Internship Report

Ricardo DA SILVA

Mention Objets Connectés ET Életronique
Filière Innovation et Intrapreunariat

University Tutors :
Mention : Clement Elvira
clement.elvira@centralesupelec.fr
Filière : Nabil Sadou
nabil.sadou@centralesupelec.fr

Enterprise Tutor :
Renaud Séguier
renaud.seguier@emobot.fr

Abstract

In this report, our primary aim is to develop a solution for the early detection of depression or mood disorders in patients, a crucial concern in today's healthcare landscape. To achieve this objective, we propose the creation of a model that leverages speech data to identify emotions based on the Circumplex Model. This model represents emotions in a two-dimensional space, with one axis corresponding to Arousal and the other to Valence.

To extract valuable information from audio recordings, we explored two distinct approaches. The first approach involves utilizing a not fine-tuned version of the Wav2Vec architecture. This architecture is currently recognized as the state-of-the-art technology for Emotion Recognition. The second approach entails employing a fine-tuned version of the Wav2Vec model. We will then investigate how to categorize the extracted features into the Circumplex Model.

In essence, our objective is to harness the power of speech analysis and advanced machine learning techniques to create a robust and accurate tool for identifying emotional patterns indicative of depression in individuals. This tool has the potential to revolutionize early diagnosis and intervention, ultimately improving the mental health care of individuals.

Contents

1 Introduction

In today's fast-paced and interconnected world, the importance of depression detection has never been more significant. Depression, a complex and often debilitating mental health condition, affects millions of individuals worldwide, transcending age, gender, and socioeconomic boundaries, and modern environments are potentially contributing to a rising prevalence of depression [1]. the Recognizing and addressing depression early is crucial, as it not only improves the quality of life for those who suffer from it but also has the potential to mitigate its widespread impact.

Machine learning (ML) technologies have been increasingly applied to analyze speech samples obtained from clinical settings or remote sources, with the aim of identifying biomarkers to enhance the diagnosis and treatment of mental disorders [2]. Psychologists have long utilized auditory and visual cues to aid in mental illness diagnosis, and these cues have proven valuable since they are inherent to human communication and easily accessible in both traditional teletherapy and clinical settings.

We could say that detecting depression and mood disorders is related to the recognition of emotions, which is a complex and multifaceted task that presents several significant challenges. These difficulties stem from the intricacies of human emotional expression and the limitations of technology and human perception. One contribution to the field of emotional research is the works of the psychologist James A. Russell [3], with his development of the Circumplex Model of Emotions. This influential model, proposed in the late 1980s, has significantly advanced our understanding of emotions and their role in human psychology. The Circumplex Model conceptualizes emotions as existing on two primary axes: valence (ranging from pleasant to unpleasant) and arousal (ranging from high to low). By organizing emotions within this framework, Russell's model provides a systematic and comprehensive view of the emotional landscape, offering valuable insights into the nature and dynamics of human emotional experiences.

In this context, the AIMAC Group (Artificial Intelligence for Multimodal Affective Computing) presents the following study, which seeks to develop an automated method for classifying the emotions of patients during their interactions with psychologists. The primary objective of this research is to categorize these emotions within the framework of the Circumplex Model using audio recordings of the patients' voices. The creation of a reliable and objective means for measuring patients' emotional states holds significant potential to assist psychologists in diagnosing individuals who may be experiencing depression or related mood disorders.

Numerous state-of-the-art Emotion Classification Algorithms are available, which are primarily designed for conducting regression tasks that involve categorizing emotions into specific emotion classes. An illustrative example of this is found in [4], where the study investigates the use of Speech self-supervised models like wav2vec 2.0 [5] and HuBERT [6]. These models have made significant strides in the domain of Automatic Speech Recognition (ASR). However, their efficacy in tasks beyond ASR has not been definitively established. In [4], the researchers delved into the potential of partially fine-tuning and fully fine-tuning pre-trained models of wav2vec 2.0 and HuBERT. Their goal was to evaluate these models' performance in three distinct non-ASR speech tasks: Speech Emotion Recognition, Speaker Verification, and Spoken Language Understanding.

Another research aimed to create a model which extract the Valence and Arousal values from data. The research in [7] focuses on the field of automatic emotion recognition (ER) and the utilization of multimodal approaches to improve performance by combining facial and vocal modalities extracted from videos for dimensional ER, specifically in the valence-arousal space. The paper introduces a joint cross-attentional model for A-V (Audio and Visual) fusion. This model extracts salient features across A-V modalities and leverages both intra and inter-modal relationships by computing cross-attention weights based on the correlation between joint feature representation and individual modalities. What we aim in our work, in contrast to this paper, is to be able to extract the Valence and Arousal values only from the speech, not needing the visual part.

2 Context of the Study

2.1 AIMAC Group

The AIMAC Group, short for "Artificial Intelligence for Multimodal Affective Computing," is at the forefront of Affective Computing, specializing in the analysis, synthesis, and monitoring of emotions. Leveraging cutting-edge tools rooted in Artificial Intelligence, such as Auto-encoders and GANs (Generative Adversarial Networks), AIMAC's expertise extends across various modalities, including image, voice, and text.

AIMAC's groundbreaking approach revolutionizes the representation of emotions by enabling the real-time tracking of an individual's emotional state. This multimodal analysis combines the use of voice, speech, context, gestures, and facial expressions to provide a comprehensive assessment of emotions. The group's current research primarily centers on stress detection and the precise identification and analysis of micro-expressions, particularly within a medical context.

The tools and methodologies developed by AIMAC are deeply rooted in Deep Learning techniques, with a strong focus on GANs, VAEs (Variational Autoencoders), CNNs (Convolutional Neural Networks), and auto-encoders. This commitment to innovation has led to the successful creation of three pioneering startups: Dynamixyz, specializing in Performance Capture; 3D Sound Labs, dedicated to Binaural Reproduction; and Immersive Therapy, focusing on the development of a Tinnitus App. AIMAC's work not only advances the field of Affective Computing but also extends its impact to practical applications and real-world solutions across various domains, contributing to a more emotionally aware and technologically sophisticated future.

The AIMAC Group is part of the IETR, which is a public research laboratory specializing in the fields of electronics and digital technologies. With its structure comprising 6 departments and 13 thematic research teams, the work of IETR is dedicated to addressing a wide range of scientific challenges, primarily associated with the digital transformation of society. However, their research also extends to environmental, ecological, energy, and healthcare transitions.

Situated in both Brittany (Rennes, Saint-Brieuc, Lannion, Coëtquidan) and Pays de la Loire (Nantes, Angers, La Roche sur Yon), the IETR brings together over 350 professionals from five different institutions and laboratory overseeing bodies (CNRS (INSIS and INS2I), CentraleSupélec, INSA Rennes, Nantes Université, and Université de Rennes). Additionally, the IETR welcomes researchers and teacher-researchers from other institutions through hosting agreements.

This research is also done with the collaboration of the entreprise Emobot. Emobot is a cutting-edge MedTech startup that specializes in the development of a revolutionary medical device designed for the detection and ongoing monitoring of mood disorders, particularly conditions like depression. Powered by advanced artificial intelligence technology, Emobot leverages any connected camera to translate facial expressions and vocal cues into valuable mood signals. This innovative technology is already making a positive impact in several nursing homes, where it continuously gathers essential data that assists healthcare professionals and caregivers in effectively managing behavioral and mood-related issues. Emobot stands as the pioneering AI-based device for behavioral and emotional monitoring in a variety of healthcare settings, including nursing homes, hospitals, and even individual homes. Notably, whenever Emobot detects a concerning signal, it promptly triggers an alert, leading to a consultation with a psychiatrist for a thorough diagnosis and appropriate intervention.

