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

School of Computer Science and Statistics

# Ethnicity, Gender,and Language: Comparing the Performance of Emotion Recognition Systems in Different Modalities (Non-verbal and Verbal) with Emphasis on Bengali Data

Deepayan Datta under the supervision of Professor Khurshid Ahmad

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MSc (Computer Science - Data Science)

# Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ____________________ Date: ____________________

# Abstract

Emotions are the windows to the nonverbal communication framework employed by the animal kingdom since inception. Humanity is not an exception to this practice. Often, throughout history, people have concentrated on the verbal aspect of the communication but recently, with the advances in Machine Learning and Artificial Intelligence, researchers have tried to delve into the non-verbal realm to try detecting the emotions, their nuances and their role in effectively understanding communication.Teaching a machine that understands only bits, to gauge the emotions of a subject is the beauty of the modern Machine learning algorithms and frameworks. The technology is still at intermediate stages and needs further analysis and research on the same. This dissertation was aimed at evaluating the performance of the different emotion recognition systems across different modalities (facial expressions, audio and text).The study also analyzed how the emotion recognition systems performed in different contexts: posed vs spontaneous, ethnicity and gender. It also analyzed how the Emotion Recognition Systems(ERs) process the physical correlates or landmarks of the modalities (Action Units for face; F0,intensity, loudness, jitter,etc for audio). Similarly how the FERs process the action units were also analyzed. Both the Affectiva and Emotient reported dominant action units(AU) very similar to what has been reported in literature. Statistically significant differences were observed between the distribution of both the facial emotion recognition and speech systems.Differences were also observed in the system outputs when compared across categories like ethnicity and gender. The findings in this dissertation corroborated several studies on posed vs spontaneity, cross-culture variations. For the speech emotion recognition systems it was observed that the inter-annotator agreement rapidly decreased when the systems were trained on databases they have not seen before.For Automatic Speech Recognition (or ASR), it has been observed that Desript outperformed Google Cloud as the preferred service where GCP showed a Word Error Rate of about 80 percent. On the other hand Descript WER rate was around 28 percent. A 3-lexicon system was used to properly extract sentiments from the transcripts and the transcripts were also analyzed across four levels with a bag of words model. The dissertation also aims to introduce a preliminary model of multi-modal emotion detection and fusion as a part of future work

# Acknowledgements

First and foremost, I would like to thank my **supervisor Professor Khurshid Ahmad** for his guidance and motivation throughout the course of the M.Sc degree. The amount of effort and dedication he has devoted towards our dissertations is beyond contemplation. It was an honour working with you, Professor and I really hope I have done justice to the trust and belief you have shown towards me.

I would also like to acknowledge the guidance provided by Professor Carl Vogel, Thank you professor for your valuable suggestions and inputs and helping us navigate the labyrinth of statistics.

To my team-mates Yatheendra Pravan, Wanying Jiang and Subishi Chemmarathil, thank you for sharing your data-sets and also sharing your valuable opinions on any topic whenever I wanted. Thank you so much for making this journey such a memorable one. I would like to express my gratitude towards Shirui Wang and Clodagh Lynch for their data-sets. I would also like to thank Rishi Malpani and Darren Scott for making the times spent at the lab so memorable.

I would like to thank my mother Mrs Kana Dutta and my father Mr Prabir Datta for always being there for me through thick and thin. Thank you for the infinite untold sacrifices you have made to make sure that I have the privilege of pursuing my dream here at Trinity College Dublin. Thank you for showering me with love and blessings always.

I would like to thank my girlfriend, Paramita for continuously motivating me and being there for me. Even with time-differences, you stayed awake and made numerous sacrifices to ensure that I never feel alone.I know I have not been able to interact properly over the past few months and I would sincerely like to apologize for that. Thank you Paro for the unconditional love and support and I would like to take this opportunity to say that I love you very much. Thank you Ruma aunty and Soumitra uncle for always being there for me.

Thank you to all my friends for tolerating me throughout the past year. Coming to a far away land on my own, I could not have made it this far without your support.

The acknowledgement cannot be ended without thanking Muvazima Mansoor. Thank you Mooz for your valuable suggestions not only in the verbal modality but also for your inputs throughout the MSc degree. I always wanted to have a sibling and I can proudly say that Dublin has introduced me to my sister. Thank you for always being there for me and taking care of your Bengali brother.

Finally, a thank you to the School of Computer Science and Statistics at Trinity College Dublin for giving me the platform to pursue my dreams.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Emotions act as the window to the inner human psyche. **Subtle changes are observed in the face and speech well before the same is communicated verbally.** Facial expression changes are observed instantaneously with the speed of light, whereas audio changes follow the speed of sound. Human beings observe all the changes in a holistic manner whereas a machine has to parse the asynchronous signals across multiple modalities separately.

Studies claim that what we perceive as emotions in the non-verbal context can actually be summarized as the movement of facial muscles, or the vibrations in the vocal cord or the larynx. It has widely been believed that though there might be subtle differences in how humans evoke or mask their emotions; the physical correlates (facial muscles, fundamental frequency of voice, etc) remain widely same across ethnicity,age and gender. Similarly there are also studies that opine that emotions are artificial constructs and depend on the physiological, cultural, ethnic and geopolitical context and to categorize emotions as invariant might be restrictive.

Widely known as the father of the facial expression analysis, Paul Ekman believed in this notion of invariance in six primary emotions: anger, happiness, sadness, disgust and fear across cultures. Later, a seventh emotion "contempt" was also brought into the fold. Following Darwinian principles, Ekman and his colleagues created the Facial Action Coding System to map the myriad of facial expressions into meaningful set of movements in the facial muscles.

Manual FACS coding was previously used to annotate the muscle movements and understanding their role in emotion detection. With the advent in Machine Learning techniques and computer vision, it has been possible to automatically extract facial features and understand their role in emotion analysis.

Similar advancements have also led to accurate mapping of the vibrations in the speech, which helps us decipher the emotions. And for the verbal modality, advance sentiment analysis techniques have also come to the fore. The applications in Emotional intelligence can be manifold: from effective understanding of patients' needs in healthcare and evaluating criminal behaviour in psychology to understanding the nuances and intricacies of emotions in fields of business and stock markets.

## 1.1 Problem Statement

Nowadays, a lot of tools and packages are widely available to evaluate "emotional intelligence" in systems. The dissertation aims to evaluate various emotion recognition systems across different modalities (face, audio and text )and understand their role and behaviour in classifying emotions. Facial Emotion Recognition softwares Emotient, Affectiva and Speech Emotion Recognition packages OpenSmile and Vokaturi have been strictly evaluated across different databases with a battery of statistical tests. Automatic Speech Recognition (or Transcription) has also been used to understand "what was said" or in other words the context of the situation. The verbal modality has been introduced in this dissertation to effective analyze and support the non-verbal modality. The paper also aims to evaluate the role of ethnicity and gender in how the emotion recognition systems perceive and categorize emotions.

Now, humans observe a **holistic view** of all the modalities while communicating. We are able to see the face, hear the tone and also discern the words, and the fused information from all the modalities are taken into consideration before determining whether a person is happy or angry or sad. **But for machine classification, information from each of the asynchronous modalities come in different speeds: facial expression changes at the speed of light, speech changes are observed at the speed of sound. Hence the dissertation also aims to have a preliminary look at the possibility of multi-modal emotion detection in machines.**

## 1.2 Research Contribution

I am part of the larger team currently working under the tutelage of Professor Khurshid Ahmad to create a multi-modal emotion detection and fusion model for semi-spontaneous speech using politicians' data. 180 videos were obtained from the work of previous students [6] and 29 videos were collected(25 were used in the dissertation) for Bengali Politicians. The videos have been processed across all modalities (facial, audio and text).Multiple Facial Emotion recognition Systems(FERs) and Speech Emotion Recognition Systems(SERs) have been analyzed. Multiple transcription services have been checked and analyzed. Previously, fundamental frequency extraction was done only using OpenSmile. **New scripting methods for extracting acoustical cues from Vokaturi and Praat have been developed**. Thorough analysis has been done using multiple lexicons and bag of words model. Groundwork for a multimodal fusion model is also being done.

A paper "Speech Emotion Recognition Systems: A cross-language, inter-racial, and cross-gender comparison" has also been submitted under the supervision of Professor Khurshid Ahmad in the "Future of Information and Communication Conference (FICC) 2023". .

# 2 Motivation and Literature Review

## 2.1 Facial Expressions :

When it comes to Facial Expressions and detection of emotions from facial expressions, Paul Ekman's works are widely regarded as some of the most important pieces ever written on this subject. Ekman, Soren and Friesen in their now famous paper "Pan-cultural effects of emotion" suggests the existence of association between the movements of our facial muscles and the "discrete primary emotions".[7]. There are two widely present views in the field of cross-cultural emotion recognition. Some opine that the movement of specific set of facial muscles are associated with the same set of emotions across people. On the other hand, there are others who subscribe to the theory that facial expressions are culture specific and learned within different cultures[8]. Ekman, a proponent of the universality or invariance of emotions , conducted four experiments on subjects from 5 literate countries (Japan,Brazil,Chile, Argentina and the USA) and 2 pre-literate cultures( both from New Guinea) and opined that "there are universal facial expressions of emotion"[8]. Or in other words, the **"universalists"** refer to the existence of a "prototype face", where the emotions can be determined just by looking at the movement of the underlying facial muscles and the same will hold true when applied to different demographics.

Now drawing inspiration from Darwin and based on this notion of "universality", Ekman and Friesen developed the Facial Action Coding System (or FACS). Now, FACS is a "behavioral coding system"[9] which intends to summarise all "observable facial actions" as some combinations of different muscular movements. The specific movements or "actions" of these muscles are classified into different **"action units"(or AUs)**. Though the action units are small in number, 7000 combinations of these action units have been observed[10]. The advent of FACS led to the boom in research involving emotions. With FACS at their disposal, researchers started looking at emotions "objectively"(Matsumoto et al).Nowadays, in most studies around the world involving emotions and facial expressions, these action units and their corresponding movements are correlated with emotions and serve as windows for looking into human emotions and psyche.

## 2.2 FACS Action Units and the association with emotions

There has been several studies correlating the movements of facial action units to emotions. Study by Du Tao Martinez et al [11] observes the presence of AU12 (lip corner puller) , AU25(parting of lips) and AU6 (cheek raiser )for happiness/joy; whereas anger is marked with the presence of AU4 (brow furrow), AU7 (lid tightener), AU10 (upper lip raiser) and AU17 (chin raiser ).

Another study by Lucey at al [12] concludes that AU12 must be present in all instances of happiness where anger must be associated with AU23 and AU24. The iMotions official website[13] categorizes happiness as a function of the action units AU6 and AU12, whereas anger is associated with the action units AU4, AU5(upper lid raiser), AU7 and AU23 (lip tightener).

Similarly, Velusamy et al[14] has noted the presence of AU12,AU6, AU26, AU10 and AU23(arranged in descending order of frequency of occurrence) for happiness whereas anger was associated with AU23, AU7, AU17, AU4 and AU2.

These differences in AUs reported by different studies might be due to the cultural,ethnic and gender biases present in the underlying databases used, as reported in the paper by [1].

## 2.3 Automatic Facial Emotion Recognition

Previously, human based manual FACS coding was used to detect facial expressions and deciphered human emotions. But this approach was manual and time consuming. More than 100 hours of training were necessary to achieve nominal competency in FACS and about an hour was needed to properly evaluate each minute of a video tape. Donato et al[15] opined the need of automating FACS for improving the speed, precision and quality of the measurements. [15] Over the last decade, rapid developments in the automation, computer vision and computing power have resulted in the advent of new tools that utilizes the full power of FACS to classify emotions. Emotion Recognition systems like FACET EMOTIENT, AFFDEX AFFECTIVA, MS Azure - all use FACS to classify emotions.

Since the 1990s, there has been a breakthrough in the number of Facial Emotion Recognition Systems which recognize the "non-verbal communication of emotions"[6] through the lens of facial muscles and their movements.

**CERT**: Developed at the University of California (San Diego), Computer Expression Recognition Toolbox(or CERT)[16] was trained on a number of databases: **Posed Expression Databases** :Ekman-Hager [15] , Cohn-Kanade [12], Man-Machine Interaction

[17] (contains both posed and spontaneous expressions,but only the posed ones were used for training ) ; **Spontaneous Expression Databases:** M3 [18] , and two datasets curated by the US government ( non-public)

CERT was evaluted on the extended Cohn-Kanade Dataset with appreciable results. The performance was also evaluated on spontaneous expressions dataset "with less success". CERT paved the way for FACET which is made available by the iMotions package for research and commercial work.

**FACET by Emotient:** Originally developed in the Machine Perception Lab by the University of California San Diego researchers[16] , Emotient has recently been acquired by Apple to further their AI driven product line. FACET leverages advanced computer vision algorithms as well as effective identification of FACS action units (AUs): to allow customers to perform intensive analysis of facial expressions.Apart from the seven Ekman defined emotions, the current versions of the SDKs have also introduced additional features for detecting frustration and confusion . As a part of the SDK, reaction of each emotional channel will be displayed in the output when subjected to the corresponding facial expression . The one with the most valence/intensity will be considered the dominant emotion at that point of time. Not only the dominant expressions and emotions, subtle or involuntary micro expressions are also detected real-time by the FACET SDK.

**Affectiva AFDEX** Currently acquired by Smart Eye, Affectiva AFFDEX is one of the pioneers in the field of Emotional AI. Available in multiple languages, the product has made excellent strides in the area of emotional and automotive AI : like detection of nuanced human emotions, understanding the emotional state of occupants in the car, speech science, etc. Based on the movement of interest points in a subject's face, AFFDEX performs emotional analysis thereby scoring emotions ( here eight emotions are considered ) in a range between 0 and 100 with a valence score ranging between -100 and 100. The product is based on the Viola-Jones algorithm for facial recognition[19]. Here, valence means the intensity of the emotions, I.e.how positive or negative the emotions are , with 0 being the neutral value.

**Cloud Based Emotion Recognition Systems** Several cloud based emotion recognition systems like the ones provided my "Microsoft Azure", "Google Cloud" and "Amazon Rekognition" are also frequently used nowadays for classifying emotions.

## 2.4 Existing Literature on Performance Comparison of FERs:

In a study by Dupre et al[20][6], 937 videos were sampled from two databases conveying six emotions (anger, happiness, sadness, dear, disgust and surprise) either in BU-4DFE( posed) or UT-DALLAS(spontaneous) data-sets. The authors compared the performance of 8 "commercially available automatic classifiers"[20] with those achieved by human judges. "Considerable variance" was observed in the recognition accuracy for the eight classifiers varying from 48 percent to 62 percent. Classification accuracy was high for the two "best performing classifiers" for posed expressions thereby exhibiting high agreement with their human counterparts. The accuracy was **found to significantly go down for spontaneous expressions** (though it was still higher than the chance level)

It was observed that humans fared better than the eight automatic classifiers with a **higher true positive rate** for **both posed and spontaneous expressions.** Among the classifiers, EMOTIENT performed the best followed by VicarVision, Neurodata Lab, Visage Technologies, Microsoft Azure, Affectiva and MorphCast with CrowdEmotion taking the last position among all of them.

Humans were found to do remarkably well in recognizing joy, surprise and sadness for the posed expressions (True Positive Rate was found to be more than 90 percent) whereas similar accuracy was recorded only for joy when the spontaneous expressions database was considered. For disgust the result was opposite where humans performed better in the spontaneous dataset than the posed one. Remarkably, the True Positive Rate for disgust was also found to be higher for both EMOTIENT and AFFECTIVA in the spontaneous database. The true positive rate for detecting anger plummeted ( less than 10 percent ) in the spontaneous expression database not only for the humans but also for the automatic classifiers (EMOTIENT, AFFECTIVA, etc. )

Looking at the association matrices, it can be observed that there was a **high agreement** between the human judges and their machine counterparts on the **posed database** which decreased when spontaneous expression database was in consideration.

Now, Stöckli et al[21] compared FACET and EMOTIENT by carrying out out two studies: The first on "standardized emotional facial expressions" and the second on "natural facial expressions". The first study focused on standard facial expression databases like the Amsterdam Dynamic Facial Expression Set (or ADFES), the Radboud Faces Database (RaFD) and the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP ) . The RAFD and WSEFEP are homogenous databases and comprise of Dutch and Polish models respectively whereas ADFES is a heterogenous database of images of Mediterranean and North European origin models. [6]

It was observed that the **a)** recognition accuracy was different for distinct emotions and EMOTIENT performed better than AFFECTIVA **b)**The performances of both the classifiers degraded when applied on the natural facial expressions database - study 2 ( performance was better for the standardized pictures database (study 1) ).

Yang et al[22] evaluated the performances of five emotion recognition softwares: Baidu Research, Amazon Rekognition, Face ++, MS Azure and Affectiva on "standardized" databases as well as "manipulated" images . A series of manipulations were done on the standardized images (simulating low-quality real-life scenarios) to create the "manipulated" images database. These manipulations include : rotation of the images,blurring the image, increasing/decreasing the brightness of the images or increasing the noise content of the image. Affectiva had the lowest accuracy out of all the five classifiers and was unable to recognize faces in 2.6 percent of the images. It was also noted that **Affectiva was highly prone to most of the manipulations.**

Now, Krumhuber et al[23] evaluated the performance of FACET EMOTIENT and human observers on posed and spontaneous databases. Like previous studies, better recognition accuracy was observed in posed expression than in spontaneous ones. Krumhuber opines that **"facial patterns were more prototypical" in posed expressions** which led to the machines performing better in posed than in spontaneous. It was also observed that FACET performed better than humans on the posed data sets; which Krumhuber again attributes to the "prototypicality" and calls the performance **"sufficiently robust".**

Apart from comparing the performance of the FERs, Bernin et al[24] also looked at the instances where the face was missed altogether by the FERs. CERT proved to be the best at face detection but surprisingly **EMOTIENT**, touted as the successor of CERT **had the worst performance along with INSIGHT**. The performance of AFFECTIVA was **also poor but it was better** than FACET EMOTIENT. On the other hand, Stockli's study saw Emotient detect faces in all the videos with Affectiva failing to recognize a face in 1 percent of the videos.

## 2.5 Possibility of cross-cultural impact

Now, there have been many studies which advocate the importance of culture in how human beings express or evoke their emotions. Lisa Barrett opines that emotions are culture-dependent and she posits that the nuances of culture must be taken into account for understanding emotions. She opines that the way we perceive emotions depends on how we are feeling at the moment and also on how we are taught to understand emotions. She further argued that the emotion observed in one individual can easily be mistaken for completely different emotions by different people. Sadness in one might be perceived as weariness or even frustration by different people.

Similarly, Lutz and White,[25] state that emotions and expressions are "socially derived and defined": and are contingent on the context in which they are expressed. It has also been pointed out that "smile is not the universal expression of happiness" and has been used to

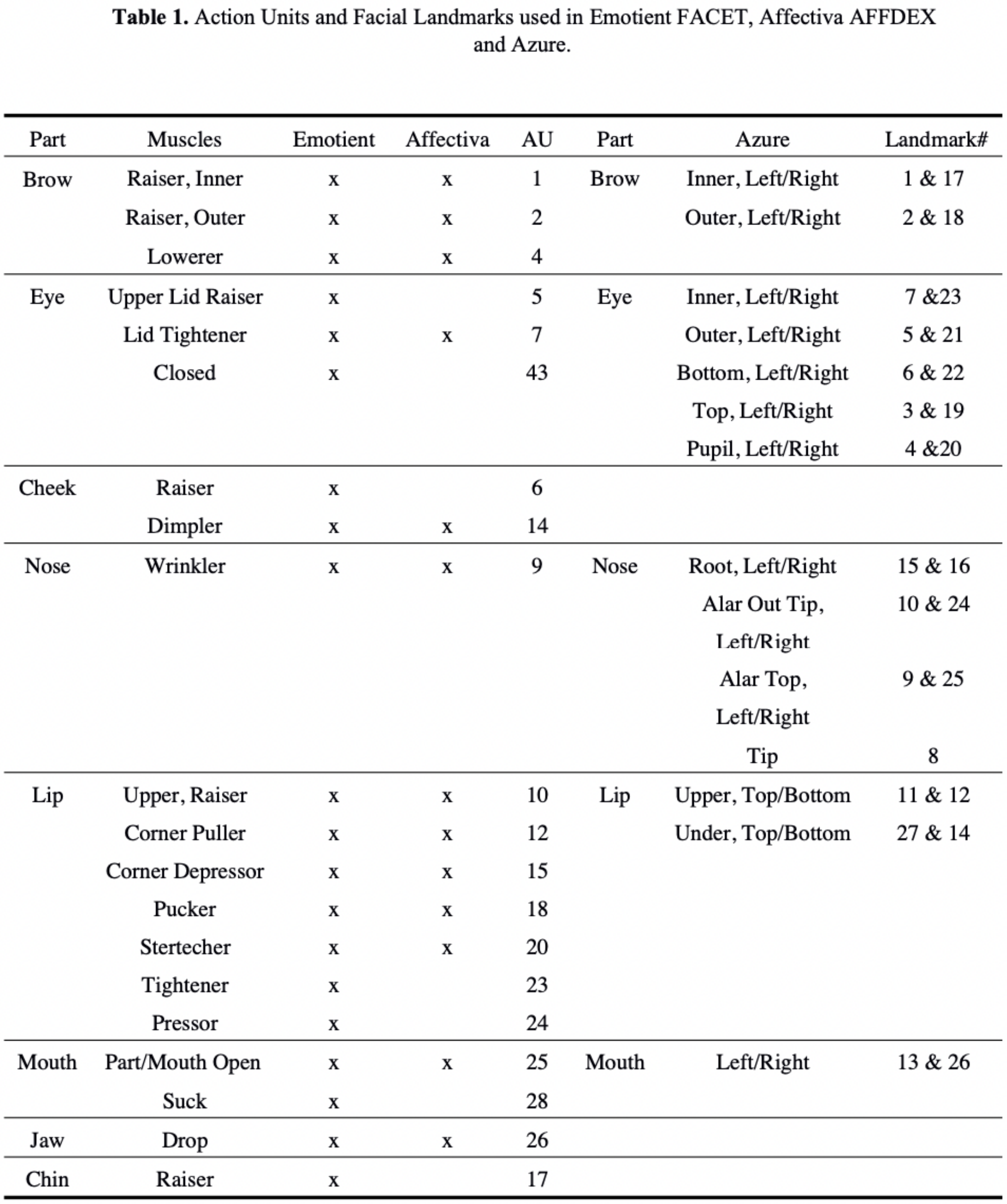


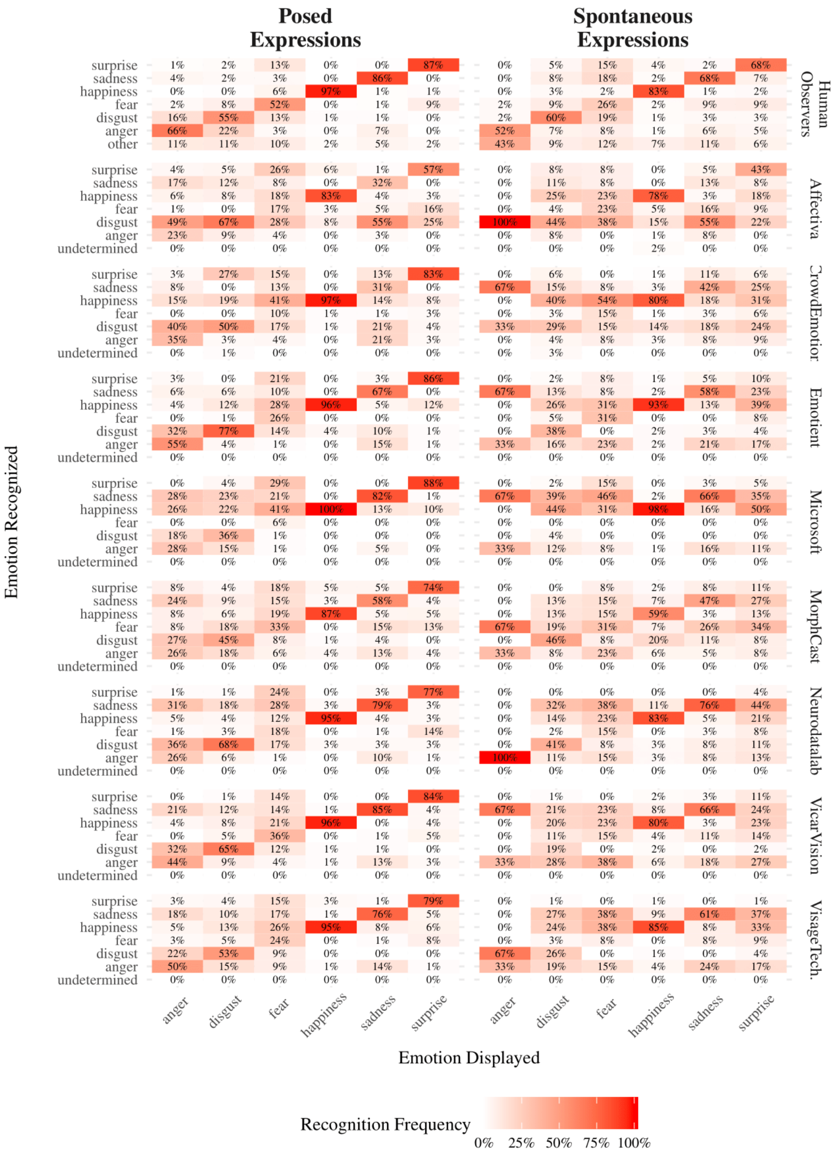
**Table 1.** Action Units and Facial Landmarks used in Emotient FACET, Affectiva AFFDEX and Azure.

