Towards the Development of Speech-Based Measures of Stress Response in Individuals

Archna Bhatia and **Toshiya Miyatsu** and **Peter Pirolli**
Institute for Human and Machine Cognition
15 SE Osceola Ave
Ocala, FL 34471

Abstract

Psychological and physiological stress in the environment can induce a different stress response in different individuals. Given the causal relationship between stress, mental health, and psychopathologies, as well as its impact on individuals' executive functioning and performance, identifying the extent of stress response in individuals can be useful for providing targeted support to those who are in need. In this paper, we identify and validate features in speech that can be used as indicators of stress response in individuals to develop speech-based measures of stress response. We evaluate effectiveness of two types of tasks used for collecting speech samples in developing stress response measures, namely Read Speech Task and Open-Ended Question Task. Participants completed these tasks, along with the verbal fluency task (an established measure of executive functioning) before and after clinically validated stress induction to see if the changes in the speech-based features are associated with the stress-induced decline in executive functioning. Further, we supplement our analyses with an extensive, external assessment of the individuals' stress tolerance in the real life to validate the usefulness of the speech-based measures in predicting meaningful outcomes outside of the experimental setting.

1 Introduction

Various psychological and physiological stress conditions, e.g., an approaching deadline, an interview not going well, a combat situation, or extreme temperatures, can have an impact on an individual due to various (maladaptive) physiological and mental processes (Yaribeygi et al., 2017; Sapolsky, 1996). Long-term exposure to stress can play a significant role in the formation and exacerbation of mental disorders, such as anxiety disorders, depression, and schizophrenia (Gomes and Grace, 2017; Yang et al., 2015; Tafet and Nemeroff, 2015; Esch et al., 2002). Stress in one's environment may result in a relatively immediate (whether short-term or long-lasting) effect on performance. For example, a condescending interviewer may lead an interviewee to not be able to respond at all or a sudden combat situation may lead an individual to make more errors. However, different individuals respond differently to the same stress conditions depending on their mental and physiological constitution, experiences, training and preparedness, among other factors. Identifying the degree of stress response in individuals would be helpful in reducing stress' impact on their health and performance both at the individual and at the community level. For example, an automatic measure of stress response can be used by an individual for self monitoring and deciding to use a management strategy of daily stress reduction exercises when needed. Similarly, community members can be supported through targeted allocation of mental health resources. An automatic measure of stress response can provide additional information about individuals' response to stress to the clinicians treating them so that appropriate and timely therapeutic support can be provided.

Previously, self-report inventories of stressors and their symptoms have been used to measure individuals' stress response (e.g., Bland et al., 2012; Tatar et al., 2018; Rushall, 1990). However, these inventories are limited in their scope (e.g., sports, school) and utilities. Further, self-report inventories have inherent problems. For example, individuals may not be fully aware of the effect stressors have on them, or they may not answer questions truthfully. Therefore, development of more objective yet accessible measures of stress response are needed.

An extensive body of research has shown the impact of stress on speech, e.g., Jackson et al. (2016); Jena and Singh (2016); Schuller et al. (2014); Giddens et al. (2013); Lierde et al. (2009); He et al. (2008); Dietrich et al. (2008); Hansen and Patil

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 192–203

(2007); Fernandez and Picard (2003); Brenner and Shipp (1988); Brenner et al. (1983) have associated certain changes in speech with exposure to stress. For example, Brenner and Shipp (1988) reported an increase in fundamental frequency, amplitude and speech rate in extreme levels of stress. They also reported changes in the energy distribution, frequency jitter and amplitude shimmer in stressed speech. Features, such as Mel-frequency Cepstrum Coefficients, have been found to be affected by emotional states including anxiety/stress to be useful for identifying or classifying these emotional states, e.g., see Vaikole et al. (2020); Dhole and Kale (2020); Tomba et al. (2018); Hansen and Patil (2007). The primary goal of the current project is to extend this work by identifying and validating a set of speech-based features that can be used as individual difference measures of stress response.

To achieve this goal, we use two continuous speech sample collection methods, namely a Read Speech Task and an Open-ended Question Task. Read speech provides much cleaner data than spontaneous speech and hence can be very useful for modeling speech related phenomena. It has been extensively used in speech processing studies, for example, see Pernkopf et al. (2009); Nakamura et al. (2008); Pruthi and Espy-Wilson (2007, 2004); Garofolo et al. (1993). Open-ended Question Task, on the other hand, provides more naturalistic data, which can complement the information available in read speech. The notion that they may provide different types of information is confirmed by works such as Schuppler (2017) which discussed the need for developing methods for different speaking styles instead of just focusing on read speech.

For the purpose of developing measures for stress response, we consider `stress response` and `stress tolerance` to be two facets of the same phenomenon, where stress response refers to how an individual responds to or is affected by stress, whereas stress tolerance refers to how tolerant an individual is to stress (i.e., how well they can still perform tasks under stress). We focus on investigating speech and identifying relevant acoustic features to develop a speech-based measure of stress response. We expect such a measure to also be informative about an individual's stress tolerance. To this end, we establish the relationship of speech features with a complex, ecologically valid stress tolerance measure based on trained judges' stress tolerance assessment of individuals as described in Section 2.3.

