IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Unveiling Emotions in Speech: A Multi-Dimensional Transformer Approach

[1]Tushar Sharma, [2]Prakash R, [3]Rajat Singhal

[1]Data Scientist, [2]Data Scientist, [3]AI Evangelist

[1]Data and A.I.,

Black Basil Technologies, Noida, India

Abstract : In this article, we talk about Speech Emotion Detection (SED), SED is the task of automatically recognizing and categorizing the emotions expressed in spoken language. The goal is to determine the emotional state of a speaker, such as happiness, anger, sadness, or frustration, from their speech patterns, such as energy, pitch, and rhythm. The most used and well known speech feature MFCC(Mel Frequency Cepstral Coefficients), despite the progress that has been made, speech emotion detection is still a challenging task. There are many factors that can affect the accuracy of emotion recognition, such as the speaker's accent, the noise level in the environment, and the emotion itself. Despite the challenges, speech emotion detection is a promising area of research. As the technology continues to improve, it will become increasingly possible to use speech emotion detection to create more natural and engaging user experiences.

IndexTerms **- acoustic features, linguistic features, visual features, timbre acoustic features, valence dimension, emotion recognition,speech processing, transformer.**

I. INTRODUCTION

This article highlights the significance of emotions in human speech and its relevance in various applications. In addition to linguistic content, speech also carries emotional cues, which are crucial for tasks like gathering emotional user behaviors in entertainment electronics, understanding the manner in which something was said in Automatic Speech Recognition, and synthesizing emotionally natural speech in text-to-speech systems. Therefore, it is important for computers in human-machine interaction applications to accurately perceive emotional states expressed in human speech.

However, detecting human emotion in an audio signal has multiple challenges. The speech emotion detection system needs an appropriate model to identify the emotions. Another critical research difficulty is to detect good vocal features in recognition of emotion in speech. These features are affected by age, gender, accent of voice and give more complications because speaking style directly affects features such as energy, pitch,frequency and so on.

The current methods for detecting emotions in speech rely on acoustic features that are mostly associated with speech recognition. These features include fundamental frequency or pitch, energy, speaking rate, and spectral coefficients like Mel-frequency cepstral coefficients (MFCCs).

The traditional features used for speech recognition do not fully capture how humans perceive emotions. These features were originally designed to decode individual sounds in speech, but emotions do not change as rapidly as these sounds. We propose a set of speech features which are designed to detect Perceptual data of speech to design an audio emotion detection system.

2. RELATED STUDY

In this section we provide an overview of speech recognition and The existing literature defines emotional categories differently. We are specifically focusing on the challenges related to emotional categorization methods, which involve accurately interpreting expressions and physiological responses. Additionally, they examine the speech features used in emotion detection and highlight the associated issues

2.1. Acoustic Features

Acoustic features in speech can be extracted from the entire speech or from smaller parts known as frames. When extracted from the entire speech, they are referred to as statistical acoustic features. These features include the arithmetic mean, standard deviation, maximum, and minimum values. One advantage of these features is that they are plentiful, which makes it quicker to apply feature selection algorithms and reduces the classification time. It is faster to apply feature selection algorithms and take less classification time. In acoustic features there are continuous, qualitative, spectral and Teager energy operator(TEO)-based features.