| Part | Muscles | Emotient | Affectiva | AU | Part | Azure | Landmark# |
|---|---|---|---|---|---|---|---|
| Brow | Raiser, Inner | x | x | 1 | Brow | Inner, Left/Right | 1 & 17 |
|  | Raiser, Outer | x | x | 2 |  | Outer, Left/Right | 2 & 18 |
|  | Lowerer | x | x | 4 |  |  |  |
| Eye | Upper Lid Raiser | x |  | 5 | Eye | Inner, Left/Right | 7 &23 |
|  | Lid Tightener | x | x | 7 |  | Outer, Left/Right | 5 & 21 |
|  | Closed | x |  | 43 |  | Bottom, Left/Right | 6 & 22 |
|  |  |  |  |  |  | Top, Left/Right | 3 & 19 |
|  |  |  |  |  |  | Pupil, Left/Right | 4 &20 |
| Cheek | Raiser | x |  | 6 |  |  |  |
|  | Dimpler | x | x | 14 |  |  |  |
| Nose | Wrinkler | x | x | 9 | Nose | Root, Left/Right | 15 & 16 |
|  |  |  |  |  |  | Alar Out Tip, Left/Right | 10 & 24 |
|  |  |  |  |  |  | Alar Top, Left/Right | 9 & 25 |
|  |  |  |  |  |  | Tip | 8 |
| Lip | Upper, Raiser | x | x | 10 | Lip | Upper, Top/Bottom | 11 & 12 |
|  | Corner Puller | x | x | 12 |  | Under, Top/Bottom | 27 & 14 |
|  | Corner Depressor | x | x | 15 |  |  |  |
|  | Pucker | x | x | 18 |  |  |  |
|  | Stertecher | x | x | 20 |  |  |  |
|  | Tightener | x |  | 23 |  |  |  |
|  | Pressor | x |  | 24 |  |  |  |
| Mouth | Part/Mouth Open | x | x | 25 | Mouth | Left/Right | 13 & 26 |
|  | Suck | x |  | 28 |  |  |  |
| Jaw | Drop | x | x | 26 |  |  |  |
| Chin | Raiser | x |  | 17 |  |  |  |

Figure 2.1: Distribution of AU units in FERs)

express different emotions in different regions of the world. In the United States of America, instances have been found where “smile” has been used to express “happiness,sadness or disgust” , whereas Klineberg[26] and LeBarre[27] have observed smile being used to express emotions of “sadness” and “uncertainty” in Japan and parts of Africa respectively

Saha et al[1] argues that cross-cultural differences exist not only in the facial expressions of emotions, but **even the Action Units(AUs) are also susceptible to "inter-culture variations"**. Significant up-tick has been observed in the performances of the rule-based classifiers (decision trees) when the rule base was changed from **"culture-independent"** to



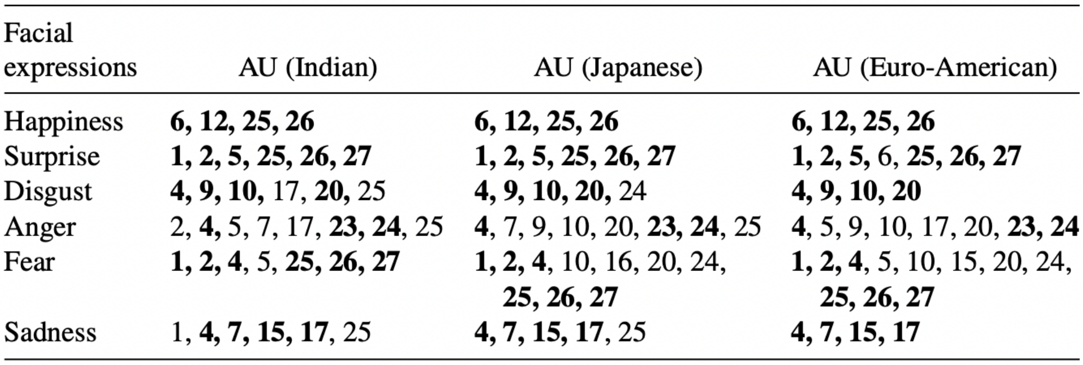
S1 Figure. Emotion confusion matrices for human observers and automatic classifiers separately by posed and spontaneous expressions. For cases with 'undetermined' confidence levels, the sum of the marginal proportion of recognized emotions can be higher than 100%.

1

Figure 2.2: Confusion matrix comparing performance of human observers and several automatic FERs as reported by [1])

### "culture-specific"

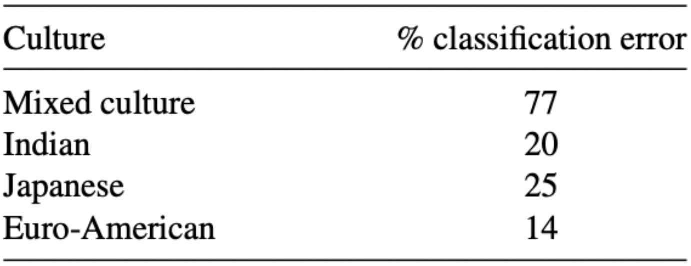
It was observed from the confusion matrices that happiness was the easily recognized emotion out of all the others. On the other hand, a **"recognition deficit"** was observed for the negative emotions: where anger was mistaken for disgust, fear for surprise, etc. The authors have also evaluated the performance of both the culture-specific decision trees



| Facial expressions | AU (Indian) | AU (Japanese) | AU (Euro-American) |
|---|---|---|---|
| Happiness | **6, 12, 25, 26** | **6, 12, 25, 26** | **6, 12, 25, 26** |
| Surprise | **1, 2, 5, 25, 26, 27** | **1, 2, 5, 25, 26, 27** | **1, 2, 5,** 6, **25, 26, 27** |
| Disgust | **4, 9, 10,** 17, **20,** 25 | **4, 9, 10, 20,** 24 | **4, 9, 10, 20** |
| Anger | 2, **4,** 5, 7, 17, **23, 24,** 25 | **4,** 7, 9, 10, 20, **23, 24,** 25 | **4,** 5, 9, 10, 17, 20, **23, 24** |
| Fear | **1, 2, 4,** 5, **25, 26, 27** | **1, 2, 4,** 10, 16, 20, 24, **25, 26, 27** | **1, 2, 4,** 5, 10, 15, 20, 24, **25, 26, 27** |
| Sadness | 1, **4, 7, 15, 17,** 25 | **4, 7, 15, 17,** 25 | **4, 7, 15, 17** |

Figure 2.3: Culture Specific Action Units (Indian, Japanese and Euro-American) [1])

(classifiers) and the mix-culture decision tree (which was run on the entire dataset ). It has been observed that the **classification error for the "mix-culture" decision tree was significantly higher than its culture-specific counterparts**. The results have been shown in figure 2.4 .



| Culture | % classification error |
|---|---|
| Mixed culture | 77 |
| Indian | 20 |
| Japanese | 25 |
| Euro-American | 14 |

Figure 2.4: Classification Errors (Culture-specific vs Mixed Culture/Culture Independent ))

## 2.6 Modality: Audio

Like their facial expression counterparts, the Speech Emotion Recognition Systems (or SERs) are also based on the "universality" theory of emotions, or in other words the "invariance" of perception and expression of emotions across ages, ethnicity, languages and genders. Researches on the "universality" branch, follow either the discrete emotion theory with their atomic constituents including the Ekman emotions like joy(happiness), anger, sadness, fear, disgust,sadness and contept - or the dimensional theory comprising of valence, arousal and dominance observed in the voice.

But there are also counter arguments to the universality/invariance theory that focuses on the "cross-linguitic and cross-culture implications" on the hypothesized universality argument. Similarly, "age" is also observed to have an impact[28][29][30] in the production of voice as the voice box undergoes significant changes with aging. Even thought the underlying physiology and mechanism of producing voice is same across demographics, there have been significant differences observed in the important physical correlates responsible for classifying and perceiving emotions. Fundamental frequency (also known as F0), considered the key metric for classifying emotions, has been observed to vary across languages and cultures. Similar varying trends have also been observed for other audio correlates like voice-onset time, intensity, jitter, shimmer etc. Moreover, differences have been observed in the physical correlates that are used in speech emotion recognition, like the fundamental frequency F0,voice onset time.

Fundamental frequency (or F0) is the sinusoidal component of the complex audio waveform with the lowest frequency(hence the name "fundamental frequency)[31]. It helps us determine the periodicity of the wave form and is perceived as "pitch" by the listener. The higher order components of the wave form are called "harmonics". These harmonics are nothing but the multiples of the fundamental frequency F0 and are also called formants (F1,F2...etc and so on).

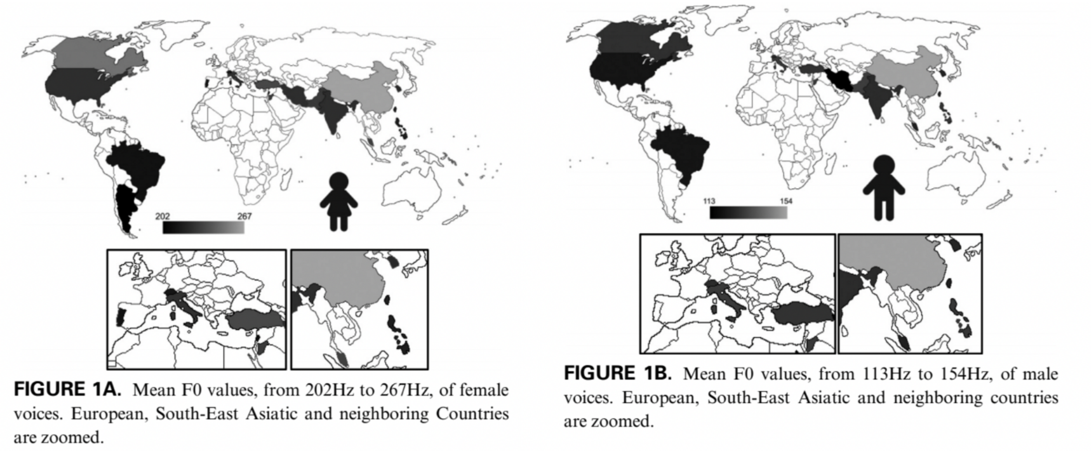


**FIGURE 1A.** Mean F0 values, from 202Hz to 267Hz, of female voices. European, South-East Asiatic and neighboring Countries are zoomed.

**FIGURE 1B.** Mean F0 values, from 113Hz to 154Hz, of male voices. European, South-East Asiatic and neighboring countries are zoomed.

Figure 2.5: F0 distribution around the world as reported by Saggio et al

Pichora and Fueller[32] analyzed 2800 stimuli ( 2 speakers x 7 emotions x 200 sentences ) to identify the acoustical features that "account for the most variance" in classifying the stimuli into the seven underlying emotional states. It was observed that mean F0 and F0 range were the most effective features for classifying the emotions, compared to other acoustical cues.Similarly, Pereira and Watson[33] state that the F0 was the "most revealing" acoustical cue where the F0 shape contour showed differences between anger and happiness. Moreover differences in F0 mean and range were also observed across the emotions.

## 2.7 Speech Emotion Recognition Softwares:

openSMILE (acronym for Speech and Music Interpretation by Large-space Extraction ) [34][35] is an open source tool-kit used in "automatic emotion recognition for affective computing". It is widely used for extracting features from the audio signal and classifying the same into a range of emotions. First introduced in a 2010 paper by the researchers(Florian Eyben, Martin Wöllmer and Björn Schuller) at the Technical University of Munich, openSMILE lets us look into both the domains of speech and music. Not only speech features like MFCC and LPC, but also features (988 acoustical features with "emobase.conf" ) extracted from music like CHROMA ( for "estimating fundamental frequencies") are taken into account.Since, audio signals are essentially continuous time-series data, it becomes difficult to recognize emotions and music genres unless we convert the same to aggregated static features like delta coefficients and moving averages.

**Vokaturi** OpenVokaturi[2] is an open source version of Vokaturi with a classification accuracy of 66.5 percent. It presents the user with the flexibility to switch languages by providing libraries in both the C and Python languages.OpenVokaturi uses a set of nine acoustical cues to classify the emotions into five emotion states. The nine cues are as follows:

- Average Pitch (expressed in semitones relative to 100 Hz (analogous to F0 mean)
- Pitch Direction
- Pitch Dynamics
- Pitch Jitter
- Average Intensity
- Intensity Dynamics
- Intensity Jitter
- Spectral Slope Spectral Jitter

The Vokaturi output is a vector of five probabilities corresponding to the five emotions, adding up to 1. Vokaturi uses the Auto-correlation method defined by Boersma(also used in Praat)[36] to effectively extract fundamental frequency or F0 from the audio signals.

**PRAAT:**

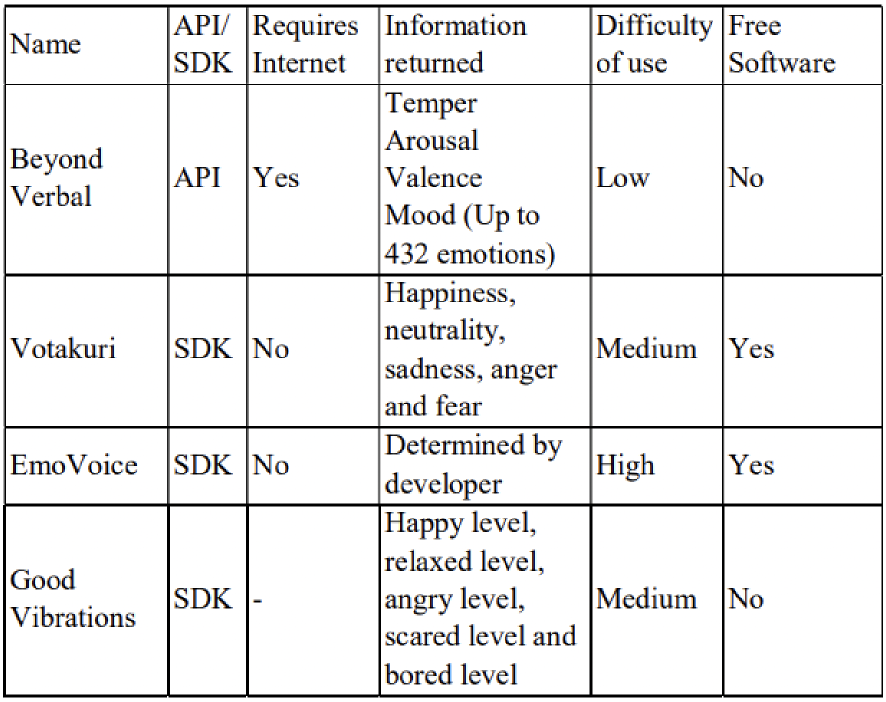
Developed by University of Amsterdam's Paul Boersma and David Weenink at the Institute of Phonetic Sciences, PRAAT is a computer program or a scripting tool that can be used for "analysing, synthesizing and manipulating speech".[36][37]. Praat has versions for the common operating systems like Mac OS, Windows, Linux, etc.

There are lots of methods for estimating F0 (used for classifying emotion). But in linguistic and phonetic studies, Praat has often been referred to as the "the de facto standard speech analysis program".[38][39]

## 2.8 Performance Comparison of SERs

Unlike its Facial Expression Recognition counterpart, there are very few studies done on the performance comparison of SERs.

A paper by Garcia et al [2] explores the different emotion recognition frameworks in audio, visual and text and compares the systems for each modality. For the audio modality, Vokaturi, Good Vibrations, EmoVoice and Beyond Verbal are compared.



| Name | API/ SDK | Requires Internet | Information returned | Difficulty of use | Free Software |
|---|---|---|---|---|---|
| Beyond Verbal | API | Yes | Temper<br>Arousal<br>Valence<br>Mood (Up to 432 emotions) | Low | No |
| Votakuri | SDK | No | Happiness, neutrality, sadness, anger and fear | Medium | Yes |
| EmoVoice | SDK | No | Determined by developer | High | Yes |
| Good Vibrations | SDK | - | Happy level, relaxed level, angry level, scared level and bored level | Medium | No |

Figure 2.6: Comparison of SERs in [2]

It has been observed that Vokaturi and EmoVoice are the tools that work without internet and are free to use ( OpenVokaturi is free for use, other versions of Vokatui require subscription ). BeyondVerbal was the easiest to use, whereas EmoVoice had the highest complexity. The authors were only able to test Vokaturi and Beyond Verbal for their study.

Multiple studies use OpenSMILE as the base for extracting acoustical features and then build machine learning models with those features for classifying emotions.

One such paper by Chen et al[40] extracts features using OpenSMILE and trains four types of machine learning learning models on the CASIA [41] corpus. The four classification models are : Random Forests, Support Vector Machines, Neural Networks and K-Nearest Neighbours. The best performance was given by the SVM model with a total accuracy of 81.11 percent, followed by the neural network at 80.56 percent. The other two models recorded accuracies of less than 60 percent.

A similar study by Felix Burkhardt, Brückl1 and Schuller leverages OpenSMILE to extract the "low-level descriptors" and "functionals" from the audio signal and train ML models on posed and spontaneous data for classifying age.

Now Ozseven et al [42] introduces a new toolkit SPAC for extracting features from the audio and compares those with attributes obtained from OpenSMILE and Praat. Mean Square Error and Mean Percentage Errors are used as the metrics for evaluation.

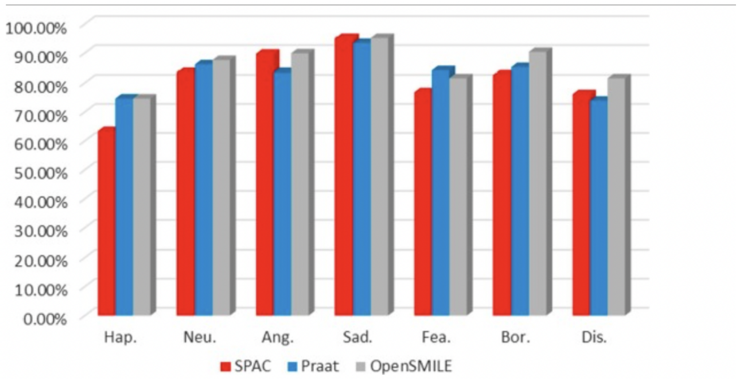


Figure 2.7: Comparison of SERs in [2]

SVM models (for emotion classification) were trained on the features obtained from each of the three systems. SVM trained on OpenSMILE features provides the most consistent output across all the emotions followed closely by PRAAT and SPAC.

## 2.9 Text Modality: Verbal Communication

Now, the first two modalities of facial expressions and audio formed the basis of the non-verbal communication framework. A third modality of text serves as the verbal counterpart and provides insights on "what was said". So a three modality face-audio-text system would not only cater to **"how something was said"** but it will also factor in **"what was said"** Since, it is so difficult to acknowledge or dismiss the presence of certain emotions at a particular timeframe, text plays a vital role in deciphering some of the complex emotional states. Text helps us in providing **context** to effectively understand what is being expressed and **increases the confidence** in the emotion recognition [43].

An interesting piece by Akhtar et al[43] proposes a "multi-task model" for extracting both sentiments(positve/negative) and emotions(anger, happiness/joy, disgust, fear, sadness or fear) from videos and consequently focuses on the "interdependence" of sentiments and emotions to **"increase the confidence"** of predicted emotions. The authors opine that information of positive/negative sentiment can help in prediction of happiness/anger (respectively) and vice versa. The authors use their MTL framework to "exploit" this sentiment-emotion interdependence "for performance improvement".

Kamboj et al [44] looked at a multimodal system involving linguistic, audio and visual modalities to uncover instances of potential "deception" in speeches of politicians from the Republican and Democratic political parties of the United States of America. For the linguistic modality(text),Linguistic Inquiry and Word Count (LIWC), semantic feature extraction, Parts-Of-Speech (POS tagging) and Unigram extraction methods were used in the above paper.

The “**invariance**” vs “**variance**” argument also holds semblance in the context of text and sentiment analysis as well. It has been suggested by Matsumoto and Assar (1992)[45] that the vocabulary of certain languages might express “emotional concepts” better than others. Harre,1986,Mesquita et al,1997 posit that the words expressing certain emotions in one culture might not convey the exact meanings when extended to a different culture.

**Impact of Gender** Differences have also been observed based on gender, where men are found to express less joy compared to women in two separate studies. The first study by Mohammad and Yang[46] observes "marked differences across genders" on the use of emotion carrying words in "work-place emails". Results from their study show men using terms from "fear-trust axis" whereas women preferring words involving joy and sadness ("joy-sadness axis"). The second study [47]considered tweets from stroke survivors and found statistically significant differences between men and women while analysing word frequencies and emotion proportions. Positive emotions like "anticipation,trust,joy" were found to be significantly higher in women whereas men showed higher proportion of negative emotions (fear, disgust and sadness) compared to women.

### Impact of Language and Culture

A study by Jackson et al[48] used "computational linguistics" to examine 24 emotion concepts ( like anger, fear, sadness, joy, grief, fear, etc ) across 2474 languages around the world. "Colexification" techniques were used and significant "cross-linguistic variability" was observed across 19 language families with geographic proximity shown to influence the cross-linguistic similarity.

Zhou et al[49] studied the impact of culture and language on self-reported emotional experiences in bilinguals (Chinese-English). There were significant differences in the self-reported emotional states of the Chinese and English monolinguals. And for the English-Chinese bilinguals, it was observed that the emotional patterns in their reportage were a mix of the patterns exhibited by their monolingual counterparts. Lindquist[50] opines that languages may "differentially impact" the experience and perception of the affective states.

## 2.10 Automatic Speech Recognition/Transcription

Accurate sentiment analysis hinges upon the accuracy of the transcription services being used for transcribing the audio/speech signals into text. This process is known as Automatic Speech Recognition (or ASR). Levis and Suvorov[51] describes ASR as an independent machine driven process of "decoding and transcribing" speech (audio signals). According to Lai et al, a typical ASR system analyzes the acoustic signals (it receives as input) using algorithms, models or patterns to produce output (generally as text ).

Davis, Biddulph and Balashek(1952)[52][53] created the first ASR sytem at the Bell Telephone Laboratories. This system was successfully able to recognize isolated digits (0 to 9) from the speech of a single speaker. Syllable recognition was made possible by Olson and Belar in 1956 with their phonetic typewriter. With the advent of Hidden Markov Models (HMMs) and Neural networks, there has been tremendous strides in the field of automatic speech recognition.