2.2 The Circumplex Model of Affection

The Circumplex Model of Emotion [3] is a comprehensive framework proposed by James A. Russell that aims to map and understand the structure of human emotions. The model posits that emotions can be organized along two primary axes: valence and arousal. Valence represents the emotional quality or pleasantness of an emotion, ranging from positive (e.g., happiness) to negative (e.g., misery). Arousal, on the other hand, describes the intensity or activation level of an emotion, spanning from low arousal (e.g., sleepiness) to high arousal (e.g., excitement or anger)

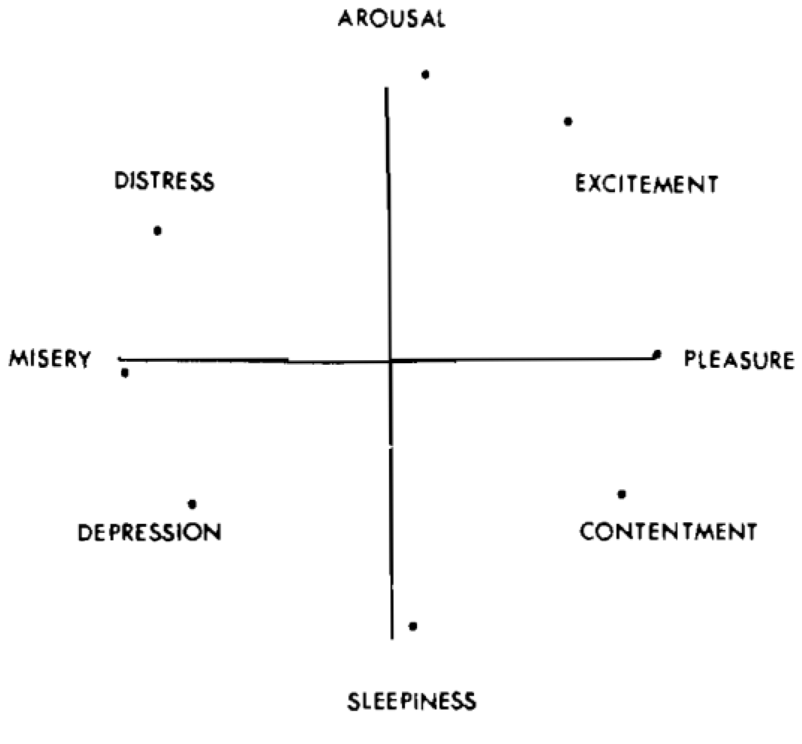


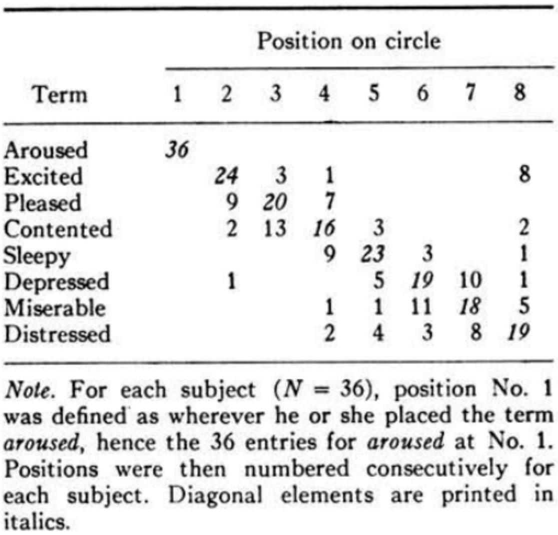
Figure 2: Circumplex Model with 8 affect concepts

In Figure 2, we can observe a graphical representation of affect's cognitive properties. Within this visualization, eight variables are situated in a two-dimensional space, resembling points on a compass, serving as a metaphor. In this spatial metaphor, the horizontal (east-west) dimension corresponds to a pleasure-displeasure continuum, while the vertical (north-south) dimension symbolizes arousal-sleep levels. The remaining four variables do not independently form dimensions but contribute to defining the quadrants of this space. For example, Excitement is precisely defined as a point in the northeast, representing high pleasure and high arousal. On the other hand, depression, as the polar opposite of Excitement, is located in the southwest. Distress and contentment likewise form a bipolar dimension in the northwest and southeast, respectively. All other affective terms can be similarly defined as vectors originating from the center of this conceptual space.

To test the model represented in Figure 2, Russel made a series of three studies taking into account the perception of layman people (not psychologists or scientists). A set of 28 words was carefully selected to represent the domain of affect. These 28 words were then subjected to scaling techniques that relied on participants' internal representations of affect, rather than their current emotional states. Two distinct scaling methods were employed: first, a technique specifically designed for variables with circular ordering. Secondly, the same 28 words underwent multidimensional scaling to derive scaling coordinates, without assuming a circular ordering. Lastly, the 28 words were unidimensionally scaled along the proposed pleasure-displeasure and degree-of-arousal dimensions. These three scaling solutions were subsequently quantitatively compared to evaluate the extent of convergence in results across these diverse methodological approaches.

As the purpose of this report does not involve delving into the exhaustive details of all the work conducted by Russel, I will provide an explanation of one of the tests mentioned earlier, known as the Circular Ordering Method. In this experiment, 36 students from the University of British Columbia were invited to categorize 28 words into 8 distinct classes, including Arousal, Excitement, Pleasure, Contentment, Sleepiness, Depression, Misery, and Distress. Subsequently, they were tasked with arranging each of these 8 classes in a circular order, ensuring that classes representing opposing emotions were diametrically opposed on the circle, while words with similar meanings were positioned closer to each other within the circle.

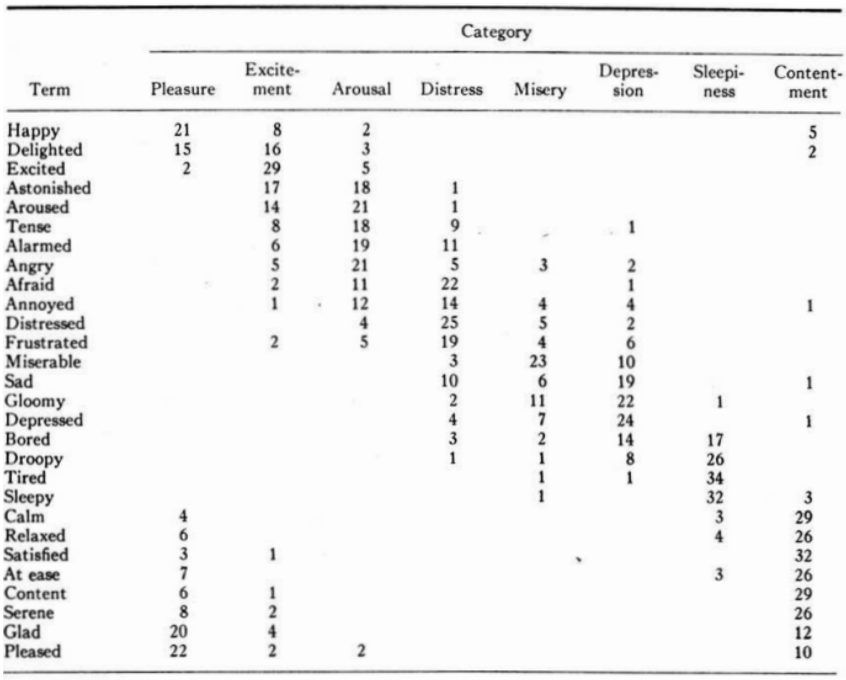
In Figure 3, 4 and 5 are the results of the circular ordering made by the students. In Figure 5, we have an estimation of the coordinates of the 28 words representing emotions into the Valence X Arousal axes. The order tests performed by Russel obtained similar positions and ordering for this set of words. When analysing the results of our model, we going to take into account this positioning of the words to verify if it's possible to create an automated way to visualise the valence and arousal values.



Term	Position on circle							
	1	2	3	4	5	6	7	8
Aroused	*36*							
Excited		*24*	3	1				8
Pleased		9	*20*	7				
Contented		2	13	*16*	3			2
Sleepy				9	*23*	3		1
Depressed		1			5	*19*	10	1
Miserable				1	1	11	*18*	5
Distressed				2	4	3	8	*19*

Note. For each subject (*N* = 36), position No. 1 was defined as wherever he or she placed the term *aroused*, hence the 36 entries for *aroused* at No. 1. Positions were then numbered consecutively for each subject. Diagonal elements are printed in italics.