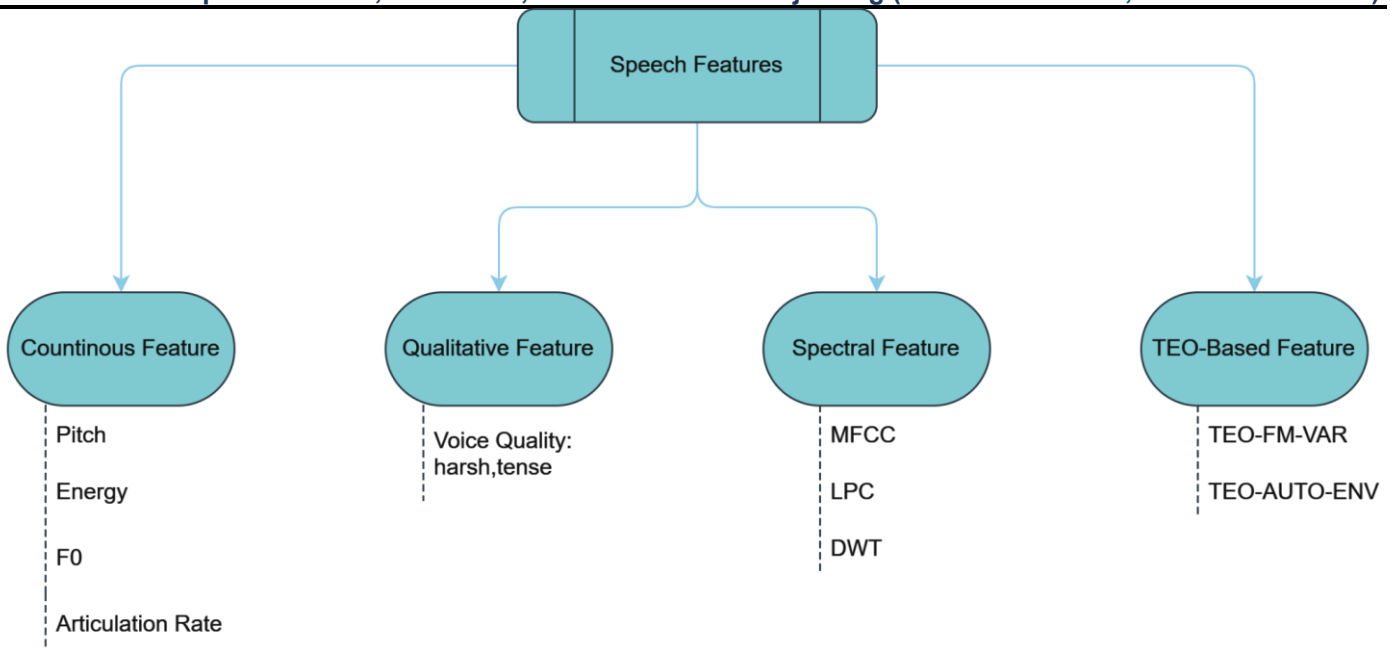

Figure 1. Categories of speech Feature

2.1.1 Continuous Feature

- **Energy**: In speech analysis, the energy feature refers to the measure of the strength or intensity of the sound signal. It represents the amount of acoustic energy present in a particular segment of speech. The energy feature is calculated by squaring the amplitude values of the speech signal within a specific time frame and summing them up. It provides information about the overall loudness or volume of the speech signal. Higher energy values typically indicate louder or more intense speech, while lower energy values indicate softer or less intense speech.

- **Pitch:** In speech analysis, the pitch feature refers to the perceived frequency of a sound or the perceived "highness" or "lowness" of a voice. It represents the fundamental frequency of the vocal fold vibrations during speech production. The pitch feature is determined by the rate at which the vocal folds vibrate, resulting in different pitch levels. Pitch is typically described in terms of Hertz (Hz) and is closely related to the perceived pitch perception of human ears. It helps in distinguishing between high-pitched voices (e.g., falsetto) and low-pitched voices (e.g., deep voice). Pitch plays a significant role in conveying various aspects of speech, including emphasis, intonation, and emotional expression.

- **Fundamental Frequency(F0):** Fundamental frequency (F0) is a key feature in speech analysis that represents the fundamental frequency of the vocal folds's vibrations during speech production. It refers to the lowest frequency component in the speech signal and is often associated with the perceived pitch of a person's voice. The vocal folds vibrate at a specific frequency, which determines the pitch of the voice. This frequency can vary depending on factors such as gender, age, and emotional state. The F0 feature provides information about the rate at which the vocal folds vibrate and helps to distinguish between high-pitched and low-pitched voices. By analyzing the F0 feature, researchers can extract valuable information about intonation, stress patterns, and emotional expression in speech. It is commonly used in various applications such as speech recognition, emotion detection, and speaker identification.

