

300COM / 303COM Declaration of originality

I Declare that This project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

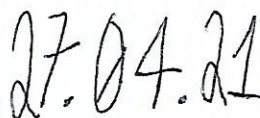
Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialize products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information, please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (<https://ethics.coventry.ac.uk/>) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: 

Date: 

Please complete all fields.

First Name:	Daniel
Last Name:	Jones
Student ID number	7929997
Ethics Application Number	P119305
1 st Supervisor Name	Beate Grawemeyer
2 nd Supervisor Name	

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

A Speech Emotion Recognition Classifier to Aid Performance Review in Learning Environments

Daniel G. Jones

jonesd37@coventry.ac.uk

Student ID: 7929997

Supervised by **Dr. Beate Grawemeyer**

ac7655@coventry.ac.uk

Submitted to the School of Computing, Engineering and Mathematics
Coventry University

Contents

1	Abstract	4
2	Acknowledgements	4
3	Introduction	5

1 Abstract

Classifying human emotion from speech is a challenging problem that has the potential to benefit various fields in different ways. One such field is in learning environments, where the application of such technology could aid in matters such as performance review / analysis or in helping children with special needs, such as autism. In this project, a speech emotion classification model has been built, trained, and deployed in the form of a graphical tool that analyses either pre-recorded audio or a live audio input and outputs the weighted predictions. The model has been developed using a 2-Dimensional Convolutional Neural Network (CNN) approach, where the training data is in the form of audio files converted into Mel Spectrograms. Through this approach, an exact prediction accuracy of 68.8% (and binary positive / negative accuracy of 94%) has been achieved on an unseen test dataset, with further personalised tests yielding promising results.

2 Acknowledgements

With thanks to my project supervisor, Dr. Beate Grawemeyer, for the initial inspiration and constant support throughout the project.

With thanks to my friend Céline Capelli of ETH Zürich for testing the application, providing useful audio data for testing, and for offering her unique perspective.

With thanks to my partner Hannah Vogt of the Pädagogische Hochschule Zürich for her love, inspiration, motivation, and for offering the insights of a primary school teacher.

3 Introduction

It can be observed that technology has yet to make a significant contribution towards aiding performance review in learning environments and indeed learning environments on the whole. This can largely be attributed to the lack of effectiveness of the existing tools, whereby much of the information obtained is neither insightful nor particularly meaningful.

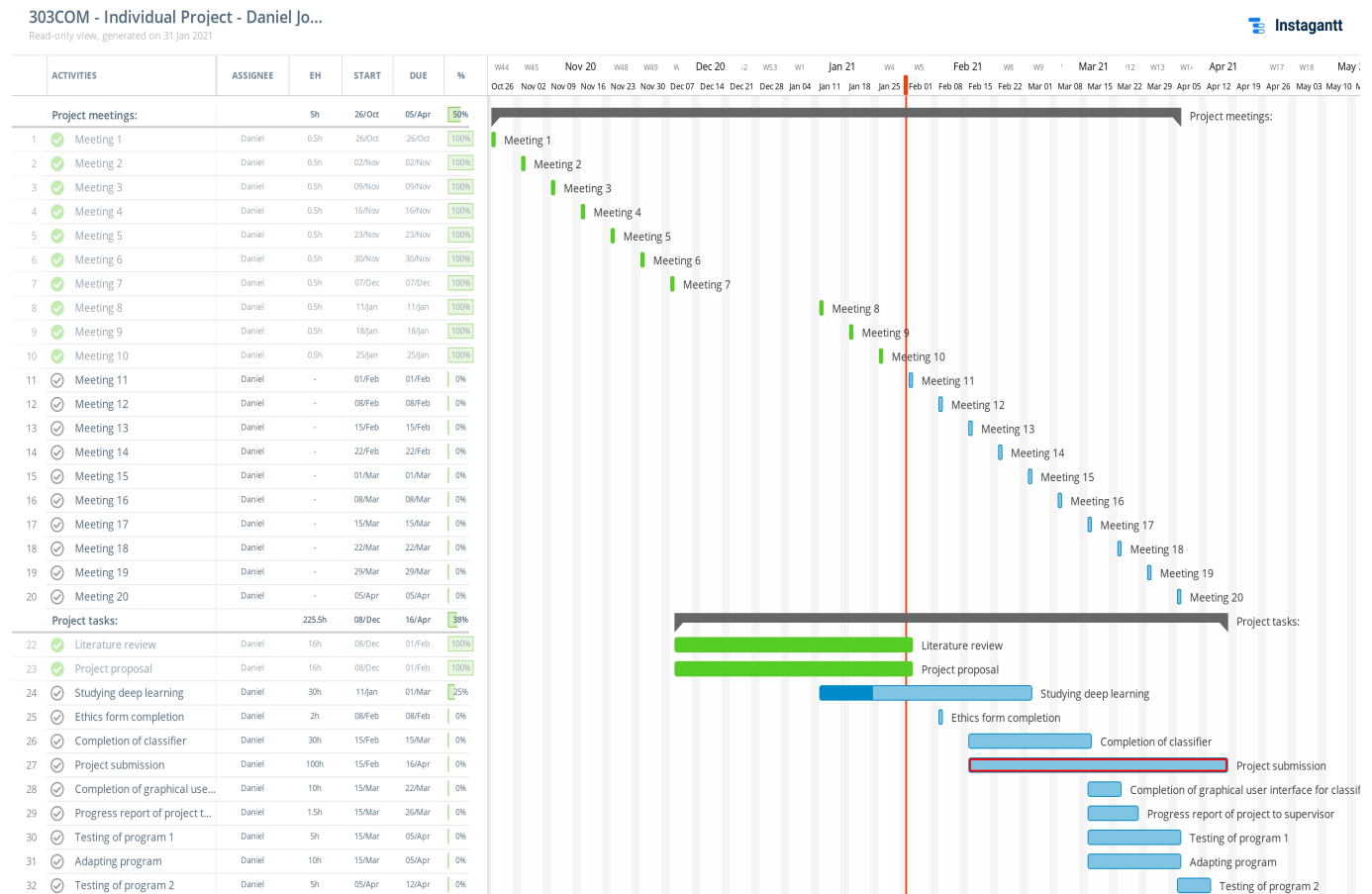
Traditional approaches to the incorporation of technology in learning environments have tended to focus more on interactivity as opposed to analysis, which has resulted in a lack of progress in the areas of performance review, analysis, and overall utility. This is due to the complexity of such environments, with many subjective metrics and mediums that, at first glance, appear to be difficult to study and quantify.