Figure 3: Frequency which each term were placed around the circle



Term	Category							
	Pleasure	Excitement	Arousal	Distress	Misery	Depression	Sleepiness	Contentment
Happy	21	8	2					5
Delighted	15	16	3					2
Excited	2	29	5					
Astonished		17	18	1				
Aroused		14	21	1				
Tense		8	18	9		1		
Alarmed		6	19	11				
Angry		5	21	5	3	2		
Afraid		2	11	22		1		
Annoyed		1	12	14	4	4		1
Distressed			4	25	5	2		
Frustrated		2	5	19	4	6		
Miserable				3	23	10		
Sad				10	6	19		1
Gloomy				2	11	22	1	
Depressed				4	7	24		1
Bored				3	2	14	17	
Droopy				1	1	8	26	
Tired					1	1	34	
Sleepy					1		32	3
Calm	4						3	29
Relaxed	6						4	26
Satisfied	3	1						32
At ease	7						3	26
Content	6	1						29
Serene	8	2						26
Glad	20	4						12
Pleased	22	2	2					10

Figure 4: Frequency of placement for the 28 words

2.3 Mood Disorder Detection

Leveraging the Circumplex Model as the guiding framework for understanding human humor and emotions, Emobot is dedicated to crafting EmoCare, an innovative diagnostic and remote monitoring software tailored to mood disorders. This groundbreaking solution serves as a multifaceted device for both diagnosing and continu-

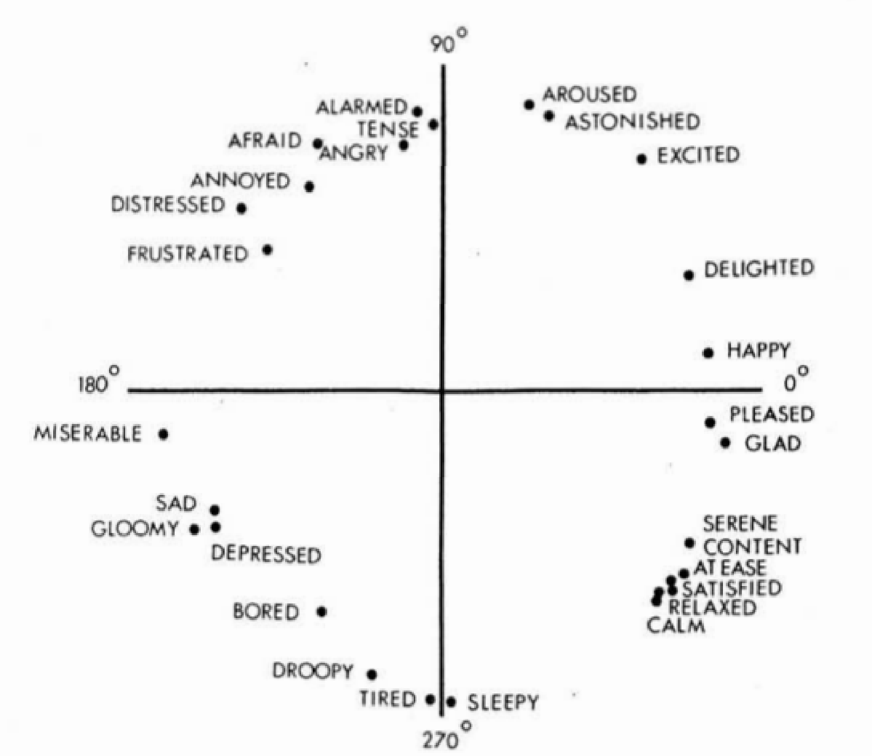


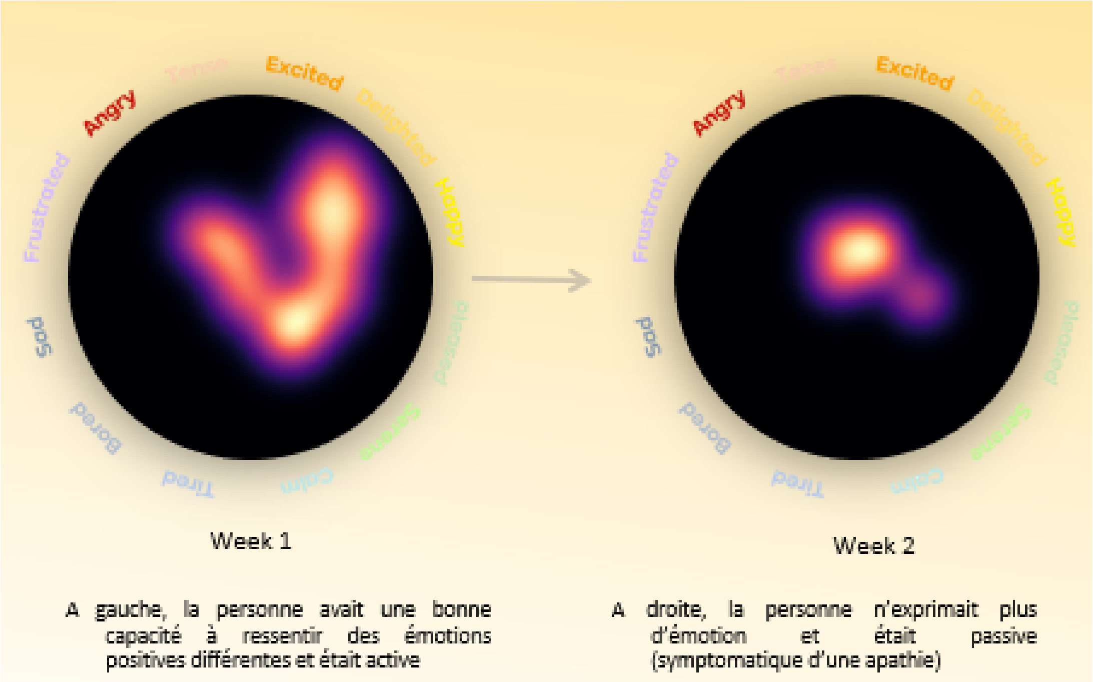
Figure 5: circular scaling coordinates of the 28 words

ously tracking mood and anxiety disorders, with a particular focus on analyzing facial expressions as indicators.

EmoCare harnesses the power of Artificial Intelligence (AI) to meticulously assess the progression of mood disorders over time. This is achieved by unobtrusively scrutinizing individuals' emotional behaviors through their facial expressions and voice. In collaboration with the AIMAC laboratory, Emobot is developing a repertoire of algorithms, signals, and digital biomarkers. These tools are engineered to automatically gauge the severity of disorders and monitor their evolving dynamics in patients.

the algorithms culminate in the creation of an "emotional heat map." This visual representation effectively encapsulates an individual's "emotional tone" throughout the day, aggregating the diverse emotions detected. For instance, one heat map may exhibit a predominance of neutral emotions and a diverse emotional spectrum for an individual with a healthy emotional tone. Conversely, another heat map may reveal lower emotional intensity and a preponderance of neutral emotions for an individual with a diminished emotional tone. In this way, EmoCare empowers healthcare providers with comprehensive insights into their patients' emotional well-being and offers the potential for timely intervention and support. In Figure 6 we can see a representation of the heat map of emotions into the Circumplex Model.

Our aim in this report, on the long term, is to be able to perform this same task,



A gauche, la personne avait une bonne capacité à ressentir des émotions positives différentes et était active

A droite, la personne n'exprimait plus d'émotion et était passive (symptomatique d'une apathie)

Figure 6: Visualization of emotion changing during two weeks in a Circumplex Model

which is to visualize the variation of the emotions of a person into the Circumplex Model, but using the voice of the person without the visual cue.

3 Methodology

3.1 Wav2Vec Model

Wav2Vec [5] is an advanced speech recognition model developed by Facebook AI Research (FAIR) that has made significant strides in the field of Automatic Speech Recognition (ASR). It represents a substantial advancement in the accuracy and efficiency of transcribing spoken language into written text. Wav2Vec has numerous applications and is poised to revolutionize various industries and domains. It leverages the principle of self-supervised learning, which means the model learns directly from the data without relying on manually transcribed audio recordings. The core of Wav2Vec is its utilization of contrastive learning and quantization-based vector quantization (VQ), where it learns to map speech data to a discrete set of codebook entries, effectively reducing the dimensionality of the input data while preserving critical information.

