Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills

Arushi
School of Science and Technology
James Cook University
Singapore
arushi@my.jcu.edu.au

Roberto Dillon
School of Science and Technology
James Cook University
Singapore
roberto.dillon@jcu.edu.au

Ai Ni Teoh
School of Psychology
James Cook University
Singapore
aini.teoh@jcu.edu.au

***Abstract*—This paper describes a work in progress Virtual Reality (VR) based serious game, integrated with an algorithmic classification model for detecting stress during public speaking in real-time by analysing the speaker's voice. The designed VR game offers real-time virtual social support/feedback for the training of public speaking skills. We developed a stress detection model that recognises the stress and an altered normal confident state based on 24 actors' voice expressions from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Three different classifiers models were constructed for each actor by extracting and identifying overall significant voice features. The results show that random forest classification using features Amplitude Envelope (AE), Root-Mean-Square (RMS) and Mel-Frequency Cepstral Coefficients (MFCCs) provides high accuracy for detection of stress, while many more features can possibly be explored.**

Keywords—Virtual Reality (VR), Voice analysis, Real-time stress detection, Public speaking stress, Machine Learning

I. Introduction

Public speaking is considered as a vital skill [1]. It has several benefits to individuals at various levels, such as rapport building and networking in professional as well as social contexts. Stress reaction is the predominant functional state of the speaker during public speaking [2]. Some level of stress is beneficial for optimal performance, however high levels of stress lead to fatigue and reduced performance. Detecting features to recognise stress in real-time during public speaking is a complex and challenging problem to be solved, as it is a phenomenon which results in various physiological changes such as increased heart rate and blood pressure as well as changes in voice/breathing patterns [3]. Stress can be reduced by participating in training programs as well as receiving social support [4].

We propose a virtual reality (VR) based game combined with a real-time stress detection model and offer feedback to the speaker for improving public speaking skills.

In this paper, our main contributions are as follows:

- This work presents the idea of a real-time stress detection model from voice, using VR to provide real-time social support/feedback for improving public speaking skills.
- We found voice features with significant differences specifically indicative of stress, such as Amplitude Envelope (AE), Root-Mean-Square (RMS) and Mel-Frequency Cepstral Coefficients (MFCCs).
- We selected the following classifiers: random forest, K-nearest neighbor (KNN) and Support Vector Machines (SVM) to test and train them with voice features indicative of stress to conceptualise the implementation. The results show that the random forest model detects stress in voice better than the other two classification models for each of 24 actors in the sample database (RAVDESS).

II. Background And Emotion Detection

VR is described as a human-to-computer interface determined to fully immerse users in a simulated environment. It provides a safe space for human-computer interactions for various purposes such as training [5]. VR has also been utilised for improving interpersonal skills [6], which includes public speaking [7]. Stress can be detected by conducting public speaking tasks in a laboratory environment [8]. This can also increase the accessibility to the training needed for public speaking skills in an immersive environment without the need to organise real audience for practice sessions.

For this purpose, a real-time voice analysis stress detection model can be developed and integrated into VR. In general, voice features can be extracted from a recorded voice sample. The time domain features such as intensity, maximum amplitude can be extracted and interpreted. Spectral features such as fundamental frequency (F_0) and formants can be obtained by converting the time-based signal into the frequency domain using Fourier Transform [9, 10]. The combination of spectral and time-domain features is known as prosodic features. These features can be used to detect emotions from speech signals. It has been reported that emotions such as anger, fear, anxiety, happiness, or surprise, which are considered to be high arousal emotions, result in increased intensity as well as a high mean F_0 [11].

Various algorithms have been used to detect affective states from voice features extracted from an audio signal by using various machine learning algorithms. Among the most common machine learning algorithms are Gaussian Mixture Model (GMM) [12, 13], Hidden Markov Model (HMM) [14], Support Vector Machines (SVM) [15], Artificial Neural Networks (ANN) [16] and Deep Neural Network (DNN) [17, 18]. However, minimal research has been carried out to analyse stress in the context of real-time public speaking [19]. Furthermore, previous studies have only attempted to discover the effect of a virtual audience on emotions such as anxiety and fear, but not stress, by utilizing VR for public speaking [20 -24]. These studies lack clear evidence on how the virtual audience was trained to detect the speaker's affective state and give feedback in VR environment. Most importantly, there is also no evidence of using machine learning algorithms for real-time audio and

978-1-6654-3886-5/21/$31.00 ©2021 IEEE

voice analysis to detect affective state such as stress during public speaking in VR.

