration of all silent pauses greater than 150 milliseconds and then averaging the log-transformed values. Log transformation was used following the results of a recent study by [19] showing that pause durations computed in real time have a long-tailed distribution that may mask an underlying bi-modal distribution evident when log time is used. The 150 milliseconds threshold was chosen conservatively to avoid treating as hesitations phonetically conditioned pauses such as the release phase in the phonation of a word-final stop consonant [20].

We experimented with three approaches of distinguishing between silent and speech segments of the audio to calculate the F0 and pause duration variables: a) semi-automated approach (semi) based on forced alignment, b) fully automated approach (auto) based on voice-to-phoneme recognition using a phoneme loop language model, and c) default functionality of Praat. For the latter, we used intensity based silence detection functionality provided by Praat (praat) with the optimal intensity threshold of -25 dB for our data determined empirically and the minimum sounding and silent interval durations set to 100 milliseconds.

3.4. Statistical Analysis

Speech characteristics are highly variable and require speaker normalization. One of the limitations of the available speech data is that, by design, the speech task was used as a stressor and therefore no speech samples were available from the relaxation periods to establish a baseline.

However, due to the fact that the study had two visits, we were able to use the speech from one of the visits (L2) for normalization. Since there were differences in physiological measures observed between the two laboratory sessions, a ratio was calculated for each of the physiological measures of interest. Thus, our statistical analysis was based on correlating the relative changes in physiologic and speech variables rather than the absolute values of these variables.

Comparisons of group means were tested using the two-tailed Student’s t-test for paired samples. Correlations were computed using the Pearson’s correlation. Associations between multiple variables were tested with linear regression. We assumed that the sex of the subjects is associated with the pitch of their voice. It is also known that the treatment condition has an effect on the subjects stress reactivity[2]. Therefore we controlled for both sex and treatment condition by including them as a covariates in the regression modeling.

4. Results

The results of comparisons between physiological and speech variables between the two laboratory visits are provided in Table 1.
Table 1. Comparisons between physiological and speech variables obtained at two laboratory sessions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>L1 (mean, sd)</th>
<th>L2 (mean, sd)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physiological variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epinephrine (ng/ml)</td>
<td>45.3(28.5)</td>
<td>34.3(19.6)</td>
<td>0.02</td>
</tr>
<tr>
<td>Norepinephrine (ng/ml)</td>
<td>373.0(99.8)</td>
<td>313.7(98.6)</td>
<td>0.004</td>
</tr>
<tr>
<td>Heart Rate (bpm)</td>
<td>78.1(11.1)</td>
<td>72.5(8.7)</td>
<td>0.01</td>
</tr>
<tr>
<td>Systolic BP (mm Hg)</td>
<td>123.5(17.6)</td>
<td>115.4(10.3)</td>
<td>0.02</td>
</tr>
<tr>
<td>Diastolic BP (mm Hg)</td>
<td>71.6(9.7)</td>
<td>68.5(7.0)</td>
<td>0.11</td>
</tr>
<tr>
<td>Semi</td>
<td>151.4(32.5)</td>
<td>150.8(36.2)</td>
<td>0.88</td>
</tr>
<tr>
<td>Auto</td>
<td>6.15(0.15)</td>
<td>6.15(0.16)</td>
<td>0.98</td>
</tr>
<tr>
<td>Praat</td>
<td>150.6(34.9)</td>
<td>150.3(36.9)</td>
<td>0.94</td>
</tr>
<tr>
<td>m-F0</td>
<td>7.94(1.88)</td>
<td>8.00(2.64)</td>
<td>0.89</td>
</tr>
<tr>
<td>v-F0</td>
<td>6.21(0.15)</td>
<td>6.26(0.15)</td>
<td>0.19</td>
</tr>
<tr>
<td>m-SPD</td>
<td>148.1(34.3)</td>
<td>148.2(36.2)</td>
<td>0.97</td>
</tr>
<tr>
<td>v-SPD</td>
<td>58.7(27.4)</td>
<td>54.6(22.5)</td>
<td>0.40</td>
</tr>
<tr>
<td>m-SPD</td>
<td>5.97(0.12)</td>
<td>6.02(0.15)</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Overall, we observe statistically significant differences between the visits on physiological variables but not the speech variables. The direction of the change in physiological variables is positive, indicating that the subjects were less stressed in the second laboratory session compared to the first session.

4.1. Relationship of speech variables to physiology

**Table 2.** Pearson correlation coefficients for associations between variables.

<table>
<thead>
<tr>
<th>Var.</th>
<th>Epi</th>
<th>NEpi</th>
<th>HR</th>
<th>SBP</th>
<th>DBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>m-F0</td>
<td>0.59**</td>
<td>0.15</td>
<td>0.22</td>
<td>-0.26</td>
<td>-0.14</td>
</tr>
<tr>
<td></td>
<td>0.61**</td>
<td>0.30</td>
<td>0.21</td>
<td>-0.10</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>0.63**</td>
<td>0.27</td>
<td>0.24</td>
<td>-0.09</td>
<td>-0.06</td>
</tr>
<tr>
<td>v-F0</td>
<td>0.51*</td>
<td>-0.01</td>
<td>-0.14</td>
<td>-0.41</td>
<td>-0.27</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>-0.13</td>
<td>-0.03</td>
<td>-0.33</td>
<td>-0.21</td>
</tr>
<tr>
<td></td>
<td>0.59*</td>
<td>0.05</td>
<td>-0.04</td>
<td>-0.39</td>
<td>-0.37</td>
</tr>
<tr>
<td>m-SPD</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.34</td>
<td>-0.53*</td>
<td>-0.57*</td>
</tr>
<tr>
<td></td>
<td>-0.10</td>
<td>-0.20</td>
<td>-0.19</td>
<td>-0.51*</td>
<td>-0.48*</td>
</tr>
<tr>
<td></td>
<td>-0.15</td>
<td>-0.13</td>
<td>-0.31</td>
<td>-0.45</td>
<td>-0.30</td>
</tr>
</tbody>
</table>