Since emotion and sentiment analysis are performed on transcribed speech, huge emphasis is generally put on refining the accuracy of the systems. Word error rate is a metric widely used to determine the accuracy of the transcriptions. Georgila et al[3] uses the Word Error Rate(or WER) to evaluate different ASR systems. Now, WER is "calculated by comparing the ASR output to the reference manual transcription".

$$\text{WER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Length of reference string}} \times 100\%$$

Figure 2.8: Word Error Rate( Georgila et al) [3]

- **Insertions** : New words introduced by the ASR system which was not there in the existing audio signal
- **Deletions** : Words present in the audio speech but have been omitted by the ASR in the transcribed output
- **Substitutions**: Words in the speech being replaced by different words by the ASR in the transcriptions

## 2.11 Performance Comparison of ASRs:

In a paper dated 2003, Moore et al[54] compared human speech recognition with automatic speech recognition systems trained using both supervised and unsupervised learning. The authors had opined that "a fantastic amount of speech"(about two to three times more than a typical human being) would be required to elevate the performance of ASR to the level of human speech recognition.

Almost 20 years later, there has been indeed a "substantial improvement"[? ] in the performance of Automatic Speech Recognition Systems. A 2018 paper by Spille et al[? ] had investigated the performance of humans (HSR) and deep-neural network (DNN) trained ASRs. It was found that **the ASR performed at par with the HSRs for "single-channel, small vocabulary tasks"** even in the presence of noise.

The performance dropped(gap of 12.3 dB) for "complex binaural scenes" where the humans could leverage spectral and spatial cues in the signals, which the machines were not able to comprehend. The performance gap reduced to only 2 dB when the positions of the speakers were supplied to the ASR system. It was also noted that HSRs were able to identify differences based on gender (by observing the differences in the fundamental frequency). On the other hand, their DNN based ASR was not able to distinguish between speakers of different gender.

As discussed in [55], the performance of the DNNs are directly proportional to the training data. Larger the amount of labeled training data available to the DNNs, better would be their performance."Retraining" or "Regularly updating" the DNNs would require tremendous amount of compute and memory resources, hence most ASR systems are **generally deployed on the cloud** instead of on-premise deployment.

A 2018 paper by Kim et al[? ] has evaluated the performance of five ASR services: Youtube, Google Cloud, Microsoft Azure, Trint and IBM Watson on noisy video-conferencing data. YouTube Captions performed the best followed by Google Cloud, Azure, Trint and Watson. The authors opine that the results "are in line" with Kepuska and Bouhouta's work [56] which showed Google ASR services to be better than Microsoft's.

The 2021 paper by Xu et al[55] has cited the word error rates (English) of the popular cloud-based speech-to-text(STT) services. Amazon Transcribe's Word Error Rate on English is reported to be 6.2 percent followed by Google Cloud at 5.6 percent, IBM Watson at 5.5 percent. Microsoft Azure's Word Error Rate is found to be the lowest at 5.1 percent. **These ASR WERs were at par with the human WER of 5.1 percent reported by the IBM/Appen study**

The paper has evaluated the WERs for French and found that Azure was robust to noises in the speech and reported the lowest ASR followed by Transcribe and Google Cloud. IBM Watson was found to be very sensitive to noise. A median WER of 14.29 percent on clean audio shot up to more than 70 percent for the noisy one.

## 2.12 Methods for Sentiment Analysis:

According to Nandwani and Verma[57], there are three types of widely used sentiment and emotion detection techniques: "lexicon based, machine learning based, and deep learning based".

Lexicon based methods do not require a labeled training set for classifying texts, whereas machine learning and deep learning methods explicitly require one[58]. Lexicon methods are found to work better for generic "less bounded" data-sets whereas machine learning methods perform considerably well for "specific domains". The lexicon based method analyzes the grammar whereas the machine learning method "fit the algorithms to the training dataset particularities".

Trinity College Dublin's Rocksteady, "a text analytic system"[59], uses the General Inquirer lexicon to perform sentiment analysis and calculate the frequency of positive and negative words in documents. GI has also been used to extract sentiment from financial data [59], movie reviews[60], fan tweets during sports matches[61] and even in Glassdoor company reviews/ [62][63]

Now, VADER and NRC lexicons were used by Mansoor et al[64] to evaluate the presence of **positive/negative** sentiments and **fear/trust** emotions (respectively) in their data-set. Time-series graphs were plotted to show the variation of sentiments and emotions post COVID-19.

Musto et al[65] evaluates the performance of four lexicons MPQA, SentiWordNet, WordNet-Affect and SenticNet on "two state-of-the-art datasets" curated from microblog posts. MPQA and SentiWordNet were found to be the best performing lexicons with WordNet-Affect performing poorly in both counts. SenticNet performed well for one dataset and failed miserably for the other one.

## 2.13 Challenges to Emotion Recognitions and the need for Multimodal emotion detection/Challenges for Labelling videos

Assessing emotions is an extremely challenging task and even during self-reportage of emotions it has been observed that the users are not able to properly express their emotions in words and there is always some element of "self-reporting error" present. (Soleymani et al)[66]. There will be instances where the users would not be able to remember the various emotions they felt while exposed to the stimuli (watching a video/clip ). In some cases, it has also been observed that the users misrepresent their feelings where a user might want to show that they were courageous where in fact they were scared[67]. To circumvent this, Soleymani proposes their "user-independent EEG" technique for classifying emotions. The approach shows "promising accuracy" on the smaller dataset and the authors opine that the performance would be scalable to a larger population as well.

Poria et al[68] uses non-invasive methods in their paper to explore two questions: whether there can be a "common framework" for multi-modal sentiment analysis and multi-modal emotion recognition and whether the three modalities (visual, audio and text) can be jointly used to improve the "performance of sentiment analysis classifiers".Using all modalities the authors clocked an accuracy of around 96.55 percent using their multi-modal model significantly outperforming the other models in the Multi-modal Opinion Utterances Dataset (MOUD Dataset) [69]. Training the model on the MOUD dataset, the authors also tried evaluating the performance of the multi-modal framework on two unseen datasets - Youtube dataset and the ICT-MMMO dataset and observed promising results.

Another study by Poria, Cambria, Hussain and Huang[70] evaluates the performance of a tri-modal (video, audio and text) system on the e-INTERFACE Dataset and reports an accuracy of 87.95 percent "outperforming the best state-of-the-art system by more than 10 percent. In this paper, the feature vectors of the three modalities were concatenated to form a single feature vector . The feature vectors were then fed to several classifiers and ten-fold cross validation was used to evaluate the performance. The Support Vector Machine (or SVM) classifer was observed to outperform the rest with a reported 60 percent reduction in error rate. Though simple, this approach of fusing tri-modal feature vectors into a single feature vector has been proven to show "significantly high accuracy".

Now, when it comes to multi-modal data fusion, two strategies are generally followed : **Decision-level fusion** and **Feature-level fusion**.

In **feature-level fusion**[70], the features/characteristics of each modality are concatenated into a "joint vector" and then classification operations are performed on the said vector. On the other hand, **decision-level fusion models** each modality independently and the classification is also performed independently. Results from each mode are combined using appropriate metrics like "expert rules" or operators like "majority votes, sums, products, and statistical weighting" to get a final multi-modal output.

# 3 Methods

## 3.1 System Diagram

A high-level system architecture diagram for multi-modal emotion detection, hypothesis testing, and data fusion has been presented in figure 3.1

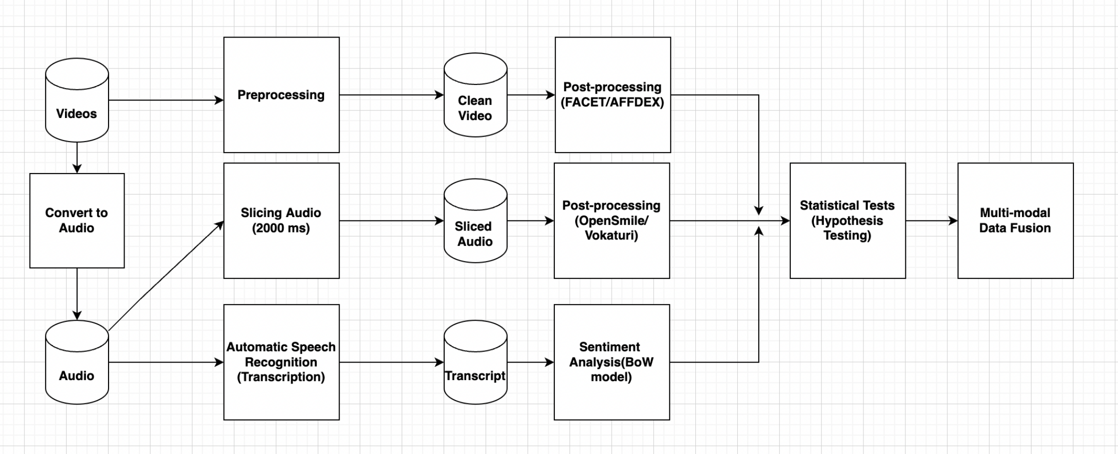

Videos
Preprocessing
Clean Video
Post-processing (FACET/AFFDEX)
Convert to Audio
Slicing Audio (2000 ms)
Sliced Audio
Post-processing (OpenSmile/ Vokaturi)
Statistical Tests (Hypothesis Testing)
Multi-modal Data Fusion
Audio
Automatic Speech Recognition (Transcription)
Transcript
Sentiment Analysis(BoW model)

Figure 3.1: System Pipeline and Architecture)

The process starts from a **MP4** video file, which requires proper processing to render it suitable for use across all the modalities in the system. Once the video is processed (section 3.3),the **video is converted to an audio file** with **.wav** format using **"FFmpeg" package** of Python. Once the video has been converted to audio, **AudioSegment** package of Python is used to slice the audio into samples of 2000 milliseconds. These samples are then fed into the OpenSmile and Vokaturi scripts for classifying emotions. The output files for the individual samples are then concatenated into a single file to get the output for the original audio file.

For transcription (automatic speech recognition), the original audio file (not the sliced samples) is fed to the ASR system to get the **transcribed output with timestamp**. Now two different analyses are done here:

First, the entire transcribed output is analyzed on four levels (lexical, syntactic, semantic, and pragmatic) using a Bag-Of-Words(BOW model). Lexicons are used to get the sentiment out of each word: positive or negative. Second, with the help of the timestamps present in the transcripts, the transcriptions are sliced into samples of 2 seconds with a python script. The slices are done at 2000 milliseconds to maintain parity with the audio signal so that the audio frames can be compared with the transcribed frames for multi-modal detection of emotion.

## 3.2 Collection of Data-set and Challenges encountered

As discussed in Chapter 2, the proposed database of politicians' speeches will be semi-posed or semi-spontaneous, having the primary attributes of both posed and spontaneous images. Data collected by previous students at the School of Computer Science and Statistics have been incorporated into the dataset.

180 videos have been adopted from the data that was collected by students of the previous cohort as a part of the ongoing study.[6] Since existing literature finds inter-culture variation in emotional expression, a new demographic (Bengal) has been introduced into the existing TCD database to check for the same. Bengali politicians from both India and Bangladesh were taken for the data-set. 25 videos of 10 Bengali politicians (speaking in English) were incorporated in the existing data-set (Please note I am a native speaker of the Bengali language).

**Hence the total number of videos that have been used in this study is 205.**

| Country | Total Videos | Female Politicians | Male Politicians | EMOTIENT frames | AFFECTIVA frames |
|---|---|---|---|---|---|
| USA | 66 | 10 | 8 | 419485 | 408055 |
| Ireland | 51 | 5 | 5 | 253319 | 257320 |
| India | 36 | 4 | 11 | 199887 | 192637 |
| Pakistan | 6 | 0 | 2 | 59553 | 58788 |
| UK | 5 | 0 | 1 | 52387 | 47068 |
| China | 15 | 2 | 4 | 65264 | 64058 |
| Japan | 3 | 0 | 2 | 35994 | 36371 |
| New Zealand | 5 | 1 | 0 | 25750 | 25128 |
| South Korea | 3 | 0 | 2 | 30457 | 30430 |
| Germany | 9 | 1 | 1 | 36372 | 36183 |
| Bangladesh | 6 | 1 | 1 | 37112 | 40325 |
| **Total** | **205** | **24** | **37** | **1215580** | **1196363** |

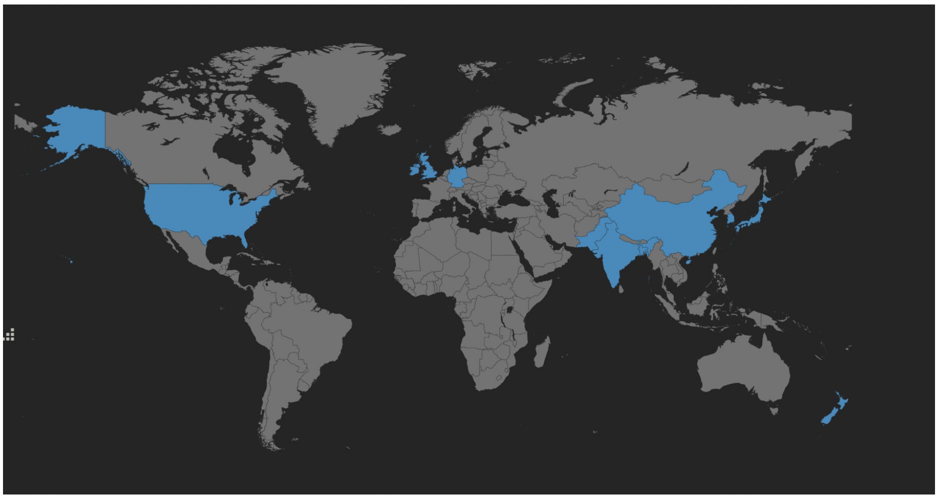


Figure 3.2: World Map: Nationalities of the politicians used in the dissertation

### Collection of Labelled data-sets and Motivation

**RAVDESS[71] :** The Ryerson Audio Visual Database of Emotional Speech and Song (or RAVDESS in short) is a collection of facial and audio expressions recorded by North American English speakers. The database consists of speeches and songs recorded by the "gender balanced" set of speakers (12 female and 12 male) expressing seven set of emotions for speech and five set of emotions for song. Now, Ravdess provides the data in three formats: visual, audio and audio-visual.

For the dissertation, 720 videos (60 videos x 12 speakers : 6 male and 6 female) have been selected from the visual modality (did not have audio) . These videos were fed into Emotient and Affectiva for processing where **2 videos faced failure** and hence the RAVDESS results in this dissertation are based on 718 videos.

**Reason for selecting the dataset:** As discussed, the data-set of the politicians is a semi-spontaneous database. As most of the literature comparing FERs evaluate the performance of the systems on posed and spontaneous expression databases, RAVDESS was introduced to act as the posed expression counterpart of our semi-spontaneous database.

**EMO-DB[72]** The Berlin Emo-DB dataset[72] is the open source gender balanced German emotional database that consists of audio recordings of ten professional speakers (five female and five males). The emotions recorded in the database are happiness, anger, boredom, anxiety, sadness, disgust and neutral. Originally containing 800 utterances, the version that was used in the dissertation contains 535 utterances.

**Reason for selecting the dataset:** Both the SER systems, OpenSmile and Vokaturi were trained on EMO-DB along with other databases. So, before checking the performance of the systems on unseen databases, it is worth evaluating their performance on a database that they both were trained on.

### SAVEE[73]

Surrey Audio-Visual Expressed Emotion Database (or SAVEE in short) consists of 4 British male actors expressing seven emotions with 480 English utterances. Sentences have been selected from the TIMIT corpus.

**Reason for selecting the dataset:** EMO-DB dataset is one which both the databases have encountered. TCD Dataset is a database which none of the systems have collected. On the other hand, only OpenSMILE has been trained on SAVEE and Vokaturi has not encountered SAVEE. It would be worth checking how the two systems behave on the SAVEE database and how inter-annotator agreement (or the Cohen's Kappa Score) reacts to the change in database.

## 3.3 Pre-processing and Challenges Encountered

The MP4 videos undergo a process of **trim and crop** before they are deemed suitable for processing via the iMotions Suite(FACET Emotient and AFFDEX Affectiva). **QuickTime Player(Mac OS)** and **Movavi Video Suite 22** are used for trimming/cropping the videos. It is done to ensure that only the subject(politician) is in the frame and nobody else. The interviewer and any person in the background must be removed to guarantee that only the politician's features get analyzed across the modalities (face, audio, and text ) and not those of any other person. Care is also taken to make sure that only those frames are selected where the person is looking directly at the camera.

Even after proper pre-processing of the videos was done, for some videos, it was observed that there was some discrepancy in how the videos got processed by the two FER systems. In one such video, it was observed that Emotient could only properly process **19 frames of the video** whereas Affectiva was able to successfully process **5283 frames.**

Whenever this issue was encountered, the video in question was further zoomed in and cropped to put more emphasis on the face. This **second level of pre-processing** resulted in a better face-detection rate for Emotient, which was able to recognize and process 4739 frames in the second run. Hence for similar instances, the **second level of pre-processing** was done, and the videos were again fed into the system.

## 3.4 Facial Emotion Recognition

### 3.4.1 Architecture of Emotient FACET

Computer Expression Recognition Toolbox( or CERT), the precursor of FACET Emotient (one of the tools that have been used for facial expression analysis in the dissertation), is a tool that can perform "real-time fully automated coding of facial expression" [16]. CERT can process live camera feed, images, and videos. CERT provides "probability estimates" for 19 AUS as well as evidence for 6 "prototypical emotions" (happiness, anger, surprise, sadness, fear, and disgust ). Yaw, pitch, roll (head orientation/head pose data ), and (x,y) coordinates of facial feature points are also provided.

The CERT pipeline can be described as follows:

- **Detection of Face:** Detection of faces in CERT is done with an extension of the Viola-Jones algorithm. [19]. Boosting algorithms like WaldBoost and GentleBoost are applied for "cascade threshold detection ".[16]. A hit rate of 80.6 percent (and 58 false alarms ) has been observed for the CERT face detector on the CMU+MIT dataset [? ]. The "largest found face" in each video frame is selected for further processing.

- **Detection of Facial Features**:

  After the face segmentation has been done, a set of 10 features: inner and outer eye corners, eye centers, nose tip, inner and outer mouth corners, mouth centers are detected on the face using "feature specific detectors". Each detector provides the "log likelihood" ratio of a feature "being present" at the coordinate(x,y).Linear regression (trained on the GENKI dataset) is used to refine the location estimates.

- **Registration of Face**:

  The facial bounding box is re-estimated at a size of 96 x 96 pixels with the help of an affline transformation. The pixels of the re-estimated patch are converted into a 2D array to be used in the next stage.

- **Extraction of Features**: The 96x96 pixel patch is "convolved" (Fast Fourier Transform (FFT)) with Gabor filters of different spatial frequencies and orientations. The filter outputs are then fused into a single feature vector for further processing in the later stages.

- **Recognition of Action Units:** The feature vector is passed as input to separate support vector machine classifiers for each Action Unit. Now the SVM classifier provides estimates of the Action Unit intensities. These AU SVMs were trained on a combination of many databases: **Posed Expression Databases** :Ekman-Hager [8] , Cohn-Kanade [12], Man-Machine Interaction [17] (contains both posed and spontaneous expressions,but only the posed ones were used for training ) ; **Spontaneous Expression Databases:** M3 [18] , and two datasets curated by the US government ( non-public)

- **Dynamics and Intensity of Expressions:** For each action unit, a continuous value (signifying the distance of the input feature vector and the Support Vector Machine's hyperplane) is provided on a frame-by-frame basis. It was observed that the CERT output values were "significantly correlated" with the facial action intensities measured by the FACS experts. This **frame-by-frame temporal information** on facial action units and their intensities proved to be a resounding success because this was extremely difficult using manual coding.

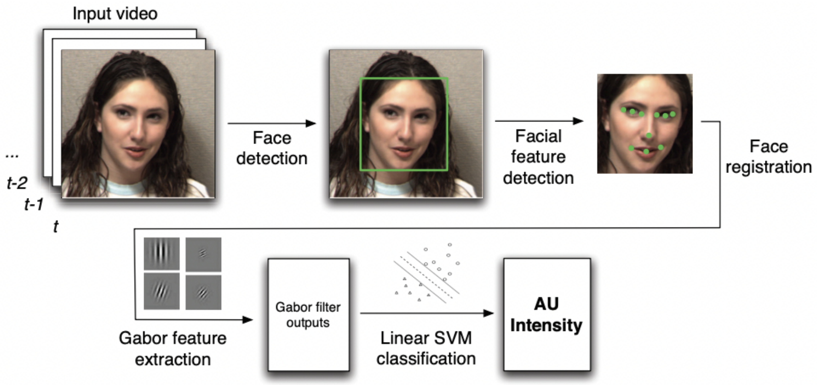


Fig. 2. Processing pipeline of the Computer Expression Recognition Toolbox (CERT) from video to expression intensity estimates.

Figure 3.3: CERT pipeline)

### Conversion of Emotion evidence into Emotion Intensities

The evidences corresponding to the emotions represent the "estimated odds "[74] of that emotion being present in a given video frame, expressed in a decimal logarithmic scale. Now, these evidences can be converted to probabilities with the following formula, where P represents the probability of the emotion to be present in the video frame under consideration.

$$P = 1/(1 + 10^{-\text{evidence score}})$$

Figure 3.4: Conversion of evidence scores to probability of presence of emotions)

If the evidence score for "anger" in a particular frame is zero, the probability of occurrence of anger at that frame is 0.5. Or in other words, it is "equally likely" that anger is present or absent in that frame. Similarly, if the evidence is 0.5, then a 76 percent probability for anger is expected in the frame. Evidences greater than 2 represent a probability of around 1 (or almost a 100 percent chance of the emotion being present in that frame). Such evidence scores are given for emotion categories as well as the Action Units. Just like emotion, higher the value of the evidence score for the AU, the higher is the probability of that action unit being activated for the frame.

### 3.4.2 Architecture of AFFDEX Affectiva

The entire AFFDEX Affectiva pipeline can be broken down into four steps

- **Detection of face and facial landmarks** : The Viola-Jones[19] face detection algorithm is used to detect faces in the input(image/video). 34 Landmarks are detected in each facial bounding box. Bounding boxes are ignored if the confidence of landmark detection falls below a certain threshold. The landmarks along with intra-ocular distance and head poses, are detected for each face.

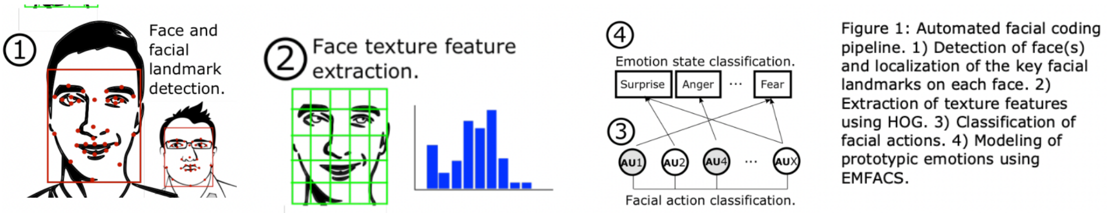

①
Face and facial landmark detection.
②
Face texture feature extraction.
④
Emotion state classification.
Surprise
Anger
···
Fear
③
AU1
AU2
AU4
···
AUX
Facial action classification.
Figure 1: Automated facial coding pipeline. 1) Detection of face(s) and localization of the key facial landmarks on each face. 2) Extraction of texture features using HOG. 3) Classification of facial actions. 4) Modeling of prototypic emotions using EMFACS.

Figure 3.5: AFFDEX Affectiva Pipeline)

- **Extraction of "face texture feature"** The facial landmarks define regions of interest in the image/frame and then "Histogram of Oriented features(HOG features) "[75] are extracted.

- **Classification of facial actions**: Each facial action is then attributed a score between 0 and 100 by Support Vector Machine (SVM Classifiers). These classifiers were trained on "10000s of manually coded facial images" that were collected from different parts of the world.

