An Evaluation of the Effect of Anxiety on Speech – Computational Prediction of Anxiety from Sustained Vowels

Alice Baird¹, Nicholas Cummins¹,², Sebastian Schnieder³,⁴, Jarek Krajewski⁴,⁵, Björn W. Schuller¹,⁶

¹ Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
² Department of Biostatistics and Health Informatics, IoPPN, King’s College London, London, UK
³ HMKW University of Applied Sciences, Berlin, Germany
⁴ Institute of Experimental Psychophysiology, Düsseldorf, Germany
⁵ Human-Technology Interaction, Rhenish University of Applied Science, Cologne, Germany
⁶ GLAM – Group on Language, Audio, & Music, Imperial College London, UK
alice.baird@informatik.uni-augsburg.de

Abstract

The current level of global uncertainty is having an implicit effect on those with a diagnosed anxiety disorder. Anxiety can impact vocal qualities, particularly as physical symptoms of anxiety include muscle tension and shortness of breath. To this end, in this study, we explore the effect of anxiety on speech - focusing on four classes of sustained vowels (sad, smiling, comfortable, and powerful) - via feature analysis and a series of regression experiments. We extract three well-known acoustic feature sets and evaluate the efficacy of machine learning for prediction of anxiety based on the Beck Anxiety Inventory (BAI) score. Of note, utilising a support vector regressor, we find that the effects of anxiety in speech appear to be stronger at higher BAI levels. Significant differences ($p < 0.05$) between test predictions of Low and High-BAI groupings support this. Furthermore, when utilising a High-BAI grouping for the prediction of standardised BAI, significantly higher results are obtained for smiling sustained vowels, of up to 0.646 Spearman’s Correlation Coefficient ($\rho$), and up to 0.592 $\rho$ with all sustained vowels. A significantly stronger (Cohens $d$ of 1.718) result than all data combined without grouping, which achieves at best 0.234 $\rho$.

Index Terms: anxiety disorders, sustained vowels, beck anxiety inventory, machine learning, wellbeing.

1. Introduction

Mental health can have a considerable impact on an individual’s general wellbeing. In modern society, the rate of diagnosis for mental disorders characterised as anxiety disorders is increasing, particularly in urban environments [1]. Anxiety disorders refer to a subgroup of disorders which range in their severity and includes disorders such as, generalised anxiety disorder (GAD), obsessive-compulsive disorder (OCD), and post-traumatic stress disorder (PTSD). The definition of GAD (henceforth, anxiety) is, excessive worry and apprehension occurring more days than not [2]. Feelings of uncertainty often exacerbate anxiety, and the current global pandemic of SARS-CoV-2 now contributes to this [3], particularly from an economic standpoint [4]. With this in mind, mechanisms to monitor and treat anxiety effectively are needed, among both general society [5], as well as for health care professionals [6].

The World Health Organisation (WHO) reports that the proportion of the global population with an anxiety disorder (as of 2015) is ca. 3.6%, and women have a higher rate of diagnosis [7]. Known physical markers include, stomach pain and shortness of breath [8]. The Beck Anxiety Inventory (BAI) [9] is one established evaluation metric for obtaining an individual’s level of anxiety. Criteria for BAI cover mental and physical characteristics and aspects which may have an effect on the vocal tract include, difficulty in breathing and feeling of choking.

Extensive and longstanding behavioural research has been made on the effect of anxiety on speech [8] and features such as, speech disturbances and varied speech-rate are amongst the characteristics of speech with which those with high anxiety tend to present [10]. Previous research suggests a redundancy in the lexical content of speech from individuals with anxiety [11]. Unlike conditions such as depression, in which research towards natural language processing approaches is becoming wide spread [12]. Additionally, acoustic aspects of speech, including disturbances and hesitations may hold meaningful information relating to anxiety [13].

