Stressed out: What speech tells us about stress

Will Paul

Golisano College of Computing & Information Science
Rochester Institute of Technology
{whp3652′, coagla′, rjb′, jmg′, lxwast′}@rit.edu, cs.rit.edu

Cecilia Ovesdotter Alm

Reynold Bailey

Joe Geigel

Linwei Wang

College of Liberal Arts
Rochester Institute of Technology

Abstract

Stress can have a negative and costly impact on people’s lives. Mitigating stress before it becomes a problem requires early, noninvasive identification and a deeper understanding of the signals of stress. To test automatic stress detection a new dataset was created with subjects completing the Stroop task under un-stressed and stressed conditions. This paper examines to what degree speech features respond to stress and if so, what features are most informative. Features were extracted from recorded speech data and trained with several classification algorithms. We explored binary classification of stressed vs. unstressed across gender and per gender, with the best results on a held-out test set improving over the majority class baseline (MCB) by 16% across genders and with 20% and 21% for the female and male subsets respectively. Overall maximum intensity emerged as the most informative feature when comparing across classification conditions. In addition, we explored leave-one subject-out classification, resulting in a 15% improvement on average considering both genders when using random forests.

Index Terms: speech as cognitive marker, stress, Stroop task, stress detection

1. Introduction

Psychological stress has been shown to have a negative effect on cognition as well as mental and physical well-being [1, 2]. It also adversely impacts productivity in the workforce. In particular, stress has been linked with impaired problem-solving [3] and poor organizational practices [4]. Understanding the underlying signals and physiological factors that indicate whether a person is stressed is the first step towards prevention.

Existing tests for stress detection often require expensive or invasive laboratory tests such as MRI scans [5] or saliva samples [1]. Research into noninvasive approaches have primarily focused on biophysical sensors such as heart rate variability [6] or galvanic skin response [7]. Speech-focused research has mostly been evaluated within the context of automatic speech recognition (ASR) systems [8, 9], without the wellness of the end user in mind. One advantage of using speech data is that it is one of the least invasive and most natural signals to collect.

This paper explores speech data collected experimentally under un-stressed and stressed conditions to determine its effectiveness for stress detection.

2. Related work

One of the first speech datasets focused on stress was the Stress Under Simulated and Actual Stress corpus [10], comprising data from 32 individuals, under different situations primarily with a closed set of words from the avionics domain. One study found that humans have trouble classifying emotional state based on single words, achieving 58% accuracy on average over a 13% baseline [11]. A computational approach attempted to discriminate medium and high stress speech from neutral, fear, and screaming [12]. They achieved just under 75% and 60% accuracy for high and medium stress respectively over a 20% baseline. While impressive this seems skewed by the straightforward fear and screaming classes. In particular the accuracy on the screaming class approached 100%.

Another stress study examined speech data from (mostly male) helicopter pilots talking to air traffic control followed by high stress speech as they were about to crash [13]. Their analysis of pitch found that across speakers only maximum pitch and to a lesser extent mean pitch were useful. An important distinction in this study was that the stress expressed in their data was, ‘closer to terror than to task-induced anxiety.’

The search and rescue corpus for stress detection used a collaborative task to gather data from 2 subjects who work together remotely, communicating via handheld transceivers [14]. Stress was induced three-quarters of the way through by introducing a time limit and an additional task. A classifier trained on data from 8 subjects (7 males, 1 female) achieved 76% accuracy, which represented an 8% improvement over the MCB (the dataset was weighted towards the unstressed case) [15]. When tested with data from held-out subjects the algorithm performed worse, even below their MCB when averaged across subjects. They found that maximum/mean intensity and mean/median pitch to be the most important features.

3. Experiment design and data

A limitation of most existing speech datasets for stress detection was that they only consider spoken data [10, 13, 14]. To address this limitation the data collection experiment was structured such that other sensors could be included. We conducted this experiment in a naturalistic work environment, specifically at a stationary desk in a relatively quiet office. We used consumer grade sensor equipment, a standard lavalier microphone and recording device, to simulate a real-world context where controlled studio recording is not possible.