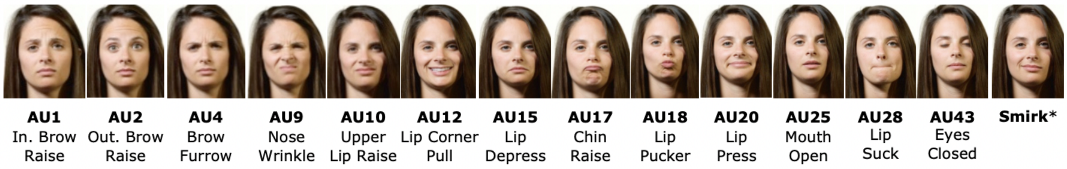


Figure 3.6: AUs detected by the SDK [2]

- **Emotion expressions**: Based on the "combinations of facial actions", the emotions (joy, anger, disgust, sadness, fear, surprise and contempt) are decided. Affectiva uses the Emotional Facial Action Coding System [76] for this purpose. Like their facial action counterparts, scores between 0 and 100 are also assigned to the emotional expressions.

## 3.5 Speech Emotion Recognition

### 3.5.1 Architecture of OpenSMILE

- **Data Memory:** Data Memory is the central component in the open-SMILE/open-EAR architecture. It manages ring-buffer storage for feature data thereby enabling "memory efficient incremental processing".
- **Data Source:** Data Source components feed the input data (wave files ) into the Data Memory with the Data Writer Sub-component.
- **Data Processor:** The brain of the OpenSMILE architecture, the data processor, reads contours or data frames from one part of the Data Memory and writes new frames to another location in the Data Memory after processing the data. Data Processor has both Data Reader and Writer sub-components for interfacing with the Data Memory.

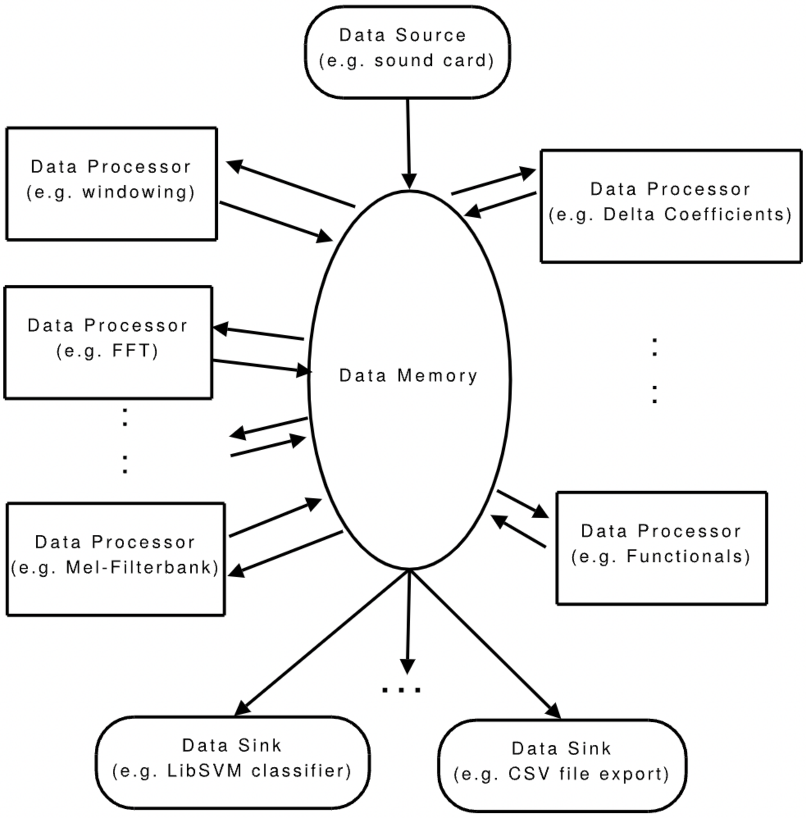


Figure 3.7: OpenSMILE architecture

- **Data Processor:** The brain of the OpenSMILE architecture, the data processor, reads contours or data frames from one part of the Data Memory and writes new frames to another location in the Data Memory after processing the data. Data

Processor has both Data Reader and Writer sub-components for interfacing with the Data Memory.

- **Data Sink:** Data Sink reads the processed data from the memory and writes the data to files or feeds the data into the classifiers for emotion classification.

### 3.5.2 Architecture of OpenVokaturi

OpenVokaturi takes into account nine acoustical cues from the sample sound and computes the probabilities for the five emotions: **happiness, anger, fear, sadness, and neutral.**

The system has been trained on two databases, EmoDB and SAVEE (768 annotated recordings in total for the five emotions. Initially trained only on EMODB with a linear discriminant analysis, later versions have switched to artificial neural networks with two hidden layers (100 nodes and 20 nodes respectively). An additional training database SAVEE has also been brought into the fold. The developers opine that the inclusion of SAVEE has improved the potential of "generalization to new sounds" significantly".

OpenVokaturi uses leave-one-out cross validation on the two databases (768 recordings in total ) and the training is repeated 3 times.Following this approach, a classification accuracy of 66.5 percent [2] has been reported on the two databases.

**Using Vokaturi:**

OpenVokaturi can be downloaded as a Python package which can be readily used to classify emotions into five possible outputs as mentioned before. Now the Python library does not the expose the nine acoustical cues which are used internally for classifying emotions. **Mimicking the Vokaturi library funtion in C, changes needed to be made directly in the Vokaturi library files to expose the nine cues along with the five emotions.**

### 3.5.3 Architecture of Praat

Praat has been introduced into this research as as third system to serve as the reference measure for comparing the acoustical cues we get from the two competing systems: openSMILE and Vokaturi. As described in the literature, Praat is considered as the gold standard for detecting the fundamental frequency and acoustical correlates.

Praat Scripting was used to generate Praat tables for the fundamental frequencies which were then converted to CSV files using Python for further analysis and comparison with OpenSmile and Vokaturi.

## 3.6 Automatic Speech Recognition(Transcription) and the challenges faced:

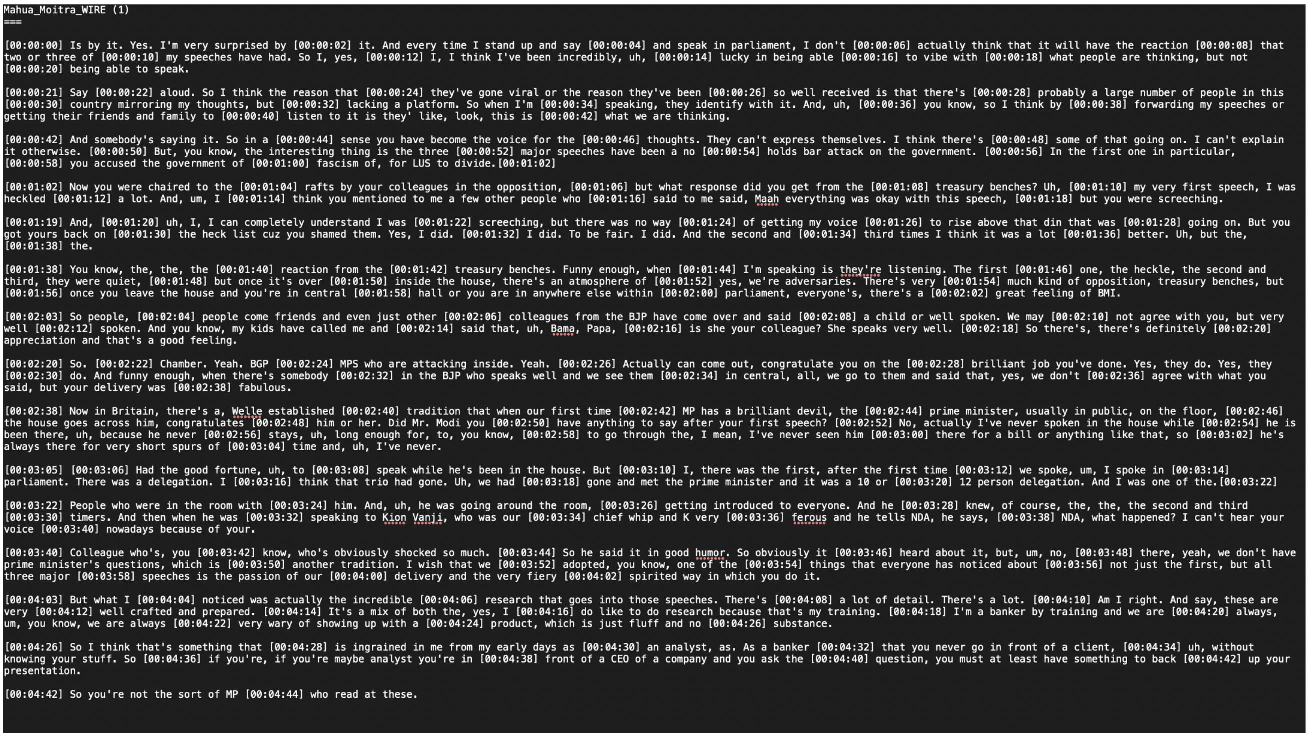
Google's **Speech-To-Text(STT) API** was initially used for transcribing the speeches of the politicians into texts. Cloud credentials were used to sign-in to the Google Console and a python script was written to call the STT API for transcription. Now, the STT API doesn't allow transcription of audio samples having duration more than a minute, unless files are kept in Google containers. For security concerns, the files were not uploaded to the cloud and instead an alternative framework was designed.

To circumvent the one minute restriction, the audio samples were split into one minute (60 seconds) intervals and then the API calls were made from the script to get the transcribed texts. The transcribed outputs were then concatenated back to get the transcription for the entire audio.

As expected, there were word-losses encountered at the interfaces(splits). But it was also observed that the Word Error Rate was very high for some videos, and in some cases, only a couple of words made it through the transcription process. As we have seen in the literature review section, Automatic Speech Recognition(ASR) is prone to noise, which might be the reason behind substantial word loss in some of the audios.

Hence, a conscious switch was made from Google's Speech-To-Text API to **Descript** for the preferred Method of transcription in the dissertation.

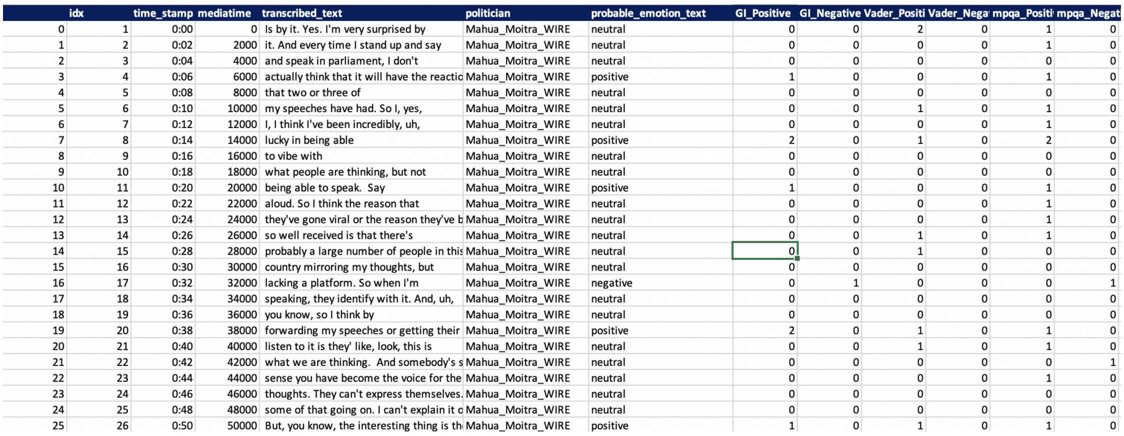
Unlike GCP, **Descript** was able to able to process the audio/video irrespective of the length. Since, this dissertation talks about the possibility of multi-modal fusions in later sections, it was imperative to have the transcriptions sliced at 2000 milliseconds(2 second intervals) to compare it with audio. Descript has the option to attach timestamps to the transcribe text and hence a two second(2000 ms) window was selected to ensure that proper timestamps are added in the final transcript.



```
Mahua_Moitra_WIRE (1)
===

[00:00:00] Is by it. Yes. I'm very surprised by [00:00:02] it. And every time I stand up and say [00:00:04] and speak in parliament, I don't [00:00:06] actually think that it will have the reaction [00:00:08] that
two or three of [00:00:10] my speeches have had. So I, yes, [00:00:12] I, I think I've been incredibly, uh, [00:00:14] lucky in being able [00:00:16] to vibe with [00:00:18] what people are thinking, but not
[00:00:20] being able to speak.

[00:00:21] Say [00:00:22] aloud. So I think the reason that [00:00:24] they've gone viral or the reason they've been [00:00:26] so well received is that there's [00:00:28] probably a large number of people in this
[00:00:30] country mirroring my thoughts, but [00:00:32] lacking a platform. So when I'm [00:00:34] speaking, they identify with it. And, uh, [00:00:36] you know, so I think by [00:00:38] forwarding my speeches or
getting their friends and family to [00:00:40] listen to it is they' like, look, this is [00:00:42] what we are thinking.

[00:00:42] And somebody's saying it. So in a [00:00:44] sense you have become the voice for the [00:00:46] thoughts. They can't express themselves. I think there's [00:00:48] some of that going on. I can't explain
it otherwise. [00:00:50] But, you know, the interesting thing is the three [00:00:52] major speeches have been a no [00:00:54] holds bar attack on the government. [00:00:56] In the first one in particular,
[00:00:58] you accused the government of [00:01:00] fascism of, for LUS to divide.[00:01:02]

[00:01:02] Now you were chaired to the [00:01:04] rafts by your colleagues in the opposition, [00:01:06] but what response did you get from the [00:01:08] treasury benches? Uh, [00:01:10] my very first speech, I was
heckled [00:01:12] a lot. And, um, I [00:01:14] think you mentioned to me a few other people who [00:01:16] said to me said, Maah everything was okay with this speech, [00:01:18] but you were screeching.

[00:01:19] And, [00:01:20] uh, I, I can completely understand I was [00:01:22] screeching, but there was no way [00:01:24] of getting my voice [00:01:26] to rise above that din that was [00:01:28] going on. But you
got yours back on [00:01:30] the heck list cuz you shamed them. Yes, I did. [00:01:32] I did. To be fair. I did. And the second and [00:01:34] third times I think it was a lot [00:01:36] better. Uh, but the,
[00:01:38] the.

[00:01:38] You know, the, the, the [00:01:40] reaction from the [00:01:42] treasury benches. Funny enough, when [00:01:44] I'm speaking is they're listening. The first [00:01:46] one, the heckle, the second and
third, they were quiet, [00:01:48] but once it's over [00:01:50] inside the house, there's an atmosphere of [00:01:52] yes, we're adversaries. There's very [00:01:54] much kind of opposition, treasury benches, but
[00:01:56] once you leave the house and you're in central [00:01:58] hall or you are in anywhere else within [00:02:00] parliament, everyone's, there's a [00:02:02] great feeling of BMI.

[00:02:03] So people, [00:02:04] people come friends and even just other [00:02:06] colleagues from the BJP have come over and said [00:02:08] a child or well spoken. We may [00:02:10] not agree with you, but very
well [00:02:12] spoken. And you know, my kids have called me and [00:02:14] said that, uh, Bama, Papa, [00:02:16] is she your colleague? She speaks very well. [00:02:18] So there's, there's definitely [00:02:20]
appreciation and that's a good feeling.

[00:02:20] So. [00:02:22] Chamber. Yeah. BGP [00:02:24] MPS who are attacking inside. Yeah. [00:02:26] Actually can come out, congratulate you on the [00:02:28] brilliant job you've done. Yes, they do. Yes, they
[00:02:30] do. And funny enough, when there's somebody [00:02:32] in the BJP who speaks well and we see them [00:02:34] in central, all, we go to them and said that, yes, we don't [00:02:36] agree with what you
said, but your delivery was [00:02:38] fabulous.

[00:02:38] Now in Britain, there's a, Welle established [00:02:40] tradition that when our first time [00:02:42] MP has a brilliant devil, the [00:02:44] prime minister, usually in public, on the floor, [00:02:46]
the house goes across him, congratulates [00:02:48] him or her. Did Mr. Modi you [00:02:50] have anything to say after your first speech? [00:02:52] No, actually I've never spoken in the house while [00:02:54] he is
been there, uh, because he never [00:02:56] stays, uh, long enough for, to, you know, [00:02:58] to go through the, I mean, I've never seen him [00:03:00] there for a bill or anything like that, so [00:03:02] he's
always there for very short spurs of [00:03:04] time and, uh, I've never.

[00:03:05] [00:03:06] Had the good fortune, uh, to [00:03:08] speak while he's been in the house. But [00:03:10] I, there was the first, after the first time [00:03:12] we spoke, um, I spoke in [00:03:14]
parliament. There was a delegation. I [00:03:16] think that trio had gone. Uh, we had [00:03:18] gone and met the prime minister and it was a 10 or [00:03:20] 12 person delegation. And I was one of the.[00:03:22]

[00:03:22] People who were in the room with [00:03:24] him. And, uh, he was going around the room, [00:03:26] getting introduced to everyone. And he [00:03:28] knew, of course, the, the, the second and third
[00:03:30] timers. And then when he was [00:03:32] speaking to Kion Vanji, who was our [00:03:34] chief whip and K very [00:03:36] ferous and he tells NDA, he says, [00:03:38] NDA, what happened? I can't hear your
voice [00:03:40] nowadays because of your.

[00:03:40] Colleague who's, you [00:03:42] know, who's obviously shocked so much. [00:03:44] So he said it in good humor. So obviously it [00:03:46] heard about it, but, um, no, [00:03:48] there, yeah, we don't have
prime minister's questions, which is [00:03:50] another tradition. I wish that we [00:03:52] adopted, you know, one of the [00:03:54] things that everyone has noticed about [00:03:56] not just the first, but all
three major [00:03:58] speeches is the passion of our [00:04:00] delivery and the very fiery [00:04:02] spirited way in which you do it.

[00:04:03] But what I [00:04:04] noticed was actually the incredible [00:04:06] research that goes into those speeches. There's [00:04:08] a lot of detail. There's a lot. [00:04:10] Am I right. And say, these are
very [00:04:12] well crafted and prepared. [00:04:14] It's a mix of both the, yes, I [00:04:16] do like to do research because that's my training. [00:04:18] I'm a banker by training and we are [00:04:20] always,
um, you know, we are always [00:04:22] very wary of showing up with a [00:04:24] product, which is just fluff and no [00:04:26] substance.

[00:04:26] So I think that's something that [00:04:28] is ingrained in me from my early days as [00:04:30] an analyst, as. As a banker [00:04:32] that you never go in front of a client, [00:04:34] uh, without
knowing your stuff. So [00:04:36] if you're, if you're maybe analyst you're in [00:04:38] front of a CEO of a company and you ask the [00:04:40] question, you must at least have something to back [00:04:42] up your
presentation.

[00:04:42] So you're not the sort of MP [00:04:44] who read at these.
```

Figure 3.8: A sample Descript Output

Now, the Descript transcriptions **were parsed using a Python script to create a time-series Python dataframe** for the entire transcript. .



| | idx | time_stamp | mediatime | transcribed_text | politician | probable_emotion_text | GI_Positive | GI_Negative | Vader_Positi | Vader_Nega | mpqa_Positi | mpqa_Negat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0:00 | 0 | Is by it. Yes. I'm very surprised by | Mahua_Moitra_WIRE | neutral | 0 | 0 | 2 | 0 | 1 | 0 |
| 1 | 2 | 0:02 | 2000 | it. And every time I stand up and say | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 3 | 0:04 | 4000 | and speak in parliament, I don't | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 4 | 0:06 | 6000 | actually think that it will have the reactio | Mahua_Moitra_WIRE | positive | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | 5 | 0:08 | 8000 | that two or three of | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 6 | 0:10 | 10000 | my speeches have had. So I, yes, | Mahua_Moitra_WIRE | neutral | 0 | 0 | 1 | 0 | 1 | 0 |
| 6 | 7 | 0:12 | 12000 | I, I think I've been incredibly, uh, | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 1 | 0 |
| 7 | 8 | 0:14 | 14000 | lucky in being able | Mahua_Moitra_WIRE | positive | 2 | 0 | 1 | 0 | 2 | 0 |
| 8 | 9 | 0:16 | 16000 | to vibe with | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 10 | 0:18 | 18000 | what people are thinking, but not | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 11 | 0:20 | 20000 | being able to speak.  Say | Mahua_Moitra_WIRE | positive | 1 | 0 | 0 | 0 | 1 | 0 |
| 11 | 12 | 0:22 | 22000 | aloud. So I think the reason that | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 1 | 0 |
| 12 | 13 | 0:24 | 24000 | they've gone viral or the reason they've b | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 1 | 0 |
| 13 | 14 | 0:26 | 26000 | so well received is that there's | Mahua_Moitra_WIRE | neutral | 0 | 0 | 1 | 0 | 1 | 0 |
| 14 | 15 | 0:28 | 28000 | probably a large number of people in this | Mahua_Moitra_WIRE | neutral | 0 | 0 | 1 | 0 | 0 | 0 |
| 15 | 16 | 0:30 | 30000 | country mirroring my thoughts, but | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 17 | 0:32 | 32000 | lacking a platform. So when I'm | Mahua_Moitra_WIRE | negative | 0 | 1 | 0 | 0 | 0 | 1 |
| 17 | 18 | 0:34 | 34000 | speaking, they identify with it. And, uh, | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 19 | 0:36 | 36000 | you know, so I think by | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 20 | 0:38 | 38000 | forwarding my speeches or getting their | Mahua_Moitra_WIRE | positive | 2 | 0 | 1 | 0 | 1 | 0 |
| 20 | 21 | 0:40 | 40000 | listen to it is they' like, look, this is | Mahua_Moitra_WIRE | neutral | 0 | 0 | 1 | 0 | 1 | 0 |
| 21 | 22 | 0:42 | 42000 | what we are thinking.  And somebody's s | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 1 |
| 22 | 23 | 0:44 | 44000 | sense you have become the voice for the | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 1 | 0 |
| 23 | 24 | 0:46 | 46000 | thoughts. They can't express themselves. | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | 25 | 0:48 | 48000 | some of that going on. I can't explain it o | Mahua_Moitra_WIRE | neutral | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 26 | 0:50 | 50000 | But, you know, the interesting thing is th | Mahua_Moitra_WIRE | positive | 1 | 0 | 1 | 0 | 1 | 0 |

Figure 3.9: Descript output parsed and converted to a time-series Python dataframe [2]

**Understanding the Context: Sentiment Analysis of the Transcripts:**

**Sentiment Analysis:**

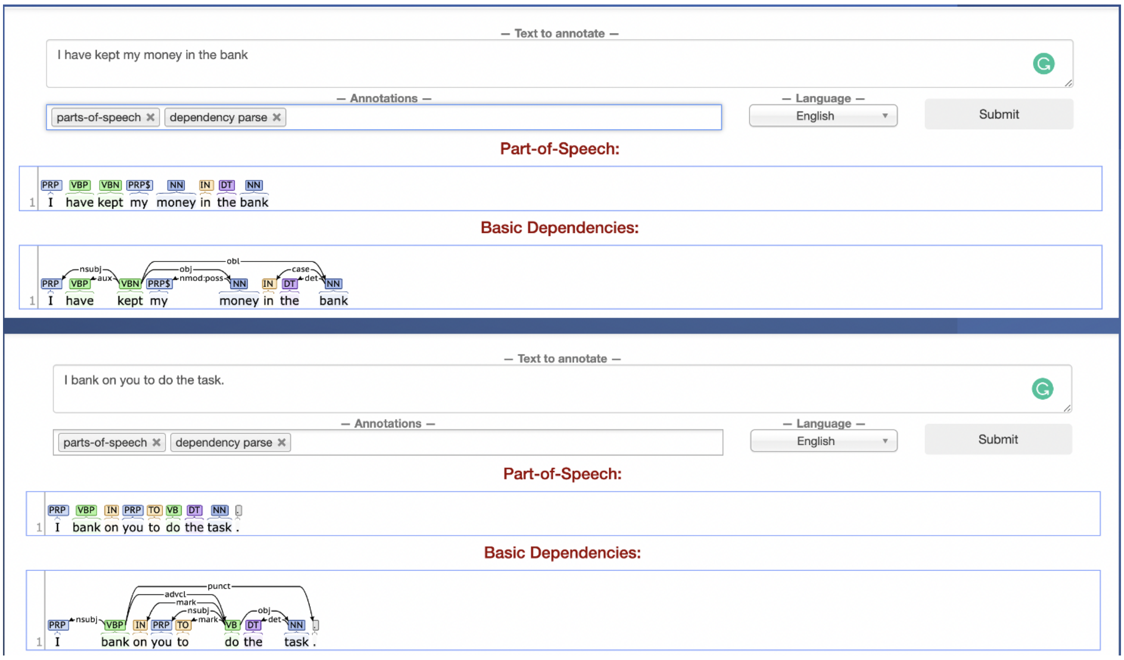
**Understanding the Corpus: An Overview:**

**Lexical Analysis** Euzenat describes lexical analysis as the ability to "segment the representation in words(or symbols) "[6].This level of the analysis splits the corpus into the basic atomic units (terms/words) and compares the frequency distribution of the words in the special language corpus with that of a General Language Corpus. Here in this dissertation, the transcriptions of the political speeches are used to create a "Political Corpus" which has been compared to the American National Corpus which has been taken as the "General language Corpus".

