



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

The Influence of Video and Audio Quality in Emotion Detection

Dylan Fitzpatrick

A Report

Presented to the University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of

B.A.(Mod.) in Computer Science

Supervisor: Khurshid Ahmad

April 2024

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Dylan Fitzpatrick

February 2, 2025

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Dylan Fitzpatrick

February 2, 2025

Abstract

This paper observes the influence of noise on emotion recognition systems. Emotion recognition is an up-and-coming area in the machine learning industry, and it is important to evaluate the robustness of these systems as they continue to develop. This study compares the performance of two Facial Emotion Recognition (FER) systems, FACET and AFFDEX, and two Speech Emotion Recognition (SER) systems, openSMILE and OpenVokaturi. A gold-standard dataset and a small dataset of Irish politicians are tested for each type of system. Gaussian noise is added to the videos at varying intensities and White noise is added to the speech signals. Image Quality Assessment (IQA) scores are found using the BRISQUE and NIQE algorithms and Speech Quality Assessment (SQA) scores are found using the NORESQA algorithm. Four methods are used to assess the influence of noise on the systems. First, the systems' accuracies are found compared to the gold-standard dataset labels. Changes in the accuracies of all systems are seen as noise increases. Then, Cohen's Kappa coefficients between the systems and the gold-standard datasets are calculated to determine if the agreements between them are more than what would be expected by chance. Both of the FER systems continuously have agreements that are more than chance, whereas the SER systems only have chance agreements as the noise levels get more intense. Spearman's Rank-Ordered Correlation Coefficient (SROCC) is subsequently calculated between the system outputs and quality scores. The correlations for each system are weak and vary depending on which dataset is being tested. Finally, Kruskal-Wallis Tests between the systems are calculated to measure similarities in their distributions. The similarities between the distributions of each system change with noise, and the change is once again different depending on the dataset. FACET proves to be the FER system that is more resilient to noise than AFFDEX. OpenVokaturi is determined to be the more robust SER system.

Acknowledgments

First of all, I would like to thank my supervisor, Professor Khurshid Ahmad, for his guidance and support throughout the past few months. He expressed a high level of confidence in me and was very helpful whenever I encountered any issues. Working with Khurshid was an absolute pleasure.

I would like to thank my family for showing me unwavering support throughout the years. In particular, I would like to thank my sister, Erica and my mother, Patricia. I could not have made it this far without them and I am eternally grateful for all of their guidance throughout my life.

I would also like to thank my friends for being a support system for me whenever times were difficult. In particular, I would like to thank my friend and coursemate, Zemyna. Throughout our time in college together, she has always lifted me up whenever I found myself struggling with my studies.

Finally, I would like to thank my partner, Jamie. He has stuck by me through thick and thin and has always made me feel as though I can achieve anything that I set my mind to. He continues to make me a better version of myself with each passing day and that is reflected in the work in this project.

DYLAN FITZPATRICK

University of Dublin, Trinity College
April 2024

Contents

Acknowledgments	iv
Chapter 1 Introduction	1
1.1 Problem Statement	2
1.2 Layout of the Report	3
Chapter 2 Background & Literature Review	4
2.1 Image Quality Assessment	4
2.2 No-Reference Image Quality Assessment	5
2.2.1 Natural Scene Statistics	5
2.2.2 BRISQUE Score	5
2.2.3 NIQE Score	5
2.2.4 Noise in Images	6
2.3 Facial Emotion Recognition	7
2.3.1 Facial Action Coding System	8
2.3.2 Image Quality and Facial Emotion Recognition	9
2.4 Speech Quality Assessment	10
2.4.1 NORESQA Score	10
2.4.2 Noise in Speech Signals	10
2.5 Speech Emotion Recognition	11
2.5.1 Classification	13
2.5.2 Speech Quality and Emotion Recognition	13
2.6 Conclusion	13
Chapter 3 Design & Implementation	14
3.1 System Overview	14
3.2 Database Selection	15
3.2.1 RAVDESS	15
3.2.2 Emo-DB	16

3.2.3	Politicians	16
3.3	Creating and Adding noise	18
3.3.1	Video	18
3.3.2	Speech	19
3.4	Quality Scoring	20
3.4.1	BRISQUE and NIQE	21
3.4.2	NORESQA	22
3.5	Facial Emotion Recognition	24
3.5.1	FACET Algorithm	24
3.5.2	AFFDEX Algorithm	26
3.5.3	FER Implementation	27
3.6	Speech Emotion Recognition	28
3.6.1	openSMILE Architecture	29
3.6.2	openSMILE Implementation	29
3.6.3	OpenVokaturi Algorithm	31
3.6.4	OpenVokaturi Implementation	32
3.7	Statistical Analysis	32
3.7.1	System Accuracies	32
3.7.2	Cohen’s Kappa	33
3.7.3	Spearman’s Rank-Ordered Correlation Coefficient	33
3.7.4	Kruskal-Wallis Test	35
Chapter 4 Evaluation		36
4.1	Video	36
4.1.1	System Accuracies	36
4.1.2	Cohen’s Kappa	38
4.1.3	Spearman’s Rank Ordered Correlation Coefficient	39
4.1.4	Kruskal-Wallis Test	40
4.2	Speech	41
4.2.1	System Accuracies	42
4.2.2	Cohen’s Kappa	45
4.2.3	Spearman’s Rank Ordered Correlation Coefficient	45
4.2.4	Kruskal-Wallis Test	46
Chapter 5 Conclusions & Future Work		49
5.1	Conclusions	49
5.2	Future Work	50

Bibliography	52
Appendices	56

List of Tables

2.1	LCC and SROCC of BRISQUE and NIQE compared to human ratings. From (Mittal et al. (2013))	6
3.1	Politician video details	17
3.2	Example of CSV file after calculating BRISQUE and NIQE	21
3.3	Average BRISQUE and NIQE scores on the RAVDESS dataset	22
3.4	Average BRISQUE and NIQE scores on the Politician dataset	23
3.5	Average NORESQA score on the Emo-DB and Politician datasets	24
3.6	Hyperparamter values for the openSMILE SVM and their MSE values	31
4.1	Cohen’s Kappa scores for the RAVDESS dataset at different qualities	38
4.2	SROCC values of FACET and AFFDEX with BRISQUE score (left) and NIQE score (right) on RAVDESS (* indicates p-value > 0.05)	39
4.3	SROCC of FACET and AFFDEX with BRISQUE score (left) and NIQE score (right) on Politician dataset (* indicates p-value > 0.05)	40
4.4	Kruskal-Wallis test on RAVDESS (* denotes p-value > 0.05)	40
4.5	Kruskal-Wallis test on Politicians (* denotes p-value > 0.05)	41
4.6	Cohen’s Kappa scores for the Emo-DB dataset at different qualities	45
4.7	SROCC of openSMILE and OpenVokaturi with NORESQA Score on Emo- DB (* indicates p-value > 0.05)	45
4.8	SROCC of openSMILE and OpenVokaturi with NORESQA Score on Politi- cian dataset (* indicates p-value > 0.05)	46
4.9	Kruskal-Wallis Test on Emo-DB (* denotes p-value > 0.05)	46
4.10	Kruskal-Wallis Test on Politicians (* denotes p-value > 0.05)	47

List of Figures

2.1	(a) Original Quality (b) Salt and Pepper noise (c) Gaussian Noise (d) Poisson Noise (e) Speckle Noise. From Owotogbe et al. (2019).	7
2.2	Muscles of Facial Expression: Lateral View. From Netter (2014)	8
2.3	Action Units 1-12. From Cohn et al. (2007)	9
2.4	A visualisation of Jitter and Shimmer	12
3.1	System Architecture	14
3.2	Gaussian Probability Distribution Functions (a) $\sigma = 0.5$ (b) $\sigma = 1$ (c) $\sigma = 2$	19
3.3	Example RAVDESS video at various qualities. (a) Original Quality. (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$	19
3.4	Gaussian Probability Distribution Functions (a) $\sigma = 0.01$ (b) $\sigma = 0.1$ (c) $\sigma = 1$	20
3.5	Speech signals with original signal (blue) and original signal + noise (yellow) (a) $\sigma = 0.01$. (b) $\sigma = 0.1$. (c) $\sigma = 1$	20
3.6	BRISQUE (blue) and NIQE (orange) score distributions on RAVDESS (a) Original Quality (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$. The primary x-axis shows the BRISQUE values and the secondary x-axis shows the NIQE values. . .	22
3.7	BRISQUE (blue) and NIQE (orange) score distributions on Politician dataset (a) Original Quality (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$. The primary x-axis shows the BRISQUE values and the secondary x-axis shows the NIQE values.	23
3.8	NORESQA score distributions on Emo-DB (a) Original Quality (b) $\sigma = 0.01$ (c) $\sigma = 0.1$ (d) $\sigma = 1$	24
3.9	NORESQA score distributions on Politicians (a) Original Quality (b) $\sigma = 0.01$ (c) $\sigma = 0.1$ (d) $\sigma = 1$	25
3.10	FACET pipeline from (Littlewort et al. (2011)).	26
3.11	AFFDEX pipeline from McDuff et al. (2016).	27
3.12	An example of FACET emotional output. Graphs from top to bottom: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness, Neutral.	27

3.13	An example of AFFDEX emotional output. Graphs from top to bottom: Anger, Contempt, Disgust, Sadness, Fear, Joy, Surprise, Engagement, Valence, Sentimentality, Confusion, Neutral.	28
3.14	Graph comparing BRISQUE (blue) and NIQE (orange) (top-left). The remaining graphs compare emotional values for FACET (blue) and AFFDEX (orange). (In order: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness and Neutral).	29
3.15	Overview of the architecture of openSMILE	30
4.1	Original Quality confusion matrices of (a) FACET and (b) AFFDEX . . .	36
4.2	Confusion matrices of (a) FACET and (b) AFFDEX for videos with Gaussian noise with $\sigma = 0.5$	37
4.3	FACET confusion matrices for videos with Gaussian noise with (a) $\sigma = 1$ and (b) $\sigma = 2$	38
4.4	Distributions of Anger for FACET (blue) and AFFDEX (orange) at the original quality (a) and at the first noise level (b)	42
4.5	Confusion matrices of (a) openSMILE (b) OpenVokaturi at the original quality	43
4.6	Confusion matrices of (a) openSMILE and (b) OpenVokaturi with white noise ($\sigma = 0.01$)	43
4.7	Confusion matrices of (a) openSMILE and (b) OpenVokaturi with white noise ($\sigma = 0.1$)	44
4.8	Confusion matrices of (a) openSMILE and (b) OpenVokaturi with white noise ($\sigma = 1$)	44
4.9	The distributions of openSMILE (blue) and OpenVokaturi (orange) at the original quality (a), the first noise level (b) and the second noise level (c). .	48
1	Correspondence with RTÉ requesting politician videos	57
2	Correspondence with an author of PAM	57
3	Correspondence with an author of NORESQA and SQAPP	58

List of Abbreviations

σ - Standard Deviation

AU - Action Unit

BRISQUE - Blind/Referenceless Image Spatial Quality Evaluator

FACS - Facial Action Coding System

FER - Facial Emotion Recognition

FR - Full Reference

IQA - Image Quality Assessment

LCC - Pearson's (Linear) Correlation Coefficient

MVG - Multivariate Gaussian

NIQE - Naturalness Image Quality Evaluator

NMR - Non Matching Reference

NORESQA - Non-matching Reference based Speech Quality Assessment

NR - No Reference

NSS - Natural Scene Statistics

PDF - Probability Distribution Function

QoE - Quality of Experience

RR - Reduced Reference

SER - Speech Emotion Recognition

SQA - Speech Quality Assessment

SROCC - Spearman's Rank-Ordered Correlation Coefficient

SVM - Support Vector Machine

Chapter 1

Introduction

In the current day, commonly referred to as the "digital age", people consume more media than ever before. As a result, people consciously expect to receive a good Quality of Experience (QoE), which is the measure of an individual's satisfaction with a product or service. The quality of the media itself is the most likely factor to influence QoE. If the quality is degraded, the consumer's experience is affected negatively. Quality assessment methods have garnered large interest as a result, as they can assist in providing measures of QoE. However, the uses of quality assessment do not end with QoE. The factors that can diminish QoE can also have an impact on software systems that take media as input. Emotion recognition systems are one such example. These systems are a huge area of interest in the machine learning industry at the moment, and it is no surprise as to why.

Emotions are an extremely important part of communication between people, acting as powerful nonverbal cues that deepen our connections. We enhance our understanding of one another through emotional signals in a variety of ways, such as through facial expressions or tone of voice. Many systems and processes have been developed with the goal of recognising a person's emotions through the analysis of facial expressions and speech signals. Machine learning allows these processes to be done entirely automatically. With the recent massive growth in the machine learning industry, further developments can be expected, with a company recently developing what they claim to be the "first AI with emotional intelligence".

The implementation of these systems can lead to many real-world benefits in a variety of fields. For example, emotion recognition systems can allow companies to analyse how their advertisements make people feel. These tools can also provide considerable benefits to the healthcare industry, where they could be implemented to assist in diagnosing disor-

ders such as depression or dementia, as well as several other possibilities. The use cases of these systems are vast and have the potential to transform many aspects of our daily lives.

This project is being undertaken within a larger corpus of studies, which evaluate the effect of various factors on the agreement between emotion recognition systems, such as ethnicity, age and gender.

1.1 Problem Statement

Automated emotion recognition systems are trained on data of pristine quality. The interest of this project lies in how these systems react when quality degradation is introduced, specifically through the introduction of noise. Noise can be introduced to a video or a speech signal in a variety of ways, many of which are beyond the average person's control, especially if they do not have the knowledge of how it occurs. If the systems' accuracies are affected significantly, this must be considered when developing the systems further, in order to improve their robustness. It is important to acknowledge discrepancies in results caused by a lack of access to high-quality recording equipment or the introduction of noise, which can go unnoticed.

The aim of this project is to take two Facial Emotion Recognition (FER) systems, FACET and AFFDEX, and compare their performance when image quality is altered. The same comparison in performance is made with two Speech Emotion Recognition (SER) systems, openSMILE and OpenVokaturi, to see how they react to the degradation of speech quality. It will also allow for the detection of algorithm bias and data bias in the systems.

The objectives of this project, in detail, are as follows:

- To analyse the influence of video quality on two pieces of Facial Emotion Recognition software: Emotient Inc's FACET and Affectiva's AFFDEX on a gold standard dataset.
- To analyse the influence of speech quality on two pieces of Speech Emotion Recognition software: audEERING's openSMILE and Vokaturi's OpenVokaturi on a gold standard dataset.
- Use a small, non-gold standard dataset of videos of politicians and see how the FER and SER systems are affected by changes in image and speech quality.

- Compare how the two FER systems are affected by changes in image quality via various statistical analyses, and perform the same comparison on the SER systems with respect to speech quality.
- Perform an evaluation of these results and come to a conclusion about how the systems are affected by changes in quality.

1.2 Layout of the Report

The chapters of the report and their contents are as follows:

- Background & Literature Review - a review of the research undertaken as part of the project and previous work done in the area.
- Design & Implementation - outlines the practical elements of the project and how the data was collected and processed.
- Evaluation - an analysis of the results obtained from the methods used in the "Design & Implementation" chapter.
- Conclusions & Future Work - Outlines any conclusions that have been made from the results and suggests what future work should be undertaken in this research area.

Chapter 2

Background & Literature Review

This chapter will outline and review existing literature within the areas of quality assessment, emotion recognition and a combination of the two. The topics covered in this chapter are Image Quality Assessment, No-Reference Image Quality Assessment, Facial Emotion Recognition, Speech Quality Assessment and Speech Emotion Recognition.

2.1 Image Quality Assessment

Image Quality Assessment (IQA) is the process of determining the quality of an image and providing a numeric value or score to represent the quality. IQA has become a large area of interest in recent times as people are now consuming visual media more than ever. As a result, Mittal et al. (2012) claims that there is a concern with video streaming services around Quality of Experience (QoE) and that one of the aims of IQA is to give users optimised QoE by using objective measures of visual quality. Image quality covers a variety of factors, such as blurring and distortion.

There are several different methods for assessing image quality, with the first that existed being known as Full-Reference (FR) IQA models. FR models calculate the IQA score of a distorted image based on the existence of an original, undistorted image that can be referenced. Many FR IQA models exist, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al. (2004)) and multiscale SSIM (Wang et al. (2003)). Further advancements in IQA led to Reduced-Reference (RR) IQA, where some, but not all, of the information to the reference image exists (Wang and Bovik (2011)). For this project, the focus is on No-Reference (NR) IQA.

2.2 No-Reference Image Quality Assessment

NR IQA is the concept of evaluating the quality of an image without the availability of any information apart from the image itself (Mittal et al. (2012)). NR IQA methods are sought for this project as they are the methods that are most applicable to real-world scenarios, where images or videos are captured and analysed in real-time. Upon researching these models, two suitable NR IQA algorithms were discovered: Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) and Naturalness Image Quality Evaluator (NIQE).