In the short-term, effects of anxiety are prominent during public speaking, particularly by social phobics. In [14], an acoustic analysis was made of parameters including pitch, loudness, and voice quality, finding that perceived and self-assessed levels of anxiety decreased in correlation with such aspects after speaking. Similarly in [15], the authors confirm the illusion of transparency effect, where speakers tend to believe the prominence of anxiety in their voice is more apparent to others. Despite much behavioural research in this area, computational approaches for monitoring and or predicting levels of anxiety are minimal. Indeed, to the best of the authors’ knowledge, this study is the first to explore prediction of anxiety from adult speech. In [16], the efficacy of machine learning to monitor the speech of children with internalising disorders (including depression and anxiety) was explored. Findings show that classical acoustic approaches utilising MFCCs, and support vector machines are effective to a high degree. To this end, acoustic feature extraction toolkits including OPENSAMPLE and DEEP SPECTRUM have shown success for predicting similar conditions including depression [17] and stress [18].

In this study, we explore features of anxiety which may be prominent in speech and evaluate the efficacy of predicting anxiety without lexical content. We utilise various emotional classes of sustained vowels (sad, smiling, comfortable, and powerful) from the Düsseldorf Anxiety Corpus (DAC) and process the data into groupings of Low and High anxiety. As well as this, we group the aforementioned symptoms from the BAI which may explicitly effect the vocal tract – implementing both brute-force and state-of-the-art features, in a conventional support vector regressor paradigm.
2. Düsseldorf Anxiety Corpus

For this study, we utilise the Düsseldorf Anxiety Corpus (DAC), collected by members of the Institute of Experimental Psychophysiology, Düsseldorf, Germany. The corpus is a dataset of individuals performing various vocal exercises, featuring 252 speakers aged 18 to 68 years old (average of 31.5 years, standard deviation of 12.3 years). The files are categorised into different types of phonations, including sustained vowels, read, and free speech. The reference data is formed by measurements which includes the Beck Anxiety Inventory (BAI) [9].

We have chosen only the sustained vowels from DAC to limit the scope of the study and explore specifically the effect of anxiety without lexical content, as this has shown in previous research to be less important for anxiety [11]. For this study, we utilise four classes of sustained vowels:

- **Sad** – a sad phonation of vowel [a] performed with low intensity and frowning face.
- **Smiling** (Smile) – a smiling phonation of vowel [a] in high intensity and smiling face.
- **Comfortable** (Comf) – a comfortable phonation of vowel [a] in comfortable intensity.
- **Powerful** (Power) – a loud phonation of vowel [a] in loud intensity.

These specific classes of sustained vowels are chosen due to their relation to anxiety literature – For example, negative emotion can often be masked as positive [19] (sad, smiling) and typically those with an anxiety disorder are less self-confident [20] (comfortable, powerful).

All speakers in DAC have been evaluated under the Beck Anxiety Inventory (BAI) questionnaire [21]. During the BAI, individuals answer a series of questions relating to their wellbeing, on a scale from 0–3. A total score of under 21 indicates low anxiety, and a score of above 36 indicates potentially concerning anxiety, and a score of above 36 indicates potentially concerning anxiety literature – For example, negative emotion can often be masked as positive [19] (sad, smiling) and typically those with an anxiety disorder are less self-confident [20] (comfortable, powerful).

An overview of the mean from results is given in Figure 2a. When evaluating pitch (F0) of the four classes of sustained vowels, we see a higher standard deviation between Low-BAI and High-BAI groupings for all classes, except smiling – particularly, for sad and comfortable, which show a smaller and medium effect size, respectively. This finding leads us to assume that lower aroused phonation types present stronger F0 variance for those with higher levels of anxiety. For the intensity of the speech signal, we see that in all cases samples of Low-BAI show strong deviation in dB, and particularly for sad and powerful which have a large and medium effect size, respectively. Additionally, we compare Low-BAI-sad and Low-BAI-power and do see a large effect size of 1.068 d, reaffirming the effect of the target vocalisation style for these sustained vowels. We also extract HNR, and see that like F0, all classes show higher mean results for the High-BAI class, aside from the smiling vocalisation. This finding is particularly significant for sad and comfortable and shows that vocal fold action is less consistent, for these classes in the High-BAI group. From a qualitative analysis of male and female speakers, we see that standard deviation in F0 appears to be larger in female speakers as shown in Figure 1, particularly for the sustained vowel class of sad. Due to this reason, in future studies, it would be valuable to explore genders independently.

### Table 1: Speaker (#) independent partitions, Train, (Dev),development, and Test. Gender (M:Male:F:Female, sustained vowel type, BAI class (Low, High), feeling of choking (No symptoms, Has symptoms), difficulty in breathing (No symptoms, Has symptoms), are reported on the audio (Instantace level).