- **Articulation Rate:** It determines the pace at which speech segments are actually produced and does not take into account speaker specific ways of conveying information, such as hesitations, pausing, emotional expression and so on. Articulation refers to making sounds. The production of sounds involves the coordinated movements of lips , tongue, teeth, palate(top of mouth) and respiratory system(lungs).

2.1.2 Qualitative Features

- **Voice quality:** The emotional content of speech is closely connected to its voice quality. Voice quality features can be grouped into four categories: voice level, voice pitch, phrase/phoneme/word boundaries, and temporal structures. However, there is ambiguity and subjectivity in describing voice quality terms like tense, harsh, and breathy. Different research studies have sparked ongoing debates about the associations between voice quality and specific emotions. For example, a tense voice is linked to anger, joy, and fear, while a lax voice is associated with sadness. Breathiness in the voice can be related to both anger and happiness, while sadness is connected to a "resonant" voice quality.

2.1.3 Spectral Features

- **MFCC:** Mel-frequency cepstral coefficients(MFCC) are a sort of acoustic feature that is used to represent the spectral characteristics of speech. They may be received through changing the speech signal to the mel scale and then taking cepstral coefficients. The mel scale is a logarithmic scale that emphasizes the frequency regions that are vital to human hearing. MFCCs are powerful in capturing the timbre, pitch, and different acoustic houses of speech. This makes them useful for a selection of speech processing responsibilities, such as speech emotion recognition. In emotion popularity, MFCCs can be used to assist classifiers differentiate between different feelings by studying the acoustic patterns of the speech.The formula used to calculate the mels for any frequency is:

$$mel(f) = 2595xlog10(1 + f/700)$$

where mel(f) is the frequency (mels) and f is the frequency (Hz).
The MFCCs are calculated using this equation:

$$\mathcal{C}\, n = \sum kn = 1(log\mathcal{S}\, k)cos[n(k - 12)\pi k]$$

where k is the number of mel cepstrum coefficients, Ŝk is the output of filterbank and Ĉn is the final MFCC coefficients.

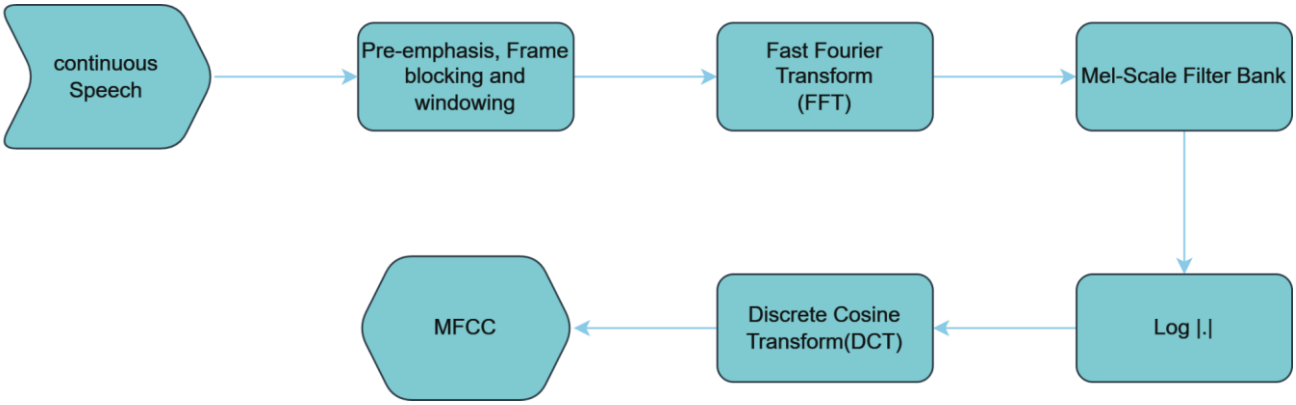

Figure 2. Block Diagram Of MFCC

- **LPC:** In LPC analysis, the speech signal is modeled as a linear combination of past samples, and the coefficients of this linear combination are called LPC coefficients. These coefficients represent the vocal tract's filter parameters, which shape the spectrum of the speech signal. LPC coefficients can provide valuable information about the unique characteristics of an individual's speech, including their vocal tract configuration and pronunciation patterns. By analyzing LPC coefficients, classifiers can identify acoustic patterns associated with different emotions, thus aiding in recognizing and distinguishing various emotional states expressed in speech.
- **DWT:** Discrete Wavelet Transform (DWT) is a signal processing technique used in various applications, including speech emotion recognition. DWT is used to analyze signals and extract features by decomposing the signal into different frequency sub-bands. DWT can be applied to the speech signal to analyze the variations in different frequency components over time. The DWT decomposes the speech signal into high-frequency and low-frequency components, providing information about the rapid changes and slow trends in the signal. These sub-bands can capture different aspects of the emotional content present in the speech.