2.2.1 Natural Scene Statistics

Both the BRISQUE score and NIQE score algorithms are based on Natural Scene Statistics (NSS). In Moorthy and Bovik (2011), the authors describe natural images as images that can be captured by a camera. They also explain that natural, undistorted images have statistical properties that are the same across different image contents. The hypothesis that they use for their approach to NR IQA is that distortion in natural images alters the natural statistical properties of images and, as a result, makes them "unnatural". They can take this "unnaturalness" and use it to relate a distorted image to perceived quality.

2.2.2 BRISQUE Score

BRISQUE score was first discussed in Mittal et al. (2012), where the algorithm is described as "an NR IQA model which utilises an NSS model framework of locally normalised luminance coefficients and evaluates 'naturalness' using the parameters of the model". The authors discuss in depth the calculation of the BRISQUE score, which involves retrieving 18 features in the image at two scales, giving a total of 36 features. These features include shape, mean and variance, which are calculated using pairwise products of neighbouring transformed luminances along four orientations: vertical, horizontal, main diagonal, and secondary diagonal. These features are then mapped to scores via regression, giving the final image quality score.

2.2.3 NIQE Score

The development of the NIQE score model is discussed in Mittal et al. (2013). The authors describe NIQE as an "opinion unaware" (OU) algorithm, meaning it does not need to be trained on databases of human-rated distorted images. They also state that the algorithm is "distortion unaware" (DU), which means that it strictly relies on exposure to

natural source images or image models rather than being trained on specific distortions. The authors compare BRISQUE and NIQE extensively, stating that the NSS features used for NIQE are similar to those seen in BRISQUE, the difference being that NIQE only uses NSS from natural images. In contrast, BRISQUE is trained on both natural and distorted images, as well as human judgments of these images. NIQE computes 36 identical NSS features from patches of an image and fits them to a Multivariate Gaussian (MVG) model, then compares this to the natural MVG model. The quality of the image is defined as the distance from the quality-aware NSS feature model and the MVG to the features in the distorted image.

	LCC (Overall)	SROCC (Overall)
BRISQUE	0.9395	0.9424
NIQE	0.9135	0.9147

Table 2.1: LCC and SROCC of BRISQUE and NIQE compared to human ratings. From (Mittal et al. (2013))

The authors evaluate both BRISQUE and NIQE using the median Spearman’s Rank Ordered Correlation Coefficient (SROCC) and the median Pearson’s (linear) Correlation Coefficient (LCC) across 1000 train-test combinations of the LIVE IQA database, a database made by the Laboratory for Image & Video Engineering at The University of Texas at Austin (Sheikh et al. (2006)). Table 2.1 shows that both of these systems perform very well when compared against human scores, justifying their use in this project.

2.2.4 Noise in Images

Owotogbe et al. (2019) state that four types of noise can be present in images: Salt and Pepper noise, Poisson noise, Speckle noise and Gaussian noise.

The authors describe each type of noise as follows:

Salt and Pepper noise originates from sudden, sharp changes in the image signal, which can be caused by faulty equipment. It appears as black and white pixels throughout the image.

Gaussian noise is a statistical noise with a Gaussian Probability Distribution Function (PDF). Gaussian noise is most commonly introduced through the acquisition or transmission of images.

Poisson noise is said to be caused by capturing an image with sensors that are not strong enough to locate statistical fluctuations in a photon measurement. Poisson noise follows

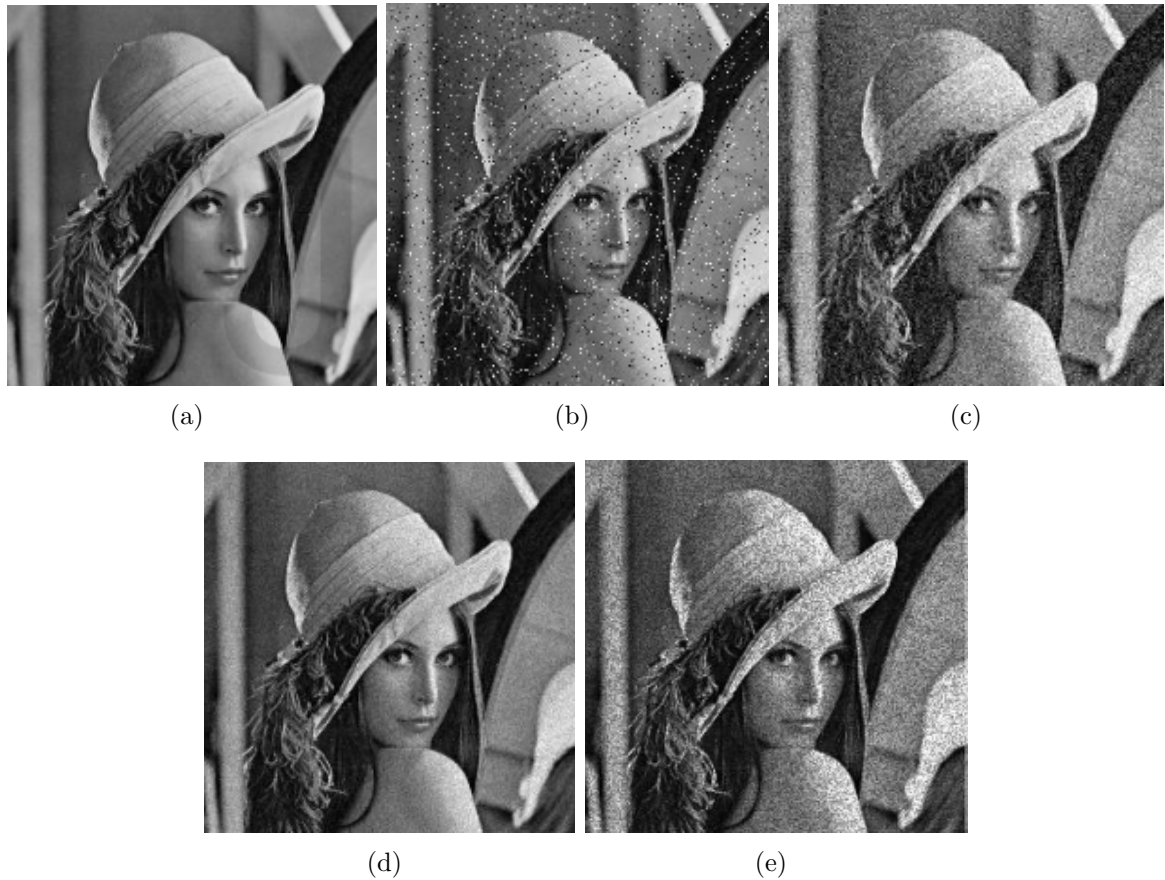


Figure 2.1: (a) Original Quality (b) Salt and Pepper noise (c) Gaussian Noise (d) Poisson Noise (e) Speckle Noise. From Owotogbe et al. (2019).

a distribution similar in nature to that of a Gaussian distribution.

Speckle noise is described as a type of granular noise. This noise can be found in many systems, such as ultrasound images and Synthetic Aperture Radar (SAR) images. It results from coherent handling of backscattered signals from numerous distributed points (Verma and Ali (2013)). Each type of noise is visualised in Figure 2.1.

By understanding the characteristics, causes and behaviours of different types of image noise, an informed decision could be made on how they could be applied in this project.

2.3 Facial Emotion Recognition

Facial Emotion Recognition (FER) is the process of using visual information of the face to attempt to classify what emotion an individual is feeling based on the movement of facial muscles. Conventional FER is split into three steps according to Ko (2018). The first step is face and facial component detection, in which the face region and facial landmarks

are detected. The second step is feature extraction and the third step is expression classification using pre-trained pattern classifiers. Ko states that deep-learning-based FER approaches reduce the reliance on pre-processing techniques by allowing learning to occur from the start of the pipeline, i.e. the input images.

2.3.1 Facial Action Coding System

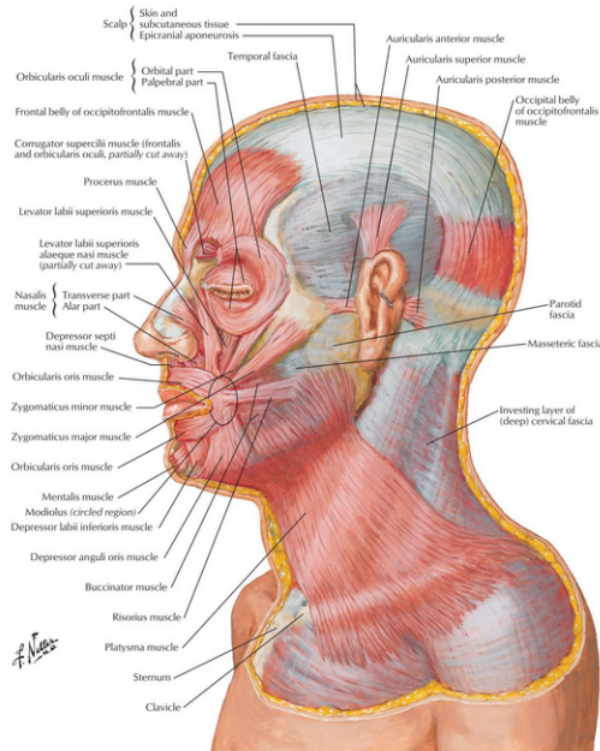


Figure 2.2: Muscles of Facial Expression: Lateral View. From Netter (2014)

Ekman and Friesen (1969) discusses the concept of emotional leakage. Their theory was that any emotional information which can be obtained through a person's words can also be obtained through analysing nonverbal behaviours. The authors then went on to develop the Facial Action Coding System (FACS), a system that describes facial expressions using Action Units (AUs) (Ekman and Friesen (1978)). They defined 44 AUs, 30 of which are related to contractions of specific facial muscles. FACS was updated in 1992, and again in 2002 (Cohn et al. (2007)). These updates provided changes to the AUs used and also the scoring system. A lateral view of the anatomy of facial muscles can be seen in Figure 2.2, giving a visual overview of the muscles that FACS is based on. These AUs can happen in combinations, and although the number of AUs is small, more than 7000 combinations have been observed according to Scherer and Ekman (1982). Figure 2.3











AU	Description	Facial muscle	Example image
1	Inner Brow Raiser	<i>Frontalis, pars medialis</i>	
2	Outer Brow Raiser	<i>Frontalis, pars lateralis</i>	
4	Brow Lowerer	<i>Corrugator supercilii, Depressor supercilii</i>	
5	Upper Lid Raiser	<i>Levator palpebrae superioris</i>	
6	Check Raiser	<i>Orbicularis oculi, pars orbitalis</i>	
7	Lid Tightener	<i>Orbicularis oculi, pars palpebralis</i>	
9	Nose Wrinkler	<i>Levator labii superioris alaquae nasi</i>	
10	Upper Lip Raiser	<i>Levator labii superioris</i>	
11	Nasolabial Deepener	<i>Zygomaticus minor</i>	
12	Lip Corner Puller	<i>Zygomaticus major</i>	

Figure 2.3: Action Units 1-12. From Cohn et al. (2007)

provides a visualisation of the first 12 AUs.

The development of FACS was a critical step in the development of FER systems, which still use versions of FACS today.

2.3.2 Image Quality and Facial Emotion Recognition

A research paper was written by Wallbott (1991) to determine if there is an impact of image quality on people's abilities to recognise emotions. In this paper, eighty judges rated 65 different images in 11 different conditions. Some examples of these conditions are undistorted, reduced spatial resolution at three scales and reduced contrast resolution at three scales. Wallbott found that, in most cases, the judges' ratings were not affected significantly by changes in picture size or contrast resolution. Wallbott found that the only time there was any significant difference in the judges' rating was when using the largest deterioration of spatial resolution. This paper serves as a good baseline of how image quality affects human ratings and can be taken into consideration when looking at how automated FER models compare. It also shows that there has been research interest in this area before.

2.4 Speech Quality Assessment

Speech Quality Assessment (SQA) is the process of determining the quality of a speech signal. There are several FR algorithms for evaluating speech quality, so they require a clean reference when performing assessments. As a result, their real-world applications are minimal. Two examples of FR SQA algorithms are PESQ (Rix et al. (2001)) and CDPAM (Manocha et al. (2021a)). Due to a lack of advancements in the field, NR SQA algorithms were hard to come by. As a result, the most suitable SQA algorithm for use in this project was found to be the NORESQA Score.

2.4.1 NORESQA Score

Non-matching Reference based Speech Quality Assessment (NORESQA) Score, is an SQA algorithm developed by Manocha et al. (2021b), which assesses the quality of speech recordings using Non-Matching References (NMRs). The authors take a test input and provide the ability to produce a relative quality score against any given reference, avoiding the need for a reference audio of the same speech signal. They do this using a neural network approach designed to detect degradation at the frame level (approximately 32ms of audio). In the paper, it is stated that the calculation is dependent on Signal-to-Noise Ratio (SNR), the ratio of signal power to noise power, and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), which is invariant to the scale of the processed signal and can be used to calculate quality. When the algorithm is compared against other SQA methods using 2 Alternative Forced Choice (2AFC) Tests to determine which system can get the most similar result to humans, FR models performed the best.

Although the authors showed that NORESQA does not provide results as accurate as existing FR models, the authors claim that NORESQA remains more useful in real-world applications, and so proves suitable for this project.

2.4.2 Noise in Speech Signals

Vaseghi (2008) defines signal noise as "any unwanted signal that interferes with the communication, measurement or processing of an information-bearing signal". There are many types of noise in signals: Acoustic noise, Electromagnetic noise, Electrostatic noise, Processing noise and "Channel Distortion, Echo, and Fading".

Vaseghi describes each of these types of noise as follows:
Acoustic noise comes from vibrating, moving or colliding sources and is the most common

kind of noise. It can be generated from many different sources such as fans, traffic and people speaking.

Electromagnetic noise is noise that is present at all frequencies but is most prevalent at radio frequencies. All electrical devices generate electromagnetic noise.

Electrostatic noise comes from voltage and can be caused by fluorescent lighting.

Processing noise comes from the analogue or digital processing of signals. An example of this is lost data packets in communication systems.

Channel Distortion, Echo and Fading come from nonoptimal characteristics within communication channels.

Vaseghi also states that, depending on the noise frequency/time characteristics, it can be classified into one of the following categories: Narrowband noise, White noise, Band-limited White noise, Coloured noise, Impulsive noise or Transient noise.

Vaseghi describes each of these classifications as follows:

Narrowband noise is noise with a narrow bandwidth.

White noise is random noise with a flat power spectrum.

Band-limited White noise is White noise with a limited bandwidth.

Coloured noise is any wideband noise with a non-flat power spectrum.

Impulsive noise has short pulses of random amplitude and duration.

Transient noise contains long-duration noise pulses.

Similarly to image noise, researching the different kinds of noise that can be found in speech signals was important to this project. A deeper understanding of these various types of noise will once again allow for informed decision-making when it comes to how noise is added to speech signals later on.

2.5 Speech Emotion Recognition

Speech Emotion Recognition (SER) is the process of classifying the emotion an individual is feeling through the analysis of their voice.

The authors of Teixeira et al. (2013) give the following examples of features that can be used for SER:

- **Fundamental Frequency (F0):** The number of times a sound wave produced by the vocal cords is repeated during a set period. It can also be defined as the number

of cycles of opening/closing of the glottis.

- **Jitter:** frequency variation from cycle to cycle.
- **Shimmer:** amplitude variation of the sound wave.
- **Harmonic to Noise Ratio (HNR):** the ratio between periodic and non-periodic components of a voiced speech segment.

Jitter and shimmer are visualised in Figure 2.4.

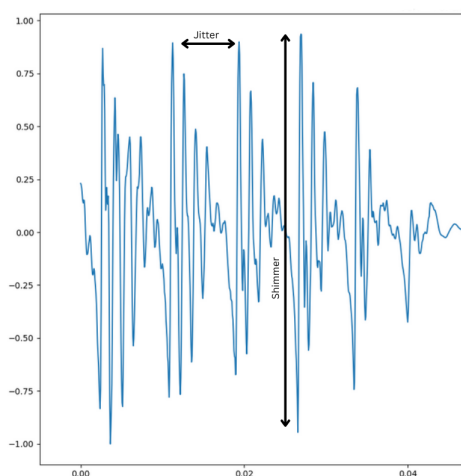


Figure 2.4: A visualisation of Jitter and Shimmer

In Swain et al. (2018), the authors explain how the consistencies seen in FER are not present as much in SER. These variations can be due to several factors, including the speaker's gender and the language spoken.

El Ayadi et al. (2011) outlines the difference between local and global features in SER. Local features are extracted from each frame, whereas global features are extracted from an utterance. The authors emphasise that it is important to differentiate between these as speech signals are not themselves stationary, but rather, the signal is said to be approximately stationary within each frame. It has been observed that global features are not great at classifying emotions with similar arousal, for example, Joy vs Anger. In addition, using global features results in the loss of temporal information. The authors state that there are four categories of speech features: Teager Energy Operator-based features (TEO), Qualitative features, Continuous features and Spectral features.