To create a representation of the audio, the Wav2Vec model undergoes a two-phase training process, each with its distinct purpose. The initial phase operates in a self-supervised manner, utilizing unlabeled data to create a high-quality representation of speech. You can liken this process to word embeddings, where the goal is to capture the most informative representation of language. However, Wav2Vec 2.0, in contrast to word embeddings, works with audio data rather than text. The second phase of training, known as supervised fine-tuning, employs labeled data to instruct the model in predicting specific words or phonemes.

The primary advantage of Wav2Vec 2.0 lies in its first training phase, where it excels in learning an exceptional speech representation. This ability enables the model to achieve remarkable results even with limited labeled data. For this part, the authors of the paper [5] initially pre-trained the model using a vast LibriVox dataset, which is a collection of free and public domain audiobooks that have been recorded by volunteers from around the world, and , subsequently, they fine-tuned it with the complete Libri Speech dataset [8], which is a dataset of approximately 1000 hours of 16kHz read English speech derived from read audiobooks from the LibriVox project.

3.1.1 Architecture and Pre-Training

In Figure 7, we can see the structure of the wav2vec model after the self-supervised learning, composed of Convolutional Neural Network layers, as well as a Transformer Layer.

In Figure 8 we can see the process which happens during the pre-training of the architecture. During the self-supervised learning, the convolutional layers are responsible to create a latent representation of the raw waveform of the audio. This latent representation will have a proportion of its features masked before being fed to the transformer. The objective of the masking is that the transformer will try

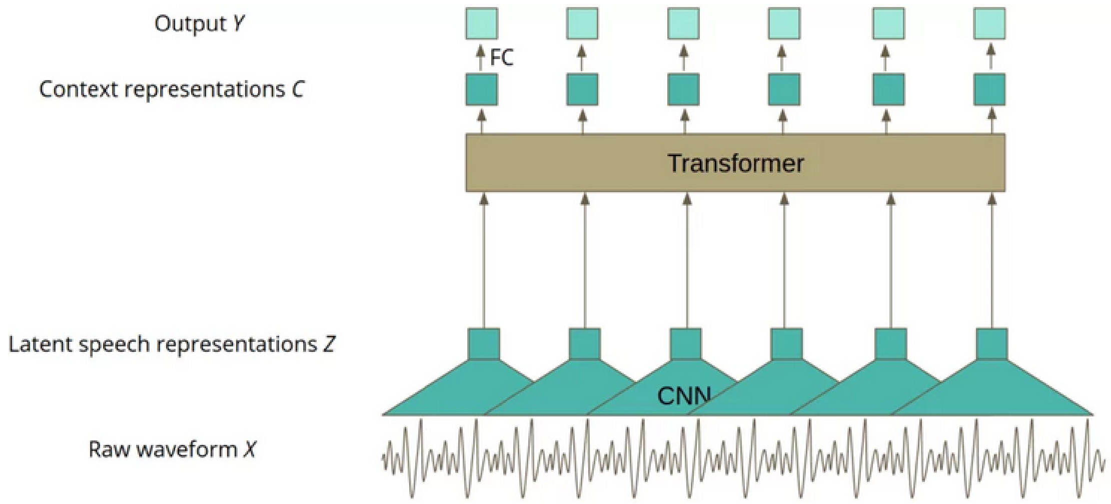

Output Y
FC
Context representations C
Transformer
Latent speech representations Z
CNN
Raw waveform X

Figure 7: Wav2Vec 2.0 model architecture

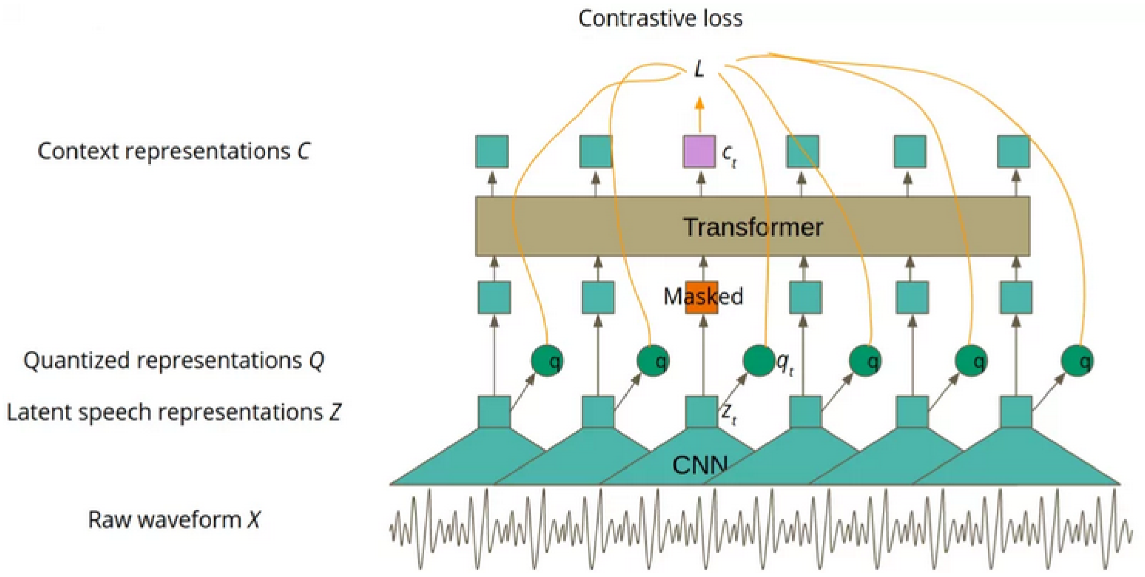

Contrastive loss
L
Context representations C
c_t
Transformer
Masked
Quantized representations Q
q
q_t
Latent speech representations Z
z_t
CNN
Raw waveform X

Figure 8: Wav2Vec 2.0 model self supervised training

to create a contextualized representation of its input, trying to guess a quantized representation of the masked features, taking into account the others quantized representations.

3.1.2 Wav2Vec Output

The Wav2Vec's output, as depicted in Figure 9, possesses a dimensional structure denoted as $t \times n$. In this context, 't' is determined by the duration of the audio. The

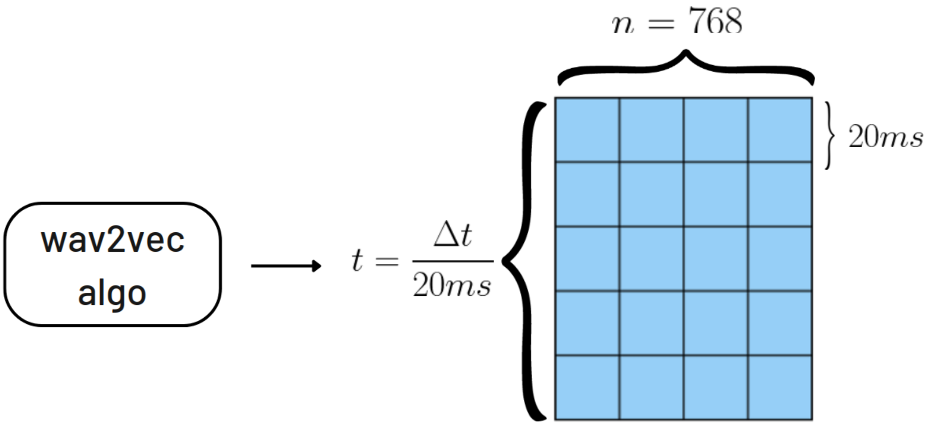


Figure 9: Wav2Vec 2.0 Base Output

Wav2Vec model segments the audio into 20-millisecond intervals, with each row in the output representing a 20-millisecond segment, summing up to a total of 't' rows, where 't' is calculated as '$t = \Delta t/20ms$'. The number of columns, denoted as 'n,' depends on the specific version of Wav2Vec used, which could be either the Base or Large version. Notably, Wav2Vec represents each 20-millisecond audio segment as a vector, and for the Base version we are employing, the vector's size is 'n = 768' elements.