### Syntactic Analysis

### Parts Of Speech Tagging: NLTK

Python's NLTK POS tagger has been used for extraction of the Parts-Of-Speech. A word can have multiple meanings in the corpus and hence can have multiple Parts-Of-Speech tags associated to it. For example, the word "bank" can be used both as a noun and a verb. When the word is used in a sentence like "I am planning to visit the bank", it has been used as a "noun". Instead if the word is used in "I bank on you" it will be considered as a verb. These nuances need to be considered for a proper syntactic analysis of the corpus.



— Text to annotate —

I have kept my money in the bank

— Annotations — parts-of-speech, dependency parse

— Language — English

Part-of-Speech:

1 I/PRP have/VBP kept/VBN my/PRP$ money/NN in/IN the/DT bank/NN

Basic Dependencies:

1 I have kept my money in the bank

— Text to annotate —

I bank on you to do the task.

— Annotations — parts-of-speech, dependency parse

— Language — English

Part-of-Speech:

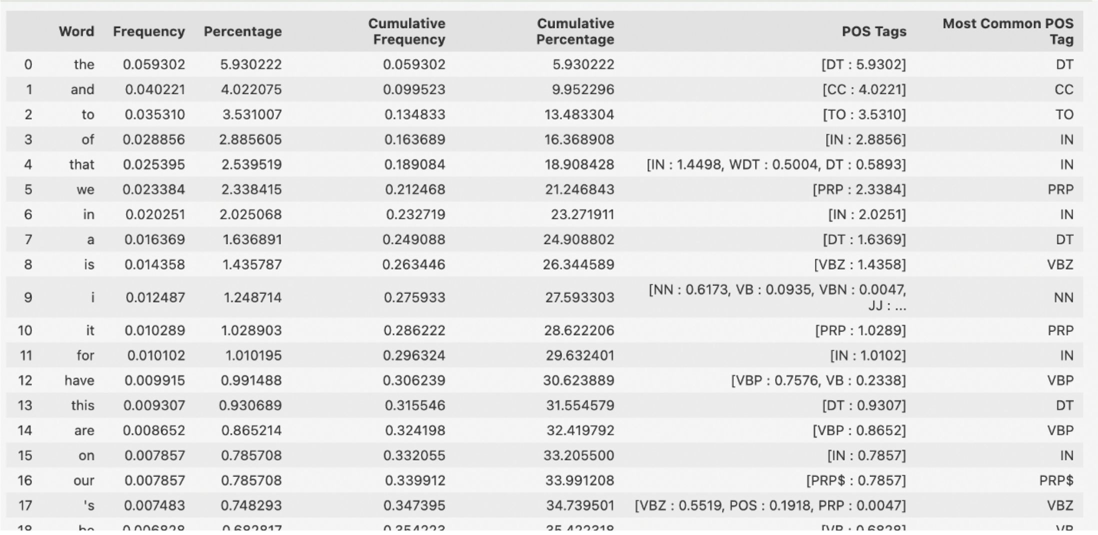
1 I/PRP bank/VBP on/IN you/PRP to/TO do/VB the/DT task/NN ./.

Basic Dependencies:

1 I bank on you to do the task .

Figure 3.10: Stanford's core NLP run showing the different usages of the word "bank" and the POS Tags [2]

Python dictionaries were created for each term to keep a track of the different POS tags and the relevent percentages for each POS tag was stored in it.



|  | Word | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | POS Tags | Most Common POS Tag |
|---|---|---|---|---|---|---|---|
| 0 | the | 0.059302 | 5.930222 | 0.059302 | 5.930222 | [DT : 5.9302] | DT |
| 1 | and | 0.040221 | 4.022075 | 0.099523 | 9.952296 | [CC : 4.0221] | CC |
| 2 | to | 0.035310 | 3.531007 | 0.134833 | 13.483304 | [TO : 3.5310] | TO |
| 3 | of | 0.028856 | 2.885605 | 0.163689 | 16.368908 | [IN : 2.8856] | IN |
| 4 | that | 0.025395 | 2.539519 | 0.189084 | 18.908428 | [IN : 1.4498, WDT : 0.5004, DT : 0.5893] | IN |
| 5 | we | 0.023384 | 2.338415 | 0.212468 | 21.246843 | [PRP : 2.3384] | PRP |
| 6 | in | 0.020251 | 2.025068 | 0.232719 | 23.271911 | [IN : 2.0251] | IN |
| 7 | a | 0.016369 | 1.636891 | 0.249088 | 24.908802 | [DT : 1.6369] | DT |
| 8 | is | 0.014358 | 1.435787 | 0.263446 | 26.344589 | [VBZ : 1.4358] | VBZ |
| 9 | i | 0.012487 | 1.248714 | 0.275933 | 27.593303 | [NN : 0.6173, VB : 0.0935, VBN : 0.0047, JJ : ... | NN |
| 10 | it | 0.010289 | 1.028903 | 0.286222 | 28.622206 | [PRP : 1.0289] | PRP |
| 11 | for | 0.010102 | 1.010195 | 0.296324 | 29.632401 | [IN : 1.0102] | IN |
| 12 | have | 0.009915 | 0.991488 | 0.306239 | 30.623889 | [VBP : 0.7576, VB : 0.2338] | VBP |
| 13 | this | 0.009307 | 0.930689 | 0.315546 | 31.554579 | [DT : 0.9307] | DT |
| 14 | are | 0.008652 | 0.865214 | 0.324198 | 32.419792 | [VBP : 0.8652] | VBP |
| 15 | on | 0.007857 | 0.785708 | 0.332055 | 33.205500 | [IN : 0.7857] | IN |
| 16 | our | 0.007857 | 0.785708 | 0.339912 | 33.991208 | [PRP$ : 0.7857] | PRP$ |
| 17 | 's | 0.007483 | 0.748293 | 0.347395 | 34.739501 | [VBZ : 0.5519, POS : 0.1918, PRP : 0.0047] | VBZ |
| 18 | be | 0.006828 | 0.682817 | 0.354223 | 35.422318 | [VB : 0.6828] | VB |

Figure 3.11: Frequency Distribution and POS Tags

**Semantic Analysis:**

- Relative frequencies of words in the Special Language Corpus (Politicians Corpus) have been calculated.
- Relative frequencies of words in the General Language Corpus (American National Corpus) have been calculated.
- Weirdness [4] is defined as the ratio of the relative frequency of the word in the special language corpus, to the relative frequency of that word in the general language corpus.

$$\text{weirdness coefficient} = \frac{f_s / N_s}{f_g / N_g}$$

Figure 3.12: Weirdnes Coefficient as defined by Ahmad et al [4] where fs = frequency of the term in the special language corpus Ns = number of terms/words in the special language corpus; fg = frequency of the terms in the general language corpus Ng = number of terms in the general language corpus

- Z-score of Weirdness has been calculated.
- Z-score of Relative Frequency in SL Corpus has been calculated
- If Weirdness-Zscore > n and Relative-Frequency>n, then Candidate Tern[n] has been defined.Calculations have been done for Candidate Terms[0],[1] and [2]

**Pragmatic Analysis** Pragmatic Layer deals with the ability to extract the meaning or the context of the representation.

**Lexicons** General Inquirer Dictionary (GI) doesn't have all the necessary positive and negative words that are needed and end up not classifying a lot of words. Negative words like ""viral", "heckle", "fluff" were not found in General Inquirer.So two other lexicons were also used for sentiment extraction and analysis.

**VADER,[77](Valence Aware Dictionary and sEntiment Reasoner)** one of the most popular lexicons in academia, was developed by researchers at the Georgia Institute of Technology, especially for social media text analytics and it includes 7520 entries with positive and negative polarities for words and emoticons. Another advantage of VADER is that it includes all the variations of the word in the lexicon, so separate stemming and lemmatization techniques are not required for VADER.

**MPQA LEXICON (UNI-PITTSBURGH)[78]**: A lexicon "gathered from several sources" [78] containing 8222 terms, the MPQA Subjectivity Lexicon has been provided by the University of Pittsburgh. The words are listed along with the "polarity (positive, negative, neutral )", "intensity" (strong/weak) and their corresponding POS tags.

**Method Adopted:** The terms are looked up in the three lexicons one after the other for sentiments. If the term holds a sentiment in one lexicon, it is not looked up in the next lexicon and the control moves on to the next term and so on. This is done to ensure consistency.

## 3.7 Statistical Tools and Hypothesis Testing

### 3.7.1 Non-Parametric Methods vs Parametric Methods

Before deciding the statistical tests that would be used for analysis, the underlying distributions of the variables in question, need to be analyzed. There are two types of statistical tests: **Parametric Test** and **Non-Parametric Test**.

The parametric tests assume that the distribution of the underlying population from which the samples have been taken is a **Normal distribution**[79]. On the other hand, **non-parametric tests do not place any constraint on the underlying distribution**.

**Checking for Normality:Quantile-Quantile (QQ Plot)**

Quantile-Quantile Plot(or QQ plot) is a visualization technique used to check whether the distribution is normally distributed or not.This has been done prior to all tests to check whether the distributions are normal or not and it has been observed that **the emotion distributions do not follow a normal distribution.**

**Statistical Family selected for the Dissertation: Family of Non-Parametric**

tests

### 3.7.2 Spearman's Rank-Order Correlation:

**Spearman's Rank-Order Correlation** is a non-parametric method for measuring the strength of correlation and linear association between two variables. It is used to evaluate the "strength" and "direction" [80] of the linear association, where strength signifies the degree of association and direction signifies a direct/positive or an inverse/negative association. Spearman correlation can also be interpreted as the Pearson Correlation evaluated on the ranks of the variables instead of the values themselves.

The strength of the correlation is denoted by the value of correlation coefficient, **rho** which varies from -1 to 1. A correlation magnitude closer to 1 denotes a strong linear association whereas values closer to 0 signify the absence or a feeble presence of association. The direction is denoted by the sign, where a positive sign implies that the variables are positively correlated: the increase of one would is observed with the increase of the other variable whereas a negative sign signifies the existence of an inverse relation: where the increase of one variable is associated with a decrease in the other variable.

It must be noted that **"Correlation implies association, but not causation"**[81]. So a high value correlation between an action unit and an emotion evidence would suggest the presence of association, but it should **not be said** that the presence of the Action Unit "**causes**" the presence of a particular emotion. Instead, the appropriate inference would be that a particular emotion is associated with the presence of one or more action units.

Following Dancey and Reidy's book on psychology and statistics,[82] the correlation coefficients in the dissertation have been segregated into five tiers or levels: **Zero (0), Weak (0 to 0.39), Moderate (0.4 to 0.69), Strong (0.7 to 0.99) and Perfect(1)**.

Spearman's correlation has been used in this disseration to check for association between the outputs of the different FERs, SERs, and also to check for association among emotion evidences and action units,audio correlates, etc. Hypotheses and further details about spearman correlation have been discussed in the following chapters.

### 3.7.3 Kruskal-Wallis Test:

Kruskal-Wallis H-Test or the Kruskal-Wallis Rank Sum Test is a non-parametric test (counterpart of the one way Analysis Of Variance (or the ANOVA test) ), used to determine whether there are statistically significant differences between two or more groups. Alternatively, it is interpreted as the test to check whether the "samples come from the same underlying distribution "[83]. It is considered an extension of the Mann-Whitney U test(which compares two independent groups)

The **null hypothesis** $H_0$ in the Kruskal-Wallis H-test is that the two distributions are similar or the samples of the two sets of observations originate from the same distribution.

The **alternate hypothesis** $H_1$ is that the two distributions are statistically different.

The null hypothesis is tested with the help of a metric called the p-value which is the probability of an event occurring given that the null hypothesis is true. Now, if the p-value is less than a threshold $\alpha$ (0.05) the null hypothesis is rejected in favour of the alternate hypothesis.

### 3.7.4 Cohen's Kappa:

Cohen's Kappa [84], used extensively in "psychological diagnoses" is a metric used to evaluate the "over-all agreement between two raters" during classification of items into categories. It was introduced to adjust the "observed proportional agreement" by taking into account the agreement that is expected "by chance "[5].Now, Cohen's Kappa is defined as the difference in observed agreement and the agreement expected by chance, divided by the maximum difference possible between agreement and agreement-by-chance.

$$\kappa = \frac{p - p_e}{1 - p_e}$$

Figure 3.13: Cohen's Kappa[5] where pe = proportion agreement by chance, p = proportion where there is an agreement between the raters/annotators

# 4 Case Studies and Results

### Hypothesis Testing

On a very high level, the null hypothesis adopted in this dissertation is that the emotion recognition systems are identical in their performance and do not show any training, testing, or architectural bias.

A sub hypothesis has also been defined, which assumes that the Emotion Recognition systems' performance is independent of the subjects' ethnicity, age, and gender.

Several statistical tests(mentioned in the Methods section) would be used to check whether the null hypothesis holds or the alternative hypothesis gets precedence.

**As mentioned in the Design and Methods section, the performances would be checked on "labelled" databases and then our dataset will be introduced.**

Labelled Dataset: RAVDESS:

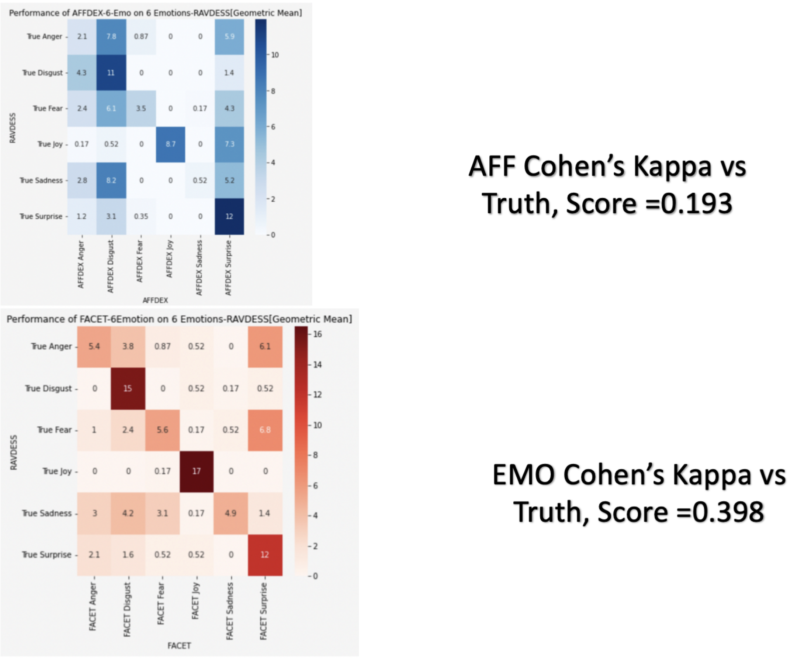
## 4.1 Facial Expressions

The facial expression outputs from both FACET Emotient and AFFECTIVA are sampled at 33 millisecond intervals, whereas the RAVDESS emotion videos are about 3-5 seconds (3000 to 5000 millisecond long). So, a proper framework needs to be adopted to correlate the iMotions output with the RAVDESS labels for association matrices.

Now, three aggregation methods have been adopted here. Outputs from both FACET and AFFDEX have been aggregated as per the below metrics and then checked with the RAVDESS data-set labels.

- **Arithmetic Mean** As suggested in the paper,[24], arithmetic means or averages of the output values for each of the emotions have been calculated and the emotion with the maximum arithmetic mean is elected as the iMotions output for that particular RAVDESS video.
- **Geometric Mean**: Since arithmetic mean might not be the best metric for values

involving probabilities, geometric mean values corresponding to each of the emotions have been calculated. The emotion with the maximum geometric mean is selected as the iMotions output for that particular video.

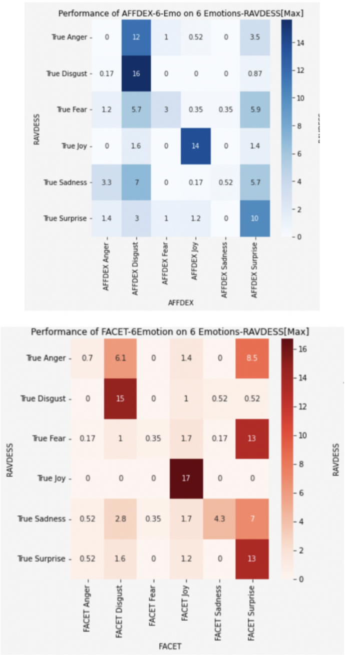


AFF Cohen's Kappa vs Truth, Score =0.193

EMO Cohen's Kappa vs Truth, Score =0.398

Figure 4.1: Cohen Kappa Scores for Affectiva vs Truth and Emotient vs Truth for Geometric mean aggregation

- **Maximum Emotion Value** The maximum values for all the emotion labels in the entire video have been taken and the emotion having the maximum value out of these 7 maximum values is chosen as the iMotions output for the video.



AFF Cohen's Kappa vs
Truth, Score =0.2419

EMO Cohen's Kappa vs
Truth, Score =0.308

Figure 4.2: Cohen Kappa Scores for Affectiva vs Truth and Emotient vs Truth for Max aggregation

**In all the above approaches, it has been noticed that the Cohen-Kappa Score for Emotient is better than that of Affectiva. So, there is a better agreement between Emotient and RAVDESS than between Affectiva and Ravdess.**

**Now there is a caveat to these association matrices because we are solely relying on the annotated labels of the videos. Since the videos are about 3-5 seconds long, there is indeed a possibility of multiple emotions getting captured within that time frame.**

Given that annotators of Ravdess videos have marked a video 'anger/angry' it is expected that the overwhelming majority of frames will clearly show anger expressed non-verbally as the most probable emotion during the entirety of the video.

Now to put this theory to the test, a threshold of 90 percent is selected for the activation level. So, percentages of frames where the emotion intensity would go beyond the 90 percent threshold would be taken into account. Had "anger" been the most probable emotion, all the frames in the videos labelled as "anger" are supposed to have a very high activation well above the threshold.

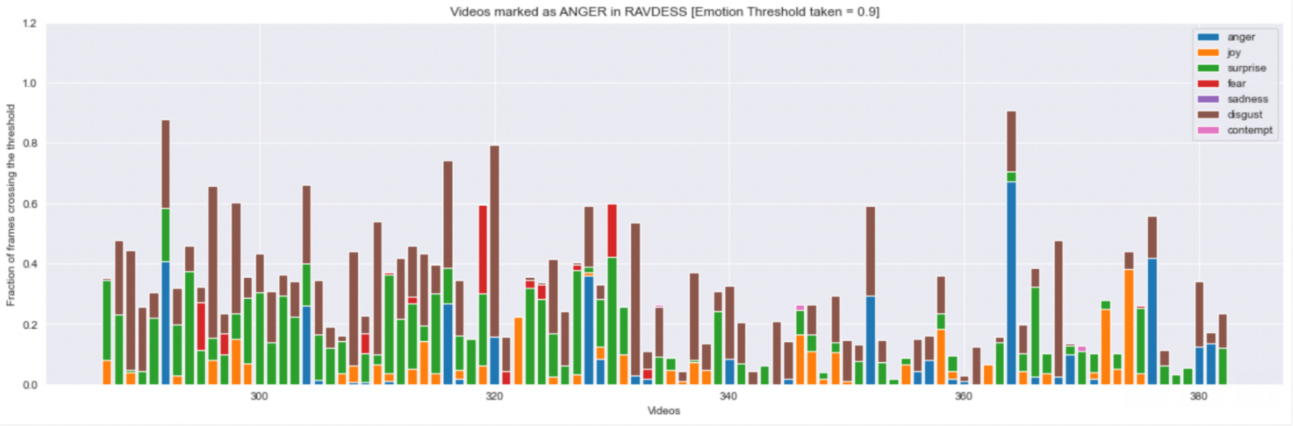


Figure 4.3: Fraction of EMOTIENT frames crossing the 90 percent threshold vs Videos marked as "ANGER" in RAVDESS annotators

But, in a bar chart of all seven emotions displayed with different colours we observe the overwhelming presence of other six emotions along with anger when plotted for videos labelled as "anger" by RAVDESS annotators. For example, if we consider the first video (Video-0) we observe that there is no frame for anger which has managed to cross the threshold but interestingly enough, 26.5 percent frames and 8 percent frames have recorded surprise and joy intensities of more than 90 percent respectively.

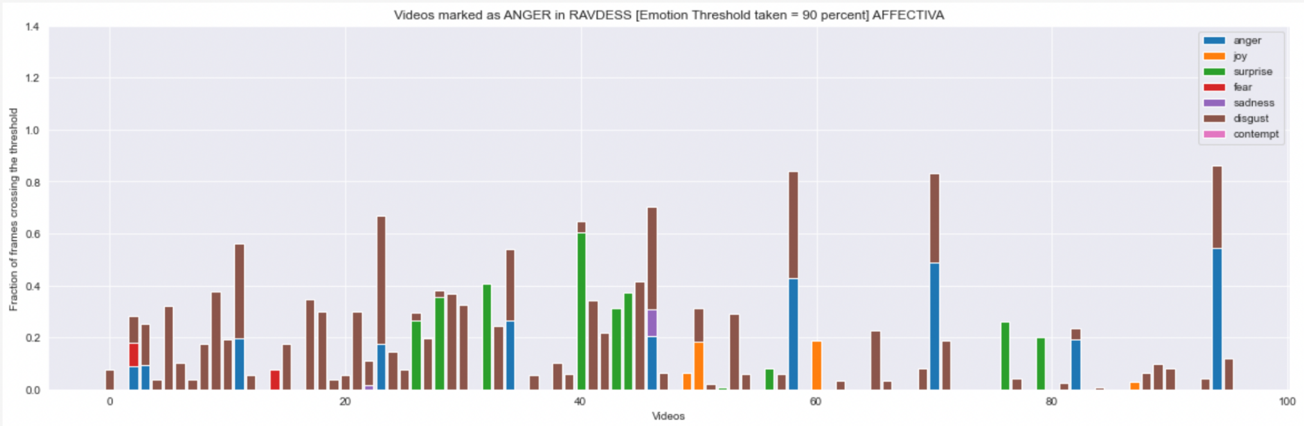


Figure 4.4: Fraction of AFFDEX frames crossing the 90 percent threshold vs Videos marked as "ANGER" in RAVDESS annotators

Similar trend is also observed in AFFDEX Affectiva where anger has been accompanied with a multitude of other emotions. Now, on the other hand, if we consider a video labelled 'happy/joyous' by the annotators and the results obtained from Facet for this video, we observe that indeed 'joy' is the most probable emotion.