<table>
<thead>
<tr>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>M/F</td>
<td>26:6</td>
<td>97:4</td>
</tr>
<tr>
<td>Inst.</td>
<td>614</td>
<td>511</td>
<td>440</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sustained Vowel Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Smile</td>
</tr>
<tr>
<td>Comf</td>
</tr>
<tr>
<td>Power</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Variables</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
by anxiety. As mentioned earlier, for evaluation, we create several subsets of the data, and we evaluate each grouping through the correlation of predicted BAI score. Utilising the the four sustained vowel classes, described previously (sad, smiling, comfortable, and powerful), experiments are performed for only those speech samples, as well as for all together. Firstly we create High-BAI and Low-BAI groupings of BAI (≤ 20: Low, ≥ 21: High), we then then perform these experiments again for groups of individuals who show (Has) symptoms relating to the BAI criteria of feeling of choking (choking) and difficulty breathing (breathing) against those who do not show these systems (No).

4.1. Acoustic features

To cover a range of well-known acoustic features, we extract hand-crafted speech-based features, as well as a state-of-the-art approach, extracting spectrogram-based deep data representations from the speech signals.

**OPENSMILE:** As a conventional and well established approach, the 6373 dimensional ComPARE feature set [22], and our 88 dimensional eGeMAPS feature set [23], are used given our experience in similar paralinguistic tasks [24, 25]. From each instance, the ComPARE and eGeMAPS acoustic features are extracted with the OPENSMILE toolkit [22]. The default parameter settings from OPENSMILE are used and due to the short duration of files (ca. 6 seconds), features are extracted as one feature vector per sample. We standardise the features by removing the mean and scaling to the unit variance for ComPARE features – for eGeMAPS, this was not beneficial.

**DEEP SPECTRUM:** Additionally, we extract a 4096 dimensional feature set of deep data-representations using the DEEP SPECTRUM toolkit [26]. DEEP SPECTRUM has shown success for similar audio- and speech-based tasks [18], and extracts features from the audio data using pre-trained convolutional neural networks (CNNs). For this study, we extract Viridis colour map spectrograms (cf. Figure 1 for colour map), using the default VGG16 pre-trained network, and as with OPENSMILE, we extract one feature vector per sample. We also apply standardisation to the DEEP SPECTRUM features.

As a brief initial step, we evaluate the effect size (Cohen’s d) between High-BAI and Low-BAI groupings of the feature sets extracted for each sustained vowel (cf. Figure 2a). Of note from this analysis, we see that DEEP SPECTRUM features appear to have consistently moderate effect sizes, larger than ComPARE and eGeMAPS, particularly for the sad and comfortable class. As this finding also seems to be reflective of our previous acoustic analysis, where sad and comfortable seem to behave similarly for F0 STD and HNR STD, this leads us to assume given the visual nature of DEEP SPECTRUM features that increased standard deviation in F0 for those with higher anxiety may be more easily captured with these features. Further to this, DEEP SPECTRUM most likely observes noise in the signal, as reflected by the high HNR for both sad and comfortable.

4.2. Training and evaluation

Given that our dataset is reasonably small (ca 3hrs), for a robust and easily reproducible approach, we choose to utilise an epsilon-support vector regressor (SVR) with a linear kernel. We split the data for training, into speaker-independent sets: training, development and test (cf. Table 1). During the development phase, we trained a series of SVR models, optimising the complexity parameters (C ∈ 10⁻⁴, 10⁻³, 10⁻², 10⁻¹, 1), and evaluating their performance on the development set. We then re-trained the model with the concatenated train and development set, and evaluate the performance on the test set. We repeat this method for each combination. Note that we report the best value for C in development for test validation.

To evaluate the results of all experiments, we utilise Spearman’s correlation coefficient (ρ) due to the ordinal nature of the raw BAI values. Additionally, we Cohen’s d is used as a measure of effect size between the predictions of results of interest. Reporting of Cohen’s d proceeds an evaluation of each test set prediction result for normality using a Shapiro-Wilk test [27], as well as two-tailed T-test, rejecting the null hypothesis at a significance level of p < 0.05.