2.1.4 TEO-Based

- **TEO-FM-VAR:** (Teager Energy Operator-based Variability). It is an acoustic feature used in speech emotion recognition to capture variations in the Teager Energy Operator (TEO) across time. The Teager Energy Operator is a signal processing technique that emphasizes the energy changes in a signal. It is often used to enhance the representation of non-stationary signals, like speech, by focusing on the abrupt changes in the signal's amplitude.

- **TEO-Auto-Env:** (Teager Energy Operator-based Autocorrelation Envelope). The Teager Energy Operator is a signal processing technique used to measure the energy variations in a signal. It is useful for emphasizing changes in the signal's amplitude and has applications in speech processing. The Teager Energy Operator is a signal processing technique used to measure the energy variations in a signal. It is useful for emphasizing changes in the signal's amplitude and has applications in speech processing.

2.2 Dimensional Approach

The dimensional technique to represent feelings uses 3 major dimensions: valence (nice or terrible), arousal (excited or apathetic), and dominance (dominant or submissive). These dimensions assist describe and categorize one of a kind emotional states based totally on their emotional intensity, positivity and dominance. The use of dimensional description of human affect defines the dependency of the categories of one another; rather than their dependency as in categorical description. Another alternative is simpler two-dimensional emotion space: arousal and valence.

