Real-time Public Speaking Anxiety Prediction Model for Oral Presentations

Real-time Public Speaking Anxiety Prediction Model for Oral Presentations

Everlyne Kimani*
kimani.e@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Timothy Bickmore
t.bickmore@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Rosalind Picard
MIT Media Lab, Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Matthew Goodwin
Holly Jimison
Northeastern University
Boston, Massachusetts, USA

ABSTRACT

Oral presentation skills are essential for most people's academic and career development. However, due to public speaking anxiety, many people find oral presentations challenging and often avoid them to the detriment of their careers. Public speaking anxiety interventions that help presenters manage their anxiety as it occurs during a presentation can help many presenters. In this paper, we present a model for assessing public speaking anxiety during a presentation—a first step towards developing real-time anxiety interventions. We present our method for ground truth data collection and the results of neural network models for real-time anxiety detection using audio data. Our results show that using an LSTM model we can predict moments of speaking anxiety during a presentation.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**; • **Human-centered computing**;

KEYWORDS

Affective computing, public speaking anxiety, speech, real-time prediction

ACM Reference Format:
Everlyne Kimani, Timothy Bickmore, Rosalind Picard, Matthew Goodwin, and Holly Jimison. 2022. Real-time Public Speaking Anxiety Prediction Model for Oral Presentations. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3536220.3563686

*Now at Toyota Research Institute

ICMI '22, Nov 07–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9389-8/22/11…$15.00
https://doi.org/10.1145/3536220.3563686

1 INTRODUCTION

Oral presentations are challenging social and cognitive tasks that often require a significant amount of training and presentation experience for presenters to feel confident and prepared. Due to the social aspect of oral presentation, many people report experiencing public speaking anxiety (PSA). PSA is the most common type of anxiety disorder, with reported prevalence rates ranging from 20 - 35% in various community samples [1, 3, 36]. Among people with social anxiety, 97% report public speaking anxiety so severe that it impairs their performance [1].

Automated systems such as the CBT conversational agent system described in [20] and other virtual reality systems [24] designed for exposure therapy could make PSA therapies accessible to more presenters. Nevertheless, even with more accessible PSA therapy, presenters still need assistance in managing their anxiety during their presentations, because they cannot always remember and perform the techniques they have been taught and have rehearsed before their presentation.

Real-time automated systems for PSA, that detect and intervene on anxiety spikes that occur during a presentation, could address some of these barriers and make oral presentations less dreadful, increase presenters' confidence, and improve the quality of presentations. To ensure that these real-time PSA management systems are effective, robust real-time PSA detection models are required.

With recent advances in sensing technologies, continuous detection of anxiety level is now possible. Prior work in affective computing has shown that audio data can be used in various emotion recognition tasks including stress [11, 43]. In this paper, we describe our efforts towards building a real-time PSA detection models that use audio data. We aim to address the research question: *Can PSA be estimated in real-time with enough accuracy to support real-time interventions?*

Given the enormous challenge of gathering ground truth labels for training real-time anxiety models, in this paper we leverage a data collection technique in which presenters do a retrospective review of their videotaped presentation, recalling and annotating their level of anxiety at each moment during their presentation. We use audio data collected from presenters in an online setting to build a continuous PSA level prediction model. We report on our data collection method, model training and evaluation results based on a set of speech features.

2 RELATED WORK

In public speaking, the quality of speech is important. A speech signal carries a mixture of information about the speaker, including their affective state [33, 40]. Advances in speech technologies and artificial intelligence have made it possible to detect stress and anxiety from speech. A meta-analysis of the literature on the detection of state and trait anxiety from auditory and visual cues reported that state anxiety (a transitory, situational response to anxiety-eliciting stimuli such as public speaking) was recognized better by judges from auditory cues than visual cues [16]. However, trait anxiety was highly correlated with the rating of visual-only cues. This suggests that momentary change in anxiety is better detected using speech signals, whereas trait anxiety (a person's proneness to experience anxiety) is better detected using non-verbal behaviors.