2.5.1 Classification

According to El Ayadi et al. (2011), SER comprises two stages: a front-end processing unit that retrieves the appropriate features from the speech data and a classifier. The authors claim that there has been no consensus on which classifier works best for SER. Examples of classifiers used in SER are Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Neural Networks, Support Vector Machines (SVM) and Multiple Classifier Systems (MCS).

It is important to be aware of how SER systems work and to be acquainted with the types of features considered when performing classification tasks. This provides a greater context for how the introduction of noise can impact the features of a speech signal.

2.5.2 Speech Quality and Emotion Recognition

Gobl and Chasaide (2003) discusses how the communication of emotions can be affected by changing the quality of a person's voice. In their paper, quality is referred to in a different sense than it is in this project. The paper states some examples of voice quality to be: harsh voice, breathy voice and whispery voice. The results of this paper show that differences in the quality of voice can take a neutral utterance and make it sound very different to if it was spoken normally. They also conclude that there is no one-to-one relationship between affect and voice quality.

2.6 Conclusion

The research undertaken in preparation for this study provided a plethora of highly relevant information that would be crucial going forward. Reviewing literature in this area also made it apparent that research into how modern emotion recognition systems react to changes in image or speech quality has not yet been conducted, highlighting the need for this study. To fill this gap in the literature, this study takes existing FER and SER systems and passes them through various levels of noise in order to see how they are affected.

A lot of highly relevant information on emotion recognition and one of the major uses of machine learning was learned through conducting the literature review. Deep insights were obtained into how IQA and SQA algorithms are created, the different types of noise in both image and speech signals and how they are introduced, how FER and SER systems have been developed over time and how previous research has been done to determine if changes in quality change human opinions on emotion recognition.

Chapter 3

Design & Implementation

This chapter will provide an in-depth view of the technologies and methods used throughout the project. It will begin with an overview of the system pipeline, followed by a discussion of how the databases were selected and pre-processed. Next, it will discuss how noise was added to the video and audio files, followed by how the IQA and SQA scores were calculated. Subsequently, the chapter will outline the systems used for emotion detection. Finally, it will demonstrate the statistical measures used to determine the influence of quality on the systems.

3.1 System Overview

This section will provide a brief overview of the pipeline that was created for this project.

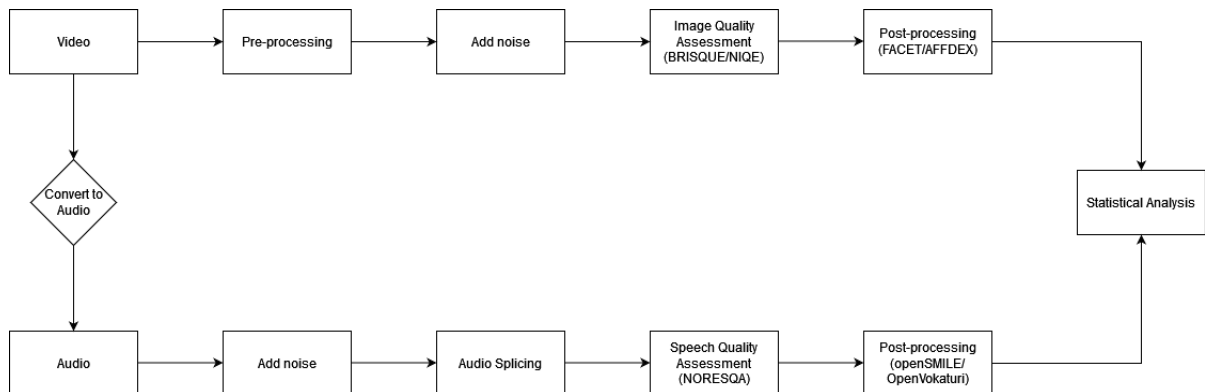


Figure 3.1: System Architecture

Figure 3.1 visualises the diagram of the pipeline used in this project. This pipeline is only relevant to any videos that are not part of the gold-standard datasets.

The pipeline begins with MP4 videos. The videos are pre-processed, if necessary, to be compatible with the two FER software systems. Gaussian noise is then added to each of the videos at varying levels of intensity. The original and noisy videos are passed through the two IQA scoring algorithms, BRISQUE score and NIQE score. The videos are then post-processed using the two FER software systems: FACET and AFFDEX.

The videos are converted from MP4 to WAV format for SER processing using Python’s soundfile library. White Gaussian noise is then added to these audio clips, and they are split into 2000ms segments. These segments are passed into the SQA algorithm, NORESQA score, and post-processed using the two SER software systems: openSMILE and OpenVokaturi.

Finally, hypothesis tests and other analyses are then performed on the video and audio files using their quality scores and the output of the emotion recognition systems.

3.2 Database Selection

This section will provide details of the gold-standard databases which were used for FER and SER, as well as the selection of the additional dataset and how it was prepared for processing.

3.2.1 RAVDESS

The gold-standard database chosen for the video analysis was the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), created by Livingstone and Russo (2018). The database consists of 24 professional actors, 12 male and 12 female, reading or singing one of two sentences at one of two levels of emotional intensity: normal or strong. The following emotions are portrayed in the dataset: *calm*, *happy*, *sad*, *angry*, *fearful*, *surprise* and *disgust*. A subset of 720 visual-only videos was used for this project, consisting of 30 videos per actor.

This dataset was chosen because the videos it provides are high quality and are all recorded in a professional environment. The creators state that each file was rated ten times by 247 individuals, and high levels of emotional validity were reported. The dataset also covers a sufficiently wide range of emotions for the comparison of the FER systems. Finally, the titles of the videos contain labels that outline what emotion is being portrayed in the video, allowing for the accuracy of the systems to be calculated. In total, there are

80,708 frames of video in this dataset, with a total length of 44:52.

3.2.2 Emo-DB

The Berlin Database of Emotional Speech (Emo-DB) is the gold standard database used for speech analysis. This database was developed by Burkhardt et al. (2005). It consists of 10 speakers, five male and five female. Each audio file in the database consists of one of these speakers reading 1 of 10 sentences in German, using 1 of 6 target emotions: *anger*, *boredom*, *disgust*, *anxiety/fear*, *happiness* and *sadness*.

This dataset was chosen as the creators performed tests on the dataset using 20 subjects in order to ensure that the human ratings of the speech files were accurate to the emotions that the speech files were attempting to portray. The dataset initially consisted of 800 files and retained only those with recognition rates greater than 80% and naturalness rates greater than 60%. Additionally, the dataset is highly compatible with the SER software used in this project as it is readily available for use in openSMILE, and it is one of the datasets used to train OpenVokaturi. Like the RAVDESS dataset, Emo-DB also provides labels on each of the files. There are 535 speech files ranging in length from 1 to 8 seconds and the dataset has an overall length of 24:47.

3.2.3 Politicians

A small dataset of 3 videos of Irish Politicians was selected for further analysis past that of the gold-standard datasets. Each video focuses on a different politician: Leo Varadkar, Mary Lou McDonald and Micheál Martin. An attempt was made to source these videos directly from RTÉ in order to obtain the videos at their original qualities; however, attempts to email them were unsuccessful (see Appendix). As a result, the videos had to be obtained from YouTube, which unfortunately means that the videos sourced were initially of lower quality.

To ensure that these videos were suitable for use with the two FER software systems, FACET and AFFDEX, the faces of the subjects had to be in the frame and visible at all times, and only one face could be visible throughout the whole video. If there are other faces in the video, the software may detect faces that don't belong to the subject, causing the results to be incorrect since the software can only detect one face at a time. If there is not a face in each frame of the video, the software will output incorrect results, which will have an impact on the analyses performed later on. Some pre-processing had to be carried out to ensure these requirements were met. Microsoft ClipChamp was the

software used to pre-process the videos.

The videos of Leo Varadkar and Micheál Martin required minimal editing, as they were the sole focus of the videos and they were directly addressing the camera in each instance. On the other hand, the video of Mary Lou McDonald contained several cuts to the audience she was addressing, so the video had to be edited to remove all of these frames. Each video was cropped so as to leave as much of the background out as possible, as it was found through experimentation that this provided improved BRISQUE scores, but had minimal effect on the NIQE scores. Table 3.1 shows the details of each of the videos.




<i>Sample Image</i>	<i>Politician Name</i>	<i>Video Title</i>	<i>Length before editing</i>	<i>Length after editing</i>
	Leo Varadkar	"Ministerial Broadcast by Taoiseach Leo Varadkar about Covid-19 pandemic"	11:27	11:17
	Mary Lou McDonald	"Mary Lou McDonald Sinn Féin Ard Fheis Presidential Address 2023"	25:08	08:00
	Micheál Martin	"Taoiseach Micheál Martin announces further Covid-19 restrictions this Christmas"	8:19	8:19

Table 3.1: Politician video details

The edited politician videos were converted to WAV format for SQA and SER. Initially, these videos were split into 5000ms segments, but it was determined that this did not provide enough data points for analysis. Instead, each video was divided into 2000ms segments for analysis. After segmenting, there were 829 files: 339 for Leo Varadkar, 240 for Mary Lou McDonald and 250 for Micheál Martin. In total the length of this dataset comes to 27:36, or 49,692 frames of video.

3.3 Creating and Adding noise

This section will discuss the process of how noise was created for images and speech signals as well as how this noise was added to the files.

3.3.1 Video

The process of determining how the video quality was degraded had multiple stages. Initially, the videos were going to be passed through deblurring software. However, software that does this for large amounts of videos free of charge could not be found. After this, an attempt was made to pass the videos through a blurring filter. It was found that blurring did not have a significant effect on either of the IQA scores and by the time an effect was seen, it was impossible to identify any facial features in the videos. Therefore, it was decided to add random Gaussian noise to each frame of each video. The decision was made to use this type of noise above others, as it emulates noise that can be caused by video acquisition and transmission, which is a common occurrence.

To add noise to the videos, the Python package NumPy, created by Harris et al. (2020) was used to generate Gaussian noise for each frame, and this noise was then added to the original videos using the OpenCV library for Python, which was developed by Bradski (2000). The noise was added by pulling random values from Gaussian Probability Distribution Functions (PDF) with mean 0 and varying standard deviations (σ), and adding these values to the original image matrix. This was done for each video frame individually, where one frame is 33ms long. Each video had noise added three times, with the level of noise incrementing each time. As a result, there are four variations of the RAVDESS dataset, the original quality as well as the three noisy qualities. The politician dataset only has three variations as at the highest level of noise the stimuli were regarded as invalid by the iMotions software package, which is used for FER.

The first level of noise to be added to the RAVDESS dataset was from a Gaussian PDF with a standard deviation of 2, as it was determined via experimentation that this level of noise was significant, but not significant enough that the facial features could not be made out. After being passed through the FER software, it was found that AFFDEX could not locate the facial landmarks, resulting in every field of each output CSV having a value of 0. In an attempt to amend this, the standard deviation was halved to 1, and once again, AFFDEX could not find the landmarks. In a final attempt, the standard deviation was reduced to 0.5, and AFFDEX was capable of recognising faces at this level. Since the data for the FACET software at all of these noise levels had already been retrieved, they

are still used despite AFFDEX not having as many, as it provides further information on how FACET’s results change with noise. The three levels of noise used for this project are three Gaussian PDFs with the standard deviations of 0.5, 1 and 2. Figure 3.2 visualises these distributions, and Figure 3.3 shows the impact of the various standard deviations on the quality of the videos compared to the original quality.

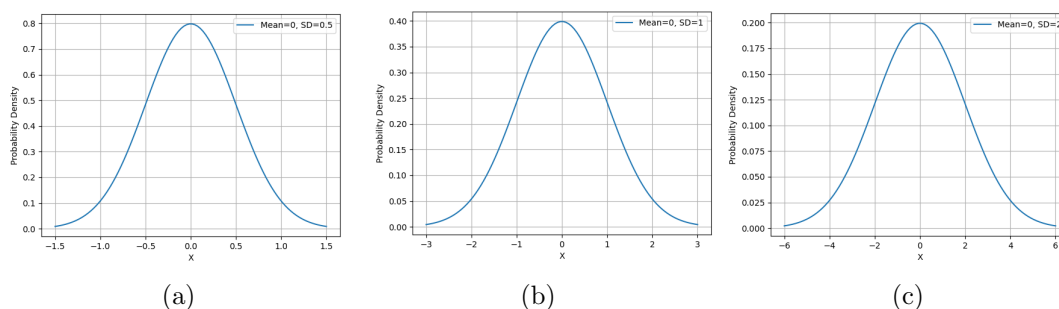


Figure 3.2: Gaussian Probability Distribution Functions (a) $\sigma = 0.5$ (b) $\sigma = 1$ (c) $\sigma = 2$

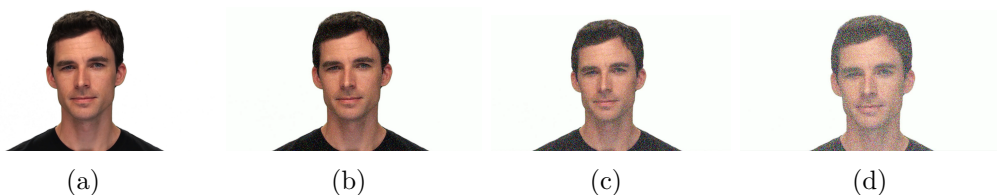


Figure 3.3: Example RAVDESS video at various qualities. (a) Original Quality. (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$

3.3.2 Speech

The process of adding noise to the speech signals was similar to how noise was added to the videos. White noise signals were generated from a variety of Gaussian PDFs with a mean of 0, and then these noisy signals were added to each of the original sound files. The Python library librosa was used for importing the sound files as signals into Python, and NumPy was once again used for generating random values from Gaussian PDFs. The noise was then able to be added to the original signals via an addition. Unlike the videos, the noise was added to the complete speech files all at once rather than being added frame by frame. Smaller standard deviations were used than what was used for the videos, as it was found that speech signals are affected much more than images when using the same standard deviation values. The standard deviations used were 0.01, 0.1 and 1. These distributions are visualised in Figure 3.4.

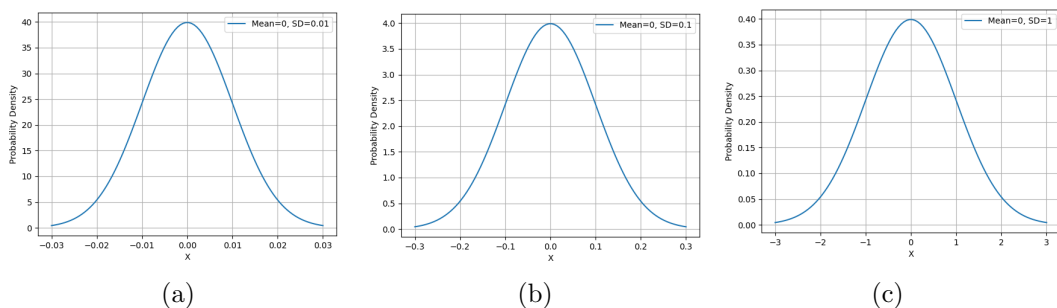


Figure 3.4: Gaussian Probability Distribution Functions (a) $\sigma = 0.01$ (b) $\sigma = 0.1$ (c) $\sigma = 1$

The process of choosing the standard deviations followed from how they were chosen for the video. A standard deviation of 2 was attempted, but the noise was highly obstructive, and no voice could be made out in the signal at all. The standard deviation of 1 also provides a very significant level of noise to the speech signals, so it was decided to use even smaller standard deviations in order to provide levels of noise which affect the speech files but still allow for the speech to be made out. A standard deviation of 0.5 was just as obstructive as the standard deviation of 1, so an even further reduction to 0.1 was made. This level of noise was less obstructive than the previous ones, and so it was retained. After this, 0.01 was decided upon as it provides a good baseline for a signal in which the speech can be made out clearly, but the presence of noise is apparent. The different noise levels are visualised in Figure 3.5 on a speech signal from the Emo-DB dataset. Both Emo-DB and the politician dataset are analysed at four levels, the original quality and the three levels of noise.

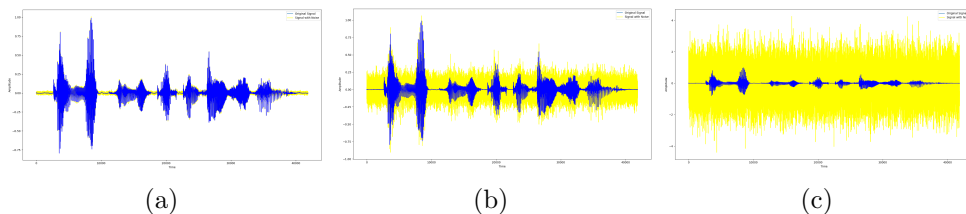


Figure 3.5: Speech signals with original signal (blue) and original signal + noise (yellow) (a) $\sigma = 0.01$. (b) $\sigma = 0.1$. (c) $\sigma = 1$.