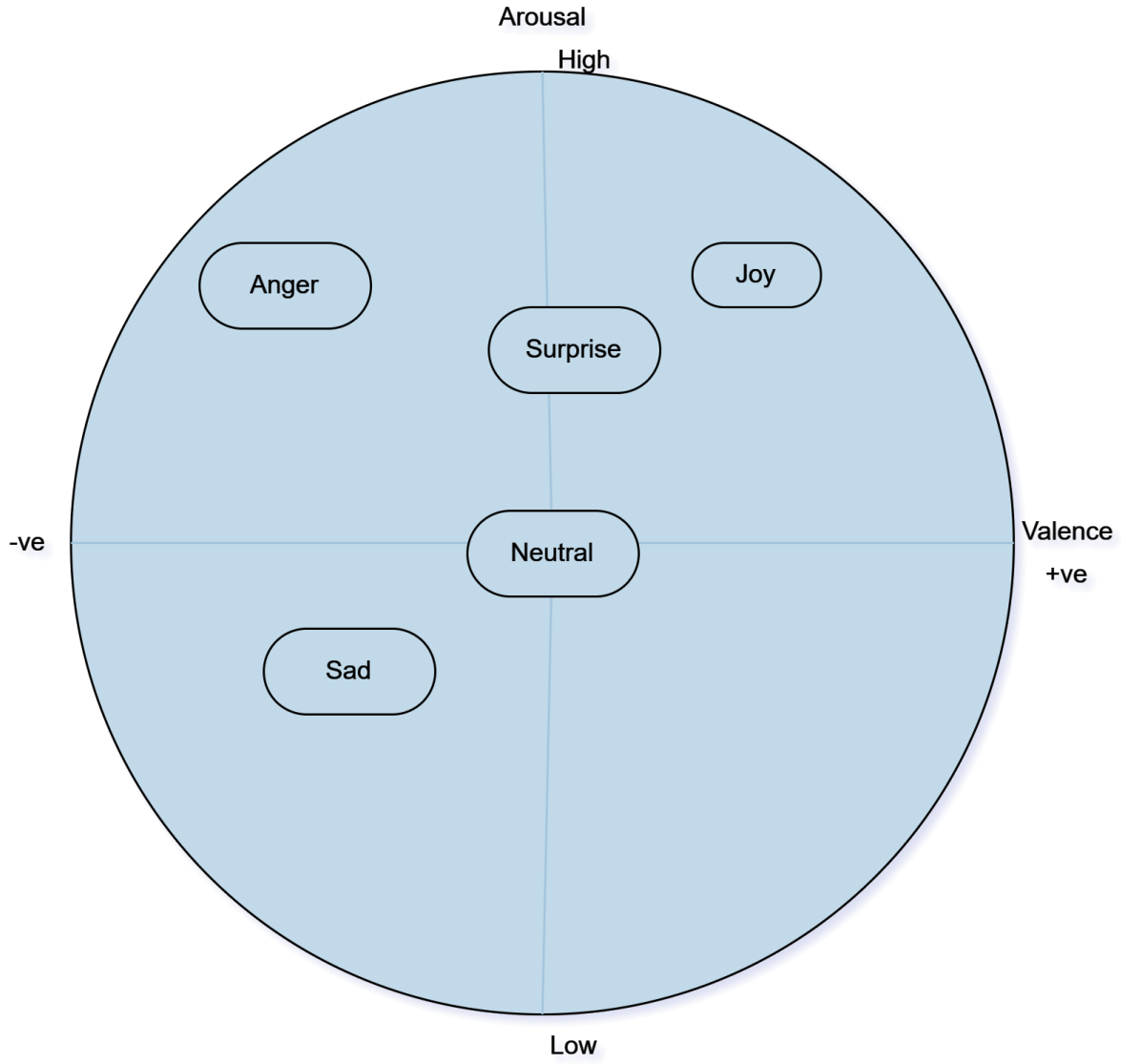

Figure 3. Representation of basic emotions in the arousal-valence dimension

2.2.1 Arousal

The arousal feature refers to a measure of the emotional intensity or activation level expressed in the speech signal. Arousal is one of the fundamental dimensions used to describe emotions and represents how excited or apathetic a person's emotional state is. To extract the arousal feature from speech, various acoustic features are analyzed, such as the energy level, pitch variations, speaking rate, and other dynamic changes in the speech signal. Higher arousal values typically indicate more intense emotions, while lower arousal values correspond to more neutral or calm emotional states.

By incorporating the arousal feature in emotion recognition systems, classifiers can differentiate between emotions with varying levels of intensity and understand the emotional dynamics expressed in the speech. Arousal, along with other emotional dimensions like valence, plays a critical role in accurately recognizing and understanding emotions conveyed through speech.

2.2.2 Valence

The valence function refers to a degree of the emotional positivity or negativity expressed inside the speech sign. Valence is one of the key dimensions used to describe emotions and represents how fine or terrible a person's emotional country is. To extract the valence characteristic from speech, numerous acoustic functions are analyzed, together with the tonal versions, pitch patterns, and spectral traits of the speech sign. Better valence values commonly suggest extra high-quality feelings, such as happiness or joy, whilst lower valence values correspond to more terrible emotions, like sadness or anger.

Via incorporating the valence function in emotion reputation systems, classifiers can distinguish among emotions with distinctive degrees of positivity or negativity and advantage insights into the emotional content of the speech. Valence, at the side of different emotional dimensions like arousal, is essential for correctly spotting and deciphering feelings conveyed through speech.

- **Timbre:** Timbre features in speech emotion recognition refer to acoustic characteristics related to the quality and texture of the speech signal. Timbre is a perceptual attribute of sound that allows us to distinguish different instruments or voices even when they are playing the same pitch and loudness. In speech, timbre features provide information about the unique tonal qualities and characteristics of the speaker's voice.

Some of the common timbre features used in speech emotion recognition include:

- **Spectral Rolloff:** Indicates the frequency below which a certain percentage of the total spectral energy lies, providing insights into the distribution of energy across different frequency bands.
- **Spectral Flux:** Measures the change in spectral magnitude between consecutive frames, indicating the rate of spectral change in the speech signal.
- **Spectral Slope**: Describes the slope of the spectral energy distribution, which can provide information about the brightness or darkness of the speech.
- **Harmonic-to-Noise Ratio (HNR)**: Represents the balance between harmonic components and noise in the speech signal.
- **Voice Break Index (VBI):** Quantifies the level of vocal instability or "breaks" in the voice, which can be indicative of emotional stress or excitement.