2.1 Effect of Anxiety on Speech

In speech production, the adjustment of the laryngeal muscles, which control the tautness, geometry, and position of the vocal folds, modulates the airflow through the glottis and vocal tract to produce sound waves [47]. Speech production studies and vocal expression theories contend that the body's physiological responses can influence the tension of the vocal fold muscles [19, 39]. Stress and anxiety can increase the tautness of vocal fold muscles and respiration rate and as a result, influences the quality and expressivity of the voice produced [23, 32, 43].

Studies assessing the relationship between acoustic features and anxiety have shown that various features are impacted by speech. For example, in one study that examined the effect of fear on speech in patients who have panic disorder with agoraphobia, pitch variability was lower in fearful than in happy speech [15]. This finding was confirmed in another study that examined the relationship between acoustic characteristics, self-ratings, and listener-ratings of public speaking [14]. Other features such as jitters and Mel Frequency Cepstral Coefficients (MFCCs) have also been shown to be affected by state anxiety and to be correlated with both self and listener/observer ratings of anxiety [12, 23]. These features can be extracted and used in machine learning models, such as one proposed by Fernandez [11] to estimate stress or anxiety in real-time.

2.2 Real-time Public Speaking Anxiety Assessment

Recent work by Nirjhar et al. [29, 30] has proposed knowledge-based and data-based models that attempt to capture the temporal trajectories of acoustic and physiological PSA assessment. The knowledge-driven models are inspired by Bodie's [4] notion of salient temporal patterns of PSA: habituation, sensitization, and escalation. In a model evaluation study [30], using data collected from 71 real-life and 232 virtual reality (VR) public speaking sessions, the researchers showed that data- and knowledge-driven models, and their combination, can reliably estimate presenter's trait anxiety with significant moderate correlation values between the actual and estimated trait anxiety scores. This work however focuses on using acoustic and physiological data to understand individual difference in trait anxiety. Although this is useful knowledge, research on how these models predict momentary state anxiety need to be done.

Real-time stress and anxiety assessment, such as stress experienced during public speaking, has received little attention. Often in building stress and anxiety detection models, researchers use protocols such as the Trier Social Stress Test Protocol (TSST) [2], the Stroop Test [38] or the Montreal Imaging Stress Task (MIST) [6] that include a 5-10-minute presentation task. Most of the studies using these protocols rely on training samples in which acute stress is assumed, based on stimulus intended to induce stress. For example, all the data collected while presenting is labeled as social-evaluation stress, and the data collected while performing an arithmetic task is labeled as cognitive stress [17, 35]. Stress and anxiety are highly subjective, and a stimulus that induces stress in one person might not induce stress in another. Particularly in the public speaking context, where presentations can last from 3 minutes to more than an hour, defining the duration and intensity of stress or anxiety is challenging. Simply labeling all data collected during a public speaking task as anxiety or stress would lead to inaccurate anxiety detection models. Additionally, given the subjective nature of stress and anxiety, asking observers to label presenters' behavior as anxious or not can result in disagreements among observers.

One viable technique for ground truth label collection is retrospective review, in which participants review their recorded presentations and annotate the start, duration and intensity of their anxiety. One prior study [46] has used a similar method of collecting moment-to-moment anxiety experience to examine how differences in adolescents' psychophysiological reactivity are related to individual differences in trait anxiety. However, this method has primary been used in the gaming context to collect player affect response during game play [26, 27] but so far has not been used to collect ground truth data for training presentation anxiety detection models.

3 SPEECH-BASED REAL-TIME PREDICTION OF PSA DURING PRESENTATION

We designed a speech-based public speaking anxiety prediction model that monitors the acceleration and decay of presenters' state anxiety continuously during a public speaking task. The model is designed to be integrated into a presentation application that supports anxious speakers by intervening when a speaker is predicted to be anxious.

3.1 Data Set

3.1.1 Participants. Thirteen participants were recruited from fliers posted on online message boards of a US university in the Northeast. Participants were required to be 18 years of age or older, speak and read English, have some college education, and have public speaking experience. Of the 13 participants recruited, 12 participants (Males = 6, Females = 6, Mean age = 29), completed the study. Of these, 6 had high speaking competence as measured by the Self-Perceived Communication Competence Scale (SPCCS) [28], while 6 had moderate levels of speaking competence.