5. Results and Discussion

Our fully-fledged results are given in Table 2. As indicated by *, there are significant difference in almost all predictions for Low-BAI vs High-BAI groupings. As well as this in most cases, High-BAI grouped results are significantly higher than Low-BAI grouped results. Although our results do vary, they suggests that the characteristics of speech, harnesses for prediction of anxiety, are stronger when anxiety is at high levels. This finding is supported by earlier discussed literature, which suggests that speech disturbances and varied speech-rate are prominent in the speech of those with high anxiety [10].

Looking closer at our BAI grouped experiments, we see High-BAI grouped anxiety predictions are stronger, with at best, 0.505 ρ for prediction of standardised BAI of all High-BAI grouped samples. Through the late-fusion of the two best results eGeMAPS and DEEP SPECTRUM, this is increased to 0.592 ρ. For the individual sustained vowels, smiling in High-BAI grouping performs best, with eGeMAPS showing up to 0.593 ρ, a result which is also improved by late-fusion up to 0.646 ρ. We see a slight moderate correlation for DEEP SPECTRUM of sad High-BAI grouping. However, this is not consistent with all feature sets. For comfortable and powerful, there are no substantial correlations, leading us to consider that these samples do not provide meaningful information for the current task.

For the grouping of Has-symptoms, or No-symptoms of

---

For BAI grouped results, we include late-fusion results taken from the mean of predictions of the two best performing feature sets.

For BAI grouped results, we include late-fusion results taken from the mean of predictions of the two best performing feature sets.

Table 2: (Dev)elopment and Test SVR results for the prediction of standardised BAI for all stressed vowel combinations selected from DAC, Reporting Spearman Correlation Coefficient (ρ) for groupings of: Low-BAI or High-BAI (Has) Symptoms or (No) Symptoms of Feeling of (Choke)ing, (Has) Symptoms or (No) Symptoms of Difficulty in (Breath)ing. For High-BAI and Has symptoms groupings * indicates significance (p < 0.05) of test predictions as compared to the equivalent Low-BAI or No symptoms grouping test prediction. For BAI grouped results, we include late-fusion results taken from the mean of predictions of the two best performing feature sets. Emphasised results show a positive ρ correlation above 0.3.

<table>
<thead>
<tr>
<th>ρ</th>
<th>Low</th>
<th>Sad</th>
<th>Smiling</th>
<th>Comfortable</th>
<th>Powerful</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAI</td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>EGE MAPS</td>
<td>-0.49</td>
<td>0.16</td>
<td>0.21</td>
<td>-0.05*</td>
<td>0.043</td>
<td>-0.012</td>
</tr>
<tr>
<td>COMPARE</td>
<td>0.25</td>
<td>-0.018</td>
<td>0.16</td>
<td>-0.241*</td>
<td>0.015</td>
<td>-0.271</td>
</tr>
<tr>
<td>DEEP SPECTRUM</td>
<td>0.104</td>
<td>0.210</td>
<td>0.05</td>
<td>0.304*</td>
<td>0.190</td>
<td>0.012</td>
</tr>
<tr>
<td>Late-fusion</td>
<td>-</td>
<td>0.194</td>
<td>-</td>
<td>0.225*</td>
<td>-</td>
<td>0.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Choke</th>
<th>No</th>
<th>Has</th>
<th>No</th>
<th>Has</th>
<th>No</th>
<th>Has</th>
<th>No</th>
<th>Has</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGE MAPS</td>
<td>0.02</td>
<td>0.190</td>
<td>0.611</td>
<td>0.050*</td>
<td>0.252</td>
<td>0.350</td>
<td>0.343</td>
<td>0.218*</td>
</tr>
<tr>
<td>COMPARE</td>
<td>-0.09</td>
<td>-0.170</td>
<td>0.110</td>
<td>0.051</td>
<td>0.309</td>
<td>0.011</td>
<td>0.223</td>
<td>0.535*</td>
</tr>
<tr>
<td>DEEP SPECTRUM</td>
<td>0.078</td>
<td>0.288</td>
<td>0.369</td>
<td>-0.397*</td>
<td>0.202</td>
<td>0.246</td>
<td>0.022</td>
<td>-0.160*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Breath</th>
<th>No</th>
<th>Has</th>
<th>No</th>
<th>Has</th>
<th>No</th>
<th>Has</th>
<th>No</th>
<th>Has</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGE MAPS</td>
<td>-0.019</td>
<td>0.120</td>
<td>-0.276</td>
<td>-0.122*</td>
<td>0.187</td>
<td>-0.253</td>
<td>0.453</td>
<td>-0.540</td>
</tr>
<tr>
<td>COMPARE</td>
<td>0.048</td>
<td>-0.139</td>
<td>0.028</td>
<td>0.363*</td>
<td>0.282</td>
<td>-0.021</td>
<td>0.357</td>
<td>0.699*</td>
</tr>
<tr>
<td>DEEP SPECTRUM</td>
<td>-0.006</td>
<td>0.357</td>
<td>0.301</td>
<td>0.284*</td>
<td>0.342</td>
<td>0.256</td>
<td>0.045</td>
<td>0.028*</td>
</tr>
</tbody>
</table>