* - indicates p-value < 0.05; ** - p-value < 0.001

The results in Table 2 show significant and strong correlations between the relative change in F0 means (m-F0) and the relative change in epinephrine concentrations (Epi) from L1 to L2. The only exception is the relationship between variability in F0 (v-F0), obtained using the fully automated approach, and epinephrine concentrations. We also observed significant correlations between the mean silent pause duration (m-SPD) and both the systolic and diastolic blood pressure measurements (SBP and DBP). However, this relationship is weaker and not significant with the Praat based method for measuring silent pause duration.

Correlations among the physiological variables show a significant but weak correlation between the relative change in heart rate (HR) and relative change in epinephrine concentrations. With the exception of relative change in F0 variability (v-F0), both semi-automated and fully automated approaches yielded similar results with respect to pitch-related measurements.

**Figure 1** Association between relative change in epinephrine concentrations and F0.

The Praat-based approach for measuring pitch resulted in better correlations with epinephrine concentrations; however, the semi and auto approaches resulted in better correlations between silent pause duration and blood pressure.

**Figure 2** Association between relative change in systolic blood pressure and mean pause duration.

4.2. Physiological correlates of change in pitch

The plot in Figure 1 shows the relationship between the relative change in epinephrine concentrations and F0 obtained with the fully automated approach. A linear regression model controlled for the treatment condition (placebo vs. bupropion with and without counseling) and the sex of the subjects showed a significant effect of the relative change in F0 ($t = 2.47, p = 0.025$). The shallow slope of the regression line indicates that relatively large changes in epinephrine concentrations are necessary to observe small changes in pitch.

4.3. Physiological correlates of change in pause times

The plot in Figure 2 illustrates the inverse relationship between relative change in mean pause duration and systolic blood pressure. Regression, with the treatment condition and sex of the subjects included as covariates in addition to the mean pause duration variable, showed that the latter had a significant effect on relative change in systolic blood pressure. The slope of this relationship was negative indicating that a
relative increase in the mean pause length was associated with a relative decrease in blood pressure.

In addition to correlation tests, we confirmed that the associations between the speech and physiologic variables were not confounded by the treatment condition and sex of the subjects. The speech variables in the regression models that also included treatment condition and sex did show significant effects while the treatment condition and sex variables did not.

A comparison between mean pause durations for the three silence detection methods showed a significant difference in means (6.16 vs. 5.97) between the intensity based silence detection and the HTK-based forced alignment (p = 0.004) and HTK-based voice-to-phoneme conversion methods (p = 0.0001), but no difference (6.16 vs. 6.20) between the latter two methods (p = 0.21).

5. Discussion

The main findings of this study include a significant association between spontaneous speech characteristics under mild stress and epinephrine concentrations, an established marker of stress reactivity. We have also demonstrated that both automatically and semi-automatically extracted speech/prosodic characteristics such as F0, variability in F0, and log-normalized mean pause duration are sensitive to even the mild stress manifestations induced in our experiments. These findings are important because measuring stress reactivity by examining the patient’s speech rather than by laboratory analysis of stress hormones concentrations is considerably less expensive and less invasive.

Our correlation results (Table 2) clearly indicate a linear association between changes in pitch and changes in epinephrine levels; however, relatively large changes in epinephrine levels are required to elicit small changes in voice pitch, as evidenced by the slope of the regression line shown in Figure 2. This is an interesting finding in the context of the group means for the variables examined in this study (Table 1) not showing significant differences between study sessions. This indicates that while we may not be able to detect a response to mild stress based on speech characteristics at the group level, we may be able to do so at the individual speaker level. Furthermore, our results indicate that using epinephrine levels is a more sensitive approach as compared to less direct methods such as heart rate and blood pressure. If we only relied on heart rate and blood pressure in this study, we would have missed the association between acoustic characteristics (m-F0 and v-F0) and stress response.

Using Praat’s intensity based silence detection for measuring the mean silent pause duration did not correlate with physiological blood pressure variables as well as using the voice-to-phoneme conversion. We believe this was due to the intensity based silence detection method not being as robust as HMM-based acoustic modeling with respect to the low-level noise in the audio signal.

The results of this study must be interpreted in the context of several limitations. The audio equipment was not optimized for speech analysis as this was not part of the original study’s data analysis plan. Also, the treatment condition (placebo vs. bupropion) may have an effect on speech characteristics independent of stress. We addressed this potential limitation by including treatment condition along with the sex of the subjects as covariates into regression models to examine the relationships between speech and physiological variables. We did not observe independent effects of the treatment condition; however, a larger study is necessary to draw more definitive conclusions.

6. Acknowledgements

We thank Elizabeth Amiot for help with data collection, and the Media Workshop Inc., Roseville MN for providing audio digitization equipment and expertise. This work was supported by the University of Minnesota Academic Health Center Faculty Research Development grant and NIH grants: #M01-RR00400 and #K23DA017307.

7. References