3.1.2 Study Protocol. In a 90-minute online study (on Zoom [48]), participants were asked to review, rehearse, and deliver a 7-minute online presentation on a pre-defined topic before an audience of three confederates. The topic was on a paper about about research-based annotation tool called texSketch [44]. At the beginning of the

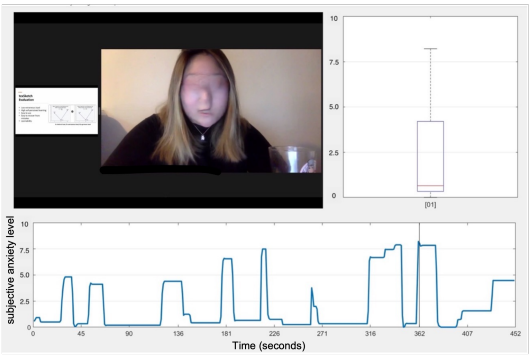

Figure 1: The line graph at the bottom shows retrospective PSA ratings over time of participant P3. The box plot on the upper right summarizes the distribution of ratings and on the top left is the video the participant viewed as they completed the ratings.

study, participants were consented and asked to complete baseline assessments (sociodemographic, SPCCS). They were then asked to relax for 5 minutes while watching a calming video. Self-reported public speaking anxiety before the presentation delivery was assessed using the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) questionnaire [37], Personal Report of Confidence as a Speaker (PRCS) [31] and the Subjective Unit of Discomfort Scale (SUDS) [45]. At the end of the presentation, participants were asked to complete an overall presentation self-rating (7-item, 7-point scale) questionnaire. One of the items in the questionnaire asked, *"How nervous were you during your presentation?"*

We used a retrospective review protocol to capture ground truth on participants' subjective anxiety during their presentation. Participants' online presentations were videotaped. Following their presentations, they were asked to watch their videotaped presentations, recall their anxiety experience, and, using a continuous affect rating application [13], record their anxiety at each second of their presentation. Anxiety was annotated on a scale of 0 (Calm and Relaxed) to 10 (Very Anxious). Figure 1, shows a box plot of one of the participant's ratings on the right and a line plot of their rating over time at the bottom.

Presentations lasted an average of *551.42 secs (SD = 92.41 secs)*. The total corpus comprised *1.83 hours* of presentation recordings and second-by-second anxiety ratings from 12 participants.

3.2 Predicting Public Speaking Anxiety Using Audio Data

3.2.1 Data Processing. Audio data was extracted from the recorded video presentation resulting in 12 wav files with a sample rate of 44KHz. We used the openSMILE toolkit [8] to identify voiced segments and extract eGeMAPS features [7], including 88 acoustic features at every second of voiced parts of the presentation. OpenSMILE has been widely used in automatic emotion recognition for affective computing [9, 18, 41] and acoustic features such as pitch, jitter, and shimmer have been shown to good predictors of stress [25, 34]. The features included statistical functionals (e.g., mean

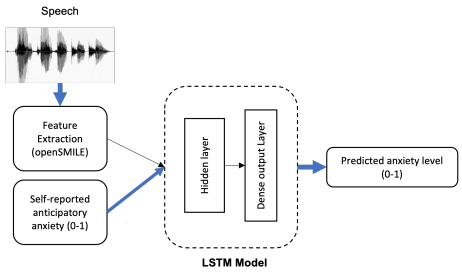

Figure 2: Real-time public speaking anxiety prediction model.

and standard deviation) of frequencies features(e.g., pitch and jitter), energy features (e.g., shimmer and harmonics-to-Noise Ratio (HNR)) and Spectral (balance) parameters (e.g., mel frequency cepstral coefficients (MFCCs)). The extracted feature columns were then normalized separately to the range 0-1. The data set was then transformed for use in a time-series anxiety forecasting task. The corresponding second-by-second anxiety labels were transformed from a scale of 0-10 to a scale of 0-1.