Prior research has shown that it is difficult to discriminate between cognitive load and stress [16]. Thus, we selected the Stroop task [17], which is used to establish cognitive load. In this task the subject is shown a color word written in a font color different than the word itself. The subject is then tasked with saying the color of the font (e.g. if the word black is written in a red-colored font, the correct response is ‘red’). The Stroop task has been shown to be a challenging task for fluent speakers, be-

---

1Zoom H1 Handy Portable Digital Recorder.
cause different parts of the brain are involved in the processing of language and color [18].

Besides the regular Stroop task, we added a second, modified Stroop task, in which stressors were introduced on top of cognitive load. Accordingly, the Stroop task was conducted in two separate trials: an unstressed version where the subjects were given unlimited time to respond and a stressed version where they had 1.5 seconds to respond and for every incorrect or late answer they lost $0.75 of their $10 bonus (see Figure 1 for a screenshot). A practice trial was first performed to familiarize the subject with the task. There was also a rest period between the following two trials to minimize the chance that the previous trial would impact the next (see Figure 2 for an overview).

Each trial had a total of 35 lexical items and was collected from 27 subjects, for a total of 1,890 data points. The subjects were graduate and undergraduate students ages 18 to 32, comprising 16 males and 11 females with 10 of the subjects being non-native English speakers. All but three participants self-reported an increase in stress between trials, and there was a general upwards trend in the stressed trial (see Figure 3). We explored classification across gender versus by gender, resulting in three data subsets: female and male, female, and male.

### 4. Methodology

The processing of the speech data was done automatically with Praat [19]. Utterance boundaries were marked through silence detection and time-aligned with the written transcripts. As the speech data from the Stroop experiment is highly structured.

Figure 1: A screenshot of the stressed version of the Stroop task used in our data collection experiment. The unstressed version did not have the progress bar or reward counter.

Figure 2: A bar graph showing the distribution of subjects' self-reported stress levels across the trials.

Figure 3: The post-experiment survey asked subjects to rate their stress level on a scale from 1 to 5 for each trial. This plot shows that in general, as expected, the subjects’ perception of their own stress levels increased in the stressed trial.

<table>
<thead>
<tr>
<th># Subjects</th>
<th>Self-Reported Stress Level (1-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstressed Trial</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Features automatically extracted from the speech data with Praat [19]. Gender was considered for pitch ranges when extracting pitch information.

Table 2: The top five features by information gain for each subset of data. Max intensity is the only feature to occur in each.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Female &amp; Male</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Int.</td>
<td>Min Int.</td>
<td>Q1 Pitch</td>
<td>Min Pitch</td>
</tr>
<tr>
<td>Mean Abs. Slope</td>
<td>Pitch Mean Abs. Slope</td>
<td>Mean Pitch</td>
<td></td>
</tr>
<tr>
<td>Time of Min Int.</td>
<td>Q1 Int.</td>
<td>Max Int.</td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td>Q3 Int.</td>
<td>Min Int.</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Features related to time that were included in the classification experiment included a time constraint to induce stress. The only features related to time that were included in the classification were related directly to pitch and intensity and were normalized to avoid encoding durational relationships. Syntactic and semantic features were also unavailable since the data consists of a fixed set of isolated color types; we leave this for future work. Disfluencies like repairs, fillers, and false starts were considered, but were too infrequent in the data to influence results.

Standardization of feature sets has improved speech classification in the past, in one study increasing accuracy by 11-18% [15]. For this reason standardization was implemented per person, gender, and across the whole dataset. These options were examined as an experimental parameter in a classification tuning phase. The standardization method implemented used the mean and standard deviation for each feature to convert each data point into z-scores, which indicate the number of standard deviations between an individual data point and the mean.