III. Ongoing Work On Proposed VR-Based Serious Game for Public Speaking

A. Virtual Environment

The VR based serious game prototype is being developed using the Unity game engine (Fig. 1). The public speaking auditorium includes a virtual audience capable of offering three different types of feedback: positive, negative, and neutral.

Fig. 1. VR Auditorium setup for the public speaking training game

The stress detection model implemented in VR will identify the excessive negative stress that hinders the optimal performance during public speaking by analysing the speaker's voice in real-time. The emotional reaction of the virtual audience will depend on the performance of the speaker, which is judged by the level of stress detected from the voice of the speaker in real-time during public speaking. To make the setup more realistic, various subsets of the audience are programmed to give different feedback at different times. The game can be played in two different modes: training mode, which provides encouraging feedback when high levels of stress are detected, and competitive mode, which assigns a score to the speaker and provides positive or negative feedback according to the speaker's performance based on detected stress level.

B. Behaviours of Virtual Audience

One group of the audience may show neutral behaviour by not showing any interest (no smiling, not paying attention or appreciation). Another subset of the audience may also appear neutral by displaying signs of laziness or boredom by turning their heads towards to left and right. Others can indicate neutral behaviour through a display of sleepiness by yawning and assuming a dull body with smiles, happiness and nodding of heads in approval.

Next, a subset of the audience starts to boo the speaker in different ways. For example, the audience could show thumbs down, make noise and angry faces between each act of booing. Furthermore, when running the game in competitive mode, there will be a visual board similar to the "rock meter" in the guitar hero [25] games, showing the performance and score of the player according to the detected stress levels. Players' scores are saved in a leaderboard for ranking purposes as well as for tracking a player's progress.

Fig. 2. A virtual audience in the front row is showing positive, negative, unattentive behaviour, respectively.

C. Machine Learning Model Development

In order to develop an initial model for the detection of stress in real-time during public speaking, we chose the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [26]. The database contains 24 professional actors (12 males and 12 females) speaking two lexically different statements in seven different emotions (calm, happy, sad, angry, fearful, surprise and disgust). Each speech expression was produced with a normal, strong and neutral level of emotional intensity. We downloaded a total of 1440 available speech files of all the actors with 60 trials per actor. While our aim is not to classify the original emotions contained in the database, it can still be meaningful to classify neutral (i.e., calm) and an altered emotional state (i.e., fearful) for acted by all actors, as the latter with normal emotional intensity can be considered, similar to stress, while the former can be considered confident, for our purposes.

D. Feature Extraction

For our work, we used the Librosa python package [27] for analysing the features of respective confident and stressed states for all 24 actors.

We began by extracting time-domain features such as amplitude envelope (AE), root-mean-square (RMS) and zero-crossing rate (ZCR). We then extracted frequency domain features such as spectral centroid (SC), spectral-bandwidth (SB) and spectral rolloff (SR). Additionally, we decided to extract spectral feature, i.e., Mel-frequency cepstral coefficients (MFCCs).

Following extraction of features, we conducted a t-Test assuming unequal variance for two samples obtained for each extracted voice feature. We formed the following hypotheses and rejection criteria, and the results are summarised in Fig. 3.

H_0: There is no significant difference between respective actor's normal calm (confident) and normal fearful (stressed) voices for the extracted feature.
H_1: There is a significant difference between respective actor's normal calm (confident) and normal fearful (stressed) voices for the extracted feature.