The rest of the paper is organized as follows: In Section 2, we describe our data and data collection procedures. In Section 3, we discuss our methodology to extract acoustic features from the speech produced by individuals in stress conditions and to prepare the performance assessment measures for evaluating the usefulness of the extracted features. In Section 4, we present our findings based on the analysis of extracted features from speech in terms of their relationship with stress conditions as well as with cognitive performance and real-life stress tolerance measures. In Section 5, we discuss our results and implications for development of speech-based measures for stress response in individuals. In addition to the features identified by our investigation to be informative of an individual's stress response, we also provide recommendations with respect to the types of tasks used to collect speech samples to extract these features. In Section 6, we conclude and briefly discuss future work. This is followed by Section 7 where we discuss a few use cases and ethical considerations for this work.

2 Data

The data used for our investigation were collected from 13 male participants. They were recruited from among the candidates going through a week-long selection assessment process at a US military unit who had provided consent prior to the selection week and remained on-site until the end of the selection week. They were provided a description of the study before obtaining their consent. The data collection was conducted the day after the selection week was over. All protocols for data collection were approved by the Institutional Review Board at the appropriate branch of the US military (where data had to be collected) as well as the Institutional Review Board at the Florida Institute for Human and Machine Cognition (where the research activity had to take place) prior to data collection.

Two types of data were collected: speech samples and selection assessment data. The speech samples were recorded in a lab setting in two stress conditions, namely `Neutral` and `Stress`. Section 2.1 provides details about stress induction and Section 2.2 provides more details about speech data collection. In addition to the speech samples, participants' scores that were assigned during the selection week by trained and experienced US mil-

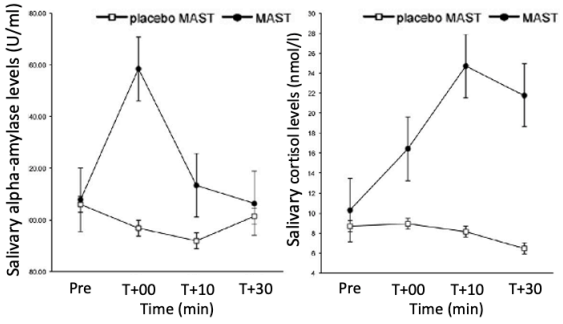

Figure 1: A demonstration of salivary gluco-corticol stress response induced by MAST (Smeets, et al., 2012)

itary assessment personnel were obtained to augment our analyses. More details about the selection assessment data are provided in Section 2.3.

2.1 Stress Induction

In order to collect speech samples from participants in the two stress conditions, `Neutral` and `Stress`, a version of the Maastricht Acute Stress Test (MAST) (Smeets et al., 2012) was used for stress induction. MAST is a clinically certified stress induction which reliably elicits gluco-corticol stress response that could be measured through an increase in established biomarkers of stress, such as alpha-amylase and cortisol, see Figure 1. MAST combines two established methods of stress induction: social stress through the Trier Social Stress Test and physiological stress through the Cold Pressor Test. In our version, participants were instructed to sit in front of a video recorder and look straight at it because their facial expression would be analyzed later. Then, they completed several rounds of hand immersion trials (HITs) and Mental Arithmetic (MA) trials. In the HITs, they were asked to submerge their hand for a set duration (see Figure 2) in a container filled with ice water which was kept at 4°C. In the MA trials, they were asked to count backwards by 17 starting from 2043, and whenever they made a mistake or took more than three seconds to say the next number, the experimenter gave them negative feedback and asked them to start over. The participants also performed a Verbal Fluency Task (see Section 2.2) during stress induction. Figure 2 shows the task sequence during stress induction.

2.2 Speech Data

Speech samples were collected from the participants through three tasks: Read Speech Task, Open-ended Question Task and Verbal Fluency Task. Participants completed these tasks under both the `Neutral` condition and the `Stress` condition. Read Speech Task and Open-ended Question Task were used to extract acoustic features from the two types of continuous speech samples, the read speech samples and the naturalistic speech samples respectively. Verbal Fluency Task, on the other hand, was used as an assessment for cognitive performance under the two stress conditions.

Read Speech Task. Participants read out loud a 243 words passage about the psychological construct 'grit' in both stress conditions. The passage was borrowed from an online blog on grit (Doyle, 2020).[1] It was modified to include all phonemes in American English to enable a rich set of analyses (including at the phonemic level).

Open-Ended Question Task. Participants were asked to speak for two minutes in response to four open-ended questions each to obtain their naturalistic speech samples in the two stress conditions. The questions focused on stimulation seeking and suppression of emotions in the `Neutral` condition and on response to distress and reappraisal of negative emotions in the `Stress` condition. The topics of these questions were derived from two well-established works regarding stress response, namely defensive reactivity (Kramer et al., 2012) and emotion regulation (Gross, 2014). Specifically, these questions sought information from participants about how they reacted to stressful situations, e.g., "Recall and describe the most recent event in which you were stressed about something. How quickly/slowly did you recover from it?".

Verbal Fluency Task. Verbal Fluency Task, a well-established measure of executive functioning (Shao et al., 2014), was used to assess participants' cognitive performance in the `Neutral` and `Stress` conditions. Participants were asked to say out loud as many words as they could remember in 1 minute that belonged to a given category. 'Body parts', 'fruits', 'words starting with A', and 'words starting with F' were used in the `Neutral` condition and 'animals' and 'words starting with C' were used in the `Stress` condition).