3.2.2 Speech-based PSA Prediction Models. A previous study [21] that explored when to nudge presenters to take a deep breath during a presentation showed that presenters want anxiety intervention when a system detects that they are getting anxious. We, therefore, performed an emotion forecasting task using the extracted speech features and the anxiety labels that captured presenters' anxiety at every second of the presentation. Emotion forecasting is the task of predicting future emotions based on a person's past and current audiovisual cues [42]. The minimum time taken by the participants to cover one slide was 30 seconds. Therefore, we designed a multivariate time-series LSTM model that would forecast PSA levels 30 seconds into the future based on 30 seconds of extracted acoustic features and passed anxiety predictions. A PSA intervention application could use the anxiety prediction model to determine if an anxiety intervention should be deployed because a presenter is predicted to be anxious in the next segment of their presentation.

We developed a speech-based prediction model that uses a Keras' long-short term memory (LSTM) model [5] with 128 memory units in the hidden layer, a rectified linear unit (RELU) activation unit, a single dense output layer (a deeply connected neural network layer) to predict presenters' anxiety 30 seconds into the future based on current and past audio features and predicted anxiety labels. The model is fit using the Adam stochastic gradient descent [22] and optimized using the mean squared error (MSE) loss function. Figure 2 visualizes the framework real-time PSA prediction model.

3.2.3 Model Evaluation. To evaluate the LSTM model, we split 12 participants' presentation data into training data (5 participants), validation data (3 participants), and test data (4 participants). We trained the model on the training data and used the validation set to fine-tune the model's performance. We then evaluated the model's performance in the test set. We used mean absolute error $MAE = \frac{1}{n}\Sigma|y - \hat{y}|$ and root squared mean error $RMSE = \sqrt{\frac{1}{n}\Sigma(y - \hat{y})^2}$ to assess and compare the performance of the LSTM model to

Table 1: Descriptive and Correlation Statistics of the Anxiety Measures

Measure	Range	Descriptive (SD - Standard Deviation, R - Range)	Correlation with Mean retrospective PSA ratings
PRCS	30 true-false questions (Scores: 0, No anxiety – 30 highest fear)	Mean = 9.33 (SD =5.91)	$\rho = 0.5, p = 0.07$
STICSA	21 items (1, Not at all – 4, Very much)	Mean = 28.83 (SD = 6.83)	$\rho = 0.5, p = 0.14$
SUDS	Single item: (0, Calm – 10, Very anxious)	Median = 4.5 (R = 1-7)	$\tau = 0.4, p = 0.07$
How nervous were you during your presentation?	Single item: (1, Not at all – 7, Very Nervous)	Median = 3 (R= 2-7)	$\tau = 0.5, p = 0.05^*$
Mean retrospective PSA ratings	Second-by-second rating (0, Calm – 10, Very anxious)	Mean = 3.41 (SD = 1.81)	$\rho = 1, p = 0.00$

a simpleRNN. The simpleRNN is a fully connected RNN where the output from the previous time step is fed to the next step. The current input and previous output are passed through a Tanh activation function.

4 RESULTS

4.1 Relationship between retrospective PSA ratings and other self-report PSA measures

The average of the mean retrospective PSA ratings of the 12 participants was 3.41. The means for presentations ranged from 1.04 to 6.07. Table 1 column three shows the means and medians of the self-report anxiety measures administered in the study. The first three measures in table 1 were collected before the presentation, and the last two, which include the retrospective PSA rating, were completed after the presentation. PRCS, STICSA, and Mean retrospective anxiety ratings are continuous normally distributed data. Therefore, we conducted a Pearson correlation test to assess the relationship between the PRCS and STICSA with mean retrospective PSA rating. SUDS and *"How nervous were you during your presentation?"* (Nervousness) are single item ordinal measures. Therefore, we conducted Kendall's coefficient of rank correlation to assess their correlation with the mean retrospective PSA ratings. The was a significant moderate positive relationship between the retrospective PSA rating and the presentation nervousness rating ($\tau = 0.5, n = 12, p = 0.05$). There was a close to significant correlation between PRCS, SUDS, and the mean retrospective PSA rating. The last column of Table 1 below summarizes the correlation results.