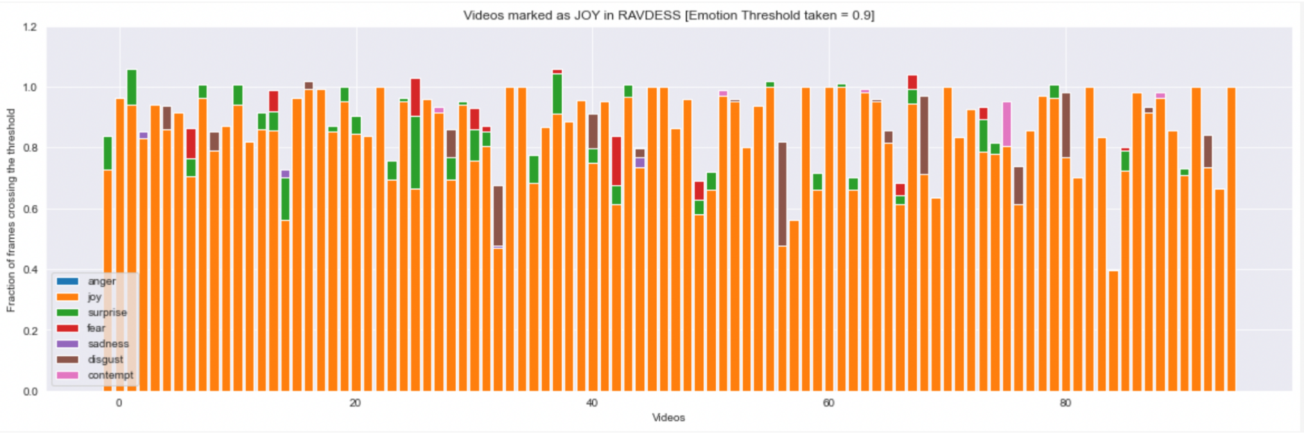


Figure 4.5: Fraction of EMOTIENT frames crossing the 90 percent threshold vs Videos marked as "JOY" in RAVDESS annotators

Figure here **But, if we process the video through Affectiva the evidence of 'joy' is not as convincing as it is for Facet. In Affectiva, we notice high evidences of surprise also cropping up for some videos.**

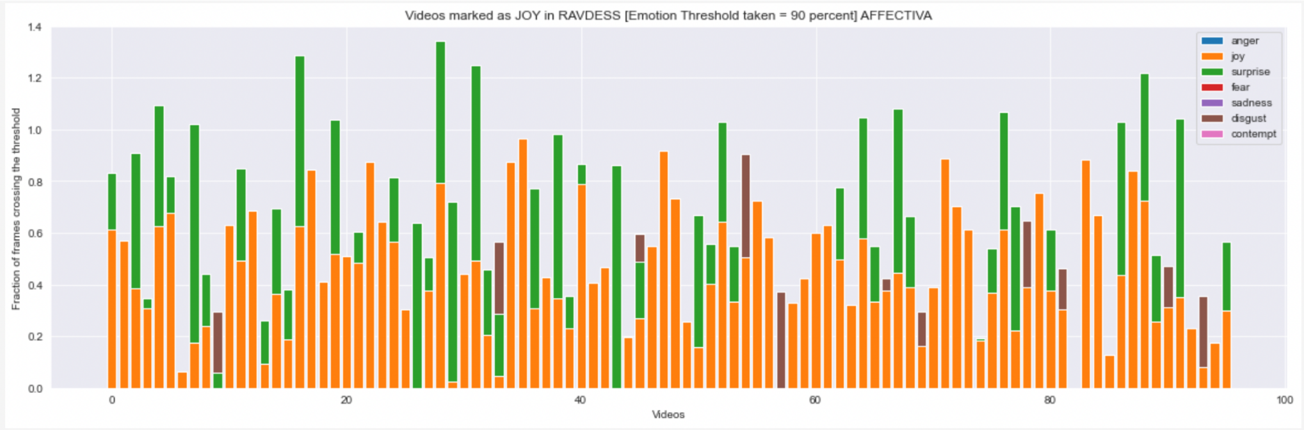


Figure 4.6: Fraction of AFFDEX frames crossing the 90 percent threshold vs Videos marked as "JOY" in RAVDESS annotators

Kruskal-Wallis THe Kruskal-Wallis test checks whether the samples from the two FERs follow the same distribution or not. Here, the null hypothesis H0 states that the outputs of the two FERs for different emotion categories follow the same distribution.

| Emotion | p-value | Remarks(p-val threshold = 0.05) |
|---|---|---|
| anger | 0 | Null Hypothesis Rejected |
| anxiety/fear | <<<0.01 | Null Hypothesis Rejected |
| happiness | <<<0.01 | Null Hypothesis Rejected |
| sadness | <<<0.01 | Null Hypothesis Rejected |
| disgust | 0 | Null Hypothesis Rejected |
| surprise | 0 | Null Hypothesis Rejected |

**Interpretation of the Kruskal-Wallis results:**

Here, it has been observed that the null Kruskal-Wallis hypothesis has been rejected for all the six emotions. This signifies that the "anger_intensity" distribution in FACET-Emotient is differenct from the "anger" distribution reported by AFFDEX-Affectiva. Now, to cross-check the Kruskal-Wallis results, histogram distributions of the emotions reported by the two FERs can be observed.

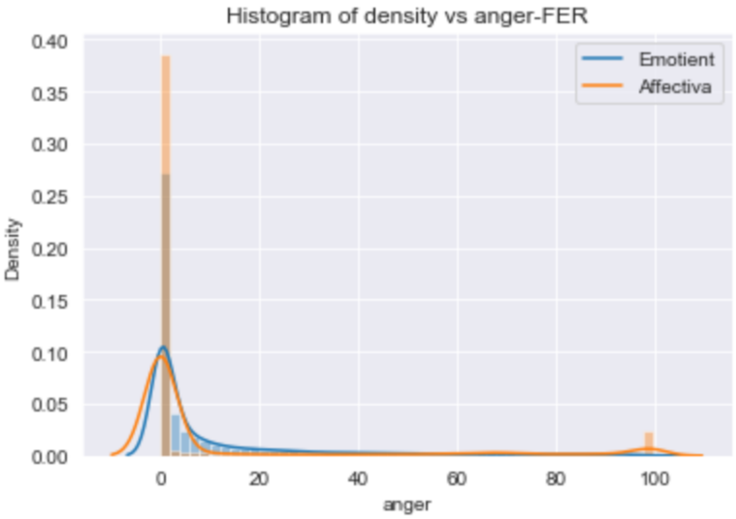


Figure 4.7: Total Anger distribution in Emotient and Affectiva

Now looking at the plot, it seems that the anger distribution in both the systems are almost similar. But the caveat to this approach of plotting histograms is that **most of the values are concentrated at the 0-10 percent bins**. As anger is one of the six RAVDESS emotions being taken in the dataset, there are lots of instances when the "anger" will not be the dominant emotion and hence would have very low values of reported anger. Now, probable emotion has been calculated for each frame by taking the emotion with the maximum intensity for that frame. Once the probable emotions have been classified for both the systems Emotient and Affectiva, a confusion matrix is created for the seven emotions.

Hence, to have a proper and **fair comparison** of the FER histogram plots, data from **10 percent activation to 100 percent activation needs to be taken.** The difference between the two distributions are clearly observed when the histograms are sliced at the 10-100 percent margin.

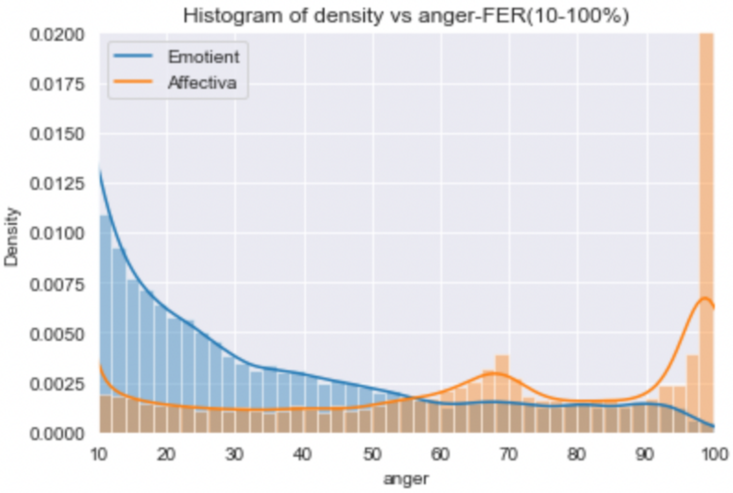


Figure 4.8: Anger distribution in Emotient and Affectiva: sliced at 10 percent

Plotting the histograms in a similar fashion for "joy" and "fear" also highlights the different distributions reported by FACET-Emotient and AFFDEX-Affectiva.

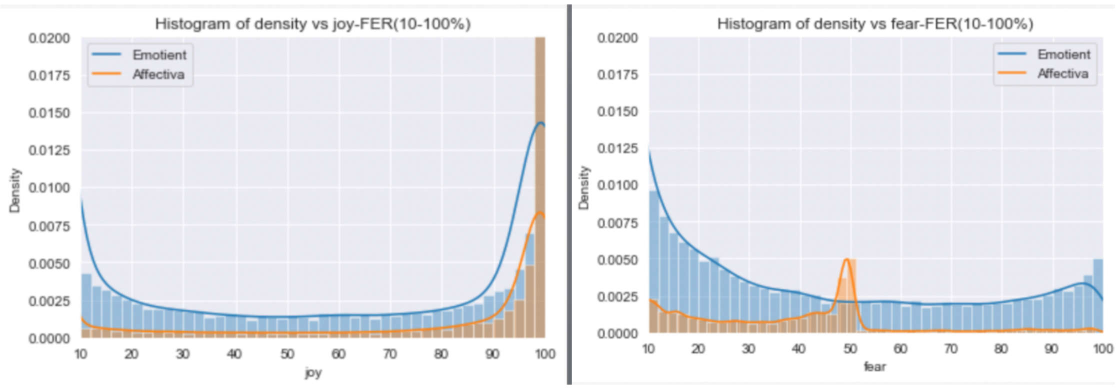


Figure 4.9: Joy and Fear

Correlation between FACET and AFFDEX:

| Emotion | rho | p-value | Correlation Strength |
|---|---|---|---|
| anger | 61.56% | 0 | Moderate |
| anxiety/fear | 38.29% | 0 | Moderate |
| happiness | 42.38% | 0 | Moderate |
| sadness | 39.23% | 0 | Moderate |
| disgust | 53.80% | 0 | Moderate |
| surprise | 40.52% | 0 | Moderate |

**Interpretation of the Spearman Rank Correlation results:**

Spearman rank correlation is the correlation between the FER anger distributions when the corresponding values have been assigned ranks .Now once,the ranks have been assigned, Spearman correlation rho is simply the slope of the linear regression line that can be fitted to the points.For example, the scatter plots of ranks of the FER-anger and FER-happiness for both the systems have been plotted. The slope for the anger-rank trend-line is 0.6157 which corresponds to the 61.56 percent correlation we observed in the correlation table. Similarly the slope for the joy-rank trend line is observed to be 0.4238, which signifies the 42.38 percent correlation observed.

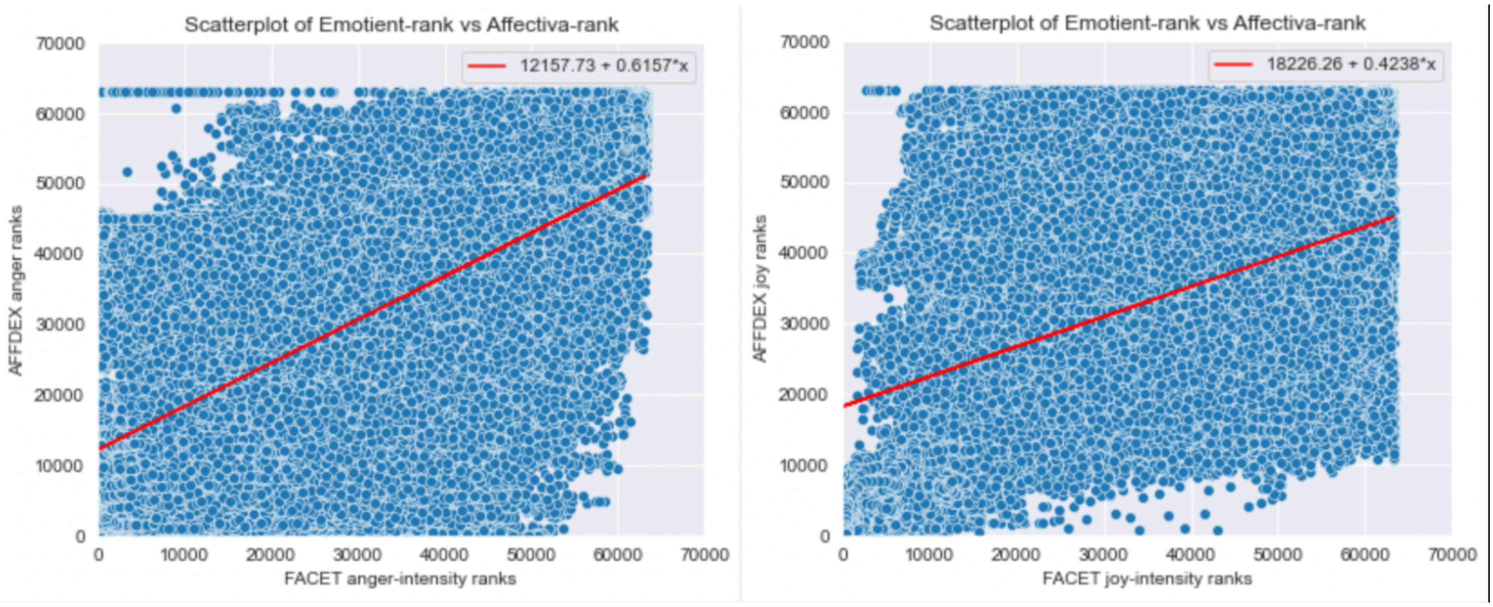


Figure 4.10: correlation-joy-anger

**Cohen-Kappa Score: Affectiva vs Emotient on a frame-by-frame basis**:

Now, RAVDESS is a posed expressions labelled data-set and the data-set that has been used in the dissertation in the later stages is a semi-spontaneous expressions data-set. Now it would not be possible to check any inter-annotation agreement with the ground truth in case of the TCD data. Or in other words, Affectiva vs Ground Truth and Emotient vs Ground Truth cannot be compared in the TCD dataset.

Hence, an Affectiva vs Emotient Inter-annotation comparison needs to be there on a frame by frame basis and that should start from RAVDESS. The Cohen'Kappa score for RAVDESS would be checked with the Conen's Kappa score for the collected dataset.

It has been observed that the systems report a 26.15 percent inter-annotator agreement between themselves which highlights the difference in the two systems.

**Kruskal Wallis Rank Sum Test on the entire dataset:**

As stated earlier, the Kruskal Wallis Test checks whether there are statistically significant differences between the systems, The null hypothesis in Kruskal Wallis H0 test assumes that

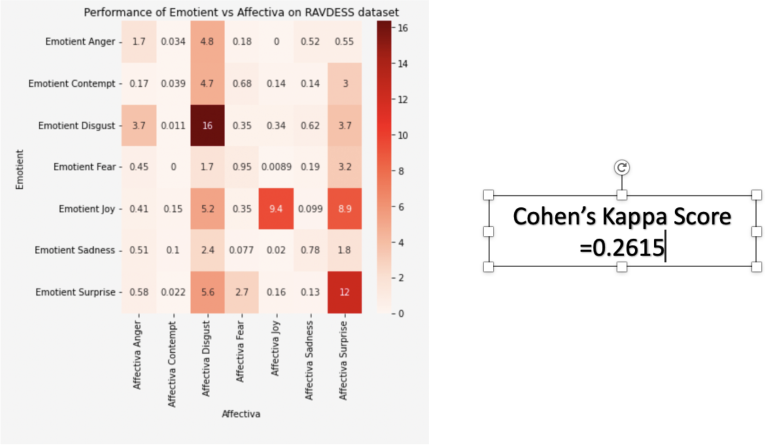


Figure 4.11: Cohen's kappa score for Emotient vs Affectiva on the RAVDESS dataset

the samples in question originate from the same underlying distribution. Or in other words, the two distributions are similar.

The alternative hypothesis would state the presence of a statistically significant difference in the distributions.

**Data-set Used: Entire Dataset**

**Statistical Test Used: Kruskal Wallis Test**

**Emotion Recognition Systems: FACET Emotient and AFFDEX Affectiva**

Here, the Kruskal Wallis test has been evaluated for both systems' emotion distributions. For each of the Ekman-defined seven emotions, the emotion-intensity in Emotient is evaluated against the Affectiva value for the corresponding emotion. If the p-value of the test is found to be less than the threshold of 0.05, the null hypothesis would be rejected in favour of the alternative hypothesis.

Table 4.1: Kruskal-Wallis Rank Sum Test Emotient-emotion intensity and Affectiva-emotion.

| Emotion | Kruskal-Wallis p-value | Null Hypothesis |
|---|---|---|
| Joy | 0 | Rejected |
| Anger | e-131 | Rejected |
| Surprise | 0 | Rejected |
| Fear | 0 | Rejected |
| Contempt | 0 | Rejected |
| Disgust | 0 | Rejected |
| Sadness | 0 | Rejected |

From the results, in the table, it has been observed that the p-values of six of the emotions: joy, surprise, disgust,fear, contempt and sadness , are zeros. For the emotion anger, the result is of the order of e-131 and hence very close to zero. So for all the emotions, the null hypothesis has been rejected and we can infer the presence of **statistically significant difference in the emotion distributions of both the systems.**

Even though the underlying distributions are different, it must be evaluated whether there is an association in the outputs of the two systems. Since both the systems are based on FACS and built on the notion of invariance and universality, the extent of association in the two systems must also be checked.

As discussed earlier that the emotion distribution in the emotion detection systems are not normal, hence the non-parametric Spearman rank order correlation is used instead of the Pearson correlation.

**Data-set Used: Entire Dataset**

**Statistical Test Used: Spearman's Rank Order Correlation**

**Emotion Recognition Systems: FACET Emotient and AFFDEX Affectiva**

Before applying Spearman's rank order correlation on the data, it must be ensured that the number of entries/samples in the two distributions are the same. A **database inner join** of the Emotient and Affectiva output has been performed with the fields: "name" (politician name) and "mediatime" (the field that signifies the frame rate of the video). After the database join, the blank frames and frames where emotion and AU units report zero activation, are removed.

Once the above pre-processing steps are completed, the Spearman correlation test is performed using the **spearmanr** method of Python's scipy.stats package.

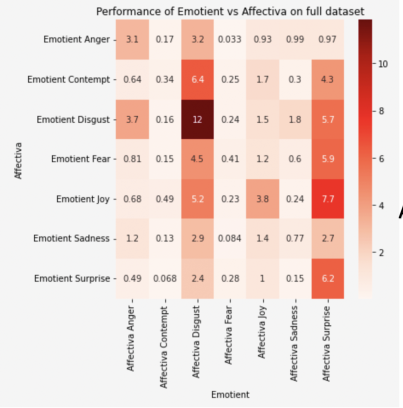
It has been observed that there is a statistically significant moderate correlation ($>= 0.4$ following Dancey's terminology mentioned in Methods ) for the emotion anger4.2 in the two systems. Surprise reports a near moderate correlation with the value of rho as 0.397. Joy, sadness and disgust report weak correlations with values of 0.287, 0.214 and 0.116. Fear has an almost zero correlation with a rho value of 0.004 whereas contempt has been found to show anti-correlation(negative correlation) with -0.124.

Table 4.2: Spearman Rank Order correlation between Emotient-emotion intensity and Affectiva-emotion.

| Emotion | Correlation Strength | p-value |
|---|---|---|
| Joy | 0.287 | 0 |
| Anger | 0.471 | 0 |
| Surprise | 0.397 | 0 |
| Fear | 0.004 | e-06 |
| Contempt | -0.124 | 0 |
| Disgust | 0.116 | 0 |
| Sadness | 0.214 | 0 |
| | | |

### Confusion Matrix and Cohen-Kappa (Inter-annotator agreement )

A 7x7 confusion matrix has been created to evaluate the inter-system agreement. Probable emotions are calculated from the outputs of the systems by taking the emotion with the maximum weightage as the emotion for that particular row. Once the probable emotions have been classified for both the systems a confusion matrix is created for the seven emotions.



**Cohen's Kappa (Inter-System Agreement adjusted for chance) = 0.118**

Figure 4.12: Cohen's kappa score for Emotient vs Affectiva on full data set

Using Python's cohen_kappa_score method of the sklearn.metrics package the Cohen's Kappa score is calculated. It has been observed that there is a 11.8 percent above chance agreement between the two systems Emotient and Affectiva for the classified emotions. This is less that what was observed for the posed data-set RAVDESS which reported a Kappa Score of 26 percent. **The results actually corroborate what has been observed in several studies as mentioned in the literature review, where the annotator agreement drops on introduction of spontaneity**

### Exploring whether there are any differences in how the Action Units are processed

FACET Emotient and AFFDEX affectiva both consider 20 action units (or AUs). In Emotient, the activation of the action units are defined in terms of the terms of AU evidences where evidences are reported for the 20 action units like, AU1,AU2, etc. On the other hand, Affectiva uses a different terminology for the action units with names corresponding to the actions like brow furrow(correspondig to AU4), cheek raiser (AU6, etc).

For both the FERs Emotient and Affectiva, Spearman rank order correlation has been performed on each of the seven emotions with the 20 action units and the results are displayed in the form of a 7x20 heatmap.

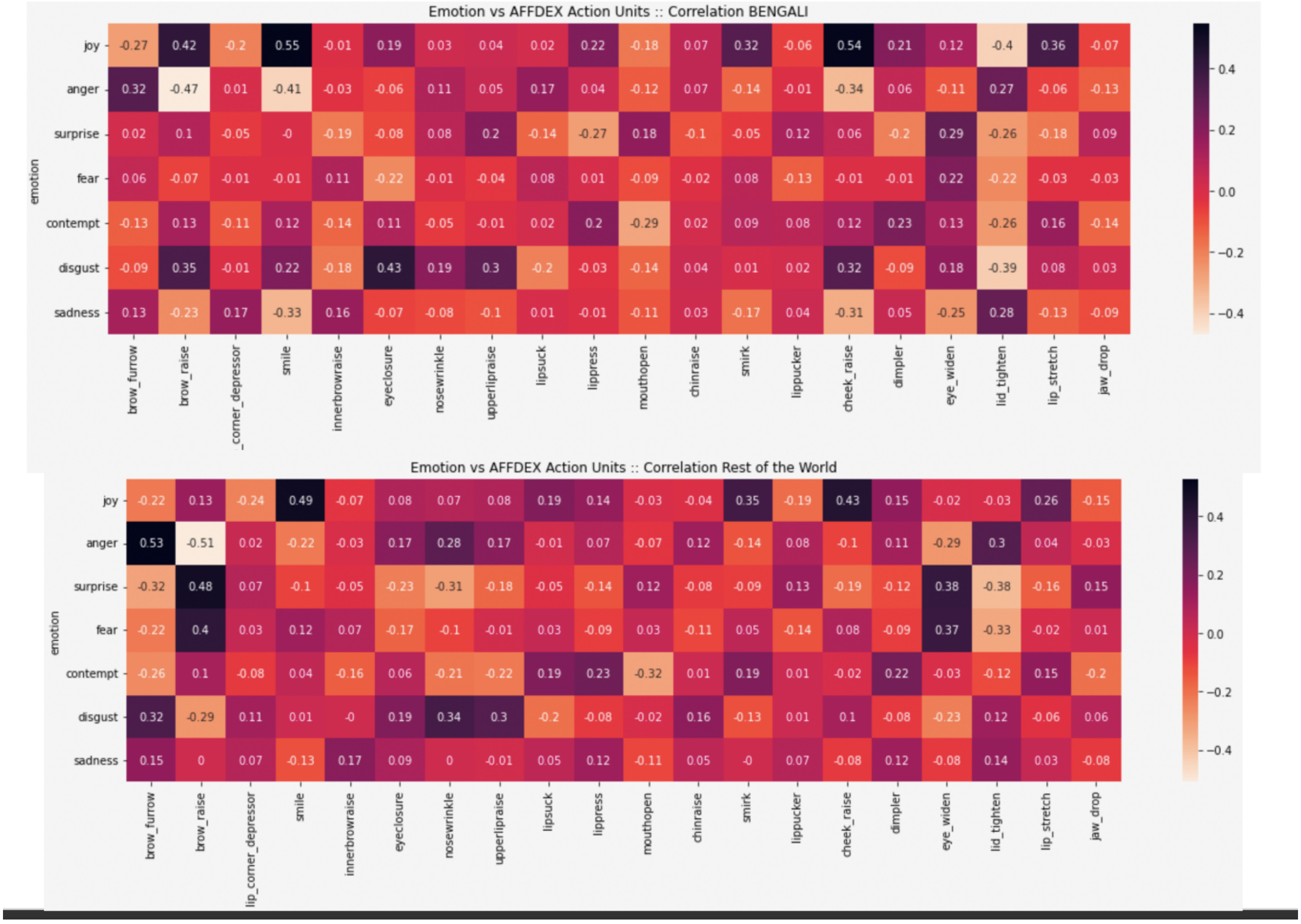


Figure 4.13: Dominant Affectiva AUs on Bengali and Rest of the World Dataset

**The darker the colour, higher is the association of the emotion with that particular action unit. Please note that the heatmap results are also summarised in the table A cut-off of 0.4 has been selected for moderate correlation and the results are compared with existing literature on the subject.**

.