3.2 Datasets

For this methodology, we used two well known datasets to evaluate our models:

IEMOCAP: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [9] database is a valuable resource in the field of emotion research, offering a rich collection of multimodal data designed to explore human emotional expression and communication. Created by the SAIL lab at the University of Southern California (USC), this dataset is a comprehensive addition to the field. IEMOCAP is extensively annotated by multiple annotators, providing categorical labels for emotions (anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state). It also includes dimensional labels for valence, activation, and dominance, allowing for a more nuanced understanding of emotional states. The audios are in English, and is performed by 10 actors, 5 males and 5 females.

EMODB: The EMODB database [10], short for the "Emotional Database," is a freely accessible resource designed to facilitate the study of emotional expression in the German language. This database was meticulously crafted by the Institute of Communication Science at the Technical University in Berlin, Germany. It features recordings from ten professional speakers, comprising an equal distribution of five males and five females. Within the EMODB database, you'll discover a total of 535 carefully curated utterances. These spoken segments encompass a broad spectrum of emotional states, making it an invaluable resource for research in affective computing and emotion recognition. Specifically, the database includes recordings of seven distinct emotions: happy, angry, anxious, fearful, bored and disgusted way as well as a neutral state.

3.3 Definition of Some Statistical Concepts

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique commonly used in statistics and data analysis. Its primary purpose is to reduce the number of variables in a dataset while preserving as much of the original data's variance as possible. PCA achieves this by transforming the original data into a new coordinate system, where the new axes, called principal components, are orthogonal and ordered by the amount of variance they explain. The first principal component accounts for the most variance in the data, the second principal component accounts for the second most, and so on. PCA is particularly useful for simplifying complex data and uncovering patterns or relationships among variables.

Explained Variance Ratio: The explained variance ratio is a metric used to quantify how much of the total variance in the data is explained by each principal component in a PCA. It provides insights into how much information is retained when reducing the dataset's dimensionality. The explained variance ratio for a particular principal component is calculated by dividing the variance of that component by the total variance of the original data. In essence, it indicates the proportion of the dataset's variance that can be attributed to that specific principal component. Typically, you'll see a scree plot or cumulative explained variance ratio to help determine how many principal components to retain in a PCA. A high cumulative explained variance ratio suggests that a relatively small number of principal components can capture most of the essential information in the data, making it an efficient way to reduce dimensionality while preserving data integrity.

Mahalanobis distance: is a measure of the distance between a point and a distribution. It takes into account the correlation between variables and the variances of those variables. This distance is a useful metric for multivariate data, where you have multiple variables that may not be independent and have varying levels of dispersion. The Mahalanobis distance considers the shape of the distribution and the correlation between variables. It is particularly useful for outlier detection, clustering, and classification tasks, as it provides a more accurate measure of distance when dealing with multivariate data than traditional Euclidean distance.

Gaussian Mixture Model (GMM): is a probabilistic model used in statistics and machine learning to represent a probability distribution over a dataset. It's particularly useful for modeling data when the underlying distribution is a combination of multiple Gaussian (normal) distributions. GMMs are a form of mixture model, where multiple component Gaussian distributions are combined to approximate the overall distribution of the data. GMMs have a wide range of applications, including clustering, density estimation, data compression, and feature extraction. They are particularly useful when dealing with complex data distributions that cannot be adequately represented by a single Gaussian distribution.

3.4 Proposed Models

3.4.1 First Proposed Model

Our initial approach involved the utilization of the Wav2Vec algorithm to process audio data, followed by an in-depth examination of the model's output in a lower-dimensional space through Principal Component Analysis (PCA). Following the PCA analysis, we aimed to devise an automated method for categorizing audio emotions within the circumplex model of affection. The schematic representation of this approach is illustrated in Figure 10. The "?" in Figure 10 is to represent the automated method, which is unknown until we can analyse and take more conclusions from the PCA of the output of the Wav2Vec model.

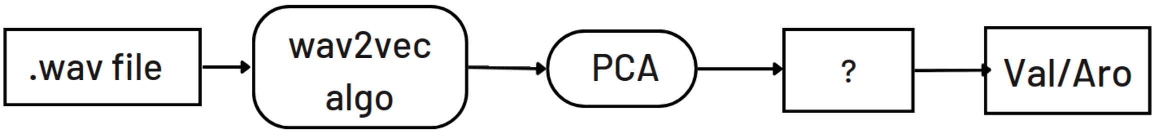


Figure 10: First Proposed Model

After applying this approach on the IEMOCAP and EMODB datasets, we noticed that the emotions didn't exhibit clear visual separation into distinct clusters within the PCA, as seen in Image 11. This outcome was, to some extent, anticipated since the Wav2Vec model was originally pretrained for speech recognition rather than emotion classification. This may suggest that the model might not be ideally suited for extracting emotional and humorous content-related information from the audio data.

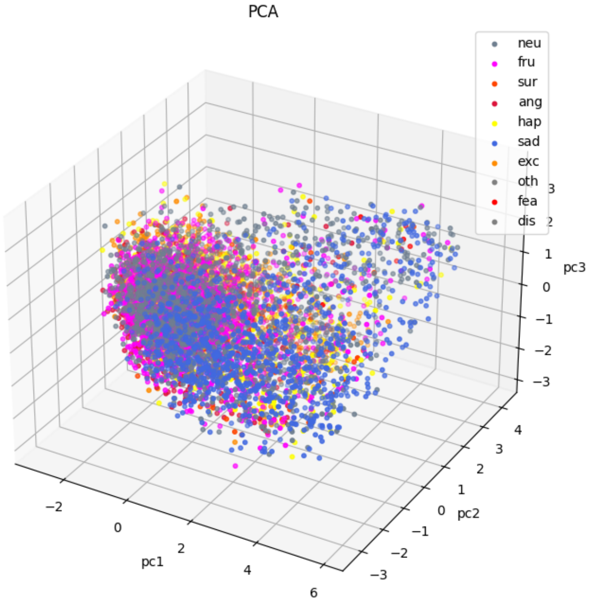


Figure 11: PCA of iemocap - First Model

3.4.2 Second Proposed Model

With this perspective in mind, we have devised a novel approach for the model. To enhance the Wav2Vec's ability to extract emotion-related information from audio recordings, we plan to fine-tune the model through an Emotion Classification Task. This entails training and evaluating a neural network model. This model will incorporate the Wav2Vec's layers along with an additional layer dedicated to the regression task, aimed at classifying the audio data into specific emotion classes .

This reimagined strategy intends to harness the power of the Wav2Vec model by tailoring it to the nuanced task of emotion classification. By pre-training the model on emotional data, we seek to enable it to better understand and interpret the emotional content present in the audio recordings.

Following the fine-tuning process, our plan is to leverage the learned weights of the Wav2Vec layer from the fine-tuned model. We aim to incorporate these optimized weights into the original version of the model, where Wav2Vec is employed to extract information from audio inputs, subsequently visualizing the output within a reduced dimensional space. In this iteration, the weights applied to the Wav2Vec model will be those fine-tuned during the initial phase.