4.2 Speech-based PSA Prediction Model Performance

We used mean absolute error (MAE) and root mean squared error (RMSE) to compare the performance of the our LSTM to a simpleRNN model in predicting future 30 seconds PSA based on current and past speech patterns and past predicted anxiety. Based on the MAE and RSME metrics, the LSTM model performed better than a simpleRNN model in forecasting presenters' PSA. Table 2 shows the aggregated performance of the model after randomized

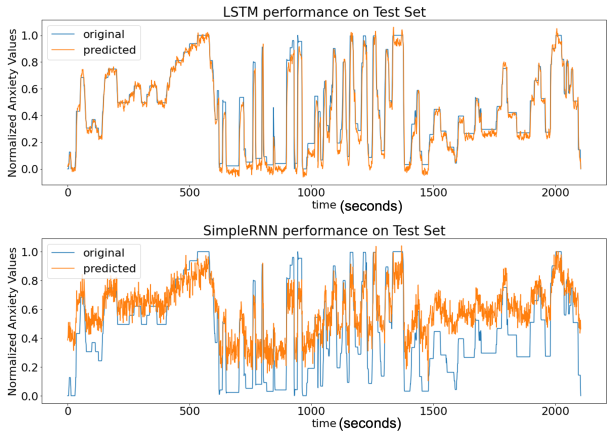

Figure 3: The top graph shows the performance of the LSTM in one trial with a test set of 4 participants data and the bottom graph shows the performance of the simpleRNN on the same test set.

5 data split trials. Figure 3 shows visualizes the prediction performance of the models. The top graph shows the performance of the LSTM in one trial with a test set of 4 participants data and the bottom graph shows the performance of the simpleRNN on the same test set.

Table 2: Speech-based Models Performance (Mean Absolute Error and Mean Root Mean Squared Error)

Model	MAE (Variance)	RSME (Variance)
LSTM	0.028 (5.48E-05)	0.042 (0.0001)
simpleRNN	0.156 (0.004)	0.183 (0.004)

5 CONCLUSION

In this work, we use speech to predict public speaking anxiety as it occurs during a presentation. We report a methodology for collecting subjective ground-truth ratings of public speaking anxiety and a

method for predicting PSA in real-time based on a model trained on this data. Using the retrospective presentation anxiety annotation methodology, we captured the variance in PSA experience during presentations. We demonstrated that our LSTM model performed better than a simpleRNN model for continuous anxiety prediction.

We found that the overall mean retrospective PSA rating was significantly moderately correlated with the self-report nervousness rating. Participants completed this nervousness rating immediately after their presentation and therefore had better recollections of their overall anxiety experience during a presentation. On the other hand, we did not find significant correlations between retrospective PSA ratings and the pre-post self-report PSA measures. Likely due to the small number of participants in our study. Another explanation could be that the self-report PSA measures used in this study (STICSA, SATI, SUDS, and PRCS) assess state anxiety at the moment of assessment and might not capture varying anxiety levels during a presentation. These measures may help evaluate an intervention's effects on anxiety, that is if an intervention leads to a reduction of PSA but may not capture continuous and dynamic PSA experienced during a presentation.

There are many advantages to using audiovisual and sensor-based tools to identify public speaking anxiety as it unfolds during a presentation, mainly because it is impossible to ask presenters to self-report their anxiety while presenting. We have shown that the continuous stress and anxiety data collecting methodology can be conducted in online presentation settings and be used in other contexts where collecting continuous affect labels for model training are of interest but interrupting the users would interfere with task performance.

The data collection study was conducted online with a confederate audience. Although this is a more realistic setting than VR settings, which have been used in recent modeling works [10], our study has some limitations. One is the small sample of presenters and the modeling data. The presentations were short (5 minutes), fully prepared, and may not represent typical presentations that presenters prepare independently.

Future work will evaluate the model with a larger dataset and live presentations of varying lengths and topics. Other inputs, such as physiological arousal, could also be included in the real-time anxiety assessment models to improve prediction reliability in different contexts. This current work lays the groundwork for expanding data collection on individual-specific factors and contexts that induce anxiety experiences and developing just-in-time PSA interventions.

REFERENCES

[1] Deborah C Beidel and Samuel M Turner. 1998. *Shy children, phobic adults: Nature and treatment of social phobia.* American Psychological Association, England.

[2] Melissa A Birkett. 2011. The Trier Social Stress Test protocol for inducing psychological stress. *JoVE (Journal of Visualized Experiments)* 56 (2011), e3238.

[3] John B Bishop, Karen W Bauer, and Elizabeth Trezise Becker. 1998. A survey of counseling needs of male and female college students. *Journal of College Student Development* 39, 2 (1998), 205–210.