Now, evaluating heat-maps for 7x20 matrices would be extremely difficult hence a cutoff threshold of 0.4 correlation has been set and the dominant Action units for both the systems

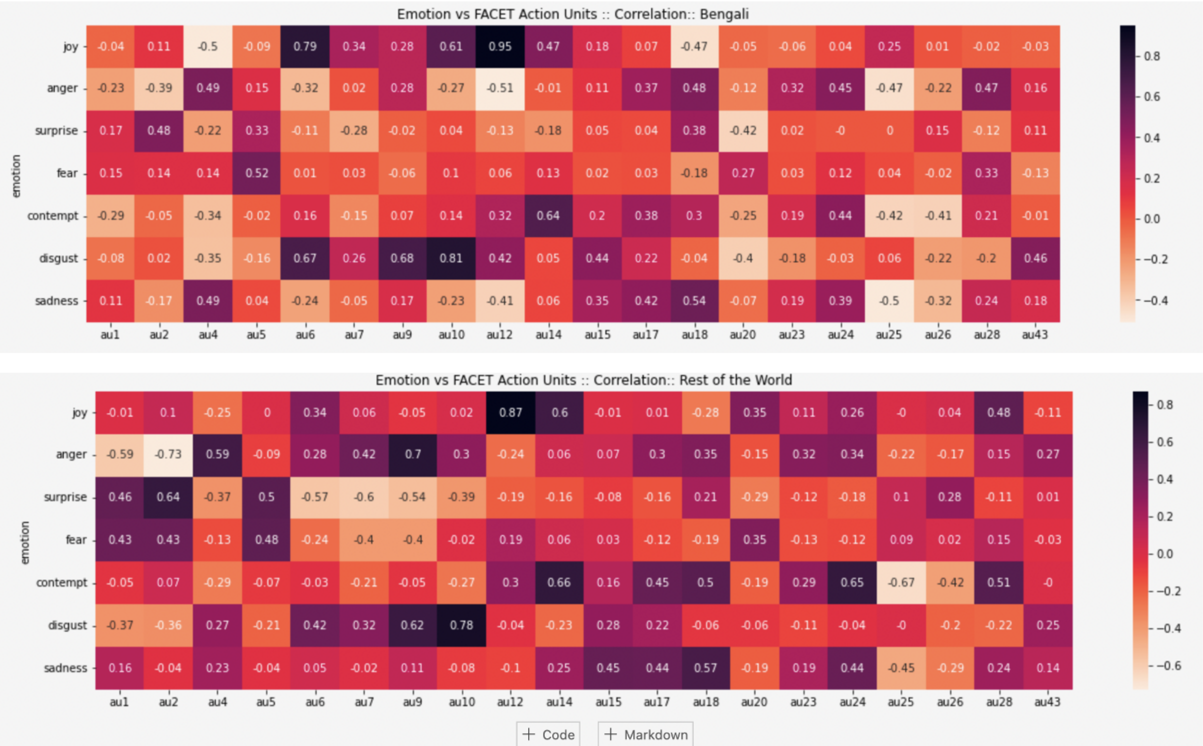


Figure 4.14: Dominant Emotient AUs on Bengali and Rest of the World Dataset

categorized in the table and compared to two prominent studies by Du, Tao, Martinex[11] and Lucey, Kahnade et al[12].

It has been observed that there are certain Action Units which are present irrespective of ethnicity and system. For example, both Affectiva and Emotient report the presence of AU12 (lip corner puller) in Bengali and rest of world dataset for happiness. Similarly, AU4 or brow furrow is present for both the system (though affectiva reported a weak correlation for Bengali and moderate for the rest of the world). This has been supported by Du Tao Martinez but not by Lucey et al.

| **Emotion** | **Bengali Emotient Threshold = 0.4** | **Rest Of the World Emotient Threshold = 0.4** | **Du Tao Martinez \cite{dutao}** | **Cohn Kanade \cite{ck}** |
|---|---|---|---|---|
| Joy | 12,6,10,14 | 12,14,28 | 12,25,6 | 12 |
| Anger | 4,18,28,24 | 9,4,7 | 4,7,10,17 | 23,24 |
| Surprise | 2 | 2,5,1 | 1,2,25,26,5 | 1,2,5 |
| Fear | 5 | 5,2,1 | 1,4,20,25,5 | 1,2,4,5 |
| Contempt | 14,24 | 14,24,28,18,17 | N/A | N/A |
| Disgust | 10,9,6,43,15,12 | 10,9,6 | 9,10,17,4,24 | 9,10 |
| Sadness | 18,15,17,24 | 18,15,17,24 | 4,17,1,6,11 | 1,4,15,11,6,5 |

The tables also corroborate existing literature that the action units are also susceptible to cross-culture variation[1] as AU28 is found to be dominant for Bengalis but not for the rest of the world. Another observation is that Affectiva did not have significant/dominant Action Units for Sadness whereas Emotient reported four such action units.

| **Emotion** | **Bengali Affectiva Threshold = 0.4** | **Rest Of the World Affectiva Threshold = 0.4** | **Du Tao Martinez \cite{dutao}** | **Cohn Kanade \cite{ck}** |
|---|---|---|---|---|
| Joy | smile,(12)<br>cheek_raise(6)<br>brow_raise(1,2) | smile(12),<br>cheek_raise(6) | 12,25,6 | 12 |
| Anger | brow furrow*<br>(4) | brow furrow<br>(4) | 4,7,10,17 | 23,24 |
| Surprise | eye_widen*(5) | brow raise(1,2),<br>eye widen*(5) | 1,2,25,26,5 | 1,2,5 |
| Fear | eye_widen*(5) | brow raise(1,2),<br>eye widen*(5) | 1,4,20,25,5 | 1,2,4,5 |
| Contempt | dimpler*(14) | lippress*(24), dimpler*(14) | N/A | N/A |
| Disgust | brow_furrow*(4),<br>nose_wrinkle*(9),<br>upper lip raise*(10) | nose_wrinkle(4)<br>upper lip raise(9)<br>brow furrow(10) | 9,10,17,4,24 | 9,10 |
| Sadness | No AU >0.2 | No AU >0.2 | 4,17,1,6,11 | 1,4,15,11,6,5 |

### Ethnicity and Gender

Table 4.3: Comparing Male and Female Distribution for Emotient and Affectiva

| Emotion | Male vs Female -Emotient p-value (Kruskal Wallis) | H0-Emotient | Male vs Female-Affectiva p-value (Kruskal Wallis) | H0-Affectiva |
|---|---|---|---|---|
| Joy | 0 | Rejected | 0 | Rejected |
| Anger | 0 | Rejected | 0 | Rejected |
| Surprise | 0 | Rejected | 0 | Rejected |
| Fear | 0 | Rejected | 0 | Rejected |
| Contempt | 0 | Rejected | 0 | Rejected |
| Disgust | 0 | Rejected | 0 | Rejected |
| Sadness | e-22 | Rejected | 0 | Rejected |

Here, it has been observed that there are "statistically significant differences" between the emotion distributions exhibited by male and female. Or in other-words Male Emotient distribution and Female Emotient Distribution are statistically different and the same goes for Affectiva.

Table 4.4: Comparing Bengali and Rest-Of-World Distribution for Emotient and Affectiva

| Emotion | Bengali vs Rest of the World -Emotient p-value (Kruskal Wallis) | H0-Emotient | Bengali vs Rest of the World -Affectiva p-value (Kruskal Wallis) | H0-Affectiva |
|---|---|---|---|---|
| Joy | 0 | Rejected | e-151 | Rejected |
| Anger | 0 | Rejected | 0 | Rejected |
| Surprise | e-308 | Rejected | e-132 | Rejected |
| Fear | 0 | Rejected | 0 | Rejected |
| Contempt | 0 | Rejected | 0 | Rejected |
| Disgust | 0 | Rejected | 0 | Rejected |
| Sadness | 0 | Rejected | 0 | Rejected |

Here, it has been observed that there are "statistically significant differences" between the emotion distributions exhibited by Bengali and the rest of the world database. Or in other-words Bengali Emotient distribution and Rest-Of-World Emotient Distribution are statistically different and the same goes for Affectiva.

## 4.2 Audio

**Speech Emotion Recognition:**

Before testing the performance of the two SERs on the TCD dataset, it is imperative that we check the performance of the systems on a common dataset they were trained on. Now, both the systems openSMILE and Vokaturi have been trained on the Berlin EMO-DB dataset.

openSMILE can detect seven emotions whereas Vokaturi provides a weighted vector of 5 emotions. To maintain parity, audios corresponding to the 5 emotions (common in both the systems) are taken from the EMO-DB data-set. Now out of the 535 labelled audio samples in the EMO-DB dataset, 408 audio samples have been selected.

| Emotion | Berlin-EmoDB Dataset | Samples in 5-emotion Emo-DB | OpenSMILE | Vokaturi |
|---|---|---|---|---|
| anger | Ärger | 127 | 143 | 103 |
| anxiety/fear | Angst | 69 | 49 | 77 |
| happiness | Freude | 71 | 74 | 99 |
| sadness | Trauer | 62 | 64 | 62 |
| neutral | neutral | 79 | 54 | 67 |
| boredom | Langeweile | 0 | 20 | 0 |
| disgust | Ekel | 0 | 4 | 0 |

Here,we observe that there were 24 instances where OpenSMILE picked "boredom" and "disgust" over the other emotions ( even though there were no instances of "boredom" and "disgust" in our labeled 5-emotion dataset. Then, both Vokaturi and openSMILE have been run on these 408 samples and the following 5x5 association matrices are generated.

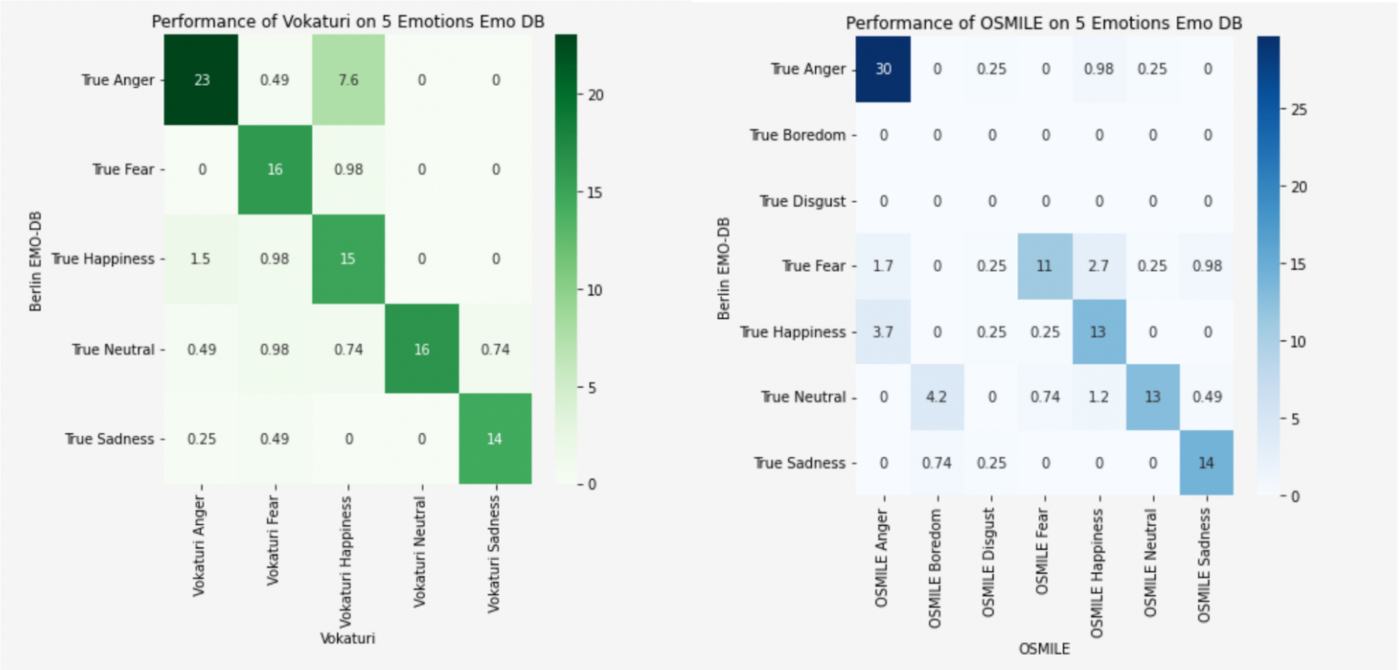


Figure 4.15: OpenSMILE and Vokaturi Outputs for 5-Emotion EMODB

Since,we have taken the audios corresponding to 5 Emotions, the True Boredom percentage and True Disgust percentage in the 5-Emotions EMO-DB dataset are nil. But, OpenSMILE categorized 0.25 percent Fear,0.25 percent happiness, 0.25 percent sadness and 0.25 percent anger samples as "disgust samples". Similarly, 4.2 percent true-neutral frames and 0.74 percent true-sadness frames were categorized as "boredom" by OpenSMILE.

Correlation between OpenSMile and Vokaturi on EMO-DB:

Table 4.5: Correlation between OpenSMile and Vokaturi on EMO-DB:

| Emotion | rho | p-value | Correlation Strength |
|---|---|---|---|
| anger | 79.7% | <<<0.01 | High |
| anxiety/fear | 64.6% | <<<0.01 | Moderate |
| happiness | 66.67% | <<<0.01 | High |
| sadness | 71.19% | <<<0.01 | High |
| neutral | 62.8% | <<<0.01 | Moderate |

THe Kruskal-Wallis test checks whether the samples from the two SERs follow the same distribution or not. Here, the null hypothesis H0 states that the outputs of the two SERs for different emotion categories follow the same distribution.

| Emotion | p-value | Remarks(p-val threshold = 0.05) |
|---|---|---|
| anger | 0.0007 | Null Hypothesis Rejected |
| anxiety/fear | <<<0.01 | Null Hypothesis Rejected |
| happiness | 0.044 | Null Hypothesis Rejected |
| sadness | <<<0.01 | Null Hypothesis Rejected |
| neutral | <<<0.01 | Null Hypothesis Rejected |

### Performance on SAVEE dataset

The performance on the SAVEE Dataset is given as follows:

| **Emotion Labels** | **Samples in 5-Emotion SAVEE** | **Vokaturi default output (5 classes)** | **OpenSMILE default output (7 classes)** | **OpenSMILE 5-class output** |
|---|---|---|---|---|
| anger | 60 | 49 | 0 | 0 |
| anxiety/fear | 60 | 59 | 0 | 0 |
| happiness | 60 | 66 | 2 | 13 |
| sadness | 60 | 52 | 46 | 123 |
| neutral | 120 | 134 | 0 | 224 |
| boredom | 0(absent in SAVEE) | 0 | 276 | 0 |
| disgust | 0(not included in 5 Emotion SAVEE) | 0 | 36 | 0 |
| surprise | 0(n.inc in 5-Emo SAVEE) | 0 | 0 | 0 |

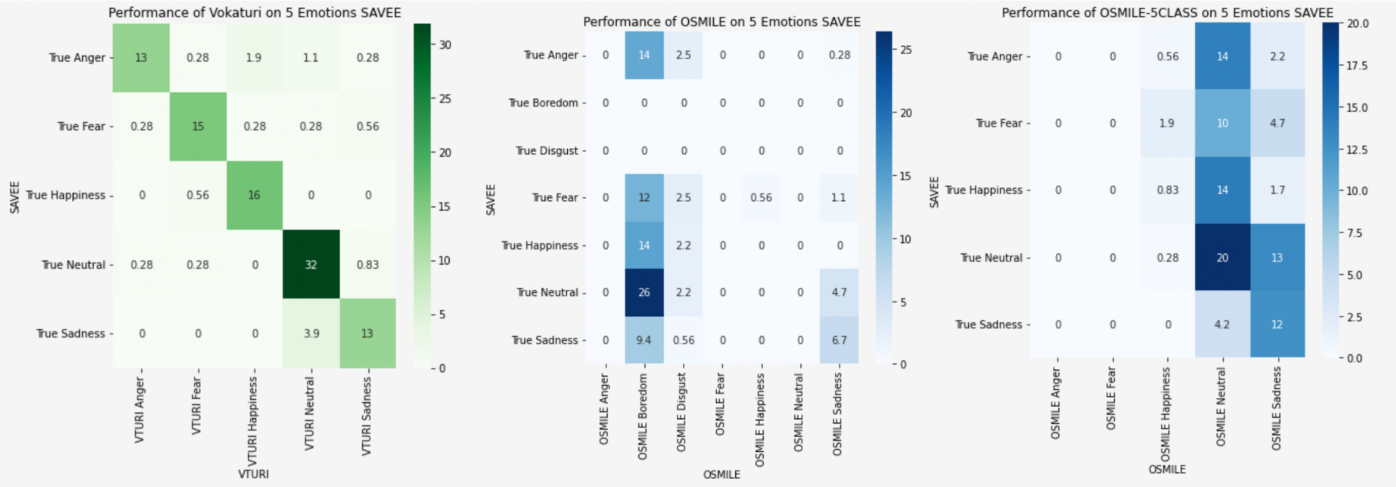


Figure 4.16: OpenSMILE and Vokaturi Outputs for 5-Emotion SAVEE

It is worth noting that even though there are no instances of "Boredom" and "Disgust" in the "5-Emotions-SAVEE" data-set. OpenSMILE reported 75.4 percent instances of Boredom. Several SAVEE samples labelled "Anger","Fear","Happiness","Neutral" and "Sadness"-got classified as "boredom" by OpenSMILE. Now, we know that openSMILE gives us the probabilities for each of the seven emotion classes and the classification output is the emotion with the highest probability.Since, we see that there more than 75 percent of the samples are getting classified as boredom, we can select 5 emotions (anger,happiness, fear,sadness and neutral) and take the maximum value as the classification output. So, from now on, we would consider two OpenSMILE outputs, the default output ( from 7 emotions) and the 5-class output. Following earlier approach, association matrix has been plotted for Truth vs OpenSMILE output(5-classes). Here, we observe that in absence of Boredom, 62.2 percent samples have been classified as neutral (neutral percentage was zero when the default OpenSMILE output was considered).There is also an uptick in the number of sadness outputs with 33.6 percent samples now marked as "sadness" compared to 12.5 percent in the default ones.

**But even when we shift to a 5 class framework, we observe that the OSMILE output is still very different from the ground truth with 0 percent anger outputs, 0 percent fear and less than 4 percent happiness outputs.**

| Emotion | rho | p-value | Correlation Strength |
|---|---|---|---|
| anger | 3.6% | 0.49 | No Correlation |
| anxiety/fear | 1.29% | 0.806 | No Correlation |
| happiness | 44.36% | <<<0.01 | Moderate |
| sadness | 47.3% | <<<0.01 | Moderate |
| neutral | 5.6% | 0.283 | No Correlation |

For the SAVEE dataset, the null hypothesis for spearman rank correlation could not be rejected for the emotions anger, fear and neutral. Statistically significant correlation is only observed for happiness and neutral.

### Data-Set Used: TCD Dataset(205 Politicians) Number of 2000 ms samples: 21709

**Kruskal-Wallis sum rank test** As discussed before, since the underlying distributions are not normal, hence non-parametric tests have been used. Like FERs, the emotion distributions produced by the two SERs are also evaluated with the Kruskal Wallis Test to check whether there are statistically significant differences between the two.

Table 4.6: Kruskal-Wallis Rank Sum Test between OpenSmile and Vokaturi.

| Emotion | Kruskal-Wallis p-value | Null Hypothesis |
|---|---|---|
| Happiness | e-89 | Rejected |
| Anger | 0 | Rejected |
| Fear | 0 | Rejected |
| Sadness | 0 | Rejected |
| Neutral | e-131 | Rejected |

Here,it has been observed that the distributions obtained from both the systems for all the five common emotions: happiness, anger, fear, sadness and neutral are statistically different. So the null hypothesis has been rejected in favour of the alternative hypothesis.

**Spearman rank order correlation:** Like its facial action counterpart, Spearman rank order correlation has also been used to evaluate the association between the two speech emotion recognition systems OpenSmile and Vokaturi.

Table 4.7: Spearman Rank Order correlation between the emotion probabilities of the two systems OpenSmile and Vokaturi

| Emotion | Spearman Correlation Strength | p-value | Null Hypothesis |
|---|---|---|---|
| happiness | 0.24 | e-103 | Rejected |
| anger | 0.24 | e-70 | Rejected |
| fear | 0.06 | e-39 | Rejected |
| sadness | 0.11 | e-13 | Rejected |
| neutral | 0.09 | e-34 | Rejected |

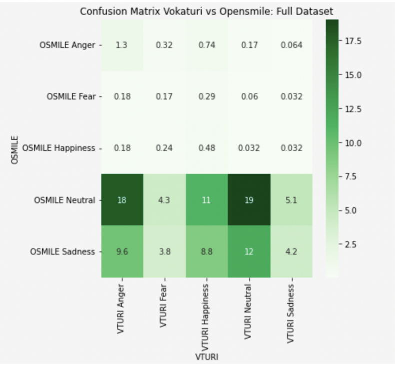
It has been observed that "happiness" values reported by both the systems have a 24 percent positive association between them. Since the strength is less than 40, the association can be classified as **weak** .Weak Associations have also been observed in the other emotions where anger reports 24 percent correlation followed by 11 percent in sadness, 9 percent in neutral and 6 percent in fear.

**Cohen-Kappa and the Confusion Matrix:**

| **Cohen's Kappa in EMODB** | **Cohen' s Kappa in SAVEE** | **Cohen's Kappa in TCD Entire Data-set** |
|---|---|---|
| 0.668 | 0.0525 | 0.034 |

Similar to the drop observed in FERs, we can also notice a huge dip in the inter-annotator agreement for the TCD data-set.

Now, both the systems were trained on EMODB, and hence **show a high agreement 66.8 percent** when tested on that database. On the other hand, SAVEE has only been



**Cohen's Kappa (Inter-System Agreement adjusted for chance) = 0.034**

Figure 4.17: Cohen's kappa score for Vokaturi vs OpenSmile on full data set

used to train Vokaturi but not OpenSmile which results in the drop of agreement on SAVEE. The Cohen's Kappa Score for the SAVEE data-set is 0.0525 or a 5.25 percent above chance inter-annotator agreement. The agreement **declines even further** when evaluated on the TCD dataset, which **both the systems have never seen before**.