Timbre features capture unique aspects of the speaker's voice that contribute to emotional expression in speech. By analyzing these timbre characteristics, classifiers can better differentiate between emotions and enhance the accuracy of speech emotion recognition systems.

The Figure below explains the two dimensions and features that encapsulate these dimensions.

Arousal	Valence
High to Low	postive to Negative
Pitch	Timbre
Speaking Rate	
Voice Quality	
Loudness	
Intensity	

Figure 4. Arousal-Valence dimension features

3.METHODOLOGY

3.1. Speech Emotion Detection System

In order to automatically detect emotions from a speech signal, an Emotion Detection System (EDS) is essential. The conventional EDS comprises three key components: initial data processing, feature retrieval, and categorization, as shown in Figure 5.

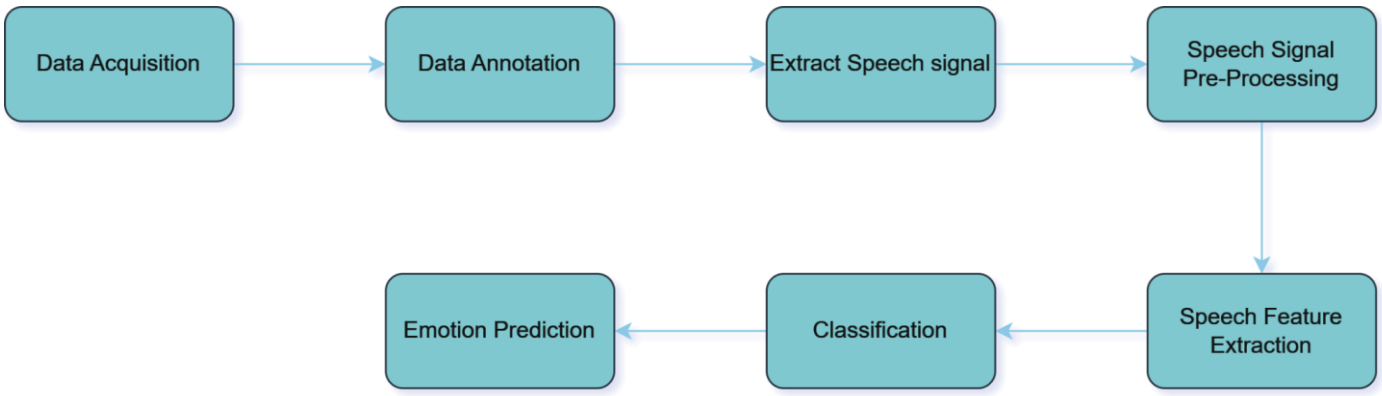

Figure 5. General speech emotion detection system.

3.1.1 Data Acquisition

In the data acquisition part, collecting audio samples containing speech utterances with associated emotional expressions from different platforms, usually extract an audio which has specific sentences with emotion and spontaneous conversation while expressing different emotions such as happiness, sadness, anger, fear and more. To ensure the effectiveness and diversity of the dataset, it is essential to gather samples from a wide range of speakers, various age groups, genders and cultural backgrounds.

3.1.2 Data Annotation

In the data annotation part, once data is collected then for better understanding of emotion we only kept vocal sound from audio and removed musical sound and noise from audio. The process of data annotations typically involves human annotators who listen to each audio sample and assign corresponding emotional labels to segments or the entire audio. Annotations can be done using categorical labels, where each emotion is assigned a discrete category. For high-quality annotations, it's important to provide guidelines to annotators to ensure consistent and accurate labeling. Inter-annotator agreement checks can be conducted to measure

the agreement among multiple annotators for the same data samples, helping identify and resolve any ambiguities or discrepancies. For this we decided to make a UI script for labeling of audio after listing particular audio and labeled them as per their emotions.