[4] Graham D Bodie. 2010. A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication education* 59, 1 (2010), 70–105.

[5] François Chollet and others. 2018. Keras: The Python Deep Learning library. Astrophysics Source Code Library, record ascl:1806.022. Article ascl:1806.022 (June 2018), ascl:1806.022 pages. ascl:1806.022

[6] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, Veronika Engert, Sonia J Lupien, and Jens C Pruessner. 2005. The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry and Neuroscience* 30, 5 (2005), 319–325.

[7] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462. https://doi.org/10.1145/1873951.1874246

[9] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-Based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 445–450. https://doi.org/10.1145/2993148.2997632

[10] Kexin Feng, Megha Yadav, Md Nazmus Sakib, Amir Behzadan, and Theodora Chaspari. 2019. Estimating Public Speaking Anxiety from Speech Signals Using Unsupervised Transfer Learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, Ottawa, ON, Canada, 1–5. https://doi.org/10.1109/GlobalSIP45357.2019.8969502

[11] Raul Fernandez. 2004. *A computational model for the automatic recognition of affect in speech.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[12] Barbara F Fuller, Yoshiyuki Horii, and Douglas A Conner. 1992. Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety. *Research in nursing & health* 15, 5 (1992), 379–389.

[13] Jeffrey Girard. 2014. CARMA: Software for Continuous Affect Rating and Media Annotation. *Journal of Open Research Software* 2 (2014). https://doi.org/10.5334/jors.ar

[14] Alexander M Goberman, Stephanie Hughes, and Todd Haydock. 2011. Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech communication* 53, 6 (2011), 867–876.

[15] Muriel A Hagenaars and Agnes van Minnen. 2005. The effect of fear on paralinguistic aspects of speech in patients with panic disorder with agoraphobia. *Journal of Anxiety Disorders* 19, 5 (2005), 521–537.

[16] Jinni A Harrigan, Kelly M Harrigan, Beverly A Sale, and Robert Rosenthal. 1996. Detecting anxiety and defensiveness from visual and auditory cues. *Journal of Personality* 64, 3 (1996), 675–709.

[17] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. CStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 493–504. https://doi.org/10.1145/2750858.2807526

[18] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 553–560. https://doi.org/10.1145/3136755.3143009

[19] Tom Johnstone, Carien M Van Reekum, Tanja Bänziger, Kathryn Hird, Kim Kirsner, and Klaus R Scherer. 2007. The effects of difficulty and gain versus loss on vocal physiology and acoustics. *Psychophysiology* 44, 5 (2007), 827–837.

[20] Everlyne Kimani, Timothy Bickmore, Ha Trinh, and Paola Pedrelli. 2019. You'll be Great: Virtual Agent-based Cognitive Restructuring to Reduce Public Speaking Anxiety. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Cambridge, UK, 641–647.

[21] Everlyne Kimani, Ameneh Shamekhi, and Timothy Bickmore. 2021. Just breathe: Towards real-time intervention for public speaking anxiety. *Smart Health* 19 (2021), 100146. https://doi.org/10.1016/j.smhl.2020.100146

[22] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. (2014). https://doi.org/10.48550/ARXIV.1412.6980

[23] Petri Laukka, Clas Linnman, Fredrik Åhs, Anna Pissiota, Örjan Frans, Vanda Faria, Åsa Michelgård, Lieuwe Appel, Mats Fredrikson, and Tomas Furmark. 2008. In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior* 32, 4 (2008), 195–214.

[24] Lan Li, Fei Yu, Dongquan Shi, Jianping Shi, Zongjun Tian, Jiquan Yang, Xingsong Wang, and Qing Jiang. 2017. Application of virtual reality technology in clinical medicine. *American journal of translational research* 9, 9 (2017), 3867.

[25] Xi Li, Jidong Tao, Michael T Johnson, Joseph Soltis, Anne Savage, Kirsten M Leong, and John D Newman. 2007. Stress and emotion classification using jitter and shimmer features. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4. IEEE, Honolulu, HI, USA, IV–1081.