This second iteration of the proposed model is visually represented in Figure 12. By integrating the fine-tuned weights, we anticipate an enhanced capability for

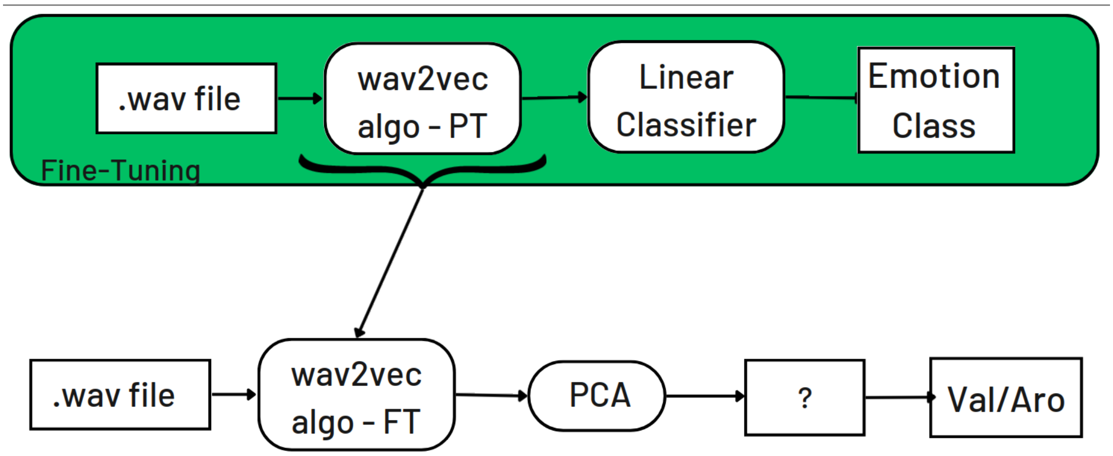


Figure 12: Second Proposed Model

this model to more accurately capture emotional content within the audio data. This refinement is expected to yield improved results in the representation and analysis of emotional states, contributing to the overall effectiveness of our approach in characterizing emotions within the circumplex model of affection.

3.4.3 Fine-Tuning Process

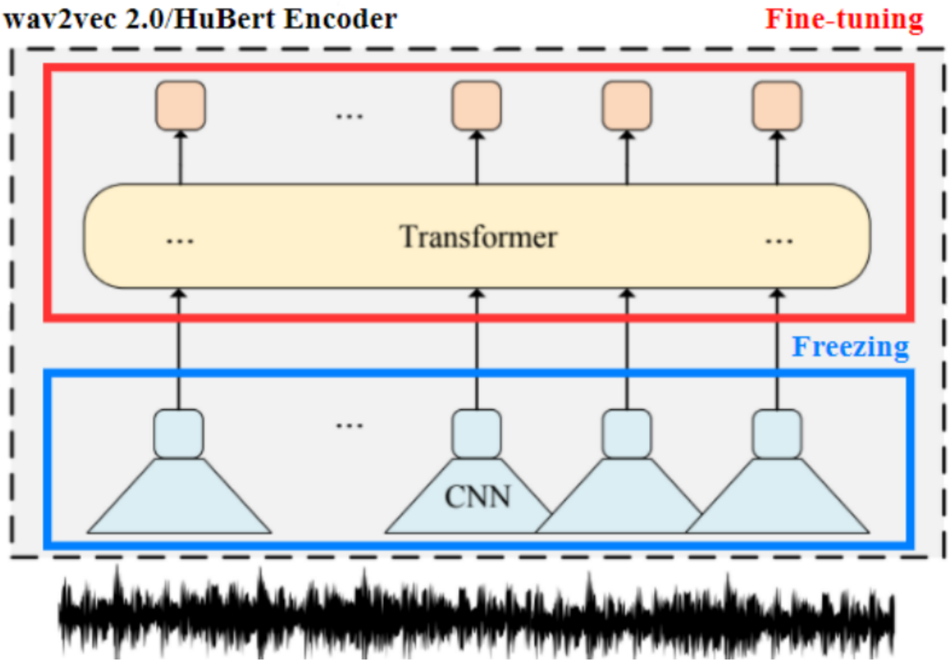


Figure 13: Partial Fine-Tuning

For the fine-tuning, we're following one of the training examples for Speech Emotion Recognition in [4]. Since the Wav2Vec 2.0 model consists of two crucial components, which are the CNN-based feature encoder and the transformer-based contextualized encoder, in the approach of the paper, we choose to keep the CNN-based feature encoder static, thereby locking all the parameters associated with these CNN blocks. The focus of our fine-tuning efforts was directed exclusively towards the transformer blocks. This approach, known as partial fine-tuning, can be likened to a domain adaptation process targeted at the uppermost layers. The primary goal here is to prevent any potential interference or detriment to the foundational CNN layers, which already possess a high level of expressiveness.This partial fine-tuning is represented in Figure 13.

With the assumption that finely-tuned Wav2Vec 2.0 models are sufficiently adept at capturing pertinent information, we simply integrated straightforward downstream adaptors, such as classifiers or decoders, directly onto the Wav2Vec architecture. This approach aims to efficiently leverage the pre-trained models' power and adapt them for specific tasks without unnecessary complexity. In the context of Speech Emotion Recognition (SER), a straightforward downstream classifier is introduced. This classifier includes two key components: average time pooling and a linear layer. The purpose of the average time pooling step is to consolidate speech data with varying time durations into a consistent representation. Subsequently, the linear layer is employed to carry out an overall classification of the entire utterance, with the aim of minimizing the cross-entropy loss.

In the training, we also implemented two separate schedulers to individually adjust the learning rates for two crucial components: the Wav2Vec 2.0 encoder and the downstream model. Both of these schedulers utilized the Adam Optimizer and employed a linear annealing strategy, which means that the learning rates were gradually adjusted based on the performance observed during the validation stage of the training process. Specifically, for tasks like Speech Emotion Recognition (SER), we initially set the fine-tuning learning rate for the Wav2Vec encoder at 10^{-5} and the downstream model's learning rate at 10^{-4}. This meticulous adjustment of learning rates allows for better fine-tuning and optimization of both components, ultimately improving the model's performance in these specific tasks.

In our experiment, we utilized the IEMOCAP dataset to train and validate a fine-tuning model in a Speaker Dependent setting. The task is to classify the data into 4 classes (Anger, Happiness, Sad and Neutral). We started the training from scratch and achieved a weighted accuracy of 76%, which aligns closely with the results reported in the research paper.

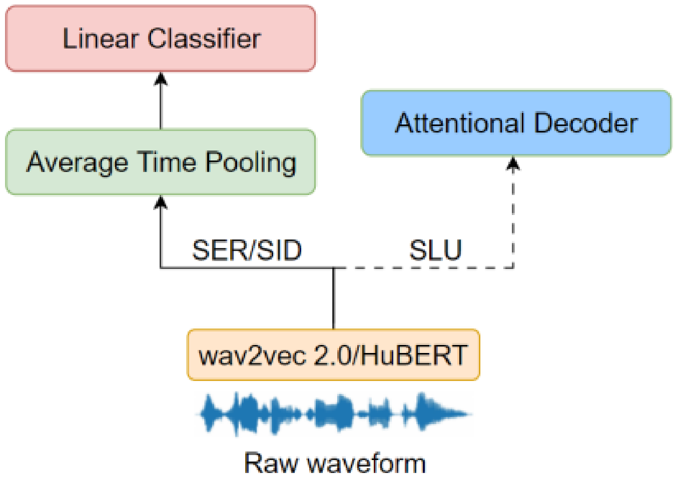


Fig. 2. Simple downstream models for SER, SID and SLU. For SER and SID, an average time pooling and a linear classifier is built over wav2vec 2.0/HuBERT. For SLU, an attentional decoder decodes intents and slots directly from the fine-tuned wav2vec2.0/HuBERT embedding.

Figure 14: Models from the paper [4]

3.4.4 Evaluating Second Model

Once we completed this particular phase, we proceeded to employ the freshly fine-tuned weights of the Wav2Vec architecture to visualize the features that had been extracted and reduced in dimension, as illustrated in Figure 12. Using the IEMOCAP dataset, our anticipation was that the four distinct classes would exhibit clear separation due to the model having been trained on this very dataset. This anticipated outcome is indeed affirmed by the observations made in Figure 15.

Hence, it becomes imperative to extend our analysis of the results of the fine tuned model using another dataset. In this case, we used the EMODB dataset. Using the Wav2Vec fine-tuned and subsequently applying a PCA (Principal Components Analysis) transformation to the resulting data, we generated a plot illustrating the explained variance ratio, as depicted in Figure 16. Analyzing this graphical representation, it becomes evident that a mere three principal components suffice to capture and represent a significant portion of the dataset's information. Remarkably, these initial three components account for more than 60% of the total data variance, emphasizing their crucial role in summarizing the underlying patterns within the data.