feeling of choking, smiling samples again performs best, with COMPARE at best 0.535 ρ. However, in this case, eGE MAPS and DEEP SPECTRUM are less able to capture the phenomena. Comfortable phonations show a strong negative correlation for the Has-symptom grouping, a finding which to a degree also appears for sad, suggesting that intensity may play a strong role in this task. When predicting standardised BAI from all samples with No-symptom of choking, we see that this is stronger than the Has-symptoms pairing. Overall, there are no strong findings from this paradigm. However, most No-symptoms grouped results perform better than Has-symptoms grouped, which suggest a need for further acoustic analysis, to observe any variation in the samples for this constellation.

For the grouping of Has-symptoms or No-symptoms of difficulty in breathing, we see that as with choking, the No-symptoms grouped results are often stronger than Has-symptoms group. However, across features sets, this is somewhat confused. For sad, for example, the No-symptoms grouping performs better with DEEP SPECTRUM, but overall, COMPARE shows slightly better results for the Has-symptoms grouping. Like all other groupings, the smiling class in the Has-symptoms grouping shows our best result with up to 0.699 ρ. COMPARE also performs best when utilising all data for the Has-symptoms grouping. This is suggesting that HNR, which may be stronger due to restricted airflow, is more easily captured by COMPARE features for individuals with this breathing symptom.

To evaluate further the degree to which highly anxious speech improves prediction accuracy, we additionally reran our experiments with all data and without any groupings (cf. Table 3). From this we find that still the High-BAI grouped with DEEP SPECTRUM and eGE MAPS results are stronger, with the best results from late-fusion being, 243 ρ for all data (a result which can be considered negligible). This result is significantly lower than the best High-BAI result with all sustained vowels, reporting a very large effect size of 1.718 d.

In general, for all scenarios, the smiling class performs best for the stronger High-BAI/Has-symptoms groupings. This finding could suggest that anxiety is more prevalent in a more facially strained stressed vowel. There is much in the literature relating to smiling and anxiety, for example, the “fooled by a smile” effect in which those who suffer from anxiety can show untrue emotional expressions [28]. Furthermore, high anxiety involves much more facial expression, and general movement, as compared to lower anxiety, with ‘non-enjoyment’ smiles being displayed frequently [29].

6. Conclusion and Future Outlook

In this study, we explored the effect of anxiety on speech. In particular, we evaluated the efficacy of predicting anxiety from adult speech for the first time and evaluated non-lexical sustained vowels as a first step. Our findings show that utilising speech-based features for prediction of anxiety is valid and that recognition of higher levels of anxiety is better. As individuals reporting high levels of BAI may need a more timely medical intervention this finding is promising. From our results, we see that smiling phonations are particularly informative for those with high anxiety. A finding related to literature which states that smiling causes an alteration of the vocal tract and can be “heard as well as seen” [30]. Additionally, those with high anxiety often overstate their emotional expression [28], possibly leading to stronger speech variance. For further studies, we hope to explore the effect of smiling phonations (and facial movement) on anxious speech further. As well as this, given the slight gender bias to our data, it would be of interest to evaluate gender independently, as we have seen that features such as STD for F0 are stronger in highly anxious female samples. Further, since our study shows promise for the presence of anxiety in speech without lexical content, it would be of interest to compare this to free speech samples.

7. Acknowledgements

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), and the Horizon 2020, EFPIA funded Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115902.
8. References