3.1.3 Extract speech signal

In the Extract speech signal part, it involves capturing and isolating acoustic features from audio to uncover the underlying emotional nuances. This process begins by converting the raw audio waveform into a spectrogram representation, which portrays the intensity of different frequency components over time. From the spectrogram, critical features such as pitch, intensity, and spectral shape are extracted. Pitch reflects the fundamental frequency of the voice, offering insights into emotional variations in tone. Intensity highlights vocal loudness, a key indicator of emotional arousal. Spectral shape details show how energy is distributed across different frequency bands, unveiling timbral shifts influenced by emotions. Additionally, temporal features like rhythm and speech rate are considered.

3.1.4 Speech signal pre-processing

In the Speech signal pre-processing, it is a pivotal phase in speech emotion detection, as it lays the groundwork for accurate and reliable emotional analysis. This initial step involves a series of techniques aimed at refining and enhancing the raw audio data. Common pre-processing steps include noise reduction, where unwanted background sounds are attenuated to ensure a clearer signal. Signal normalization is then employed to bring the audio to a consistent level, reducing potential bias caused by volume variations. Subsequently, the speech signal is segmented into smaller units, such as frames or windows, facilitating the analysis of emotional changes over time. Overall, speech signal pre-processing acts as a critical preparatory phase, enhancing the effectiveness of subsequent emotion recognition algorithms by ensuring that the underlying emotional cues are effectively captured and represented.

3.1.5 Speech Feature Extraction

In the Speech feature extraction, it constitutes an important step in the emotion detection process, encompassing the conversion of raw speech signals into compact yet information-rich representations. This process involves the calculation of diverse acoustic features that encapsulate crucial aspects of the audio, enabling the extraction of emotional nuances. Mel-frequency cepstral coefficients (MFCCs) are widely utilized, quantifying the spectral characteristics of the speech signal by capturing its frequency content. Spectral features detail the distribution of energy across different frequency bands, while statistical measures like mean and variance encapsulate temporal patterns. These extracted features collectively form a representation of speech signal, serving as input for emotional classification algorithms.

3.1.6 Classification

In the realm of Speech Emotion Detection systems, a pivotal consideration pertains to the classification model. Numerous researchers have delved into a variety of classifiers, encompassing methods like Hidden Markov Models (HMM), the Gaussian Mixture Model (GMM), K-nearest neighbor (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), in the pursuit of addressing the challenge of speech emotion recognition. However, a consensus has yet to emerge regarding the definitive choice of the most potent classifier for speech emotion classification. Each classification model boasts its unique strengths and shortcomings. Hence, the selection of the appropriate classification model should be contingent upon the specific problem at hand.

Speech emotion recognition is a pivotal task in understanding human communication and interaction. The nuances and complexities of emotional expression require sophisticated models that can capture subtle patterns and relationships in audio data. One such powerful architecture that has gained immense popularity in various natural language processing tasks is the Transformer architecture. In this exposition, we delve into the architecture of a speech emotion recognition model enhanced by Transformer encoders and explore its advantages in comprehending and interpreting emotional cues from speech.

In many recent studies, recurrent neural networks (RNNs) have been used for the SER task .Ruben and Gloria investigated the discriminative capabilities of RNNs in SER using low-level acoustic features of speech signals. RNNs are known to be useful in sequential data. Generally, deep neural networks (DNNs) use different parameters at each layer, but RNNs share the same parameters through all steps. However, they inadequately cover long context information because of the gradient vanishing problem. To solve this problem, a long short-term memory (LSTM) RNN was proposed ,which consisted of recurrently connected memory blocks. In this research different transformer architectures were evaluated for the SED task. We used arousal and valence features to train and evaluate the transformers model. We extracted the features using the librosa framework. In addition, emotions were recognized in verbal speech sounds. For better understanding we use a 2 dimensional approach, In which features are separated into arousal and valence for classification of emotion.

3.2 Architecture

In the field of Speech Emotion Detection (SED), recent advancements have seen the integration of state-of-the-art transformer-based models, to process acoustic features like arousal and valence for emotion classification within spoken language. This research leverages the capabilities of transformer models to capture intricate patterns and contextual information in speech, enhancing the accuracy of emotion detection. By inputting arousal and valence features, these models offer a holistic understanding of emotional states, enabling the classification of emotions such as happiness, sadness, anger, and surprise, among others, with a higher degree of precision.