3.4 Quality Scoring

This section will outline how each of the quality scoring algorithms was implemented.

3.4.1 BRISQUE and NIQE

The BRISQUE and NIQE scores are calculated for each frame of each video. The BRISQUE scores were calculated first, using the OpenCV library. The frame number, timestamp and BRISQUE score were retrieved for each video frame and exported to a CSV file for each video. The frame number was obtained using a count and the timestamp was obtained through OpenCV. After this, the NIQE score was calculated using the Python library scikit-video and the scores were appended to the same CSV file containing the BRISQUE scores.

The scikit-video library has not been supported since 2017, meaning it uses some outdated data types from the NumPy library. As a result, the code had to be amended locally. In order to allow for the code to execute, all of the files that used these outdated types had to be adjusted to use regular Python data types. As well as this, they use the *imresize()* function from the SciPy library for image resizing, which has been deprecated. This code also had to be changed to use resizing functionality from the Python Imaging Library (PIL), also known as Pillow. The functionality of the code remained the exact same despite these adjustments.

The table containing the frame numbers, timestamps, BRISQUE scores and NIQE scores is visualised in Table 3.2.

FRAME NUMBER	TIMESTAMP	BRISQUE SCORE	NIQE SCORE
1	0	50.2433	10.79332
2	33.36667	52.31662	11.07712
3	66.73333	50.23013	11.00507

Table 3.2: Example of CSV file after calculating BRISQUE and NIQE

The collection of these scores was a very lengthy process. It took approximately 10 hours to calculate both BRISQUE and NIQE on the RAVDESS dataset for a single quality level. In total, it took approximately 80 hours to retrieve all of the scores on both datasets. It is highly important to note that a lower BRISQUE and NIQE score indicates better video quality. For example, a score of 10 indicates better quality than a score of 15. Table 3.3 shows the average BRISQUE and NIQE scores on the RAVDESS dataset and Figure 3.6 shows the distributions of the scores on the dataset at each quality. Table 3.4 shows the average BRISQUE and NIQE scores on the Politician dataset and 3.7 shows the distributions of the scores. The graphs show that the NIQE score is much less reactive to Gaussian noise than the BRISQUE score.

	Original	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$
BRISQUE	40.56	47.56	91.87	100
NIQE	11.31	11.57	12.67	16.85

Table 3.3: Average BRISQUE and NIQE scores on the RAVDESS dataset

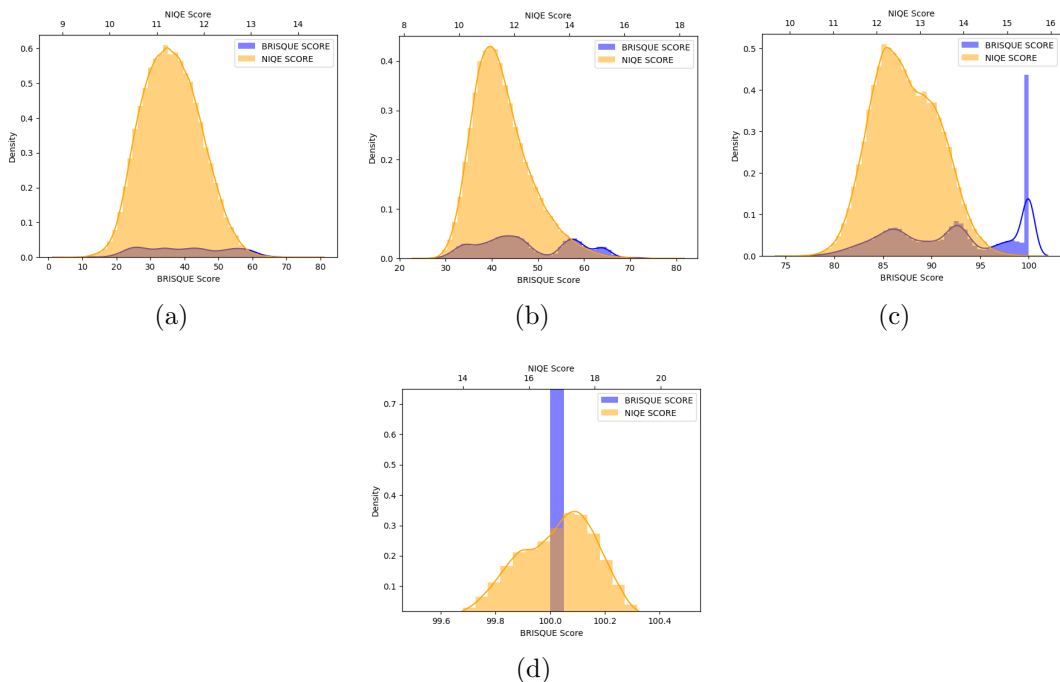


Figure 3.6: BRISQUE (blue) and NIQE (orange) score distributions on RAVDESS (a) Original Quality (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$. The primary x-axis shows the BRISQUE values and the secondary x-axis shows the NIQE values.

3.4.2 NORESQA

Before deciding on the use of NORESQA, several SQA models were researched, and two promising models claiming to be NR were found, specifically SQAPP (No-Reference Speech Quality Assessment Via Pairwise Preference), created by Manocha et al. (2022) and PAM (Prompting Audio-Language Models for Audio Quality Assessment) by Deshmukh et al. (2024). Unfortunately, the code for these models was not available at the time of carrying out work in the early stages of this project. The authors of both papers were contacted via email (see Appendix). However, neither could share the algorithms for use in this project. The code for SQAPP was made in collaboration with a company, so the authors were not at liberty to provide the algorithm. As for PAM, the code has been released, but it was released at a later stage of undertaking work for this project and due to time constraints, it was not used.

	Original	$\sigma = 0.5$	$\sigma = 1$
BRISQUE	62.99	50.43	63.25
NIQE	10.96	15.96	18.36

Table 3.4: Average BRISQUE and NIQE scores on the Politician dataset

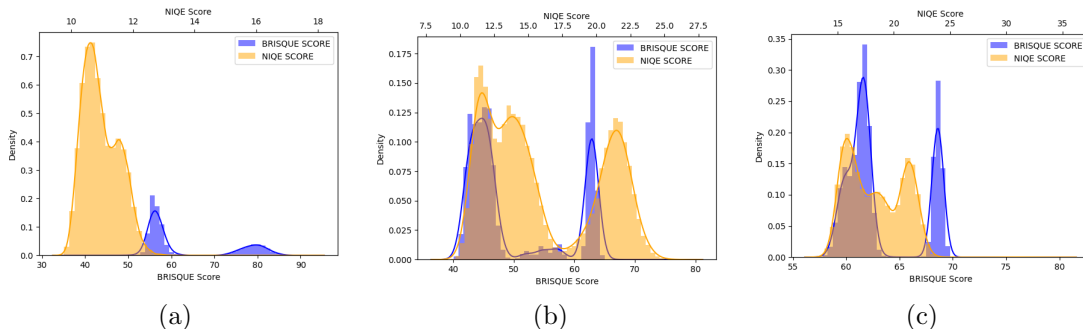


Figure 3.7: BRISQUE (blue) and NIQE (orange) score distributions on Politician dataset (a) Original Quality (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$. The primary x-axis shows the BRISQUE values and the secondary x-axis shows the NIQE values.

As a result, the NORESQA score was decided as the SQA algorithm to use. NORESQA is not an NR algorithm, but rather relies on Non-Matching References (NMR), so a set of NMRs had to be found. The NMRs used for this project come from the Deep Speech Enhancement Challenge (Dubey et al. (2023)), which is the fifth edition of the Deep Noise Suppression (DNS) Challenge, which is run by Microsoft yearly. The goal of this challenge was to create models for joint dereverberation, denoising and suppressing interfering voices. Specifically, the "Track 1 Headset Clean speech" emotional speech dataset was used. This dataset contains 188 clean speech files, which are all 1 second long and consist of a person speaking in one of a subset of emotions, such as crying, yelling and laughter. Each speech file in the datasets was given a NORESQA score relative to each of the 188 files. The 188 NORESQA scores were added to a CSV file for each of the speech files, and the average was taken from each file and placed into a new CSV file, which contained the averages for all of the files per level of noise. The score calculation process was also quite lengthy, taking approximately five hours per dataset for one level of quality. Similarly to the IQA scores, a lower NORESQA score indicates better quality. Table 3.5 shows the average NORESQA scores at each quality on both of the datasets. Figure 3.8 shows the distribution of the NORESQA score on the Emo-DB dataset and Figure 3.7 shows the distribution on the Politician dataset.

	Original	$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 1$
Emo-DB	11.09	11.69	19.33	15.43
Politician	11.53	11.52	16.11	14.52

Table 3.5: Average NORESQA score on the Emo-DB and Politician datasets

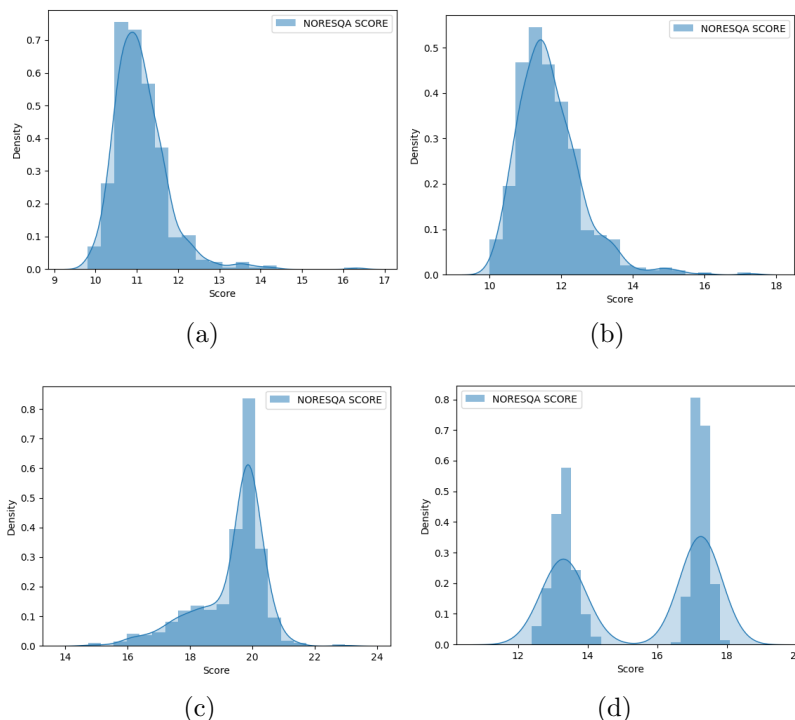


Figure 3.8: NORESQA score distributions on Emo-DB (a) Original Quality (b) $\sigma = 0.01$ (c) $\sigma = 0.1$ (d) $\sigma = 1$

3.5 Facial Emotion Recognition

This section will outline the algorithms and implementation of the two FER algorithms used as part of this project: Emotient’s FACET and Affectiva Inc.’s AFFDEX.

3.5.1 FACET Algorithm

The FACET module is based on the FACET (formerly CERT) algorithm (Littlewort et al. (2011)). This algorithm estimates 19 AUs and provides probability estimates for six emotions: *happiness*, *sadness*, *surprise*, *anger*, *disgust* and *fear*. In addition, the intensity of posed smiles, head orientation, and the Cartesian coordinates of 10 facial feature points are estimated. The pipeline of the algorithm is as follows:

- **Detecting the Face** - The authors trained the detector by using an extension of the "Viola-Jones approach" (Fasel et al. (2005), Viola and Jones (2004)). They also

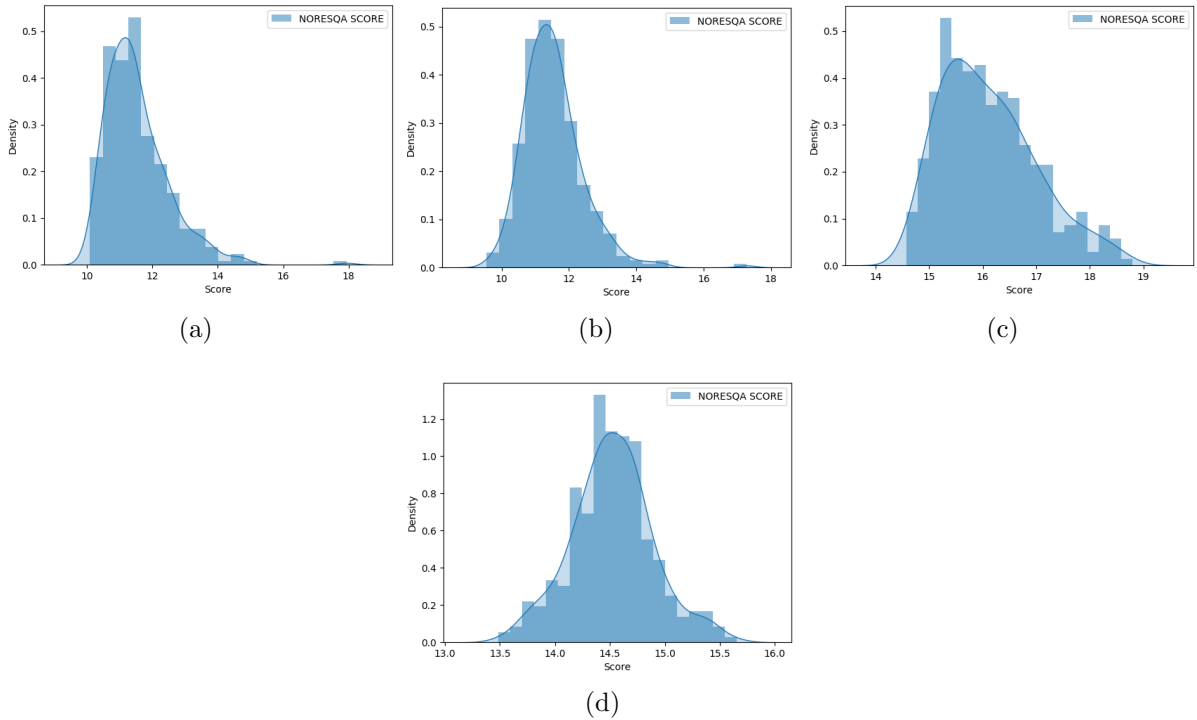


Figure 3.9: NORESQA score distributions on Politicians (a) Original Quality (b) $\sigma = 0.01$ (c) $\sigma = 0.1$ (d) $\sigma = 1$

use GentleBoost (Friedman et al. (2000)) as a boosting algorithm and WaldBoost (Sochman and Matas (2005)) for automatic cascade threshold detection.

- **Detecting the Facial Features** - After the face is initially segmented, a set of 10 facial features is detected using feature-specific detectors. Each facial feature detector is trained using the GentleBoost algorithm and outputs the log-likelihood ratio that a specific feature is present at a specific location to its absence. The posterior probability is then estimated using the likelihood term and a feature-specific prior over locations within the face. Linear regression is used to refine the location estimates, with the regressor being trained on the GENKI dataset.
- **Registering the Face** - The ten facial feature positions are used to re-estimate the face patch using an affine transformation. The pixels are then extracted into a 2D array.
- **Extracting Features** - The face patch is convolved using 72 complex-valued Gabor filters. The magnitudes of the filter are then combined into a feature vector.
- **AU Recognition** - The feature vector is passed into a linear SVM, which estimates the AU intensities.

- **Expression Intensity and Dynamics** - FACET outputs a continuous value for each AU for each video frame. These values consist of the distance of the feature vector to the separating hyperplane of the SVM for each AU. The bounding box is ignored if the confidence of the facial landmark detection is below a certain threshold.

The FACET pipeline is visualised in Figure 3.10.

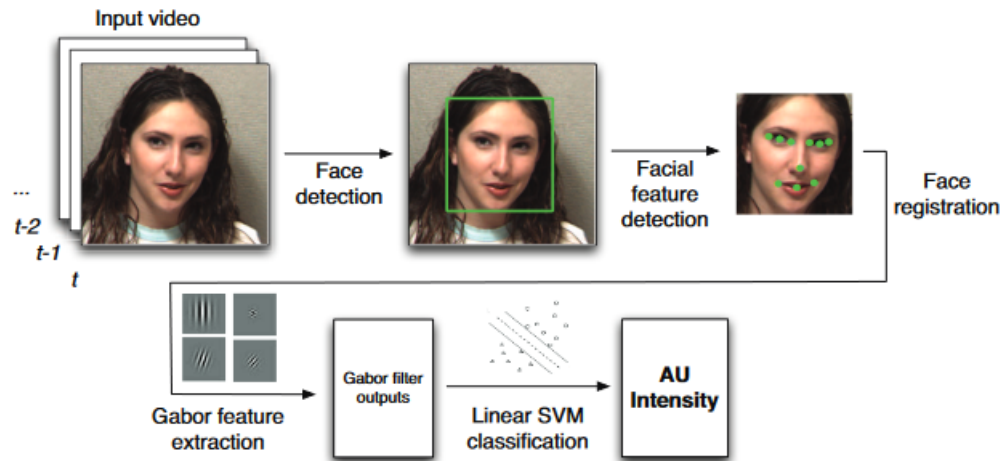


Figure 3.10: FACET pipeline from (Littlewort et al. (2011)).