In this project, speech emotion has been identified and selected as a suitable metric for analysis. Speech emotion is a language agnostic metric that transcends many of the barriers posed by the more obvious metrics such as spoken / body language or facial expressions.

Despite the many nuances of speech, such as tonality or pitch, there are clear patterns that can be detected when studying how emotions are conveyed. A spectrogram is a visual representation of the spectrum of frequencies across a signal as a function of time. Consider the following two spectrograms of two different men saying a short sentence in a blatantly angry manner.

Gantt Chart

In order to illustrate the schedule of the project, I have produced a Gantt chart, shown below.



[Link to full size image](#)

[Link to latest edition of the chart](#)

5. Initial Literature Review

The first set of literature has been selected with development, methodologies, and comparisons between Speech Emotion Recognition (SER) technologies in mind. From the perspective of a beginner, starting with such papers makes sense; it enables one to develop an initial understanding of the current methods and strategies used to tackle the problem of SER.

Automatic Speech Emotion Recognition Using Machine Learning

The first paper, Automatic Speech Emotion Recognition Using Machine Learning (2019), takes a comparative look at SER systems. The underlying method to extract and process the speech signals (using Mel-frequency cepstrum coefficients and modulation spectral features in conjunction with feature selection) remains the same whilst the machine learning paradigms differ. The paper first looks at a recurrent neural network approach and goes on to compare it against both multivariate linear regression and support vector machine approaches. The paper makes the claim that such approaches were selected due to their existing popularity in the field.

The initial criticism of the paper is that it fails to consider the convolutional neural network (CNN) approach, despite its notable popularity and differences to the other methods. However, the paper is still concise and informative enough to warrant a full review. One positive quality that immediately stands out is how clear the information is presented, both in terms of looks and in use of language. An additional outstanding quality of this paper is the lack of pre-requisite knowledge required of the reader; many similar such papers assume at least a fundamental understanding and generally a lot more, whilst this paper describes most of the necessary knowledge required to grasp the content.

The paper starts by explaining the importance behind SER technology, the main argument can be reduced to how effective and convenient it is to obtain information pertaining to the emotional state of an individual in comparison to other methods (such as facial expressions and physiological signals). This argument can be applied to the idea of using such technology in learning environments based on the documented effectiveness of the results in similar scenarios. The paper even uses learning environments as an example of a potential application of such technology: “a teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment.” (Kerkeni et al., 2019, p. 2).