Feature selection was also tested as an experimental parameter during a classification tuning phase. This was done to automatically eliminate low-variance features and to compare feature importance outside of the context of a given classifier. In addition to including all features, two Scikit-learn [20] feature selection algorithms were tested: information gain from an entropy-based decision tree and a linear SVM-based approach [21]. The top five features for each data subset based on information gain are listed in Table 2.

Four machine learning algorithms were explored to model this binary classification task, including random forest, k-neighbors, decision tree, and naive Bayes, all using Scikit-learn’s implementations [20]. Each data subset (female, male, and both) was split randomly into 80% train and 20% test, while...
5. Results

In the first classification scenario, the k-neighbors classifiers had more consistent performance across data subsets (see Figure 4). As expected the simple naive Bayes classifier had the worst accuracy across the board. Somewhat surprisingly the decision tree slightly outperforms the random forest classifier in two data subsets. A visualization of the decision tree for the female and male dataset is shown in Figure 6. In addition, standardizing features with z-scores slightly improved results across data subsets, regardless of the data subset used to calculate them, though in the end the best performing classifiers all used z-scores calculated across the entire data subset. Feature selection experimentation did not show consistent behavior; it varied slightly between algorithm, inclusion/exclusion, and dataset.

The best results per subset of data are as follows: k-neighbors for the whole dataset with an accuracy of 66%, k-neighbors for the male dataset with an accuracy of 71%, and random forest for the female dataset with an accuracy of 70% (see Table 4 for confusion matrices for each). All demonstrate substantial improvement over the baseline. The experimentation also illustrates the challenge of combining speech data across genders, as even after standardization the features show wide variation. See Table 5 for detailed precision and recall.

Average subject accuracy per algorithm from the LOSOCV experiment is shown in Figure 5. The random forest algorithm performed best for each data subset and maintained similar accuracy scores to the previous experiment. This suggests that it is the more robust approach for new speakers. This also shows that even though k-neighbors consistently performed well on the held-out test set, it does substantially worse on subjects it had not seen before, implying it overfits and does not generalize well to new speakers. Both the decision tree and naive Bayes classifiers experienced a less prominent drop in improvement (15% to 10% and 8% to 5% respectively on the female and male data).

Interestingly, regardless of approach the LOSOCV models consistently excel and struggle with the same subjects. Specifically there were 5 subjects that none of the classifiers improved over the MCB and there were 5 that did so by 18% or more (see Table 3 for the 3 best and worst). Misclassification could be because the subjects were not actually stressed (see subjects E and F who did not report an increase in stress for the stressed trial), but that is not the whole story as subject D reported the largest increase of stress in the dataset, yet was classified incorrectly across genders. This also might indicate that individuals react differently to stress in a way that standardization can not counteract.

A potential limitation in the dataset was that across both Stroop trials (unstressed vs. stressed) most subjects did not make any mistakes in naming the font colors. A few subjects...
features, including maximum intensity as the root node. This classifier achieved a 15% improvement over the MCB on the held-out test set. This points to the importance of intensity.

Figure 6: The decision tree created for both genders. Left nodes are accepted when the parent condition is true and right when it is false. S means the input is classified as stressed, U unstressed.

Table 4: Confusion matrices (columns are predicted and rows are gold standard) for the algorithm with the best performance on each held-out test set.

Table 3: The 3 best (A-C) and worst (D-F) performing subjects by average accuracy improvement over MCB from the LOSOCV experiment, along with self-reported stress levels (1-5) for the unstressed trial (U), and the stressed trial (S).