Reject Criteria: *Reject H_0 if p-value < 0.05*

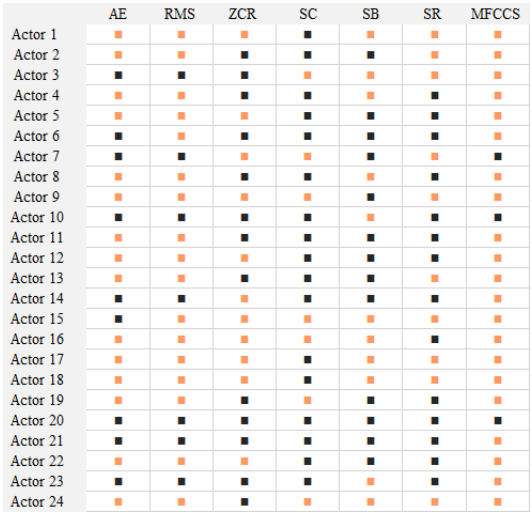

Fig. 3. The orange boxes indicate significant difference (p-value <0.05) between respective actor's normal calm (confident) and normal fearful (stressed) voice's extracted features and black boxes show non-significant difference (p-value >0.05) for the same, where AE = Amplitude Envelope, RMS = Root-Mean-Square, ZCR = Zero-Crossing Rate, SC = Spectral Centroid, SB = Spectral-bandwidth, SR = Spectral rolloff and MFCCs = Mel-Frequency Cepstral Coefficients.

Results (Fig. 4.) indicate that Actor 1, Actor 15, Actor 16, Actor 17 and Actor 18 are a few of the actors in the dataset who showed significant differences between their normal calm (confident) and normal fearful (stressed) voice's extracted features for AE, RMS, MFCCs, ZCR and SB except for Actor 9 and Actor 24. While most of the other actors have a significant difference between four or less extracted voice features.

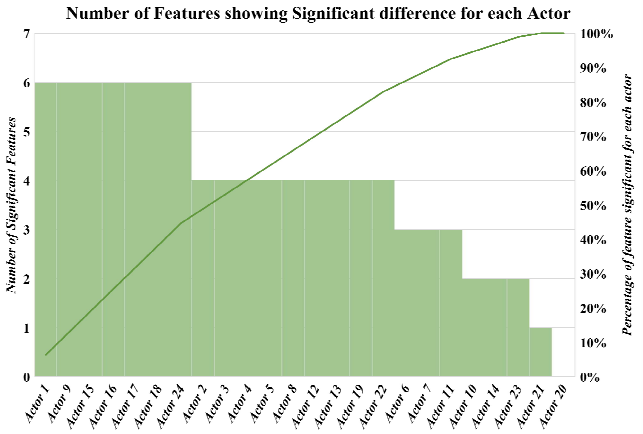

Fig. 4. Total number of features showed a significant difference for each actor

E. Feature Selection

Fig. 5. indicate that MFCCs, RMS and AE showed a significant difference highest number of times. Furthermore, these three features were also significant for Actor 1, Actor 15, Actor 16, Actor 17 and Actor 18, who showed the most significant difference between their normal calm (confident) and normal fearful (stressed) vocal expressions. Therefore, we selected the MFCCs, RMS, AE to train and test various algorithms.

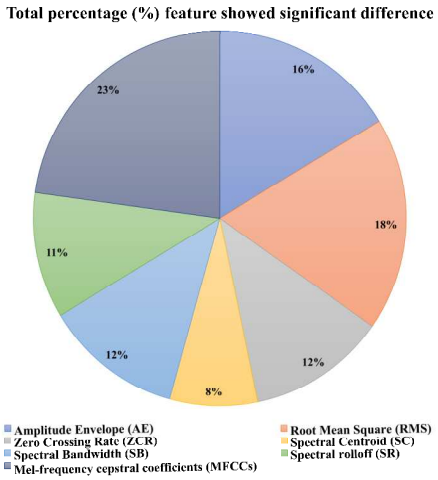

Fig. 5. MFCCs = 23%, RMS = 18%, AE = 16% showed significant difference percentage respectively

F. Training and testing of model development

Different machine learning algorithms exist for the classification of emotions and affective states from voice. We decided to begin our conceptualisation and work with Random Forest classification algorithms, K-nearest neighbor (KNN) algorithm and Support Vector Machines (SVM). We kept 70% of the data for training and 30% for testing for each actor's voice data sample.