With that in mind, the results of this model using EMODB are prominently featured in Figure 17. This figure prominently demonstrates that within the three-dimensional (3D) space, discernible clusters of emotions are visually apparent. These clusters represent a compelling visual representation of the model's ability to capture

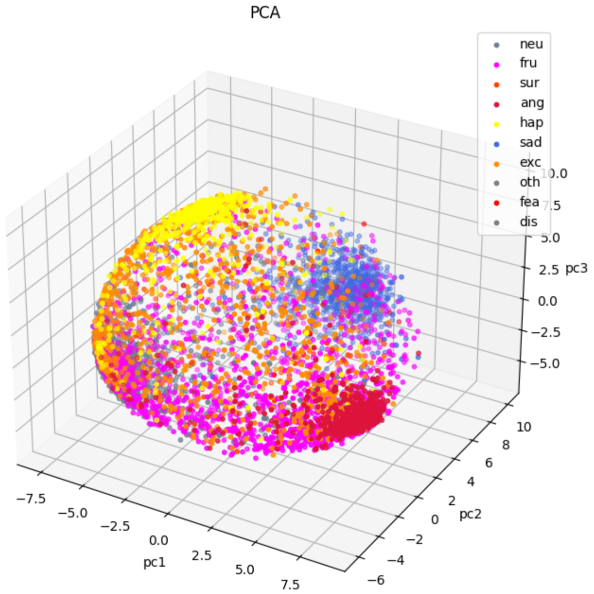


Figure 15: PCA of IEMOCAP - Second Model

and differentiate various emotional states.

In Figure 17, a noteworthy observation emerges as emotions such as Anger and Boredom, which exhibit considerable separation along the Arousal axis in the Circumplex Model (as illustrated in Figure 5), also exhibit distinct and well-separated clusters in the 3D reduced feature space obtained through fine-tuning the Wav2Vec model. Conversely, emotions like Sadness and Boredom, which are relatively close to each other on the Circumplex Model, manifest as closely positioned clusters within the 3D feature space generated by Wav2Vec fine-tuning.

To provide quantitative validation for this observation, we opted to represent these emotion clusters as Gaussian Distributions using a Gaussian Mixture Model and subsequently calculate the Mahalanobis distance between these clusters. Figures 18, 19, and 20 showcase the Gaussian Distributions for select emotions in a 2D pca reduced space from the 3D representation, allowing us to visualize the relative distances between these emotional clusters. This approach offers an objective means to evaluate the spatial relationships and separations between different emotional states, as we can observe that the mahalanobis distance is bigger between Anger and Boredom than it is between Sadness and Boredom and between Anger and Fear.

An unexpected outcome that emerged from our analysis was the proximity of clusters representing Anger and Happiness, as evident in Figure 21. This result

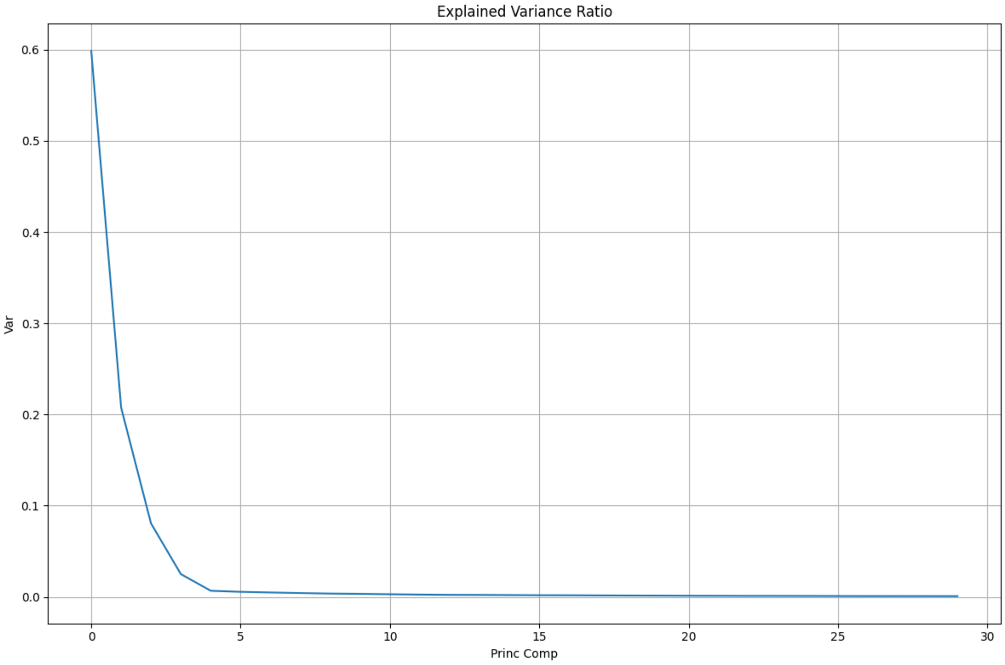


Figure 16: Explained Variance Ratio for EMODB

diverges from the anticipated findings, as per the Circumplex Model illustrated in Figure 5, where Anger and Happiness should exhibit significant separation along both the Arousal and Valence axes. An explanation for this unexpected result may lie in the characteristics of the EMODB dataset. It appears that the fine-tuned model encounters challenges in accurately classifying Happy audio samples within this dataset.

Upon conducting a more in-depth investigation, we turned our attention to the classifier model employed during the fine-tuning process, which categorizes audio samples into four emotions, including Happy. Our findings revealed that the classifier model frequently struggled to correctly classify audio samples from EMODB as Happy, instead tending to misclassify them as Anger. This classifier's limitations may, in part, account for the observed proximity of Anger and Happy clusters within the 3D feature space.