[26] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, San Antonio, TX,

USA, 158–163.
[27] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2021. The pixels and sounds of emotion: General-purpose representations of arousal in games. *IEEE Transactions on Affective Computing* 12 (2021), 1–1.
[28] James C. McCroskey and Linda L. McCroskey. 1988. Self-report as an approach to measuring communication competence. *Communication Research Reports* 5, 2 (1988), 108–113. https://doi.org/10.1080/08824098809359810
[29] Ehsanul Haque Nirjhar, Amir Behzadan, and Theodora Chaspari. 2020. Exploring bio-behavioral signal trajectories of state anxiety during public speaking. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Barcelona, Spain, 1294–1298.
[30] Ehsanul Haque Nirjhar, Amir H. Behzadan, and Theodora Chaspari. 2021. Knowledge- and Data-Driven Models of Multimodal Trajectories of Public Speaking Anxiety in Real and Virtual Settings. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 712–716. https://doi.org/10.1145/3462244.3479964
[31] Gordon L Paul. 1966. *Insight vs. desensitization in psychotherapy: An experiment in anxiety reduction.* Stanford University Press, California, USA. 402–403 pages.
[32] Silke Paulmann, Desire Furnes, Anne Ming Bøkenes, and Philip J Cozzolino. 2016. How psychological stress affects emotional prosody. *Plos one* 11, 11 (2016), e0165022.
[33] Rosalind W Picard. 2000. *Affective computing.* MIT press, London, UK.
[34] Katarzyna Pisanski and Piotr Sorokowski. 2021. Human stress detection: cortisol levels in stressed speakers predict voice-based judgments of stress. *Perception* 50, 1 (2021), 80–87.
[35] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, et al. 2011. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks*. IEEE, Chicago, IL, USA, 97–108.
[36] Alec Pollard and Gibson Henderson. 1988. Four types of social phobia in a community sample. *Journal of Nervous and Mental Disease* 176, 7 (1988).
[37] Melissa J Ree, Davina French, Colin MacLeod, and Vance Locke. 2008. Distinguishing cognitive and somatic dimensions of state and trait anxiety: Development and validation of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA). *Behavioural and Cognitive Psychotherapy* 36, 3 (2008), 313–332.
[38] Patrice Renaud and Jean-Pierre Blondin. 1997. The stress of Stroop performance: Physiological and emotional responses to color–word interference, task pacing, and pacing speed. *International Journal of Psychophysiology* 27, 2 (1997), 87–97.
[39] Klaus R Scherer. 1986. Vocal affect expression: a review and a model for future research. *Psychological bulletin* 99, 2 (1986), 143.
[40] Klaus R Scherer and Agnes Moors. 2019. The emotion process: Event appraisal and component differentiation. *Annual review of psychology* 70 (2019), 719–745.
[41] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication* 53, 9-10 (2011), 1062–1087.
[42] Sadat Shahriar and Yelin Kim. 2019. Audio-Visual Emotion Forecasting: Characterizing and Predicting Future Emotion Using Deep Learning. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*. IEEE, Lille, France, 1–7. https://doi.org/10.1109/FG.2019.8756599
[43] George M Slavich, Sara Taylor, and Rosalind W Picard. 2019. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress* 22, 4 (2019), 408–413.
[44] Hariharan Subramonyam, Colleen Seifert, Priti Shah, and Eytan Adar. 2020. *TexSketch: Active Diagramming through Pen-and-Ink Annotations.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376155
[45] Barry Tanner. 2012. Validity of Global Physical and Emotional SUDS. *Applied psychophysiology and biofeedback* 37 (03 2012), 31–4. https://doi.org/10.1007/s10484-011-9174-x
[46] Xiao Yang, Nilam Ram, Jessica P Lougheed, Peter Molenaar, and Tom Hollenstein. 2019. Adolescents' emotion system dynamics: Network-based analysis of physiological and emotional experience. *Developmental psychology* 55, 9 (2019), 1982.
[47] Zhaoyan Zhang. 2016. Mechanics of human voice production and control. *The journal of the acoustical society of america* 140, 4 (2016), 2614–2635.
[48] Zoom. 2021. Video Conferencing, Web Conferencing, Webinars, Screen Sharing. https://zoom.us/. (2021). Accessed: 2021-01-11.