3.5.2 AFFDEX Algorithm

The AFFDEX module is based on Affectiva Inc's AFFDEX algorithm (McDuff et al. (2016)). The pipeline of the algorithm is as follows:

- **Detecting the Face and Facial Landmarks** - The Viola-Jones algorithm is used. Landmark detection is applied to each facial bounding box, and 34 facial landmarks are identified.
- **Face Texture Feature Extraction** - Histograms of Oriented Gradient features (HOG) are extracted from the area defined by the facial landmark points.
- **Facial Action Classification** - SVM classifiers score each facial action from 0 to 100.
- **Emotion Expression Modelling** - Combinations of facial actions are used to express emotions. The emotions are given a score from 0 to 100.

Figure 3.11 visualises the AFFDEX pipeline.

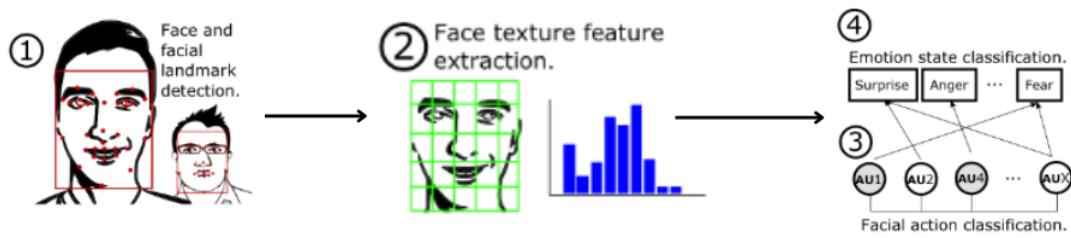


Figure 3.11: AFFDEX pipeline from McDuff et al. (2016).

3.5.3 FER Implementation

iMotions (www.imotions.com) is a software package containing the two FER software systems: FACET and AFFDEX. The first step to using this software is adding a "Face Recording" stimulus. Then, the software takes the videos as input and first checks to see if the stimuli are valid. Once the validity of the videos is checked, post-processing can be performed for each video in FACET and AFFDEX. Post-processing was first done for FACET, followed by AFFDEX on each collection of videos. Once the post-processing was completed, the results for each video were exported into CSV files so that they could be used for further analysis. Two separate CSV files were produced for each video, one for FACET and one for AFFDEX. Examples of FACET and AFFDEX output can be seen in Figure 3.12 and Figure 3.13, respectively.

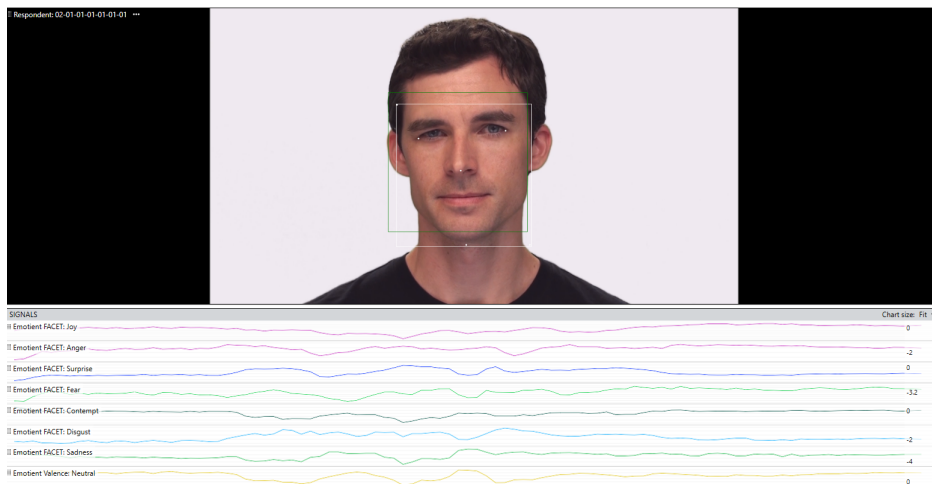


Figure 3.12: An example of FACET emotional output. Graphs from top to bottom: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness, Neutral.

After the CSV files had been retrieved, some manipulation had to be performed. The files contain a variety of data on each video, including the location of various landmark points of the face and the degrees to which each AU is activated. As the interest was

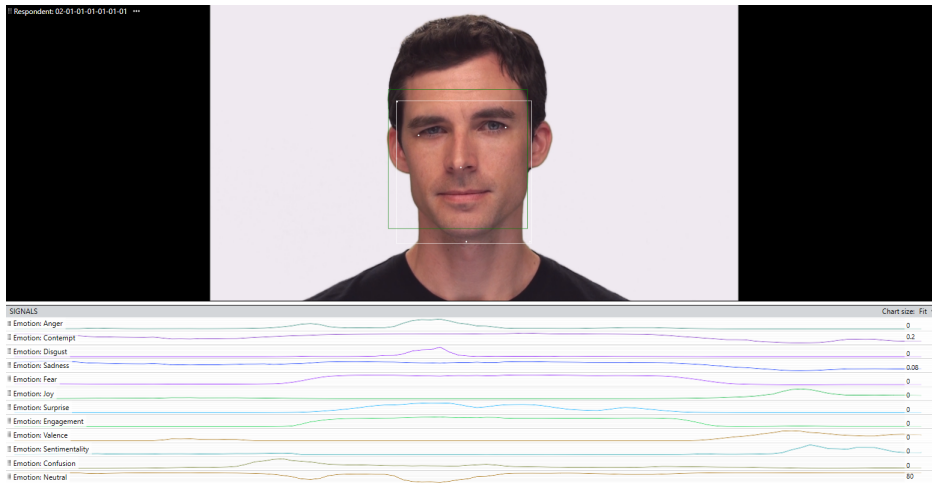


Figure 3.13: An example of AFFDEX emotional output. Graphs from top to bottom: Anger, Contempt, Disgust, Sadness, Fear, Joy, Surprise, Engagement, Valence, Sentimentality, Confusion, Neutral.

only in the emotion values, only these were extracted from the files. Only the values of the emotions that FACET and AFFDEX share were extracted since the objective is to compare the systems. The emotions that the two systems share are: *Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness* and *Neutral*. The output values for each system were manipulated so that they would be in the same interval of 0 to 1. Originally, FACET outputs the log-likelihood values for each emotion, and AFFDEX outputs the values in a range of 0 to 100. To get these values to be in the desired interval, each FACET value had to be passed through the equation $1/(1 + 10^{-score})$, which was obtained from (Warnick et al. (2021)). As an example, if the original log-likelihood value were -0.64, the probability value would be 0.81. For AFFDEX, the values simply had to be divided by 100. These adjusted emotional values were then appended to the CSV files containing the BRISQUE and NIQE scores, meaning all of the key details needed for further analysis were in one place. The manipulation of CSV files was performed using the Python library pandas (Team (2020), Wes McKinney (2010)). An example of a graphical comparison between the two systems on one of the RAVDESS videos can be seen in Figure 3.14. It is notable that FACET did not locate the X and Y coordinates of pupils in any of the videos across both datasets at any quality level.

3.6 Speech Emotion Recognition

This section will outline the implementation of the two SER software used in this project: audEERING’s openSMILE and Vokaturi’s OpenVokaturi.

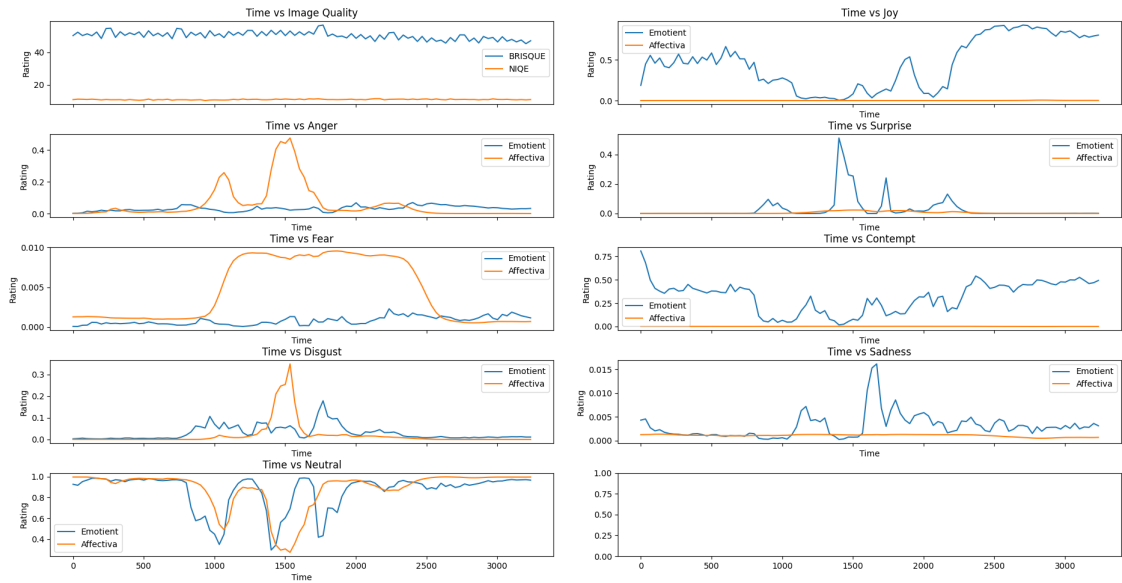


Figure 3.14: Graph comparing BRISQUE (blue) and NIQE (orange) (top-left). The remaining graphs compare emotional values for FACET (blue) and AFFDEX (orange). (In order: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness and Neutral).

3.6.1 openSMILE Architecture

OpenSMILE (Open-source Speech and Music Interpretation by Large-space Extraction) is a toolkit for audio feature extraction and classification of speech and music signals. The basic architecture of openSMILE is as follows:

- **Data Memory** - serves as the central link between the Data Source, Data Processor and Data Sink Components.
- **Data Source** - produces data frames and writes to the Data Memory.
- **Data Processor** - reads the frames, applies the data to an algorithm and writes the modified data to a different location in Data Memory.
- **Data Sink** - reads the frame and interprets the data or passes it to some external source.

This architecture is visualised in Figure 3.15.

3.6.2 openSMILE Implementation

Unlike other emotion recognition software discussed in this report, openSMILE does not come pre-trained, instead, it provides the feature sets that can be used for training. The

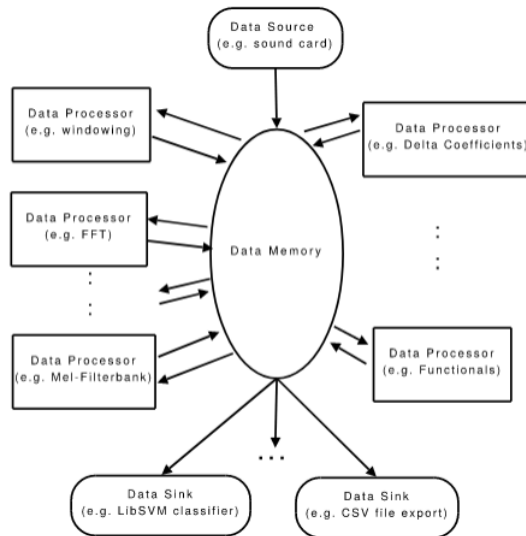


Figure 3.15: Overview of the architecture of openSMILE

Python version of openSMILE was used in this project. openSMILE detects seven emotions: *anger*, *boredom*, *disgust*, *fear*, *sadness*, *happiness* and *neutral*.

Using openSMILE, a Support Vector Machine (SVM) was trained on the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al. (2015)) Version 2.0, which contains 88 features. The reason that this feature set was used despite many being available is that, through experimentation, it was found that this feature set provided the most accurate initial results. As well as this, the high accuracy of the SVM showed that it was a suitable classification model to use. The SVM was implemented using the Python package scikit-learn, by Pedregosa et al. (2011). The original quality Emo-DB dataset was used to train the SVM, which is also one of the datasets used to train OpenVokaturi. The dataset was split randomly into 80% train and 20% test. This means that 80% of the data was used for training the SVM, and the other 20% was used for testing the model. Using this method, it was found that the model was 65% accurate.

K-Fold cross-validation was used with five splits to determine the best hyperparameter, C , for the SVM. K-Fold cross-validation with five folds takes the dataset, shuffles it randomly and splits the dataset into five groups. It cycles through the groups, using one as the test set and the other four as the training sets. It does this five times, using a different group as the test set each time. The values tested were: 0.001, 0.01, 0.1, 0.5, 1, 5, 10, 25, 50, and 100. The scoring method used was negative mean squared error (MSE). With this method of scoring, the C value with the lowest MSE is taken to be the value which has the least error on the validation sets. Table 3.6 shows all of the C values and

C	MSE
0.001	1.89
0.01	1.75
0.1	1.8
1	1.87
5	1.88
10	1.88
25	1.88
50	1.88
100	1.88

Table 3.6: Hyperparameter values for the openSMILE SVM and their MSE values

their scores. 0.01 was chosen for C, as it had the lowest MSE value.

Once the SVM was trained, it was used to predict the probabilities of Emo-DB at each noise level. The probabilities could be found by using the scikit-learn function `svm.predict_proba()`. The same SVM was then used to perform predictions on the segmented politician speech files. The probabilities of each emotion for each file were exported to the same CSV file that contained the NORESQA scores to be used for analysis later on.

3.6.3 OpenVokaturi Algorithm

OpenVokaturi is trained on two datasets, Emo-DB and SAVEE, and detects five emotions: *happiness*, *anger*, *fear*, *sadness* and *neutral*. The system measures nine acoustic features, which are used to compute emotion probabilities using a neural network with three levels of linear connections. The input to the network has nine nodes containing modified strengths of nine features. The nine features are:

- Average Pitch (relative to 100Hz)
- Pitch dynamics
- Average intensity
- Intensity dynamics
- Spectral slope
- Pitch direction
- Pitch Jitter

- Intensity Jitter
- Spectral Jitter

The strengths are modified by subtracting the means of all of the five emotion classes and dividing them by their standard deviations. These strengths are then passed to the first layer of 100 nodes, all with a bias and a weight to each of the nine previous nodes. The next hidden layer consists of 20 nodes, each with a weight to each of the previous 100 nodes. The output layer contains five nodes, one per emotion, each with a bias and a weight to each of the previous 20 nodes. The output is turned into probabilities via a softmax transformation. The probability of each class is proportional to its exponentiated output value.

3.6.4 OpenVokaturi Implementation

OpenVokaturi was implemented using Python. As the system had already been trained, the implementation was much more straightforward than that of openSMILE. Each quality level of Emo-DB was passed into the system, as well as each quality of the segmented politician speech files. OpenVokaturi outputs the probabilities of each of the emotions, and the classification of emotion is determined by which probability is the highest. The probabilities of each emotion are extracted to the same CSV file as the openSMILE probabilities to be used for analysis. OpenVokaturi determined that there was "not enough sonorancy to determine emotions" at the final level of noise for Emo-DB. The same message was given for the last two noise levels of the politician dataset.

3.7 Statistical Analysis

This section will provide an overview of the statistical methods used to evaluate the performance of each of the systems.

3.7.1 System Accuracies

The first step of the analysis was to determine the accuracy of each system on the gold-standard datasets. To do so, the emotion of each video is determined via labels present in the titles of the gold-standard dataset files. For example, suppose the title of a RAVDESS video is "02-01-03-01-01-01". In that case, it can be determined by looking in the third position that the emotion being portrayed is "Fearful", as 03 corresponds to this emotion. In each CSV file, the column with the maximum values is identified to determine which

emotion was classified by each FER system. Videos in RAVDESS that are labelled "calm" are not considered, as neither FACET nor AFFDEX can detect this emotion. Next, the outputs of each system are compared to the labels of each video, and confusion matrices are made to visualise the results. This process is repeated at each degradation of video quality.

The accuracy of the speech systems is determined by analysing each file in Emo-DB, extracting the emotion from the title, and determining if it matches the results received from openSMILE and OpenVokaturi. Suppose the video title is "03a01Fa". In that case, it can be determined from the sixth character in the title that the emotion is "happiness," as F corresponds to the German word "Freude," which translates to "happiness". Confusion matrices are made for these results in the same way they were for the videos. The confusion matrices, as well as the overall accuracies, were implemented using the Python library scikit-learn.

3.7.2 Cohen's Kappa

Cohen's Kappa measures inter-annotator agreement. It is calculated using the following equation: $\kappa = (p_o - p_e)/(1 - p_e)$. p_o is the empirical probability of agreement on the label assigned to any sample, and p_e is the expected agreement when both annotators assign labels randomly (Cohen (1960)).