Table 5: The precision (P) and recall (R) of the best models for each data subset, for the stressed and unstressed trials.

mentioned that to make the task easier they let the screen go out of focus so the word became a blob of color. Another possible reason for the lack of mistakes is that 10 of the subjects were non-native English speakers, which depending on their fluency, might make the task easier as they might not internalize the words’ meaning in the same way. Yet, many realistic contexts will involve non-native speakers, so their inclusion is important. Overall, it will be useful to continue to explore different methods to gather similar data. Also, in this study, the feature standardization process considered for the full data subsets; future work could explore the impact of only considering features between the three data subsets, with one feature showing up in the top 5 of all three (maximum intensity). This limited overlap could reflect that pitch is notably variable in females (even after standardization) such that its effectiveness is limited, while for men it seems as useful as intensity measures. This matches the findings of two other studies [13, 15], whose datasets were weighted towards males.

Some spectral and formant features were useful for another stress detection research study [15] (though less important than pitch and intensity), so it may be worthwhile exploring its effect on our data. That said, another study warns that spectral features often encode too much phonetic content and end up causing the model to overfit to the content [22].

The LOSOCV experiment illustrates the challenge of speaker variability, which is in line with the results of [15]. In some cases subjects have the opposite response to the two trials. This may be due to a variable outside of stress or that the subjects are not actually stressed. This is an issue with relying on task-induced labels in general and in the future using biophysical sensors to help verify actual stress levels may help to address this. Even so, it is likely that people can respond differently to the same pressures and any detection system will need to handle this variability.

Moving forward one aspect to explore may be personalized, adaptive models that use the speaker’s differences to its advantage, rather than aiming to remove them through standardization. This method has been used to some success in the biophysical and medical domain using for example personalized feature mapping [23] or continuously updating neural networks [24]. Another approach is to aggregate data from a number of multimodal sensors, the thought being that even if one modality does not work for a particular individual some other modality in the system will. Since biophysical and behavioral data was collected for this research alongside speech data both avenues are available for future work.

6. Conclusion and future work

A new dataset was created and was classified with promising improvements over MCB, achieving good results even with an approach as simple as the decision tree in Figure 6. K-neighbors tended to do best on random held-out test, but failed to work well with data from previously unseen speakers in the LOSOCV experiment. The random forest classifier on the other hand maintained stable accuracy rates in both cases, showing that it is probably a more robust approach.

These results demonstrate that pitch and intensity features can be useful for predicting stress. There was some flux in the top features between the three data subsets, with one feature showing up in the top 5 of all three (maximum intensity). This limited overlap could reflect that pitch is notably variable in females (even after standardization) such that its effectiveness is limited, while for men it seems as useful as intensity measures. This matches the findings of two other studies [13, 15], whose datasets were weighted towards males.

Some spectral and formant features were useful for another stress detection research study [15] (though less important than pitch and intensity), so it may be worthwhile exploring its effect on our data. That said, another study warns that spectral features often encode too much phonetic content and end up causing the model to overfit to the content [22].

The LOSOCV experiment illustrates the challenge of speaker variability, which is in line with the results of [15]. In some cases subjects have the opposite response to the two trials. This may be due to a variable outside of stress or that the subjects are not actually stressed. This is an issue with relying on task-induced labels in general and in the future using biophysical sensors to help verify actual stress levels may help to address this. Even so, it is likely that people can respond differently to the same pressures and any detection system will need to handle this variability.

Moving forward one aspect to explore may be personalized, adaptive models that use the speaker’s differences to its advantage, rather than aiming to remove them through standardization. This method has been used to some success in the biophysical and medical domain using for example personalized feature mapping [23] or continuously updating neural networks [24]. Another approach is to aggregate data from a number of multimodal sensors, the thought being that even if one modality does not work for a particular individual some other modality in the system will. Since biophysical and behavioral data was collected for this research alongside speech data both avenues are available for future work.

7. Acknowledgment

Thanks to the group and fellow researchers Brendan John, Va-sudev Bethamcherla, Taylor Kilroy, and Kritihka Sairamesh. This work was supported by a Golisano College of Computing and Information Sciences Kodak Endowed Chair Fund Health Information Technology Strategic Initiative Grant.
8. References