[1] https://www.aceable.com/blog/aceable-essay-on-grit/

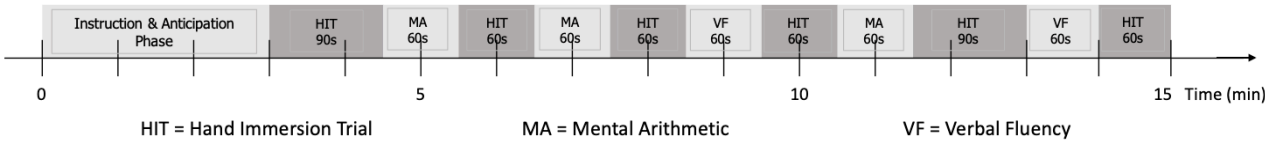

Figure 2: A schematic illustration of the task sequence within the version of MAST employed in the current study

2.3 Selection Assessment Data

In addition to the participants' responses to Verbal Fluency Task used for assessing their cognitive performance, we also obtained selection assessment data from the host military organization. Specifically, these data were scores assigned by trained assessment personnel based on their observations of attributes demonstrated by the participants in a given task, such as teamwork, leadership, and stress tolerance. The data consisted of 61 scores from 17 tasks performed by the participants during the week-long selection assessment.

3 Methodology

3.1 Feature Extraction from Speech Samples

In order to identify indicators of stress response in the speech signal, we extracted a number of acoustic features from the speech samples collected from the participants under the two stress conditions. We used Librosa (McFee et al., 2020, 2015), a Python library for audio and music analysis, to extract the acoustic features from the speech samples. Both time domain and frequency domain features were extracted. The time domain features included Amplitude Envelope, Root Mean Square Energy and Zero Crossing Rate. The frequency domain features included Magnitude Spectrum, Short-time Fourier transform Spectrum, and 13 Mel-frequency Cepstrum Coefficients (MFCCs) as well as their first and second derivatives. Many of these time domain and frequency domain features have been found to be affected by stress (Hansen and Patil, 2007; Fernandez and Picard, 2003). Features were extracted from speech using overlapping frames with frame size of 1024 and hop size of 512 waveform samples. The mean of the feature values extracted for each frame in a speech sample was used as the extracted feature from the speech sample for the analyses discussed in Section 4. Additionally, duration of the signal was also taken as a feature from Read Speech Task because the rate of speech has previously been found to be affected by stress (Wikibooks, 2018; Brenner and Shipp, 1988; Brenner et al., 1983). Thus, a total of 45 features from Read Speech Task and 44 features from Open-ended Question Task were extracted from the speech samples of participants corresponding to each of the stress conditions.

3.2 Performance Score Based on Verbal Fluency Task

The speech data collected from Verbal Fluency Task were transcribed using speech_recognition, a Python client for the Google Speech-to-Text API (Google, 2019). The generated transcriptions were manually checked for any errors by one of the authors of this paper. Although the transcription had a high accuracy ($> 90\%$), manual checking of errors and manual counting of the total number of words recalled by participants in the transcriptions was necessary because the transcription often recognized two components of a multiword item as two words (e.g., 'dragon' and 'fruits' for 'dragon fruits', 'polar' and 'bear' for 'polar bear').[2] The transcriptions were then used to compute the performance metric 'Word Recall', the total number of words recalled by the participant in one minute.

Given the different base rates for different categories (i.e., the semantic categories of 'body parts', 'fruits' and 'animals', and the phonetic categories of 'words starting with A', 'words starting with F' and 'words starting with C'), we first normalized (Z-score) the performance metric 'Word Recall' for each of the categories. We, then, computed the mean of the normalized word recall for the four `Neutral` categories (viz. 'body parts', 'fruits', 'words starting with A', and 'words starting with F') and for the two `Stress` categories (viz. 'animals' and 'words starting with C'), and subtracted the resulting score in the `Neutral` condition from the score in the `Stress` condition for each participant. Thus obtained score represented the change in cognitive performance due to stress (in *SD*) relative to other participants.

[2]Inspired by the relevance of the error rates in mental health contexts, since speech-to-text systems' performance may also decline as the speech is affected under stress, in our future investigations, we may examine the error rate also as a feature for an individual's stress response.

3.3 Stress Tolerance Score Based on Selection Assessment Data

Of the 61 selection assessment scores assigned by the host agency's experienced personnel based on participant attributes demonstrated in 17 tasks performed during selection week (see Section 2.3), seven were on stress tolerance. Given that these scores came from a diverse set of tasks (e.g., a team building exercise) we computed a correlation (Pearson's Correlation) among all of them to see if these scores converged to measure the same construct (i.e., stress tolerance). These seven stress tolerance scores showed reasonable convergence, $r(7) = .41$. Thus, an average of these stress tolerance scores was taken as a complex, ecologically valid measure of the participants' stress tolerance.

4 Analysis and Findings

One can expect that the features in speech indicative of an individual's stress response can also be taken to indicate that the corresponding speech was produced in a `Stress` condition. Hence, one strategy to identify potentially relevant features associated with stress response from among the 45 extracted features (44 in case of Open-ended Question Task) is to find the ones that can distinguish between the `Neutral` and the `Stress` conditions. Hence, we performed a paired sample t-test on our data where the extracted features in the two stress conditions were taken as the two sets of observations pre- and post- stress induction.[3] We found that a number of speech features, extracted from Read Speech Task and Open-ended Question Task, showed statistically significant difference between the means in the observations corresponding to the two stress conditions, as shown in Table 1.

We found that duration and a time domain feature Zero Crossing Rate extracted from Read Speech Task showed significant difference between the two stress conditions with $p < .005$. [4] The rest of the features that showed significant difference with varying p-values ($p < .01$ or $p < .05$) were all associated with the frequency domain features MFCCs or their derivatives.

[3]We plan to collect more data to increase sample size to confirm our findings and increase reliability of the results. With the larger dataset in the future, for a more robust set of significant features, Bonferroni or similar corrections may be applied to the multiple comparisons.