Cohen's Kappa is calculated using scikit-learn. The result is a value between -1 and 1. A result of 1 means complete agreement between the annotators, whereas 0 or below indicates no agreement between the annotators beyond what would be expected by chance. The score is calculated between each system and the labels of the gold-standard datasets at each of the qualities. This allows for a comparison to be made as to how each system agrees with the gold standard labels and how the comparisons are affected as noise is added. This will give us additional information to what is provided by analysing the accuracies of the systems, as Cohen's Kappa outlines if the agreement is beyond what is expected by chance.

3.7.3 Spearman's Rank-Ordered Correlation Coefficient

Spearman's Rank-Ordered Correlation Coefficient (SROCC) is calculated for each video and speech file. It was decided that this metric would be used over Pearson's Correlation Coefficient, which assumes normality of the data, due to the nature of how BRISQUE and NIQE scores react differently to noise, and so this metric provides a much fairer com-

parison between the two systems. It can also be seen from looking at the distributions of the scores in the previous sections that the data is not always normal.

SROCC, also known as Spearman’s rho, is a non-parametric rank statistic which measures the strength and direction of the relationship between two variables (Hauke and Kossowski (2011)). SROCC is non-parametric, meaning that it does not make any assumptions about the frequency distribution of the data. SROCC uses the ranks of the data rather than the data itself (Gauthier (2001)), making it suitable for comparing the BRISQUE and NIQE scores, as they have different distributions. SROCC is calculated on both of the FER systems for each emotion on each dataset. Each emotion value per frame is correlated with the BRISQUE score for each frame. The same is then done for the NIQE scores. FACET and AFFDEX can both return NaN values for emotions at times, which can cause issues for calculating SROCC. NaN stands for ”Not a Number” and is how programming languages represent a numerical value that is undefined. In order to clean the data to allow the SROCC values to be calculated without error, any rows in the CSV files that contained a NaN value were removed entirely.

The SROCC is then calculated for both of the SER systems. The correlation is calculated between each emotion score in each system and the NORESQA score for each speech file.

An SROCC value of -1 indicates a perfectly negative correlation, +1 indicates a perfectly positive correlation, and 0 indicates no correlation. The stats module of the SciPy Python library (Virtanen et al. (2020)) was used to calculate SROCC, as well as the p-values associated with them.

The null hypothesis and alternative hypothesis for each emotion X are as follows:

Null hypothesis (H_0): There is no correlation between emotion X and image/speech quality.

Alternative hypothesis (H_A): There is a correlation between emotion X and image/speech quality.

Where X is one of *Joy*, *Anger*, *Surprise*, *Fear*, *Contempt*, *Disgust*, *Sadness* or *Neutral* for the FER systems and one of *Happiness*, *Fear*, *Sadness*, *Anger* or *Neutral* for the SER systems.

An alpha value of 0.05 is chosen for this hypothesis test, meaning that there is a 5% significance level. If any of the tests return a p-value above 0.05, the result is considered to be statistically insignificant and so there is a failure to reject the null hypothesis. This means that there is not enough evidence to say that there is a correlation between emotion X and image or speech quality. If a p-value below 0.05 is encountered, the null hypothesis is rejected.

3.7.4 Kruskal-Wallis Test

The Kruskal-Wallis Test determines if there is a significant difference between the distributions of two or more independent groups (Kruskal and Wallis (1952)). It is a non-parametric version of the one-way Analysis of variance (ANOVA) test, which compares the means of two or more independent groups.

The tests are calculated between the emotional values of each system at varying qualities. For example, it is calculated between the Joy values of FACET and AFFDEX at the original quality, then is calculated again at the first level of noise. The results of this test will show whether the distributions of the systems differ more or less when noise is added.

The null hypothesis and alternative hypothesis for each emotion X are as follows:

Null hypothesis (H_0): The distribution of emotion X across the two systems is equal.

Alternative hypothesis (H_A): One of the distributions of emotion X across the two systems is not equal.

An alpha value of 0.05 is used again. If a p-value above 0.05 is encountered, it means that there is not enough evidence to suggest that the distributions of emotion X between the two systems are not equal.

This test is again calculated using the stats module of the SciPy library, and the variable *nan-policy* is set to "omit", meaning NaNs are omitted when performing the calculation. The magnitude of the results will help determine which emotions have the most significant differences across groups.

Chapter 4

Evaluation

This chapter will outline the results found from performing the statistical analyses discussed in the previous chapter: Systems Accuracies, Cohen’s Kappa, Spearman’s Rank-order Correlation Coefficient and the Kruskal-Wallis Tests. First, the four analyses will be discussed on the FER systems, followed by an analysis of the SER systems.

4.1 Video

This section will outline the results of performing the statistical analyses on the FER systems.

4.1.1 System Accuracies

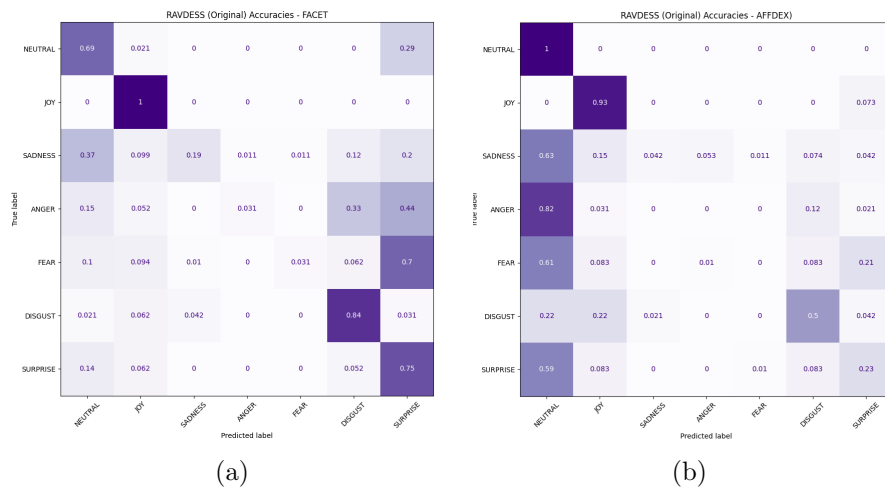


Figure 4.1: Original Quality confusion matrices of (a) FACET and (b) AFFDEX

Figure 4.1 shows the confusion matrices of FACET and AFFDEX accuracies at the original quality. The horizontal axis of each matrix represents the predicted label, and the vertical axis represents the 'true' label. A darker square corresponds to more predictions, and the range of values is 0 to 1, with 0 representing 0% and 1 representing 100%. The two systems begin with differing accuracies for each of the emotions. For example, FACET starts with a lower neutral accuracy than AFFDEX, whereas AFFDEX initially has a lower Joy value. The overall accuracy of FACET is 48.9%, and the overall accuracy of AFFDEX is 33.8%, showing that FACET is initially the more accurate system.

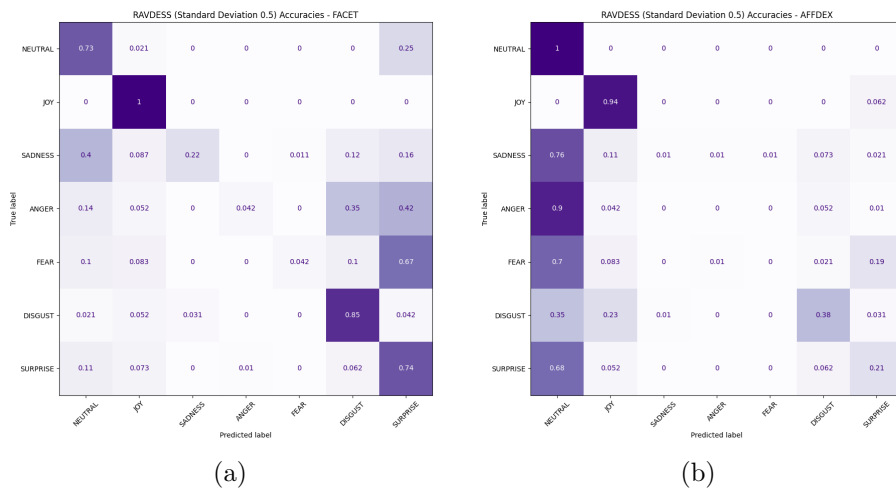


Figure 4.2: Confusion matrices of (a) FACET and (b) AFFDEX for videos with Gaussian noise with $\sigma = 0.5$

Figure 4.2 shows the confusion matrices of FACET and AFFDEX accuracies at the first level of noise, where the standard deviation is 0.5. The matrices show changes in the accuracy of both systems. The accuracies per emotion for each system sometimes are in opposite directions, such as for Surprise, which increases by 2% for FACET and decreases by 2% for AFFDEX. FACET's overall accuracy at the first noise level increases by 1.1% to 50%, and AFFDEX decreases by 2.8% to 31%. Once again FACET is the more accurate FER system.

Figure 4.3 shows the confusion matrices for FACET at the final two levels of noise. AFFDEX is not present as it could not recognise faces at these noise levels. Further changes in accuracy are seen; for example, Joy has perfect accuracy at the original quality as well as the first and second noise levels and then drops to 97% at the final noise level. In other instances, the accuracy increases, such as Neutral. The overall accuracy of FACET for the second level of noise is 49%, which is a slight decline of 1% from the

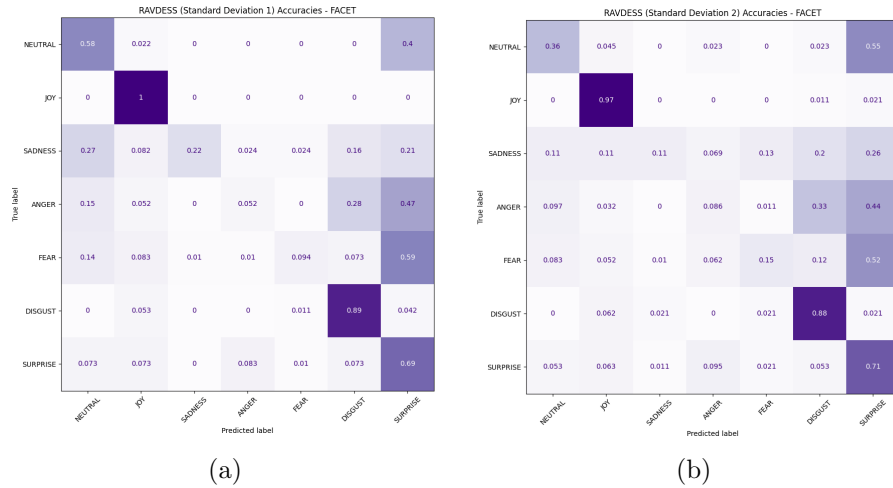


Figure 4.3: FACET confusion matrices for videos with Gaussian noise with (a) $\sigma = 1$ and (b) $\sigma = 2$

first level of noise and ever so slightly higher than the original quality. For the final level of noise, a decrease of 2.2% is seen, which gives an accuracy of 46.8%, which is the first time the system’s accuracy decreases below the original quality level.

It is clear from looking at the confusion matrices and observing the overall accuracies that reducing the quality of videos does have an effect on the overall accuracy of the systems as well as the accuracies of the individual emotions within the systems. It is apparent that FACET is much better at withstanding noise and retaining accuracy when compared to AFFDEX.

4.1.2 Cohen’s Kappa

	<i>FACET</i>	<i>AFFDEX</i>
Original	0.345	0.215
SD = 0.5	0.357	0.196
SD = 1	0.358	N/A
SD = 2	0.32	N/A

Table 4.1: Cohen’s Kappa scores for the RAVDESS dataset at different qualities

Table 4.1 shows Cohen’s Kappa scores on each system compared with RAVDESS at various qualities.

The table shows that the agreement between FACET and RAVDESS increases slightly up until the final level of noise, where it drops below the original agreement. For AFFDEX,

when noise is added, the agreement drops instantly. A difference in how FACET and AFFDEX react to noise can be seen when, in the first level of noise, the score for FACET increased whereas AFFDEX decreased. These results follow closely from the system accuracies.

This shows that adding noise to videos has an impact on the agreement between what emotion each system detects and the emotional labels of the RAVDESS dataset. A value below 0 is never encountered, so it can be concluded that the agreements are more than what would be expected to be seen by chance.

4.1.3 Spearman’s Rank Ordered Correlation Coefficient

It is important to recall when looking at the correlations that a lower IQA score indicates better quality. This means that when a positive correlation is encountered, there is an increase in the emotion detection values as the IQA scores increase, i.e. the emotion is detected more as image quality worsens and vice versa. A negative correlation indicates the opposite: a decrease in the detection of an emotion as image quality worsens and vice versa.

	<i>FACET</i>	<i>AFFDEX</i>		<i>FACET</i>	<i>AFFDEX</i>
JOY	0.166	0.036	JOY	0.133	0.063
ANGER	0.327	-0.087	ANGER	0.284	0.006
SURPRISE	0.162	-0.016	SURPRISE	0.154	0.086
FEAR	0.322	-0.037	FEAR	0.337	0.052
CONTEMPT	0.31	-0.08	CONTEMPT	0.267	-0.043
DISGUST	0.181	0.027	DISGUST	0.149	0.019
SADNESS	0.319	-0.112	SADNESS	0.319	-0.022
NEUTRAL	0.04	-0.003 *	NEUTRAL	0.036	-0.076

Table 4.2: SROCC values of FACET and AFFDEX with BRISQUE score (left) and NIQE score (right) on RAVDESS (* indicates p-value > 0.05)

Table 4.2 shows the SROCC results of the RAVDESS dataset for both systems with BRISQUE and NIQE. There were 315,173 data points for FACET and 157,820 for AFFDEX. The only circumstance in which the null hypothesis is rejected is the SROCC of BRISQUE with AFFDEX Neutral. FACET has stronger correlations than AFFDEX in all cases. A few correlations can be seen in opposite directions for each system; FACET has all positive correlations, whereas exactly half of the correlations for AFFDEX are negative.

Table 4.3 shows the SROCC for the politician videos. There are 148,023 data points

	<i>FACET</i>	<i>AFFDEX</i>		<i>FACET</i>	<i>AFFDEX</i>
JOY	-0.065	-0.403	JOY	0.143	-0.082
ANGER	0.196	0.04	ANGER	0.277	-0.143
SURPRISE	0.438	0.119	SURPRISE	0.259	-0.245
FEAR	0.291	0.09	FEAR	0.328	-0.216
CONTEMPT	-0.112	0.03	CONTEMPT	0.111	-0.032
DISGUST	-0.198	0.316	DISGUST	0.043	-0.051
SADNESS	-0.05	-0.098	SADNESS	0.144	-0.178
NEUTRAL	-0.404	-0.202	NEUTRAL	-0.201	0.183

Table 4.3: SROCC of FACET and AFFDEX with BRISQUE score (left) and NIQE score (right) on Politician dataset (* indicates p-value > 0.05)

for FACET and 99,372 for AFFDEX. The null hypothesis is rejected in every case, so there is enough evidence to support that a correlation exists in all circumstances. The correlations can again be seen in opposite directions, with the correlations for NIQE being opposite between the two systems in all cases. Unlike the SROCC values for RAVDESS, FACET does not always have a stronger correlation, with AFFDEX being stronger in five cases. It can also be seen that not all of the FACET correlations are positive, as is the case for the RAVDESS dataset, with six of the correlations being negative.

From the data, it can be concluded that there is a correlation between the IQA scores and the rate at which each emotion is detected. Most of the time, FACET has a stronger correlation with both IQA scores than AFFDEX. The two datasets show varying results, having different correlation magnitudes and directions in some cases. Overall, the correlations are weak to moderate, with the highest correlation seen from either system being 0.438.

4.1.4 Kruskal-Wallis Test

	<i>ORIGINAL</i>	<i>Std Dev = 0.5</i>
JOY	5374.73	3999.93
ANGER	39149.84	28087.5
SURPRISE	13068.72	9623.18
FEAR	26414.31	21106.3
CONTEMPT	6423.16	3833.68
DISGUST	1837.65	483.76
SADNESS	34510.6	27975.04
NEUTRAL	2992.54	248.25

Table 4.4: Kruskal-Wallis test on RAVDESS (* denotes p-value > 0.05)

	<i>ORIGINAL</i>	<i>Std Dev = 0.5</i>
JOY	7096.6	80498.28
ANGER	46057.9	3298.9
SURPRISE	7821.51	31373.83
FEAR	18492.96	51488.84
CONTEMPT	3198.59	77823.27
DISGUST	2142.37	82940.85
SADNESS	255.42	75141.63
NEUTRAL	16265.27	34568.62

Table 4.5: Kruskal-Wallis test on Politicians (* denotes p-value > 0.05)

Table 4.4 shows the results of the Kruskal-Wallis Tests on the RAVDESS dataset.

Every p-value is below 0.05, so the null hypothesis is rejected in all cases, as there is sufficient evidence to show a difference in the distributions of each emotion between the two systems at both qualities.

A larger value here corresponds to a greater difference between the distributions of the two systems. It can be seen when looking at the RAVDESS dataset that the difference in the distributions is larger for each emotion when looking at the original quality videos than when looking at the noisy videos. Figure 4.4 visualises these differences for Anger, showing how the distributions become more similar when noise is added.