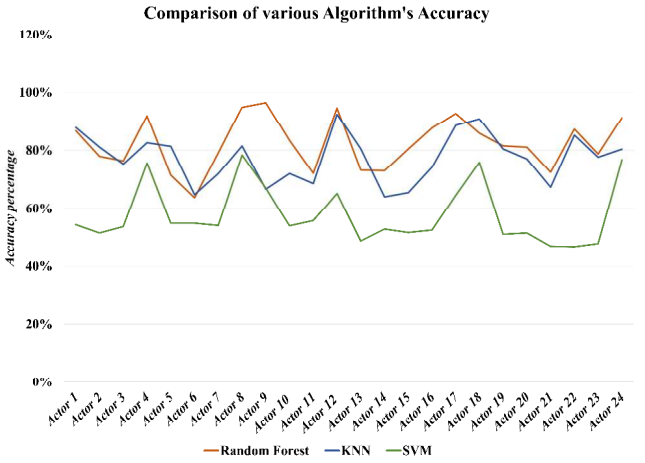

Fig. 6. Comparison of random forest, KNN and SVM for detection of stress levels for each of the actor. Overall accuracy of random forest = 82%, KNN = 77.39% , SVM =57.80%

IV. Ongoing Results

In our ongoing work, up until now, we have developed the basic VR game set-up for public speaking, which includes a virtual audience able to provide social support/feedback. To develop a stress detection model, we also explored the RAVDESS database, began with extracting various time-domain features such as zero-crossing rate (ZCR), root-mean-square (RMS) and amplitude envelope (AE). We further extracted frequency domain features such as spectral centroid (SC), spectral bandwidth (SB) and spectral rolloff (SR). Spectral features such as Mel-frequency cepstral coefficients (MFCCs) were also extracted later for all the 24 actor's voice samples for neutral calm voice (confident) and normal fearful voice

(stressed). After the extraction of voice features for all the 24 actors, we performed t-Test assuming unequal variance. The results indicate that different features showed significant differences for each actor's vocal expression. However, the most common voice features which showed significant difference were root-mean-square (RMS), amplitude envelope (AE) and Mel-frequency cepstral coefficients (MFCCs). We next selected few algorithms such as random forest, KNN and SVM to train and test the classifiers for detection and classification of stress from extracted voice features. Furthermore, since we also played with some of the voice features, the accuracy percentage changed a little bit. The results showed that random forest (accuracy 82%) outperformed KNN (accuracy 77.39%) and SVM (accuracy 57.80%) with an increased accuracy of 5% and 24% more for detection of stress with features including RMS, AE and MFCCs.

V. Conclusion and Future Work

This paper proposes a serious VR game integrated with a real-time stress detection model by analysing voice during public speaking to provide real-time virtual social support/feedback for training public speaking skills. The developed stress detection model will be integrated into the VR game's virtual audience using reinforcement learning techniques. Based on the stress levels detected from the analysis of specific audio features in the speaker's voice, which is done without any attempt at understanding the content and meaning of the actual speech, the virtual audience in the game will provide social support (in training mode) and appropriate, realistic feedback (in competitive mode) to train users and make them feel more comfortable in a real public speaking situation. Additionally, the project outcomes would also contribute to advancing research on integrating emotion detection models into the designs of intelligent and automated virtual agents in VR. We are currently setting up additional experiments to work with more actor speeches in order to carefully study the most relevant features for the classification of the stressed and neutral confident speech.

References

[1] L. Schreiber, "informative speaking," The Public Speaking Project (Ed.). Public speaking. The virtual text, pp. 15-1, 2011.

[2] M. Koroleva, A. Bakhchina, I. Shyshalov, S. B. Parin, and S. A. Polevaia, "Influence Of The Context Of Public Speaking On Human Functional State," *International Journal of Psychophysiology,* vol. 94, no. 2, pp. 230-231, 2014, doi: 10.1016/j.ijpsycho.2014.08.901.

[3] L. F. Droppleman and D. M. McNair, "An experimental analog of public speaking," *Journal of Consulting and Clinical Psychology,* vol. 36, no. 1, p. 91, 1971.

[4] S. S. Fredrick, M. K. Demaray, C. K. Malecki, and N. B. Dorio, "Can social support buffer the association between depression and suicidal ideation in adolescent boys and girls?," *Psychology in the Schools,* vol. 55, no. 5, pp. 490-505, 2018.

[5] A. Nijholt, "Breaking Fresh Ground in Human-Media Interaction Research," *Frontiers in ICT,* vol. 2014, 2014.