[4]In this paper, we have specified a p value of $< .005$ to indicate high significance but these denote the cases where the p values are very close to though slightly greater than $< .001$.

Read Speech	Open-ended Question
Duration***	-
Zero Crossing Rate***	Zero Crossing Rate*
MFCC2**	MFCC2***
Delta MFCC4*	-
Delta Delta MFCC4*	-
-	Delta Delta MFCC12*
-	Delta Delta MFCC11*
-	Delta Delta MFCC2*
-	Delta Delta MFCC10*

Table 1: Speech features that show statistically significant difference between the `Neutral` and `Stress` conditions for the two speech tasks, Read Speech Task and Open-ended Question Task. *** indicates $p < .005$, ** indicates $p < .01$, * indicates $p < .05$

Some features showed significant difference when extracted from either of the two speech tasks. For example, the time domain feature Zero Crossing Rate showed significant difference when extracted from the Open-ended Question Task speech samples as well, although with comparatively less significance than when extracted from the Read Speech Task speech samples. In contrast, the frequency domain feature MFCC2 showed more significance when extracted from the Open-ended Question Task samples than when extracted from the Read Speech Task samples.

The frequency domain features Delta MFCC4 (first derivative of MFCC4) and Delta Delta MFCC4 (second derivative of MFCC4) showed significant difference between the means for observations in the two stress conditions when they were extracted from Read Speech Task, but not when they were extracted from Open-ended Question Task. Similarly, second derivatives of MFCC12, MFCC11, MFCC2 and MFCC10 showed significant difference in the two stress conditions when they were extracted from Open-ended Question Task but not when they were extracted from Read Speech Task.

While more features extracted from the speech samples collected for Open-ended Question Task showed significant difference in the means between the two stress conditions, the features extracted from the Read Speech Task speech samples showed a higher significant difference between the two conditions, in general.

We then correlated the extracted features with Verbal Fluency Task-based scores representing the change in cognitive performance due to stress to

provide a measurement of stress response in participants (performance score), as described in Section 3.2. Similarly, we correlated these acoustic features with the selection assessment-based scores that provided a measurement of stress tolerance in participants (stress tolerance score), as described in Section 2.3. Table 2 presents the correlations (Pearson's Correlation Coefficients) of the acoustic features with the performance score (left) and the stress tolerance score (right) for the two speech tasks. We explored these correlations to test two hypotheses as follows. First, some of the acoustic features show significant correlations with stress response/stress tolerance. Second, these features overlap with features identified to be differential between the two stress conditions based on inferential statistics (e.g., the paired sample t-test above).

In regards to the hypothesis about the relationship between acoustic features and performance scores/stress tolerance scores, we found that there were a number of acoustic features in both speech tasks that showed high to moderate correlations with performance scores as well as with stress tolerance scores as shown in Table 2. Many of these were found to be highly significant with p-values $< .005$ as in the case of acoustic feature Delta Delta MFCC7 for correlations with performance scores for Open-ended Question Task, and acoustic features magnitude spectrum for correlations with stress tolerance scores for Read Speech Task and MFCC11 for correlations with stress tolerance scores for Open-ended Question Task. Additionally, there were other features with which also correlations were significant with p-values of $< .01$ or $< .05$. For example, Delta Delta MFCC1 showed strong correlations with performance scores for Open-ended Question Task with p-values $< .01$. Similarly, the duration feature, representing rate of speech, was found to be moderately correlated with performance scores for Read Speech Task with p-values $< .05$. Delta Delta MFCC13 was moderately correlated with stress tolerance in Read Speech Task with p-values $< .05$, and so were features MFCC1, MFCC8 and MFCC9 in Open-ended Question Task. The full set of correlations corresponding to all extracted features are provided in Appendix A.

In regards to the second hypothesis about the highly differential features for stress conditions to also be correlated with the performance and the stress tolerance scores, we found that there is some overlap between the two sets of features. However, not all features that significantly distinguished the stress conditions were also strongly/moderately correlated with the performance and the stress tolerance scores for the two speech tasks. Duration extracted from Read Speech Task was found to be significantly differential for stress conditions and it showed moderate correlation with the performance score in Read Speech Task which was also significant. However, it was not found to be correlated with stress tolerance (with $r(13) = .184$ for Read Speech Task and .313 for Open-ended Question Task). Zero Crossing Rate, on the other hand, was found to be significantly differential for the stress conditions for both Read Speech Task and Open-ended Question Task, but it did not show significant correlations with the performance and the stress tolerance scores for either of the tasks. Similarly, MFCC2, while significantly differential for stress conditions for both the tasks, showed low correlations with the performance and the stress tolerance scores. Delta MFCC4 and Delta Delta MFCC4 were significantly differential for Read Speech Task but showed low correlations with both the scores for both the tasks. Delta Delta MFCC12, Delta Delta MFCC11, Delta Delta MFCC2 and Delta Delta MFCC10 were significantly differential for Open-ended Question Task, but showed low correlations with both the scores for both the tasks.

While exploring the above two hypotheses, we found a subset of acoustic features (e.g., duration, magnitude spectrum, some of the MFCCs or their derivatives) that showed strong to moderate correlations with the performance score (stress response) or the stress tolerance score, answering our first question setforth in Section 1. Next, we explored the effectiveness of speech tasks in indicating an individual's stress response to answer the second question setforth in Section 1.