The authors then describe their system from a top-down perspective, this includes which emotions are in scope as well as the decision making process behind which algorithms are selected. It appears that the researchers were restricted by their datasets as to which emotions ended up in-scope. Whilst this does not affect the content of this particular paper, a more careful consideration will have to be made when applying this to learning environments (as per this project).

Delving further into the paper, the authors describe how the different algorithms are applied to solve the problem, which is achieved by mathematical notation and descriptions of important variables. This is of course a necessity for such a paper, although the notation is clear enough for even those with weaker mathematical backgrounds to gain insight. The paper concludes by displaying and contrasting between the results obtained from the

experiments; the main conclusion is that SER produces the best recognition results with a limited dataset whilst RNN performs better when a higher volume of data is available. Such results will prove vital when comparing against and assessing the effectiveness of our SER system.

A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition

Due to the notable omission of a CNN approach in the first reviewed paper, this paper has primarily been selected to gain practical insight into such an approach. The paper starts by defining the motivation behind it, using similar arguments to the previously reviewed paper. One particular area of interest is the point about “extracting hidden information” (Kwon, S. 2020, p. 1), where the author refers to uncovering previously undetected, new information using a CNN approach. At first glance, it appears that the author looks at the issue more through the lens of human computer interaction (HCI), contrasting with the previous paper’s theoretical approach.

An immediate positive point of the paper is that it describes its approach clearly and in plenty of detail, though unlike the first paper, this paper requires the reader to possess fundamental knowledge concerning machine learning terminology and methods. The paper goes on to discuss the proposed methodology, which is essentially a model comprising of input layers (taking in 2D speech signal spectrograms), several convolutional layers flattened down, and two fully connected layers with the softmax classifier applied at the end. Compared to the first paper, this approach seems easier and requires a lot less pre-processing work applied to a given dataset (since the audio just needs transforming into a 2d spectrograph). The main downside to this paper is that it lacks technical and theoretical insight, providing only a single equation unrelated to the classifier itself.

Cutting to the end results of the paper, given the clear lack of training applied to the model, the results achieved are not bad. Although lower than one might expect given some of the cutting-edge results achieved through CNNs on other projects; the lack of training surely impacts this. In conclusion, the paper has provided an insightful and inspiring approach to be considered for this project. As with the prior paper, the end results will be interesting to compare against in the end.

This next section of literature focuses more on the learning environment side of the project. The purpose of this is to justify the motivations behind the project and to refine and further develop the research questions.

Factors Affecting Technology Integration in the Classroom

This paper provides solid reasoning on the critical factors affecting the integration of technology in learning environments. The paper consists of 5 key reasons as to why classrooms benefit less from technology than many other environments: poor infrastructure, inadequate technology, lack of sufficient, effective professional development, low self-efficacy, and teacher perceptions. Each reason stated in the paper is insightful, but the points regarding inadequate technology and teacher perceptions are of the most relevance to the project. As previously mentioned, one of the goals of this project would be to put forth a valid and viable application of technology in learning environments (as to build some interest around the idea). The paper states that teachers “perceive the effort needed to learn the new technology and practicality of it as a significant consideration in whether they use it or not” (Harrell, S., & Bynum, Y. 2018. p. 15), so in order for this to work, the educators must be convinced of the technology.

Modeling learners’ cognitive and affective states to scaffold SRL in open-ended learning environments

Whilst this paper largely covers an area unrelated to the project, the points regarding relationships between cognitive and affective states are interesting and will be of use to the analysis stage of the project. The paper goes on to describe an interesting correlation between two affective states, boredom and delight, and academic performance. As one might expect, an exposed state of boredom significantly decreases cognitive and subsequently academic performance, as documented by the paper. With these statistics in mind, prioritising the prediction of the states of boredom and delight from speech will be areas of particular interest to focus training the model on.

6. Bibliography

- Harrell, S., & Bynum, Y. (2018). Factors affecting technology integration in the classroom. *Alabama Journal of Educational Leadership*, 5, 12-18
- Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M.A. & Cleder, C. (2019). Automatic Speech Emotion Recognition Using Machine Learning. In *Social media and machine learning*. IntechOpen. <https://doi.org/10.5772/intechopen.84856>.
- Kwon, S. (2020). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1), 183. <http://doi.org/10.3390/s20010183>.
- Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018, July). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In *Proceedings of the 26th conference on user modeling, adaptation and personalization* (pp. 131-138). <https://doi.org/10.1145/3209219.3209241>