### Looking at the cross-cultural impact on the two Speech Emotion recognition systems

**Ethnicity** Using Quantile-Quantile plots, the distributions were checked for "normality" and it was found out that the distributions are not normally distributed.Hence,like before, Kruskal-Wallis Rank Sum tests were performed on the distributions. For both OpenSmile and Vokaturi, the **Bengali** and **Rest of the World** distributions were compared to evaluate the presence of any difference between them.

Here the null hypothesis for OpenSmile or H0-OpenSmile assumes that there are no statistically significant difference between the Bengali and the rest-of-the-world emotion distributions. The tests have been performed for the seven opensmile emotions and the null hypothesis was rejected in favour of the alternative hypothesis in all cases except sadness in Vokaturi and fear in OpenSmile.

| **Emotion** | **Bengali vs Rest-Of-World p-value(Kruskal)** | **H0-Vokaturi** | **Bengali vs Rest-Of-World p-value(Kruskal)** | **H0-OpenSmile** |
|---|---|---|---|---|
| Happiness | e-5 | Rejected | e-21 | Rejected |
| Anger | e-87 | Rejected | e-7 | Rejected |
| Fear | e-13 | Rejected | 0.08 | **Not Rejected** |
| Sadness | 0.08 | **Not Rejected** | e-17 | Rejected |
| Neutral | e-7 | Rejected | e-133 | Rejected |
| Boredom | - | - | 0.0001 | Rejected |
| Disgust | - | - | e-12 | Rejected |

### Gender

For both OpenSmile and Vokaturi, the female and male distributions were compared using Kruskal Wallis Test to evaluate the presence of any difference between them.

Here the null hypothesis for OpenSmile or H0-OpenSmile assumes that there are no statistically significant difference between the male and female emotion distributions. The tests have been performed for the seven opensmile emotions and the null hypothesis was rejected in favour of the alternative hypothesis in all cases.

| **Emotion** | **Female vsMale p-value(Kruskal)** | **H0-Vokaturi** | **Female vs Male p-value(Kruskal)** | **H0-OpenSmile** |
|---|---|---|---|---|
| Happiness | 0 | Rejected | e-191 | Rejected |
| Anger | e-32 | Rejected | e-143 | Rejected |
| Fear | e-183 | Rejected | e-82 | Rejected |
| Sadness | e-08 | Rejected | e-200 | Rejected |
| Neutral | 0 | Rejected | e-24 | Rejected |
| Boredom | - | - | e-72 | Rejected |
| Disgust | - | - | e-20 | Rejected |

Similarly, the null hypothesis for Vokaturi or H0-Vokaturi assumes that there are no statistically significant difference between the male and female emotion distributions. For Vokaturi as well, the null hypothesis was rejected in favour of the alternative hypothesis across all the five Vokaturi emotions.

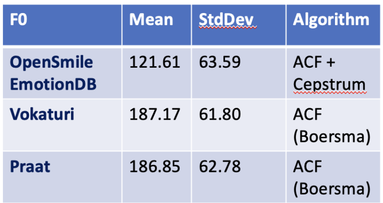
### Comparing the Fundamental Frequency Distributions of OpenSmile and Vokaturi

Table 4.8: F0 comparisons across Praat, Opensmile and Vokaturi(* signifies significance) on the Bengali dataset

| **Comparisons** | **Kruskal H0** | **Spearman** |
|---|---|---|
| Praat F0 vs Vokaturi F0 | Not Rejected (p-val = 0.74) | 0.95* |
| Praat F0 vs OpenSmile F0 | Rejected | 0.691* |
| Vokaturi F0 vs OpenSmile F0 | Rejected | 0.692* |

As discussed in the literature review section, Praat is generally considered the "gold standard" when measuring the fundamental frequency. It can be observed that in Table 4.8 that the null hypothesis has not been rejected (p-value very high) for Praat F0 vs Vokaturi F0 which means that the F0 distribution in Praat and Vokaturi are similar. Alternatively, the null hypothesis has been rejected for Praat vs OpenSmile and Vokaturi vs OpenSmile.

For Spearman Rank Order correlation, a 95 percent near perfect association is observed between Praat and Vokaturi F0s while OpenSmile exhibits about 69 percent association with both Praat and Vokaturi F0s.



| F0 | Mean | StdDev | Algorithm |
|---|---|---|---|
| OpenSmile EmotionDB | 121.61 | 63.59 | ACF + Cepstrum |
| Vokaturi | 187.17 | 61.80 | ACF (Boersma) |
| Praat | 186.85 | 62.78 | ACF (Boersma) |

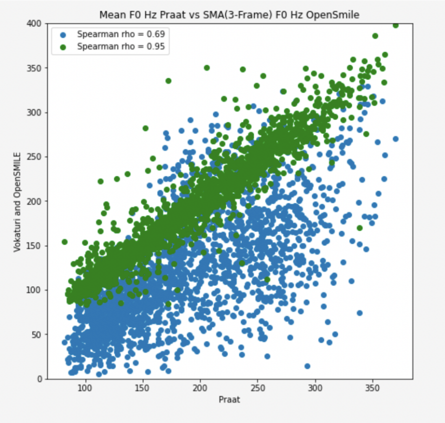


Figure 4.18: Descriptive Statistics for F0 and F0 plots for OpenSmile, Vokaturi and Praat

So, it has been observed that not only the emotion distributions are different, but the differences in the two systems also persist at how they measure and record the fundamental frequency F0. It has been observed that OpenSMile "emodb" config uses a combination of AutoCorrelation Function and Cepstrum methods, whereas both Praat and Vokaturi use Boersma's autocorrelation.

## 4.3 Text

### Automatic Speech Recognition

As discussed in Chapter 2, Word Error Rate (or WER) is one of the most important metrics used to evaluate the accuracy of transcriptions. It has already been stated that because of the poor performance with Google Speech API, a pivot was made to involve the software "Descript" for transcription. Even after using "en-IN" language code(India),the Word Error rate, when evaluated with Python's Jiwer package against the reference translation, comes out to be 80.6 percent with only 101 words being reported out of possible 340 words. On the other hand, Descript reported a WER of 28.8 percent.

| ASR | Word Error Rate | No. of Words in Output (340 words in reference) |
|---|---|---|
| Google Cloud | 80.6% | 101 |
| Descript | 28.8% | 344 |

**The problem is exacerbated when the entire corpus is evaluated and Google Cloud returns only 6471 words compared to 10474 words by Descript.**

Verdict: Hence, Descript was chosen for ASR and the later analysis has been done using output transcribed with Descript.

Following the process discussed in methods, the time-series dataframe with the transcriptions and the sentiments were created using the GI lexicon.

| Total | +ve Words GI | -ve Words GI | +ve Words Vader | -ve Words Vader | +ve Words MPQA | -ve Words MPQA | +ve Words (3-Lexicon System) | -ve Words (3-Lexicon System) |
|---|---|---|---|---|---|---|---|---|
| 10474 | 455 | 213 | 511 | 209 | 701 | 491 | **943** | **591** |

Now, even after the use of WordNet Lemmatizer and Porter Stemmer and trying various permutations and combinations, it was observed that a lot of negative words ended up without proper classification because they were not present in GI (examples:"viral","heckle","fluff",etc ). So, a decision was taken to **update the look-up dictionary in the dissertation**. As mentioned in Chapter 3, two new lexicons VADER and MPQA were added and each individual word was searched in the lexicons one after the other(to ensure consistency and prevent multiple counting ).

It has been observed that the three-lexicon system has increased the number of words we classify as "positive", "negative" and would be an important milestone for the multimodal fusion in later sections and future work.

**Analyzing the Corpus with a Bag of Words Model** Similar to the descriptive statistics taken for the non-verbal modalities, an in-depth analysis of the politicians' transcribed corpus has also been done for verbal modality

As discussed in the methods section, a bag of words model is used to extract the sentiments and analyze the transcriptions on four levels: Lexical, Syntactic, Semantic and Pragmatic Level.

On a Lexical level, the relative frequencies of the words have been found and compared with the relative frequencies in the General Language corpus. Like before Spearman rank order correlation has been used and a **52.72 statistically significant percent correlation** has been observed in the frequencies of the words.

On the syntactic level, the Parts-Of-Speech tagging has been done for the entire Bengali politicians' corpus. Open Class Words and Closed Class words have been segregated for the entire corpus and also for the individual transcripts. Here, the NLTK POS tags NN (Singular Common Noun), NNS (Plural Common Noun) , NNP (Singular Proper Noun), NNPS (Plural Proper Noun ), JJ (Adjective or Numeral), JJS (Superlative Adjective) and

RB(Adverbs) have been marked as Open Class Words. The rest of the words have been classified as clossed classed words.

Using a Chi-Square Goodness of Fit test, it has been checked whether the OCW-CCW distribution in the special language corpus follows the OCW-CCW distribution of the General Language corpus (ANC)

Table 4.9: Comparing OCW and CCW in Special Language Corpus and General Language Corpus

| Corpus | OCW | CCW |
|---|---|---|
| Bengali Politicians Corpus | 37.139 | 62.851 |
| ANC Corpus | 37.84 | 62.16 |

The null-hypothesis could not be rejected here. On the semantic level, the "weirdness" ratio has been defined for the terms in the corpus. As discussed before (in Methods) weirdness is the ratio of the relative frequency of the candidate terms in the special-language corpus(politicians' data) to that of the same terms in the general language corpus. A weirdness value of 1 indicates that the relative frequency of the word in the two corpora is same. Now, the primary focus here is on the terms with weirdness more than 1 (i.e. **the terms that are unique to the corpus** )

Since the corpus is curated from speeches of the politicians, the content is supposed to be rich in political terms. 17 candidate terms have been found in the corpus where the z-score of weirdness and the z-score of relative frequency were greater than 0 (the mean value ) (Please refer to 4.10 for details ). As expected, terms like **"elections", "speeches", "communist", "elections", "literacy","mp"** (member of parliament) have been observed . Since the politicians highly use references to the country and state, words like **"india" and "bengal"** have also found their place in the list of candidate terms. Hesitation words like "uh" and "um" are also found in the list of the candidate terms.

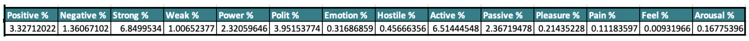
Since, English is the **L2 language** for the speakers, a high percentage of hesitation markers or filler words are observed in the transcriptions leading to their presence in the list of candidate terms.

Table 4.10: Semantic Analysis: Candidate terms in the Special Language Corpus

| Candidate Terms[ weirdness z-score >0 and relative frequency z-score>0] |
|---|
| uh ,um , india , elections , three , two , parliament , railway , trains ,<br>communist, five , speeches , railways , literacy , four , mp , bengal |

### Pragmatic Analysis

Following the pragmatic analysis, the distribution of positive negative, strong, weak, active, passive words as well as emotion words are retrieved. Please note that the below figure represents the pragmatic analysis done only with the General Inquirer Lexicon on the Bengali dataset(Bengali politicians speaking in English)



| Positive % | Negative % | Strong % | Weak % | Power % | Polit % | Emotion % | Hostile % | Active % | Passive % | Pleasure % | Pain % | Feel % | Arousal % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.32712022 | 1.36067102 | 6.8499534 | 1.00652377 | 2.32059646 | 3.95153774 | 0.31686859 | 0.45666356 | 6.51444548 | 2.36719478 | 0.21435228 | 0.11183597 | 0.00931966 | 0.16775396 |

Figure 4.19: Corpus Overview

# 5 Conclusion

This dissertation was aimed at evaluating the performance of the different emotion recognition systems across different modalities (facial expressions, audio and text).The study also analyzed how the emotion recognition systems performed in different contexts: posed vs spontaneous, ethnicity and gender. It was also analyzed how the ERs process the physical correlates or landmarks of the modalities (Action Units for face, F0,intensity, loudness,jitter,etc for audio).

For posed vs spontaneity argument, the Cohen-Kappa score or the inter-annotator agreement between the systems was found to be lower in the semi-spontaneous dataset compared to the posed data-set. This result corroborated the findings of Krumhuber[23],Stockli [21] and Dupre[20].

Similarly how the FERs process the action units were also analyzed. Both Affectiva and Emotient reported dominant action units(AU) very similar to what has been reported in literature. [11] [1][12] Some cross-cultural variations were also observed in the dominant action units, as cited by Saha et al.[1]

Statistically significant differences were observed between the distribution of emotions produced by Emotient, with that of Affectiva[6]. Statistical differences were also observed when the individual system outputs were compared across categories like ethnicity and gender.

It was observed that the Cohen-Kappa agreement rapidly plummeted when the SERs were evaluated on databases that they were not trained on. The agreement of about 67 percent dropped to 5 and 3 percent for the SAVEE and TCD datasets. OpenSMILE also reported a very high percentage of neutral when the 5-emotion framework was applied. When the classifications were done on the 7-emotion framework, very high percentage of boredom samples were reported. Majority of the frames that Vokaturi was classifying as other emotions were all clubbed as neutral/sadness by OpenSMILE.

Now, moving on to the text modality, it has been observed that Desript outperformed Google Cloud as the preferred service for ASR where GCP showed a Word Error Rate of about 80 percent. On the other hand Descript WER rate was around 28 percent. A

3-lexicon system was used to properly extract sentiments from the transcripts and the transcripts were also analyzed across four levels with a bag of words model.

# 6 Future Work

As mentioned before, this dissertation is part of an ongoing study to build a truly robust multi-modal emotion detection sytem where all the three modalities: face, audio and text would be fused. The fourth modality, head pose estimation can also be brought under the fold.

Now, since facial expression changes are observed at 33 millisecond intervals, speech samples are at 2000 milliseconds,**a common reference frame needs to be decided to fuse all the asynchronous outputs.** If the facial expression signals can be aggregated at 2000 milliseconds with a **delay line**, the two modalities face and speech can be compared with each other.Since Descript has also given us the flexibility to slice transcriptions into timestamps of our choice, a truly multi-modal would be possible with this approach.

Suggested by Clodagh[63], a truly unique and robust approach would be to use machine learning driven polarity based sentiment analyzers to get a probability of sentiment or emotion. Then the output probabilities of all the three systems can be given as input to a machine learning to model to get a probability value for the said emotion.

It has also been observed (but not included in the results in this dissertation )that the correlates in one modality also have a weak to moderate correlation with those of another modality for emotions like joy and anger. Bearing this in mind, the correlates can be used as features to a machine learning model for classification.

# Bibliography

[1] C. Saha, W. Ahmed, S. Mitra, D. Mazumdar, and S. Mitra, "Facial expressions: A cross-cultural study," in *Emotion Recognition*, pp. 69–87, Wiley-Blackwell, 2015.

[2] J. M. Garcia-Garcia, V. M. Penichet, and M. D. Lozano, "Emotion detection: a technology review," in *Proceedings of the XVIII international conference on human computer interaction*, pp. 1–8, 2017.

[3] K. Georgila, J. Henderson, and O. Lemon, "Learning user simulations for information state update dialogue systems," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[4] K. Ahmad, L. Gillam, L. Tostevin, *et al.*, "University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder).," in *TREC*, pp. 1–8, 1999.

[5] M. Bland, "Cohen's kappa," *University of York Department of Health Sciences http://wwwusers. york. ac. uk/~ mb55/msc/clinimet/week4/kappash2. pdf.[Accessed February 13 2014]*, 2008.

[6] K. Ahmad, S. Wang, C. Vogel, P. Jain, O. O'Neill, and B. H. Sufi, "Comparing the performance of facial emotion recognition systems on real-life videos: Gender, ethnicity and age," in *Proceedings of the Future Technologies Conference*, pp. 193–210, Springer, 2021.

[7] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[8] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*, University of Nebraska Press, 1971.

[9] E. L. Rosenberg and P. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2020.

[10] Y.-l. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[11] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[12] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*, pp. 46–53, IEEE, 2000.

[13] B. Farnsworth, "Facial action coding system (facs) - a visual guidebook." `https://imotions.com/blog/facial-action-coding-system/`. Accessed: 2022-08-17.

[14] S. Velusamy, H. Kannan, B. Anand, A. Sharma, and B. Navathe, "A method to infer emotions from facial action units," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2028–2031, IEEE, 2011.

[15] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[16] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 298–305, IEEE, 2011.

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*, pp. 5–pp, IEEE, 2005.

[18] M. Frank, M. Bartlett, and J. Movellan, "The m3 database of spontaneous emotion expression (university of buffalo)," *In pres*, 2010.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, Ieee, 2001.

[20] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown, "A performance comparison of eight commercially available automatic classifiers for facial affect recognition," *PloS one*, vol. 15, no. 4, p. e0231968, 2020.

[21] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with affdex and facet: A validation study," *Behavior research methods*, vol. 50, no. 4, pp. 1446–1460, 2018.

[22] K. Yang, C. Wang, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets," *The visual computer*, vol. 37, no. 6, pp. 1447–1466, 2021.

[23] E. G. Krumhuber, D. Küster, S. Namba, D. Shah, and M. G. Calvo, "Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis.," *Emotion*, vol. 21, no. 2, p. 447, 2021.

[24] A. Bernin, L. Müller, S. Ghose, K. von Luck, C. Grecos, Q. Wang, and F. Vogt, "Towards more robust automatic facial expression recognition in smart environments," in *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 37–44, 2017.

[25] C. Lutz and G. M. White, "The anthropology of emotions," *Annual review of anthropology*, pp. 405–436, 1986.

[26] O. Klineberg, "Emotional behavior.," 1940.

[27] W. La Barre, *The cultural basis of emotions and gestures*. Ardent Media, 1947.

[28] S. N. Awan, "The aging female voice: acoustic and respiratory data," *Clinical linguistics & phonetics*, vol. 20, no. 2-3, pp. 171–180, 2006.

[29] R. T. Sataloff, D. C. Rosen, M. Hawkshaw, and J. R. Spiegel, "The aging adult voice," *Journal of Voice*, vol. 11, no. 2, pp. 156–160, 1997.

[30] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *Journal of Voice*, vol. 32, no. 5, pp. 644–e1, 2018.

[31] J. Harrigan, R. Rosenthal, and K. Scherer, *New handbook of methods in nonverbal behavior research*. Oxford University Press, 2008.

[32] M. K. Pichora-Fuller, K. Dupuis, and P. Van Lieshout, "Importance of f0 for predicting vocal emotion categorization," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3401–3401, 2016.

[33] C. Pereira and C. I. Watson, "Some acoustic characteristics of emotion.," in *ICSLP*, 1998.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pp. 1–6, IEEE, 2009.

[36] P. Boersma *et al.*, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, pp. 97–110, Citeseer, 1993.

[37] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[38] S. Strömbergsson, "Today's most frequently used f0 estimation methods, and their accuracy in estimating male and female pitch in clean speech.," in *INTERSPEECH*, pp. 525–529, Dresden, 2016.

[39] P. Martin, "Multi methods pitch tracking," in *Speech Prosody 2012*, 2012.

[40] H. Chen, Z. Liu, X. Kang, S. Nishide, and F. Ren, "Investigating voice features for speech emotion recognition based on four kinds of machine learning methods," in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 195–199, 2019.

[41] W. Bao, Y. Li, M. Gu, M. Yang, H. Li, L. Chao, and J. Tao, "Building a chinese natural emotional audio-visual database," in *2014 12th International conference on signal processing (ICSP)*, pp. 583–587, IEEE, 2014.

[42] T. Özseven and M. Düğenci, "Speech acoustic (spac): A novel tool for speech feature extraction and classification," *Applied Acoustics*, vol. 136, pp. 1–8, 2018.

[43] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," *arXiv preprint arXiv:1905.05812*, 2019.

[44] M. Kamboj, C. Hessler, P. Asnani, K. Riani, and M. Abouelenien, "Multimodal political deception detection," *IEEE MultiMedia*, vol. 28, no. 1, pp. 94–102, 2021.

[45] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis.," *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[46] S. M. Mohammad *et al.*, "Tracking sentiment in mail: How genders differ on emotional axes," *arXiv preprint arXiv:1309.6347*, 2013.

[47] A. Garcia-Rudolph, S. Laxe, J. Saurí, M. B. Guitart, *et al.*, "Stroke survivors on twitter: sentiment and topic analysis from a gender perspective," *Journal of medical Internet research*, vol. 21, no. 8, p. e14077, 2019.

[48] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 366, no. 6472, pp. 1517–1522, 2019.

[49] C. Zhou, J.-M. Dewaele, C. M. Ochs, and J. De Leersnyder, "The role of language and cultural engagement in emotional fit with culture: an experiment comparing chinese-english bilinguals to british and chinese monolinguals," *Affective Science*, vol. 2, no. 2, pp. 128–141, 2021.

[50] K. A. Lindquist, "Language and emotion: introduction to the special issue," *Affective Science*, vol. 2, no. 2, pp. 91–98, 2021.

[51] J. Levis and R. Suvorov, "Automatic speech recognition," *The encyclopedia of applied linguistics*, 2012.

[52] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, vol. 1, p. 67, 2005.

[53] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[54] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[55] B. Xu, C. Tao, Z. Feng, Y. Raqui, and S. Ranwez, "A benchmarking on cloud based speech-to-text services for french speech and background noise effect," *arXiv preprint arXiv:2105.03409*, 2021.

[56] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (microsoft api, google api and cmu sphinx)," *Int. J. Eng. Res. Appl*, vol. 7, no. 03, pp. 20–24, 2017.

[57] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–19, 2021.

[58] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in facebook and its application to e-learning," *Computers in human behavior*, vol. 31, pp. 527–541, 2014.

[59] J. A. Cook and K. Ahmad, "Behaviour and markets: the interaction between sentiment analysis and ethical values?," in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 551–558, Springer, 2015.

[60] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[61] A. Agarwal and D. Toshniwal, "Application of lexicon based approach in sentiment analysis for short tweets," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 189–193, IEEE, 2018.

[62] A. Moniz and F. d. Jong, "Sentiment analysis and the impact of employee satisfaction on firm earnings," in *European conference on information retrieval*, pp. 519–527, Springer, 2014.

[63] C. Lynch, *Exploration of a Multimodal Emotion Recognition System and the Physical Correlates of Emotion*. PhD thesis, University of Dublin, 2021.

[64] M. Mansoor, K. Gurumurthy, V. Prasad, *et al.*, "Global sentiment analysis of covid-19 tweets over time," *arXiv preprint arXiv:2010.14234*, 2020.

[65] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts.," in *DART@ AI* IA*, pp. 59–68, Citeseer, 2014.

[66] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[67] K. R. Scherer, "Studying the emotion-antecedent appraisal process: An expert system approach," *Cognition & Emotion*, vol. 7, no. 3-4, pp. 325–355, 1993.

[68] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 439–448, 2016.

[69] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[70] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.

[71] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, 2018.

[72] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of german emotional speech.," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[73] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[74] M. Beringer, F. Spohn, A. Hildebrandt, J. Wacker, and G. Recio, "Reliability and validity of machine vision for the assessment of facial expressions," *Cognitive Systems Research*, vol. 56, pp. 119–132, 2019.

[75] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, Ieee, 2005.

[76] W. V. Friesen, P. Ekman, *et al.*, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983.

[77] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.

[78] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts.," in *DART@ AI* IA*, pp. 59–68, Citeseer, 2014.

[79] A. Kaur and R. Kumar, "Comparative analysis of parametric and non-parametric tests," *Journal of computer and mathematical sciences*, vol. 6, no. 6, pp. 336–342, 2015.

[80] A. M. Alsaqr, "Remarks on the use of pearson's and spearman's correlation coefficients in assessing relationships in ophthalmic data," *African Vision and Eye Health*, vol. 80, no. 1, p. 10, 2021.

[81] N. Altman and M. Krzywinski, "Points of significance: Association, correlation and causation.," *Nature methods*, vol. 12, no. 10, 2015.

[82] C. P. Dancey and J. Reidy, *Statistics without maths for psychology*. Pearson education, 2007.

[83] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[84] T. O. Kvålseth, "Note on cohen's kappa," *Psychological reports*, vol. 65, no. 1, pp. 223–226, 1989.

# A1 Appendix