Based on the correlations in Table 2, we found that Open-ended Question Task resulted in more acoustic features than Read Speech Task that showed strong correlations with performance scores (Delta Delta MFCC7 and Delta Delta MFCC1) as well as stress tolerance scores (MFCC11, MFCC1, MFCC8 and MFCC9) that were significant. However, in Read Speech Task also, we found a few acoustic features that were not identified by Open-ended Question Task but still showed strong correlations with the stress tolerance scores (magnitude spectrum and Delta Delta

Feature	Performance Score		Stress Tolerance Score	
	RST	OQT	RST	OQT
Duration related features:				
Duration (Speech Rate)	-.628*	-		-
Time domain features:				
Zero Crossing Rate	.546			
Frequency domain features:				
Magnitude Spectrum			-.753***	
Short-time Fourier Transform Spectrum	-.502			.515
MFCC1				.617*
MFCC7		.524		
MFCC8				-.583*
MFCC9				-.554*
MFCC11		.516		-.735***
Delta Delta MFCC1		-.710**		
Delta Delta MFCC7		.759***		
Delta Delta MFCC8		.519		
Delta Delta MFCC10				.534
Delta Delta MFCC13			.589*	

Table 2: Moderate to high correlations (> .5) between the difference scores (`Stress - Neutral`) from among the 45 features extracted from Read Speech Task (RST) and from Open-ended Question Task (OQT) and the stress-induced change in Verbal Fluency Task performance (left) and the stress tolerance score from the selection assessment data (right). The duration feature was dropped for OQT since the task duration itself was set to two minutes. *** indicates $p < .005$, ** indicates $p < .01$, * indicates $p < .05$

.

MFCC13) that were significant. Other features were also identified in these tasks that showed moderate correlations with the performance scores (e.g., Zero Crossing Rate and Short-time Fourier Transform spectrum in Read Speech Task, and duration, MFCC7, Delta Delta MFCC8 and MFCC11 in Open-ended Question Task) as well as with the stress tolerance scores (e.g., Delta Delta MFCC10 and Short-time Fourier Transform Spectrum in Open-ended Question Task) but they were not found to be significant. Thus, we found that the two tasks provided complementary information in terms of features that were strongly correlated with stress response/stress tolerance with significance.

5 Towards Building a Speech-Based Measure of Stress Response/Tolerance

In Section 4, we identified a number of acoustic features that strongly/moderately correlated with Verbal Fluency Task-based performance scores and external assessment-based stress tolerance scores. We found that duration (speech rate) and frequency domain features, such as magnitude spectrum and some of the MFCCs or their derivatives, showed strong to moderate correlations with stress response/stress tolerance that were significant. Time domain features, on the other hand, did not show significant correlations with either stress response or stress tolerance. Although this finding needs to be confirmed with larger data, this may be taken to indicate the effectiveness of duration and frequency domain features over time domain features in a speech-based measure for stress response/tolerance. It should be noted that some of these duration and frequency domain features, e.g., speech rate, MFCCs and their derivatives have also been found to be indicative of stress in prior works on stress detection from the speech signal, e.g., Vaikole et al. (2020); Dhole and Kale (2020); Tomba et al. (2018); Brenner and Shipp (1988), to name a few. Also, opportunities to validate stress response or any predictive features against rich, ecologically valid datasets like the Selection Assessment Data used in our experiments are rare, and it is encouraging to see significant correlations between some of these features and the stress tolerance score based on the Selection Assessment Data (Table 2).

We compared the two speech collection tasks, Read Speech Task and Open-ended Question Task, for their effectiveness in providing useful informa-

tion in speech features for stress response/tolerance. The findings showed that these two tasks provided complementary information in that a different set of features correlated with stress response/tolerance when samples were collected through the two tasks. This makes sense given the fact that Read Speech Task provides cleaner speech samples whereas Open-ended Question Task provides more naturalistic speech. This suggests that a speech-based measure for stress response/tolerance would benefit from using both these tasks for data collection. However, if only one task needs be used in order to minimize time, effort, and other resources expended in data collection, Open-ended Question Task should be used since Open-ended Question Task samples led to a larger number of correlated acoustic features which showed strong correlations with high significance values.

Another finding from Section 4 was that the features helpful in distinguishing between the stress conditions may not necessarily be the features that indicate stress response/tolerance. This may reflect that certain aspects of an individual's speech, while being affected by environmental stressors, may not involve the same mechanism as performance decline. However, there may be other aspects of speech that are involved in this way and hence correlated with stress response/tolerance. This may suggest that all speech features are not equal when identifying their relationship with stress response (performance decline) or stress tolerance. Environmental stress' effect on an individual's speech does not imply an effect on their cognitive performance to the same extent or in the same way. This finding supports a recommendation that while speech features employed for stress detection can be potential candidates for stress response/tolerance prediction, they do not provide an exhaustive set of features useful for such predictions and hence further features should be explored while developing measures for stress response/tolerance.

The flipside of the above finding is that a number of features that did not show significant difference between the `Neutral` and `Stress` conditions nevertheless showed a good correlation with the stress-induced change in executive functioning (i.e., Verbal Fluency Task) and with the external assessment of stress tolerance. To provide further support for the recommendation made above, this finding may suggest that subtle differences that do not reach significance in distinguishing between the `Neutral` and `Stress` conditions may still be useful as an indicator for stress response. Alternatively, this mismatch between the features that appear more frequently in the stress conditions and the features that predict stress response/tolerance elsewhere could be due to the difference in the type of stress between our stress induction method and the stress experienced by the participants during the selection week (for the Selection Assessment Data) or the small sample size. Hence, for development of a reliable speech-based measure for stress response/tolerance, further exploration would be useful to test the mismatch hypothesis above, and with larger datasets.