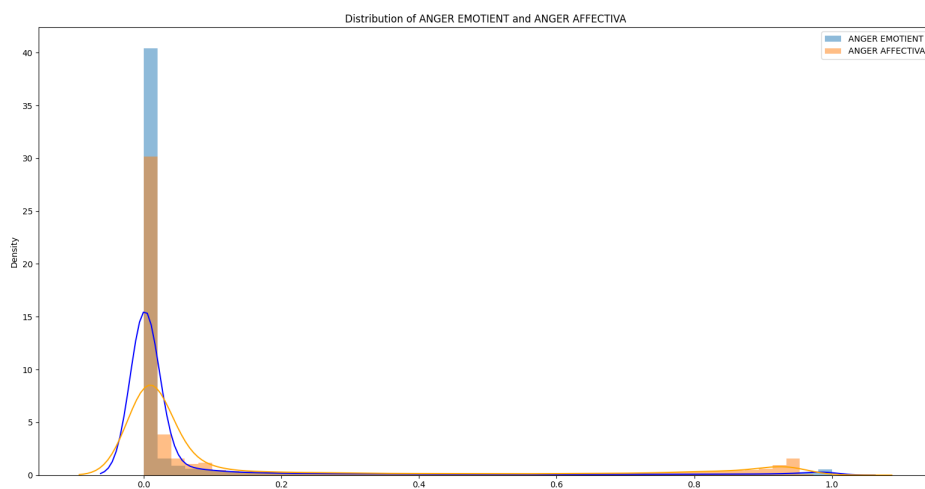
Table 4.5 shows the Kruskal-Wallis test results on the politician dataset. Again, all of the p-values are below 0.05, so the null hypothesis is rejected in all cases.

In all cases but Anger, the difference in the distributions between the two systems is greater when noise is added, which is the opposite of what is seen on the RAVDESS dataset.

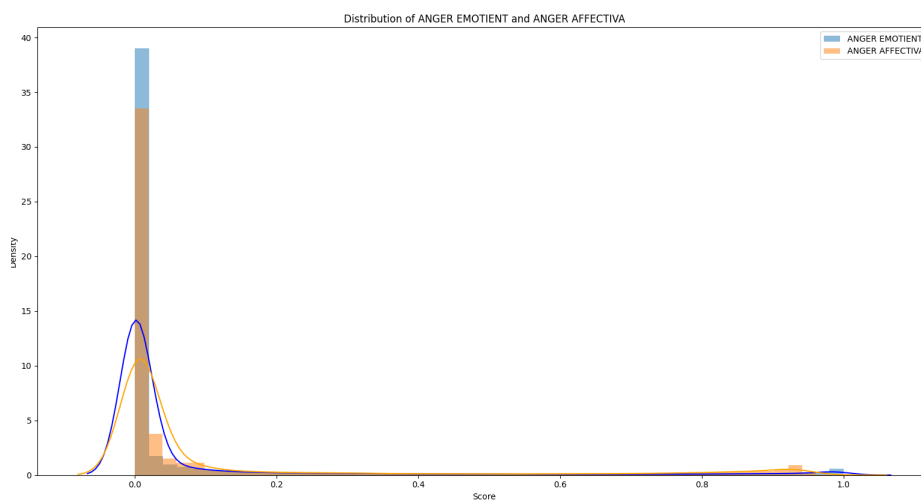
It can be seen from the results that adding noise affects the agreements between FACET and AFFDEX. The effect is different depending on the dataset used.

4.2 Speech

This section will outline the results of performing the statistical analyses on the SER systems.



(a)



(b)

Figure 4.4: Distributions of Anger for FACET (blue) and AFFDEX (orange) at the original quality (a) and at the first noise level (b)

4.2.1 System Accuracies

Figure 4.5 shows the confusion matrices of openSMILE and OpenVokatURI on the Emo-DB dataset at the original quality. Both systems start with high accuracy, with openSMILE’s lowest emotion accuracy being 75% for Happiness and OpenVokatURI’s being 74% for Anger. The overall accuracy of openSMILE at the original quality is 86.5%, and OpenVokatURI is 64.7%

Figure 4.6 shows the confusion matrices of openSMILE and OpenVokatURI at the first level of noise. openSMILE’s accuracy is affected much more than OpenVokatURI’s, with only Boredom and Disgust remaining at high accuracies. The overall accuracy of openS-

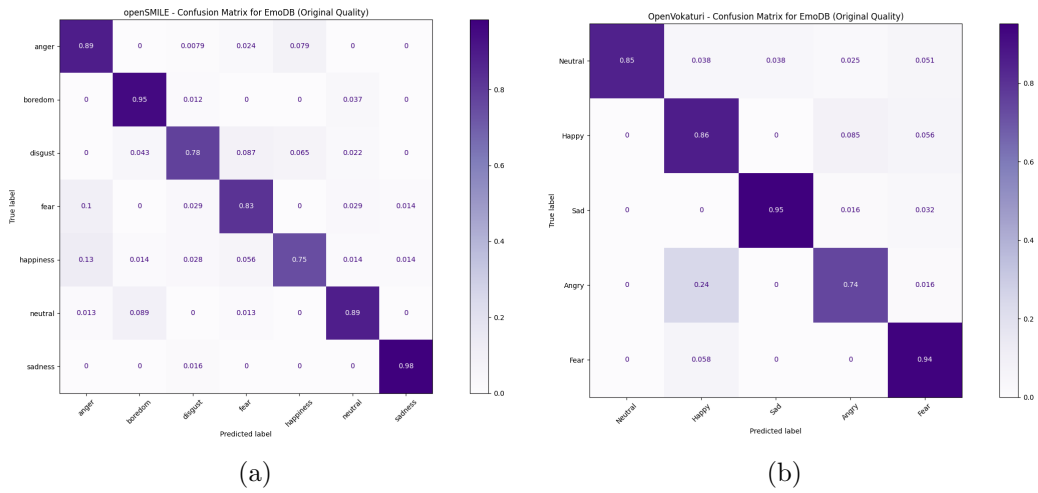


Figure 4.5: Confusion matrices of (a) openSMILE (b) OpenVokaturi at the original quality



Figure 4.6: Confusion matrices of (a) openSMILE and (b) OpenVokaturi with white noise ($\sigma = 0.01$)

MILE decreases by 57.9% to 28.6%, and OpenVokaturi decreases by 10.7% to 54% at the first level of noise.

Figure 4.7 shows the confusion matrices of openSMILE and OpenVokaturi at the second level of noise. openSMILE tends to categorise emotions as either Anger, Disgust or Fear, whereas OpenVokaturi is much more spread out in its predictions. The overall accuracy of openSMILE is 22.1%, a decrease of 6.5%. OpenVokaturi's overall accuracy is 28.8%, a decrease of 25.2%.

Figure 4.8 shows the confusion matrices of openSMILE and OpenVokaturi at the final

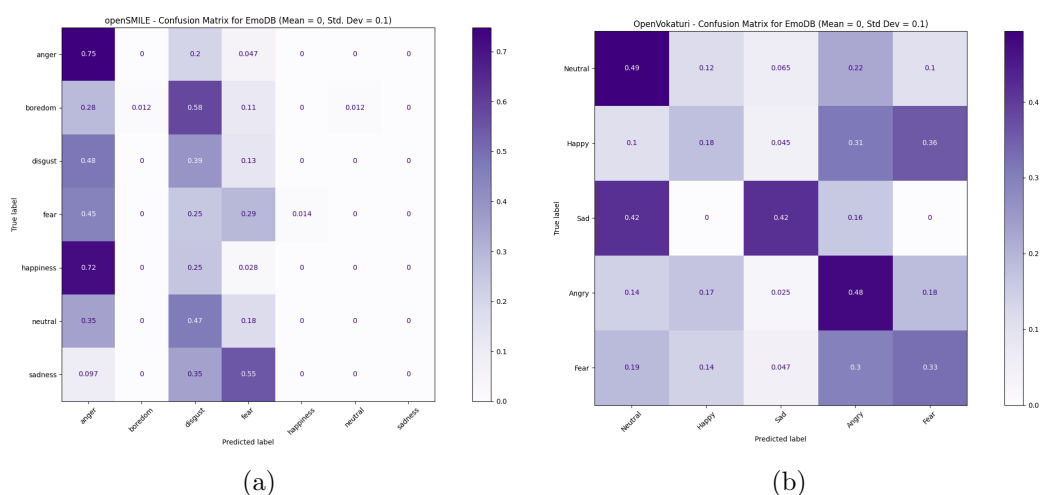


Figure 4.7: Confusion matrices of (a) openSMILE and (b) OpenVokaturi with white noise ($\sigma = 0.1$)

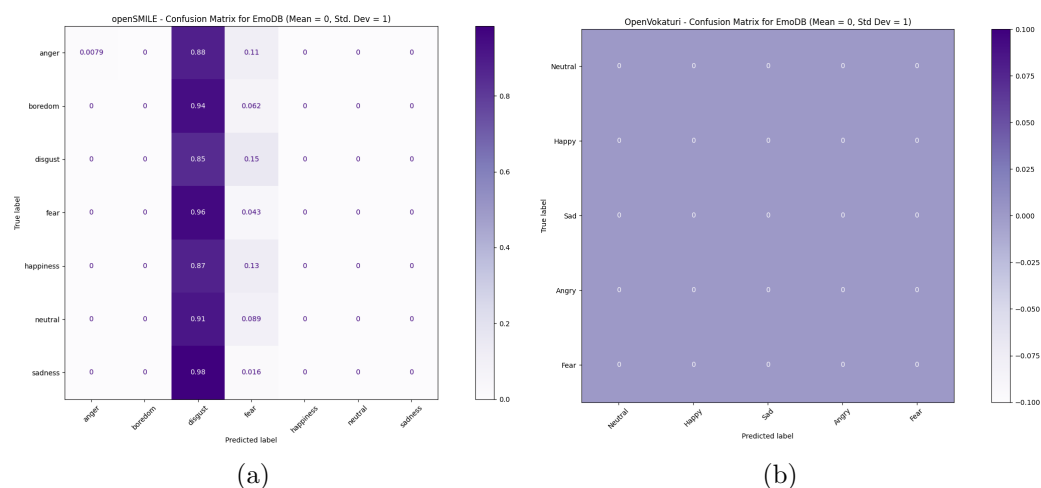


Figure 4.8: Confusion matrices of (a) openSMILE and (b) OpenVokaturi with white noise ($\sigma = 1$)

level of noise. openSMILE detects the vast majority of emotions as Disgust. In contrast, OpenVokaturi does not attempt to classify any emotions as there was "not enough sonorancy to determine emotions". The overall accuracy of openSMILE at the final level of noise is 8.4%, a decrease of 13.7%.

It can be seen that the presence of noise significantly affects the accuracy of both SER systems. Both systems see a consistent decrease in accuracy as the noise level increases. OpenVokaturi seems to be more resilient to noise as its accuracy does not decrease as sharply as openSMILE's when noise is added.

4.2.2 Cohen’s Kappa

	<i>openSMILE</i>	<i>OpenVokaturi</i>
Original	0.77	0.26
Std Dev = 0.01	0.3	0.1
Std Dev = 0.1	0.114	-0.04
Std Dev = 1	-0.01	N/A

Table 4.6: Cohen’s Kappa scores for the Emo-DB dataset at different qualities

Table 4.6 shows the Cohen’s Kappa correlations on each SER system on Emo-DB. Speech files that convey emotions that OpenVokaturi cannot recognise, Boredom and Disgust, are not considered when performing the calculations for OpenVokaturi.

OpenSMILE has a higher agreement with Emo-DB than OpenVokaturi at the original quality. Both systems decrease in accuracy as the quality worsens, reflecting what is seen in the system accuracies. openSMILE has a value below 0 at the final level of noise and OpenVokaturi is below 0 at the second level. It can be determined from looking at the results that as higher levels of noise are added, the agreements of the systems with the labels of Emo-DB are only expected to be by chance.

4.2.3 Spearman’s Rank Ordered Correlation Coefficient

	<i>openSMILE</i>	<i>OpenVokaturi</i>
Happiness	-0.296	-0.08
Fear	0.177	0.019 *
Sadness	-0.032 *	-0.097
Anger	-0.129	0.052
Neutral	-0.528	-0.089
Boredom	-0.475	N/A
Disgust	0.48	N/A

Table 4.7: SROCC of openSMILE and OpenVokaturi with NORESQA Score on Emo-DB (* indicates p-value > 0.05)

Table 4.7 shows the SROCC of the two SER systems on Emo-DB. Boredom and Disgust are labelled as "N/A" for OpenVokaturi as it does not detect these emotions. openSMILE has 2,140 data points, and OpenVokaturi has 1,586. The null hypothesis is rejected in all but two cases: openSMILE Sadness and OpenVokaturi Fear. The values in the table show that the correlation direction of Anger is the opposite for each system. In

	<i>openSMILE</i>	<i>OpenVokaturi</i>
Happiness	-0.363	-0.08 *
Fear	0.553	0.019 *
Sadness	-0.026 *	-0.1
Anger	-0.049 *	0.076 *
Neutral	-0.498	-0.096
Boredom	-0.56	N/A
Disgust	-0.447	N/A

Table 4.8: SROCC of openSMILE and OpenVokaturi with NORESQA Score on Politician dataset (* indicates p-value > 0.05)

all but one case, openSMILE has stronger correlations than OpenVokaturi.

Table 4.8 shows the SROCC of the SER systems on the politician dataset. There were 1,328 data points for openSMILE and 637 for OpenVokaturi. There is a failure to reject the null hypothesis twice for openSMILE, for Sadness and Anger, and three times for OpenVokaturi, for Happiness, Fear and Anger. openSMILE tends to have stronger correlations than OpenVokaturi, except for Anger. Once again, Anger is in opposite directions for each system.

Looking at this data, it can be seen that the correlations between the two systems are different. There are times when the correlations are opposite and they tend to have different magnitudes, with openSMILE tending to be stronger. It cannot be concluded that there is any correlation between NORESQA and openSMILE Sadness or NORESQA and OpenVokaturi Fear, as there is a failure to reject the null hypothesis in both datasets. Speech quality affects each system’s detection of Anger oppositely, with there being a negative correlation for openSMILE and a positive correlation for OpenVokaturi.

4.2.4 Kruskal-Wallis Test

	<i>Original</i>	<i>Std Dev = 0.01</i>	<i>Std Dev = 0.1</i>
Happiness	10.8	29.82	98.23
Fear	146.98	172.17	201.72
Sadness	38.83	53.12	229.28
Anger	35.4	29.22	25.32
Neutral	68.28	32.85	70.68

Table 4.9: Kruskal-Wallis Test on Emo-DB (* denotes p-value > 0.05)

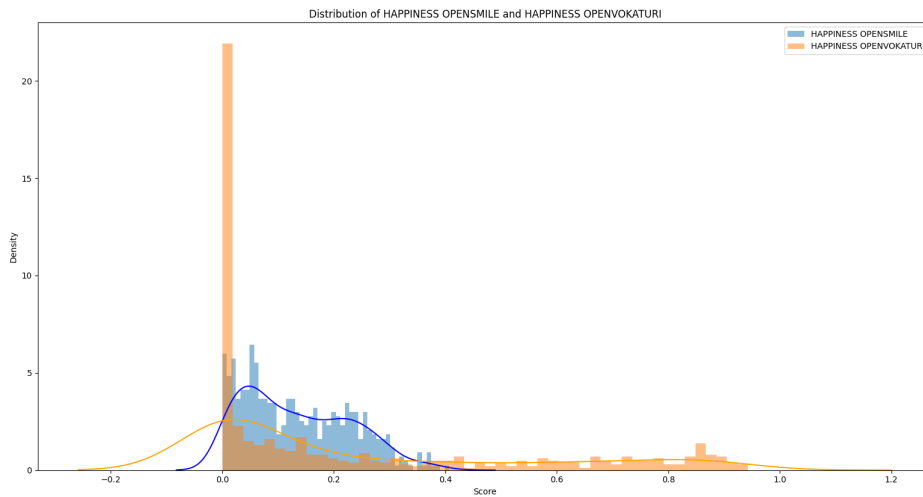
	<i>Original</i>	<i>Std Dev = 0.01</i>
Happiness	487.63	483.99
Fear	432.91	468.91
Sadness	0.007 *	24.71
Anger	453.225	439.67
Neutral	268	317.1

Table 4.10: Kruskal-Wallis Test on Politicians (* denotes p-value > 0.05)

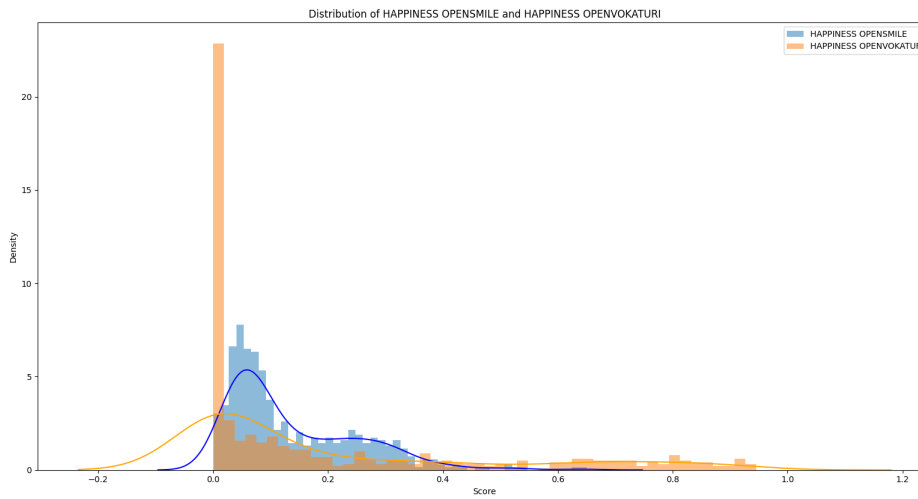
Table 4.9 shows the results of the Kruskal-Wallis Tests on openSMILE and OpenVokaturi. The null hypothesis is rejected in every case, so it can be concluded that there is a difference in the distributions between the two systems at each noise level. For each emotion, there tends to be an increase in the difference between the distributions as noise is added, except for Anger and Neutral. Anger shows a decrease in the difference between the distributions and Neutral decreases at the first level before increasing again. Figure 4.9 shows the distributions of Happiness for both systems to visualise the distributions becoming less similar as noise is added.