[6] M. Schmid Mast, E. P. Kleinlogel, B. Tur, and M. Bachmann, "The Future Of Interpersonal Skills Development: Immersive Virtual Reality Training With Virtual Humans," *Human Resource Development Quarterly,* vol. 29, no. 2, pp. 125-141, 2018, doi: 10.1002/hrdq.21307.

[7] M. Chollet and S. Scherer, "Perception of virtual audiences," *IEEE computer graphics and applications,* vol. 37, no. 4, pp. 50-59, 2017.

[8] D. Jezova, N. Hlavacova, I. Dicko, P. Solarikova, and I. Brezina, "Psychosocial Stress Based On Public Speech In Humans: Is There A Real Life/Laboratory Setting Cross-Adaptation?," *Stress,* vol. 19, no. 4, pp. 429-433, 2016, doi: 10.1080/10253890.2016.1203416.

[9] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handbook of affective sciences,* pp. 433-456, 2003.

[10] K. R. Scherer, "Expression of Emotion in Voice and Music," *Journal of Voice,* 1995.

[11] A. Kappas, U. Hess, and K. R. Scherer, "Voice and emotion," *Fundamentals of nonverbal behavior,* vol. 200, 1991.

[12] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features And Classifiers For Emotion Recognition From Speech: A Survey From 2000 To 2011," *Artificial Intelligence Review,* vol. 43, no. 2, pp. 155-177, 2015.

[13] I. J. Tashev, Z. Q. Wang, K. Godin, and Ieee, "Speech Emotion Recognition based on Gaussian Mixture Models and Deep Neural Networks," presented at the 2017 Information Theory and Applications Workshop, 2017.

[14] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Detection of stress and emotion in speech using traditional and FFT based log energy features," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, 2003, vol. 3: IEEE, pp. 1619-1623.

[15] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged Support Vector Machines For Emotion Recognition From Speech," *Knowledge-Based Systems,* vol. 184, p. 104886, 2019/11/15/ 2019.

[16] S. Agarwalla and K. K. Sarma, "Machine learning based sample extraction for automatic speech recognition using dialectal Assamese speech," *Neural Networks,* vol. 78, pp. 97-111, 2016.

[17] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks For Large-Vocabulary Speech Recognition," *Ieee Transactions On Audio, Speech, And Language Processing,* vol. 20, no. 1, pp. 30-42, 2011.

[18] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A New Approach Of Audio Emotion Recognition," *Expert Systems With Applications,* vol. 41, no. 13, pp. 5858-5869, 2014, doi: 10.1016/j.eswa.2014.03.026.

[19] T. Pfister and P. Robinson, "Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis," *IEEE Transactions on Affective Computing,* vol. 2, no. 2, pp. 66-78, 2011, doi: 10.1109/T-AFFC.2011.8.

[20] M. Slater, D.-P. Pertaub, and A. Steed, "Public speaking in virtual reality: Facing an audience of avatars," *IEEE Computer Graphics and Applications,* vol. 19, no. 2, pp. 6-9, 1999.

[21] Poeschl, S. (2017). Virtual reality training for public speaking—a QUEST-VR framework validation. *Frontiers in ICT*, *4*, 13. (22)

[22] El-Yamri, M., Romero-Hernandez, A., Gonzalez-Riojo, M., & Manero, B. (2019). Designing a VR game for public speaking based on speakers features: a case study. *Smart Learning Environments*, *6*(1), 12. (25)

[23] Chollet, M., Stefanov, K., Prendinger, H. and Scherer, S. Public speaking training with a multimodal interactive virtual audience framework - Demonstration. City, 2015.

[24] Chollet, M., Wörtwein, T., Morency, L.-P., Shapiro, A. and Scherer, S. Exploring Feedback Strategies To Improve Public Speaking: An Interactive Virtual Audience Framework. In Proceedings of the Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Osaka, Japan, 2015). Association for Computing Machinery, [insert City of Publication],[insert 2015 of Publication].

[25] Guitar Hero - Wikipedia. (2021). Retrieved 29 May 2021, from https://en.wikipedia.org/wiki/Guitar_Hero

[26] Livingstone, S. R. and Russo, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13, 5 (2018), e0196391.

[27] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, 2015.