6 Conclusion and Future Work

The main contribution of this paper is to present a proof of concept identifying potential language markers for stress response which we plan to extend and refine in the future with more focused and larger trials. Specifically, we have explored the usefulness of specific acoustic features extracted from speech produced in two stress conditions for their relationship with stress response and stress tolerance. We identified duration and a number of frequency domain features that are significantly correlated with stress response or tolerance. In the future, we plan to extract further acoustic features to test a more extensive list of features as potential candidates for involvement in the development of the speech-based measure.

We also tested the effectiveness of two continuous speech sample collection tasks, Read Speech Task and Open-ended Question Task, in developing speech-based measures for stress response and tolerance. We found that both tasks provided complementary information in the speech features and hence it can be beneficial to use both these tasks for data collection. However, if one of the tasks needs to be selected, Open-ended Question Task would provide more information that has consequences for stress response/tolerance.

Since this study involved only 13 participants, in order to confirm the current results and develop a more reliable and robust measure of stress response/tolerance, we plan to extend it further by increasing the sample size. The participants in the current study belong to a very particular population, young males aspiring to serve in a military unit. One can reasonably assume that this special population tends to have high stress tolerance. Testing

general population with the current experimental design might reveal greater range of stress response scores. Hence, training on the general population may lead to greater predictive validity.

Additionally, in this study, we focused on correlations as a measure of the relationship between extracted acoustic features and stress response/tolerance. In our future work, we plan to use the identified acoustic features to develop and train machine learning models to predict the stress response/tolerance scores. Additionally, Open-ended Question Task's capability of extracting semantics-based features, such as sentiment and topic, will also be used to develop robust models for stress response and stress tolerance predictions.

7 Possible Use Cases and Ethical Considerations

The set of features identified in the current paper has a few possible use cases. First, as stress plays a central role in the development of psychopathology, it is possible that those who show greater response to stress are more likely to develop psychopathology. An assessment based on these features may be used to identify those who are at a greater risk to develop psychopathology later in life. These features could be used to build a real-time sensor to detect signs of stress response. Such a sensor could automatically identify when people are experiencing a heightened sense of stress and help appropriate parties to reach out for mitigation.

Privacy. Building such real-time sensors that can detect stress or stress response in individuals based on the speech/language they produce, however, calls for ethical considerations. For example, as the smart home assistance devices (Amazon's Alexa or Google home) become increasingly common in households, the companies that operate the devices can easily collect the speech data and detect stress. This information could be used for commercial and other purposes (e.g., showing an advertisement for vacation or spa upon detecting stress) without users' consent. In general, the previous discussions of privacy concerns regarding speech data have focused on the identifying information in the speech signal and the semantic content of the speech. The possible use case of the current study could result in information about one's emotional state also to be collected by third parties without consent.

To mitigate some of these undesired consequences, free public programs could be planned that educate users of technology what capabilities modern technologies have, how they can be used both for the good and for sabotage depending on who controls it, what privacy preferences are available to the users and how they can choose these preferences in a more informed manner to be able to benefit from the capabilities without being sabotaged. Developers of such technologies also have a responsibility to ensure easy access and ease of selection of preferences related to users' privacy.

Equality. In order to develop a fully automated measure for stress response expanding on the approach illustrated in this paper, one needs to rely on the speech-to-text systems' output to compute an individual's performance score. However, automatic speech recognition systems may perform differently for different social groups (Koenecke et al., 2020) or populations with different mental health conditions (Miner et al., 2020), for example. Hence, an automatic measure for stress response that uses speech-to-text transcriptions may not work as well for certain populations as it would for others. This could result in unintended biases against individuals belonging to certain social/clinical groups through misdiagnoses or misclassifications. To overcome such unintended results, one needs to account for the differences in performance of the components used to develop the automated measure of stress response for different populations. Involving individuals from different populations while developing such automated systems, e.g., by training systems on data obtained from them, can be helpful.

Study Ethics Statement. Approval of the oversight military Institutional Review Board (IRB) was obtained prior to starting the study. Informed consent was obtained from all study participants. The IRB protocol was followed without exception during performance of this research. The findings from the collected data are reported in aggregated forms and no identifying information is released.

Acknowledgements

This work was supported by DARPA/AFRL Contract FA8650-19-C-7944 within the DARPA Measuring Biological Aptitude (MBA) program. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of DARPA, AFRL or the U.S. Gov-

ernment. Authors thank Vanessa Oviedo, Shelby Greene, and Joel Schooler for their assistance with data collection, and the anonymous reviewers for their feedback. Authors also thank colleagues Timothy Broderick, Arash Mahyari, Ian Perera, Kurtis Gruters, Ursula Schwuttke and Vandana Puri for discussions and/or feedback. This document was approved by DARPA for Public Release, Distribution Unlimited.

References

Helen W Bland, Bridget F. Melton, Paul Welle, and Lauren Bigham. 2012. Stress Tolerance: New Challenges for Millennial College Students. *College Student Journal*, 46(2):362–375.

Malcolm Brenner and Thomas Shipp. 1988. Voice Stress Analysis. NASA. Langley Research Center, Mental-State Estimation 1987.

Malcolm Brenner, Thomas Shipp, E.T. Doherty, and P. Morrissey. 1983. Voice Measures of Psychological Stress: Laboratory and Field Data. In Ingo R. Titze and Ronald C. Scherer, editors, *Vocal Fold Physiology, Biomechanics, Acoustics, and Phonatory Control*, pages 239–248. The Denver Center for the Performing Arts, Denver.