Table 4.10 shows the Kruskal-Wallis Tests on the segmented politician speech files. Only two qualities are shown, as OpenVokaturi could not detect emotions past this as there was "not enough sonorancy" in the speech files. There is a failure to reject the null hypothesis for Sadness at the original quality, therefore there is insufficient evidence to support that the two systems have different distributions. Both increases and decreases in the difference between the distributions are seen. Happiness and Anger show decreases in difference, whereas the rest show increases.

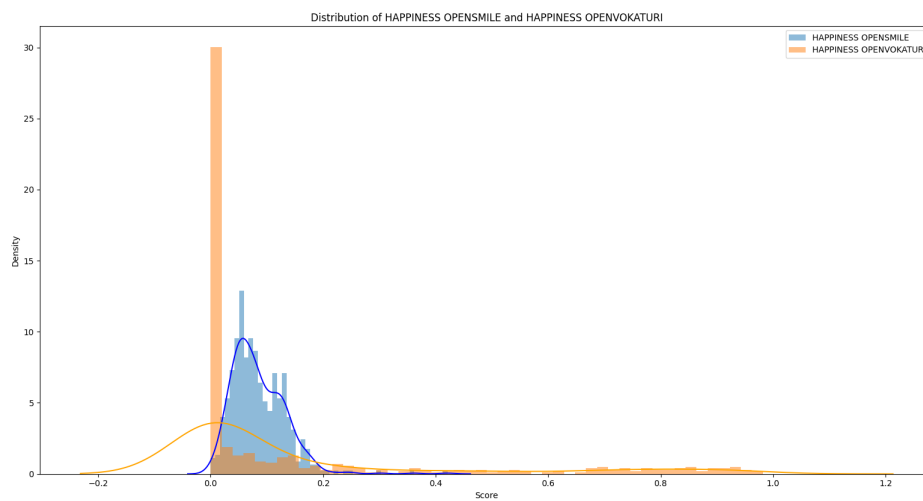
It can be determined that, overall, the addition of noise has an impact on the similarity between the distributions of the systems. There is only one instance in which the null hypothesis fails to be rejected. The results of the test are somewhat different between the two datasets.



(a)



(b)



(c)

Figure 4.9: The distributions of openSMILE (blue) and OpenVokaturi (orange) at the original quality (a), the first noise level (b) and the second noise level (c).

Chapter 5

Conclusions & Future Work

This section will outline the conclusions reached from conducting this project, reflect on the project and profile any future work that can be done to make further developments in this research area.

5.1 Conclusions

This project analysed two FER systems, FACET and AFFDEX, as well as two SER systems, openSMILE and OpenVokaturi, and compared them to evaluate how each system is affected by the presence of noise across two datasets. This was done by analysing System Accuracies, Cohen's Kappa correlations, SROCC and Kruskal-Wallis Tests.

It was discovered that FACET is better at withstanding noise than AFFDEX, as AFFDEX could not recognise faces past the first noise level. At the original quality and the first level of noise, FACET was more accurate than AFFDEX. A slight increase was seen for FACET accuracies up until the final level of noise, whereas AFFDEX experienced a drop in accuracy at the first level of noise. Using Cohen's Kappa, it was seen that for both systems, at each of the noise levels, the agreement between them and the RAVDESS labels was considered to be more than what would be expected by chance. FACET tended to have stronger SROCC results between its emotions and the IQA scores than AFFDEX. In many instances, the correlations between each of the two systems were seen in opposite directions. The correlations of the systems with both IQA scores range between weak and moderate. For the RAVDESS dataset, the two systems had more similar distributions when noise was added, whereas, for the politician dataset, there was a decrease in similarity in all cases but Anger.

It can be concluded that the addition of noise to the videos has varying effects on each of the systems. Overall, FACET can be determined as the system that is more robust when it comes to noise due to its ability to detect faces at much more intense levels of noise than AFFDEX can and its consistently higher accuracy levels.

For the SER systems, it was found that OpenVokaturi retained higher accuracies for longer as more intense levels of noise were added than openSMILE. The Cohen's Kappa correlations showed that at the highest level of noise for each system, it was determined that the agreement between the systems and the gold-standard labels could not be beyond what is to be expected by chance. openSMILE tended to have stronger SROCC results, but it was found that there was not enough evidence to prove that there was a correlation between NORESQA and openSMILE's Sadness, as well as NORESQA and OpenVokaturi's Fear for either of the datasets. The Kruskal-Wallis test demonstrated that the agreement between the systems varied depending on the emotion and the dataset used; for Emo-DB, Neutral fluctuated between higher and lower differences, Anger showed a decrease in difference and the rest showed increases in difference. For the politician dataset, two emotions, Happiness and Anger, showed decreases in difference, and the remaining three showed increases.

It can be concluded that OpenVokaturi is the SER system that is more robust to noise. While it starts out at a lower accuracy, its accuracy remains higher than openSMILE's when noise is added.

5.2 Future Work

Based on the results of this study, there are many areas in which future work can be conducted in order to gain more information on this research area and help to provide further details on the way the systems handle noise.

As discussed previously, many types of noise can be present in both images and speech signals. Work should be undertaken to consider these different kinds of noise to see how they compare and if the influence they have on the systems is different. This will provide further information on what sources of noise can affect emotion recognition systems the most, allowing for further developments of these systems so that the risks of inaccurate emotion detection can be minimised.

Research should be done to see if there is a correlation between image noise and the detection of specific AUs. This can give insight into why there is a difference in correlation and accuracy with each emotion. The two FER systems can be compared to see if each of them reacts differently when it comes to the detection of AUs at various levels of noise, as this can provide further insight into what specific areas of the face are impacted the most by noise. Focusing on specific AUs may also provide explanations for why the different datasets used in this project output such different results.

Work should be undertaken on the AFFDEX algorithm to improve its ability to detect facial landmarks. AFFDEX already showed lower accuracy than FACET at both of the noise levels it was considered for and so if AFFDEX is to compete with FACET, the algorithm should be looked at and improved upon.

A significant limitation of this project was the lack of NR SQA models. With the recent development of the NR SQA algorithm, PAM, and the inevitable further developments that will come from this area, studies similar to this one using these new NR SQA algorithms should be undertaken, as it will provide results that can be compared more similarly to those of the SER systems.

Another limitation of this project was the time constraints that arose from the length of time it took to calculate quality scores. Further studies should be conducted on a wider range of datasets to see if the results are consistent.

Bibliography

- Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Cohn, J. F., Ambadar, Z., and Ekman, P. (2007). Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221.
- Deshmukh, S., Alharthi, D., Elizalde, B., Gamper, H., Ismail, M. A., Singh, R., Raj, B., and Wang, H. (2024). Pam: Prompting audio-language models for audio quality assessment. *arXiv preprint arXiv:2402.00282*.
- Dubey, H., Aazami, A., Gopal, V., Naderi, B., Braun, S., Cutler, R., Gamper, H., Golestaneh, M., and Aichner, R. (2023). Icassp 2023 deep noise suppression challenge. In *ICASSP*.
- Ekman, P. and Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106.
- Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.
- El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

- Fasel, I., Fortenberry, B., and Movellan, J. (2005). A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98(1):182–210.
- Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.
- Gauthier, T. D. (2001). Detecting trends using spearman’s rank correlation coefficient. *Environmental forensics*, 2(4):359–362.
- Gobl, C. and Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1-2):189–212.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.
- Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson’s and spearman’s correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.
- Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2).
- Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.
- Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011). The computer expression recognition toolbox (cert). In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 298–305.
- Livingstone, S. R. and Russo, F. A. (2018). 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):e0196391.
- Manocha, P., Jin, Z., and Finkelstein, A. (2022). Sqapp: No-reference speech quality assessment via pairwise preference. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 891–895. IEEE.

- Manocha, P., Jin, Z., Zhang, R., and Finkelstein, A. (2021a). Cdpm: Contrastive learning for perceptual audio similarity. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200. IEEE.
- Manocha, P., Xu, B., and Kumar, A. (2021b). Noresqa: A framework for speech quality assessment using non-matching references. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22363–22378. Curran Associates, Inc.
- McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., and Kaliouby, R. e. (2016). Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726.
- Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.
- Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.
- Moorthy, A. K. and Bovik, A. C. (2011). Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364.
- Netter, F. H. (2014). *Atlas of human anatomy, Professional Edition E-Book: including NetterReference. com Access with full downloadable image Bank*. Elsevier health sciences.
- Owotogbe, J., Ibiyemi, T., and Adu, B. (2019). A comprehensive review on various types of noise in image processing. *International Journal of Scientific & Engineering Research*, 10(11):388–393.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

- Scherer, K. R. and Ekman, P. (1982). Handbook of methods in nonverbal behavior research. (*No Title*).
- Sheikh, H. R., Sabir, M. F., and Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451.
- Sochman, J. and Matas, J. (2005). Waldboost-learning for time constrained sequential detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 150–156. IEEE.
- Swain, M., Routray, A., and Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21:93–120.
- Team, T. P. D. (2020). pandas-dev/pandas: Pandas.
- Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112–1122.
- Vaseghi, S. V. (2008). *Advanced digital signal processing and noise reduction*. John Wiley & Sons.
- Verma, R. and Ali, J. (2013). A comparative study of various types of image noise and efficient noise removal techniques. *International Journal of advanced research in computer science and software engineering*, 3(10).
- Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57:137–154.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.
- Wallbott, H. G. (1991). The robustness of communication of emotion via facial expression: Emotion recognition from photographs with deteriorated pictorial quality. *European Journal of Social Psychology*, 21(1):89–98.

- Wang, Z. and Bovik, A. C. (2011). Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.
- Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.
- Warnick, B. J., Davis, B. C., Allison, T. H., and Anglin, A. H. (2021). Express yourself: Facial expression of happiness, anger, fear, and sadness in funding pitches. *Journal of Business Venturing*, 36(4):106109.
- Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Appendix A - Email Correspondences

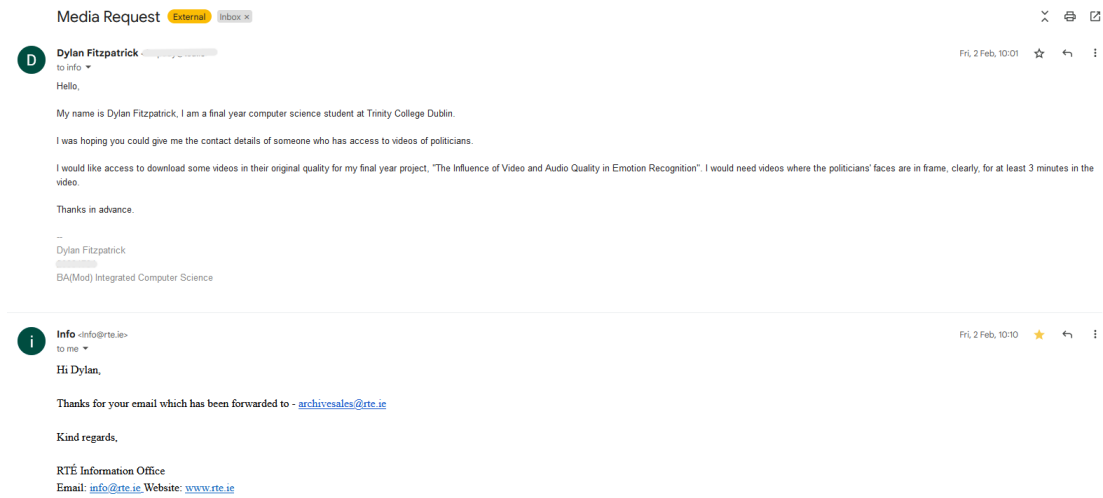


Figure 1: Correspondence with RTÉ requesting politician videos

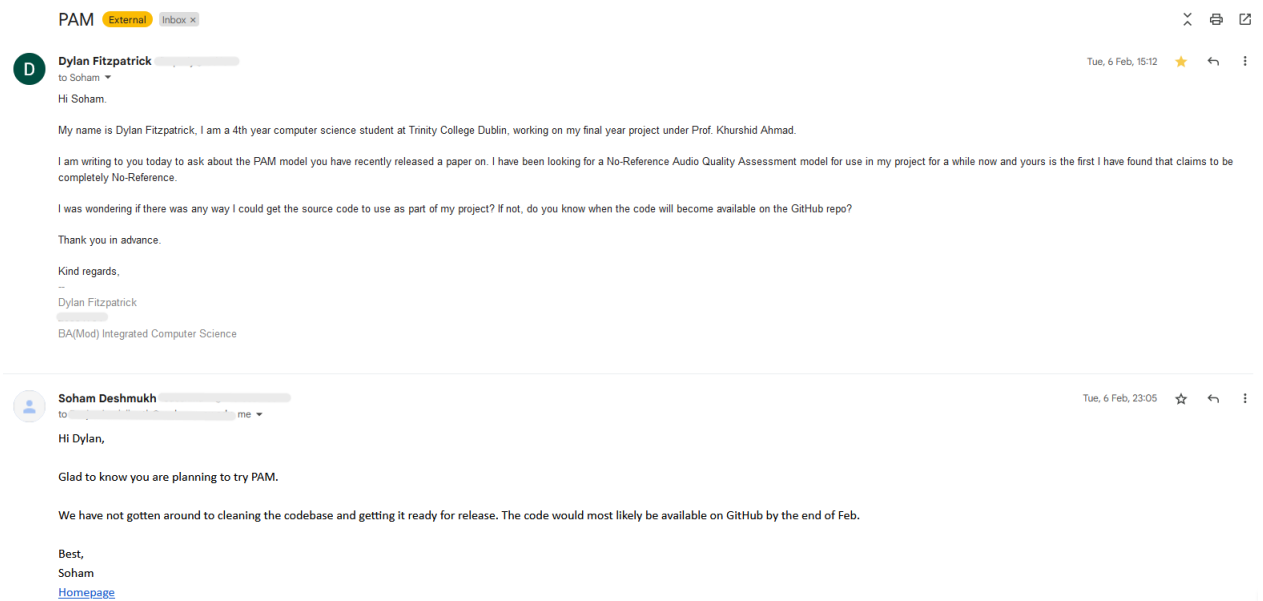


Figure 2: Correspondence with an author of PAM

NORESQA and SQAPP request External Inbox x



Dylan Fitzpatrick

Thu, 1 Feb, 15:14 ☆ ↶ ⋮

Hi Pranay,

My name is Dylan Fitzpatrick, I am a final year student of Computer Science at Trinity College Dublin, working under Prof. Khurshid Ahmad.

I am writing to you today about your NORESQA and SQAPP programs. I am undertaking my final year project titled "The Influence of Video and Audio Quality in Emotion Recognition" and I was hoping I could receive some help.

I was hoping you could supply me with the non-matching reference audio that you used for NORESQA, as I am unable to locate any myself.

I was also hoping you could provide me with the program for SQAPP, as I would like to use both of these systems for analysing audio quality.

Thank you in advance

Kind Regards,

--

Dylan Fitzpatrick

BA(Mod) Integrated Computer Science



Pranay Manocha

to me

Fri, 2 Feb, 03:48 ☆ ↶ ⋮

Hey Dylan,

Thanks so much for your email. The noresqa (and noresqa-mos) source code is open-sourced here:

<https://github.com/facebookresearch/noresqa>

As for SQAPP, it was done in collaboration with Adobe so I would not be able to release the code for that. Please let me know if you have more questions I can answer.

Pranay

Figure 3: Correspondence with an author of NORESQA and SQAPP