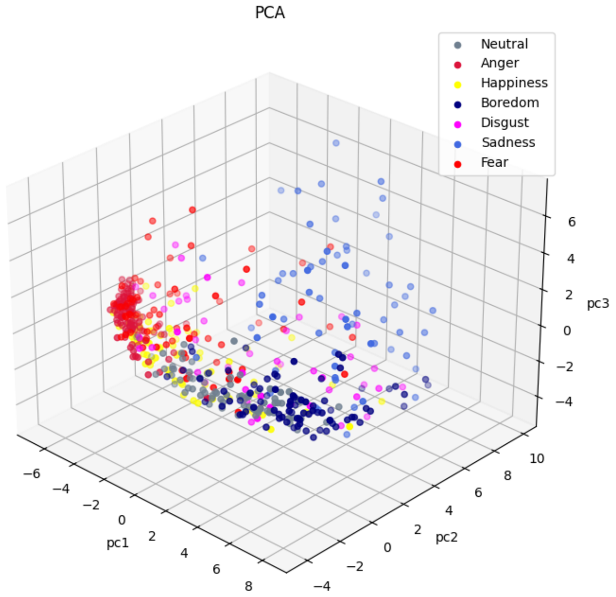


Figure 17: PCA of EMODB - Second Model

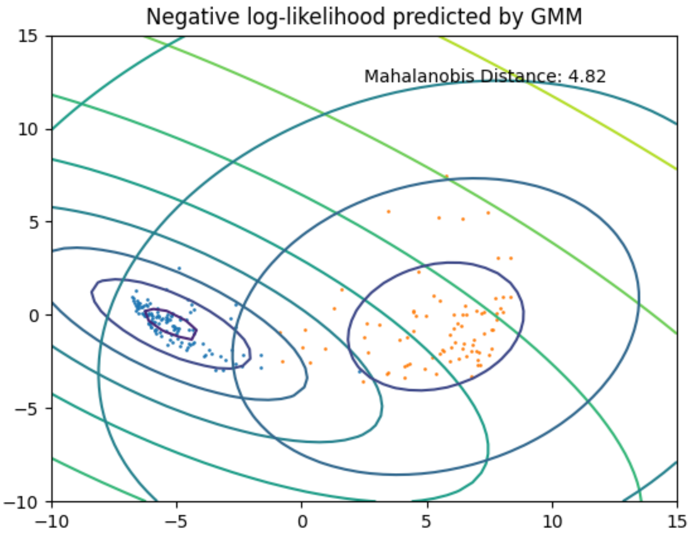


Figure 18: Gaussian Distribution - Anger X Boredom

Another intriguing observation that emerged from our analysis of Figure 17 is

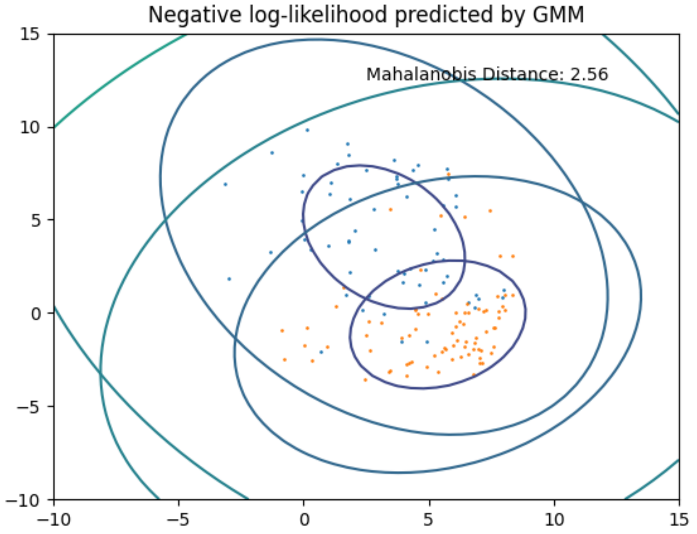


Figure 19: Gaussian Distribution - Sadness X Boredom

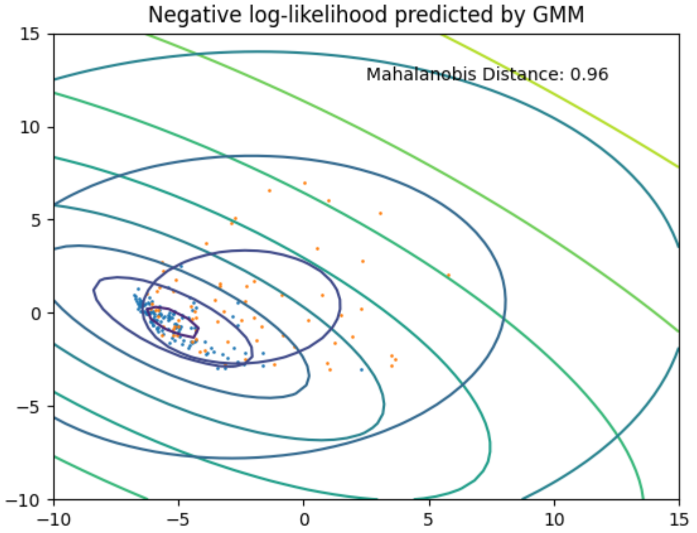


Figure 20: Gaussian Distribution - Anger X Fear

the semblance of a circular order in the positioning of emotion clusters, somewhat echoing the circular arrangement depicted in Figure 5. To further substantiate this observation, we employed the centroids of the Gaussian Distributions as a reference point. We transformed their Cartesian coordinates into polar coordinates and subsequently organized the corresponding angles in an ascending order, mirroring the

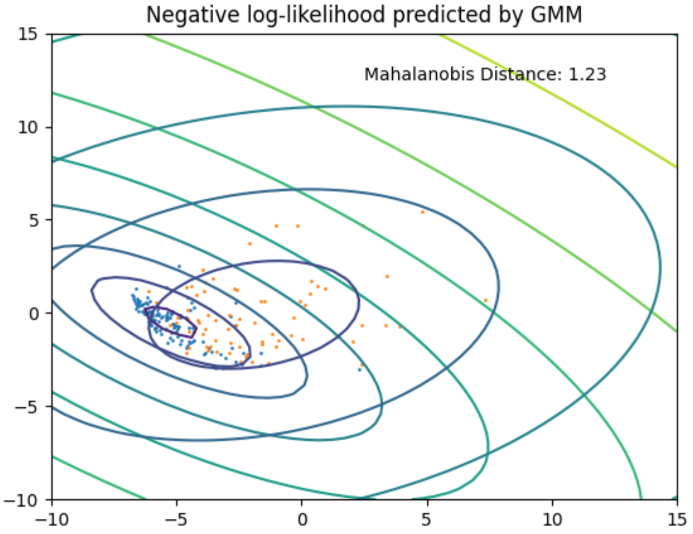


Figure 21: Gaussian Distribution - Anger X Happy

crescent arrangement found in the Circumplex Model. The polar coordinates are shown on the Table 1.

Centroid	Radius	Angle (degrees)
(Sad)	5.19	51.65
(Disgust)	1.77	67.37
(Fear)	2.50	172.78
(Happy)	1.91	183.12
(Anger)	5.23	185.23
(Neutral)	4.81	322.68
(Boredom)	5.40	353.36

Table 1: Centroid of Gaussian Distributions on EMODB

As we can see on Table 1, the circular order of the emotions would be like: **Sad, Disgust, Fear, Happy, Anger, Neutral, Boredom**. If we take into account Figure 5 and follow a ***counterclockwise*** direction, we can order these emotions, starting from Sad, as: **Sad, Disgust, Fear (Afraid), Anger (Angry), Happy, Boredom (Bored)**. Disgust is not represented on Figure 5, but since we can define this emotion has low valence and somewhat high arousal, it's fair to put Disgust between Sad and Fear. And, as we can see, with the exception of the emotion Happy, which was already stated that the model have problems to classify in this dataset, the order of the emotions are the same as in the Circumplex model.

4 Conclusion

In this study, we presented two approaches for extracting Valence and Arousal coordinates from speech. The first approach utilized a not fine tuned version of the Wav2Vec model, while the second involved fine-tuning the Wav2Vec model through an emotion classification task. The latter approach exhibited superior results, as the fine-tuned Wav2Vec model effectively learned to extract emotional information from audio data.

Upon scrutinizing the output of the fine-tuned model using the EMODB dataset, a compelling observation emerged. Specifically, when considering the Arousal axis within the Circumplex Model, as opposed to the Valence axis, emotions with contrasting Arousal values exhibited greater separation. This suggests that determining Arousal from speech may be more feasible, whereas extracting Valence could necessitate alternative sources, such as video or text data derived from audio.

Furthermore, an intriguing finding revealed a partial alignment between the circular order of emotions in our model and the Circumplex Model. This alignment indicates the promise of our model in effectively capturing Arousal and Valence from speech.

Nonetheless, a problem arose concerning audio samples associated with the "Happy" emotion. The model tended to place them closer to "Anger" and "Fear" audio samples than expected based on their Arousal position within the Circumplex Model. This discrepancy might stem from the particular characteristics of the EMODB dataset. It was noted that the model struggled to recognize the "Happy" emotion within this dataset, which could be due to language differences. Notably, our model was fine-tuned with English audio data, while EMODB employs German audio.

For future research endeavors, it would be valuable to assess the performance of our fine-tuned model with diverse datasets to ascertain whether the circular ordering of emotions remains consistent beyond the confines of the EMODB dataset, and analysing these different datasets, find a method to automatically put the model result into the circumplex model. This exploration could shed light on the generalizability of our model and its capacity to extract emotional dimensions from speech across various linguistic and cultural contexts.

5 Bibliography

References

[1] Brandon H Hidaka. Depression as a disease of modernity: explanations for increasing prevalence. *J Affect Disord*, 140(3):205–214, January 2012.

[2] Zhe Sage Chen, Prathamesh, Kulkarni, Isaac R. Galatzer-Levy, Benedetta Bigio, Carla Nasca, and Yu Zhang. Modern views of machine learning for precision psychiatry, 2022.

[3] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980.

[4] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *CoRR*, abs/2111.02735, 2021.

[5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[6] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447, 2021.

[7] R Gnana Praveen, Eric Granger, and Patrick Cardinal. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention, 2022.

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.

[10] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. A database of german emotional speech. volume 5, pages 1517–1520, 09 2005.