N.P. Dhole and S.N. Kale. 2020. Stress Detection in Speech Signal Using Machine Learning and AI. In Debabala Swain, Prasant Kumar Pattnaik, and Pradeep K. Gupta, editors, *Machine Learning and Information Processing. Advances in Intelligent Systems and Computing*, volume 1101. Springer, Singapore.

Maria Dietrich, Katherine Verdolini Abbott, Jackie Gartner-Schmidt, and Clark A. Rosen. 2008. The Frequency of Perceived Stress, Anxiety, and Depression in Patients with Common Pathologies Affecting Voice. *Journal of Voice*, 22(4):472–488.

Krista Doyle. 2020. Aceable Field Notes: An Essay on Grit.

Tobias Esch, George B. Stefano, Gregory L. Fricchione, and Herbert Benson. 2002. The Role of Stress in Neurodegenerative Diseases and Mental Disorders. *Neuroendocrinology Letters*, 23(3):199–208.

Raul Fernandez and Rosalind W. Picard. 2003. Modeling Drivers' Speech under Stress. *Speech Communication*, 40(1-2):145–159.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and Denise S. Pallett. 1993. DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Technical Report N.

Cheryl L. Giddens, Kirk W. Barron, Jennifer Byrd-Cravenand Keith F. Clark, and A. Scott Winter. 2013. Vocal Indices of Stress: A Review. *Journal of Voice*, 27(3):390.E21–390.E29.

Felipe V. Gomes and Anthony A. Grace. 2017. Adolescent stress as a driving factor for schizophrenia development—a basic science perspective. *Schizophrenia Bulletin*, 43(3):486–489.

Google. 2019. Python Client for Cloud Speech API. Google Cloud Client Libraries for google-cloud-speech.

James J. Gross. 2014. Emotion Regulation: Conceptual and Empirical Foundations. In James J. Gross, editor, *Handbook of Emotion Regulation*, pages 3–20. The Guilford Press.

John H.L. Hansen and Sanjay A. Patil. 2007. Speech under Stress: Analysis, Modeling and Recognition. In Müller C., editor, *Speaker Classification I. Lecture Notes in Computer Science*, volume 4343, pages 108–137. Springer, Berlin, Heidelberg.

Ling He, Margaret Lech, Sheeraz Memon, and Nicholas Allen. 2008. Recognition of Stress in Speech using Wavelet Analysis and Teager Energy Operator. In *Proceedings of INTERSPEECH*, pages 605–608, Brisbane, Australia.

Eric S. Jackson, Mark Tiede, Deryk Beal, and D.H. Whalen. 2016. The Impact of Social-Cognitive Stress on Speech Variability, Determinism, and Stability in Adults Who Do and Do Not Stutter. *Journal of Speech, Language, and Hearing Research*, 59(6):1295–1314.

Bhagyalaxmi Jena and Sudhansu Sekhar Singh. 2016. Psychological Stress Speech Analysis: A Review. *International Journal of Engineering Research & Technology*, 4(28):1–4.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial Disparities in Automated Speech Recognition. In *Proceedings of the National Academy of Sciences*, volume 117, pages 7684–7689.

Mark D. Kramer, Christopher Patrick, and Robert F. Krueger. 2012. Delineating Physiologic Defensive Reactivity in the Domain of Self-Report: Phenotypic and Etiologic Structure of Dispositional Fear. *Psychological Medicine*, 42(6):1305–1320.

Kristiane Van Lierde, S. Van Heule, S. De Ley, E. Mertens, and S. Claeys. 2009. Effect of Psychological Stress on Female Vocal Quality: A Multiparameter Approach. *Folia Phoniatr Logop*, 61(2):105–111.

Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana,

Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, Keunwoo Choi, viktorandreevichmorozov, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Adam Weiss, Matt Vollrath, and Taewoon Kim. 2020. librosa/librosa: 0.8.0 (version 0.8.0). Librosa 0.8.0.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–24.

Adam S. Miner, Albert Haque, Jason A. Fries, Scott L. Fleming, Denise E. Wilfley, G. Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A. Arnow, W. Stewart Agras, Li Fei-Fei, and Nigam H. Shah. 2020. Assessing the Accuracy of Automatic Speech Recognition for Psychotherapy. *NPJ Digital Medicine*, 3(1):1–8.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between Acoustic Characteristics of Spontaneous and Read Speech and their Effects on Speech Recognition Performance. *Computer Speech Language*, 22(2):171–184.

Franz Pernkopf, Tuan Van Pham, and Jeff A. Bilmes. 2009. Broad Phonetic Classification using Discriminative Bayesian Networks. *Speech Communication*, 51(2):151–166.

Tarun Pruthi and Carol Y. Espy-Wilson. 2004. Acoustic Parameters for Automatic Detection of Nasal Manner. *Speech Communication*, 43(3):225–239.

Tarun Pruthi and Carol Y. Espy-Wilson. 2007. Acoustic Parameters for the Automatic Detection of Vowel Nasalization. In *INTERSPEECH*, pages 1925–1928, Antwerp, Belgium.

Brent S. Rushall. 1990. A Tool for Measuring Stress Tolerance in Elite Athletes. *Journal of Applied Sport Psychology*, 2(1):51–66.

Robert M. Sapolsky. 1996. Why Stress is Bad for Your Brain. *Science*, 273(5276):749–750.

Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Sebastian Schnieder. 2014. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive Physical Load. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Barbara Schuppler. 2017. Rethinking classification results based on read speech, or: why improvements do not always transfer to other speaking styles. *International Journal of Speech Technology*, 20:699–713.

Zeshu Shao, Esther Janse, Karina Visser, and Antje S. Meyer. 2014. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, 5:772.

Tom Smeets, Sandra Cornelisse, Conny W.E.M. Quaedflieg, Thomas Meyer, Marko Jelicic, and Harald Merckelbach. 2012. Introducing the Maastricht Acute Stress Test (MAST): A quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. *Psychoneuroendocrinology*, 37(12):1998–2008.

Gustavo E. Tafet and Charles M. Nemeroff. 2015. The Links Between Stress and Depression: Psychoneuroendocrinological, Genetic, and Environmental Interactions. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 28(2):77–88.

Arkun Tatar, Gaye Saltukoğlu, and Ercan Özmen. 2018. Development of a Self Report Stress Scale Using Item Response Theory-I: Item Selection, Formation of Factor Structure and Examination of Its Psychometric Properties. *Archives of Neuropsychiatry*, 55(2):161–170.

Kevin Tomba, Joel Dumoulin, Elena Mugellini, Omar Abou Khaled, and Salah Hawila. 2018. Stress Detection Through Speech Analysis. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE)*, volume 1, pages 394–398.

S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, and S. Dhas. 2020. Stress Detection through Speech Analysis using Machine Learning. *International Journal of Creative Research Thoughts (IJCRT)*, 8(5).

Wikibooks. 2018. Speech-Language Pathology/Stuttering/Stress-Related Changes.

Longfei Yang, Yinghao Zhao, Yicun Wang, Lei Liu, Xingyi Zhang, Bingjin Li, and Ranji Cui. 2015. The Effects of Psychological Stress on Depression. *Current neuropharmacology*, 13(4):494–504.

Habib Yaribeygi, Yunes Panahi, Hedayat Sahraei, Thomas P. Johnston, and Amirhossein Sahebkar. 2017. The Impact of Stress on Body Function: A Review. *EXCLI Journal*, 16:1057–1072.

A Full Set of Correlations of Difference Scores for Extracted Acoustic Features with the Performance Score and the Stress Tolerance Score

Feature	Performance Score		Stress Tolerance Score	
	RST	OQT	RST	OQT
Duration related features:				
Duration (Speech Rate)	**-.628***	-	.184	-
Time domain features:				
Amplitude Envelope	-.336	-.342	-.094	.330
Root Mean Square Energy	-.381	-.410	-.055	.377
Zero Crossing Rate	**.546**	-.115	-.470	.359
Frequency domain features:				
Magnitude Spectrum	.050	.337	**-.753****	-.492
Short-time Fourier Transform Spectrum	**-.502**	-.271	-.140	**.515**
MFCC1	-.282	-.341	-.247	**.617***
MFCC2	.017	-.297	-.217	.405
MFCC3	.019	.138	.030	-.300
MFCC4	-.464	.134	.223	.202
MFCC5	-.447	-.207	.099	-.094
MFCC6	.167	-.245	.293	-.125
MFCC7	-.155	**.524**	-.050	-.486
MFCC8	-.272	.432	-.451	**-.583***
MFCC9	-.423	.048	.132	**-.554***
MFCC10	.153	-.343	-.009	-.206
MFCC11	-.209	**.516**	-.037	**-.735****
MFCC12	.128	.288	.104	-.387
MFCC13	.411	-.250	-.167	.363
Delta MFCC1	.063	-.279	-.052	.360
Delta MFCC2	.061	.124	-.156	.028
Delta MFCC3	-.037	.125	-.240	-.028
Delta MFCC4	-.349	-.191	-.015	.090
Delta MFCC5	-.055	.221	-.308	.064
Delta MFCC6	-.268	.358	.016	-.151
Delta MFCC7	-.401	.439	.051	-.187
Delta MFCC8	-.159	.086	.025	-.205
Delta MFCC9	-.149	-.280	-.169	.014
Delta MFCC10	.469	-.261	-.413	.046
Delta MFCC11	.329	.129	-.474	-.177
Delta MFCC12	-.181	-.350	-.480	.203
Delta MFCC13	-.073	-.197	-.406	.068
Delta Delta MFCC1	-.183	**-.710****	.129	.322
Delta Delta MFCC2	-.293	-.122	-.057	-.118
Delta Delta MFCC3	-.003	.323	-.060	-.129
Delta Delta MFCC4	.187	.090	-.320	-.112
Delta Delta MFCC5	-.413	-.069	-.115	0.177
Delta Delta MFCC6	.256	.531	.059	-.199
Delta Delta MFCC7	.094	**.759*****	.038	-.090
Delta Delta MFCC8	-.163	**.519**	-.133	-.137
Delta Delta MFCC9	-.037	-.153	-.225	-.020
Delta Delta MFCC10	.353	-.459	.200	**.534**
Delta Delta MFCC11	.301	-.258	.097	.186
Delta Delta MFCC12	-.008	-.094	.293	-.015
Delta Delta MFCC13	-.456	.350	**.589***	-.155

Table 3: Correlations between the difference scores (`Stress - Neutral`) of the 45 features extracted from Read Speech Task (RST) and Open-ended Question Task (OQT) and the stress-induced change in Verbal Fluency Task performance (left) and the stress tolerance score from the selection assessment data (right). The duration feature was dropped for OQT since the task duration itself was set to two minutes. Moderate to high correlations (> .5) are indicated with the bold font. *** indicates $p < .005$, ** indicates $p < .01$, * indicates $p < .05$

.