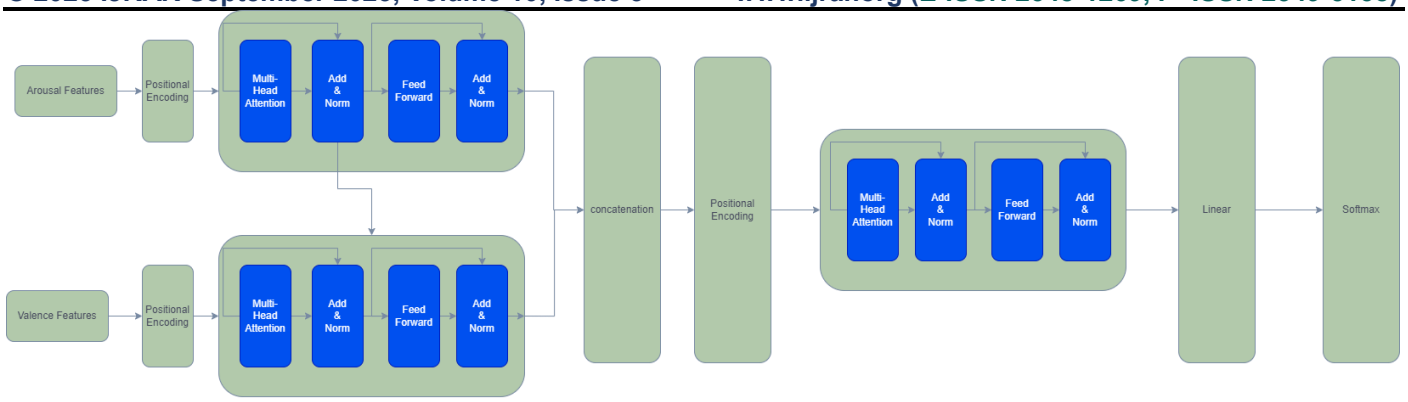

Figure 6. Architecture of transformer-based model

The proposed architecture is built upon the Transformer framework. The innovation here lies in adapting the Transformer's self-attention mechanism to the realm of speech emotion recognition. The model comprises three major components: the Arousal Transformer, the Valence Transformer, and the Final Transformer.

- **Features:** In this we use a 2 dimensional approach for selecting features. One dimension is arousal and second is valence. These dimensions help psychologists and researchers categorize and understand various emotional experiences in a more nuanced way.
    - **Arousal:** Arousal refers to the level of physiological and psychological activation or stimulation associated with an emotion. It ranges from low arousal (calm and relaxed) to high arousal (excited or agitated). For example, feeling calm has low arousal, while feeling anxious has high arousal.
    - **Valence:** Valence refers to the emotional positivity or negativity of an emotion. It ranges from negative (unpleasant or sad) to positive (pleasant or happy). For example, happiness has a positive valence, while sadness has a negative valence.

- **Positional Encoding**: Before understanding self-attention, it's important to note that transformers don't inherently have any understanding of the order of the input data (e.g., sequence of words in Speech). To give the model information about the positions of speech data in the sequence, a positional encoding is added to the input embeddings. This encoding provides a sense of relative positions, which is crucial for capturing sequential information.

- **Self-Attention Mechanism:** The self-attention mechanism allows each word (or element) in the input sequence to consider the relationships with all other words.For each word in the input sequence, self-attention computes a weighted sum of the values, where the weights are determined by the similarity between the word and other words. The similarity is calculated as the dot product between the query of the current word and the key of the other words, followed by a softmax operation to ensure the weights are normalized. This is achieved through three learnable weight matrices:
    - **Query Matrix (Q):** This matrix projects the input into a space where it can be compared to other elements in the sequence.
    - **Key Matrix (K):** Similar to the query matrix, this projects the input for comparison.
    - **Value Matrix (V):** This matrix holds the actual information to be used in the output.

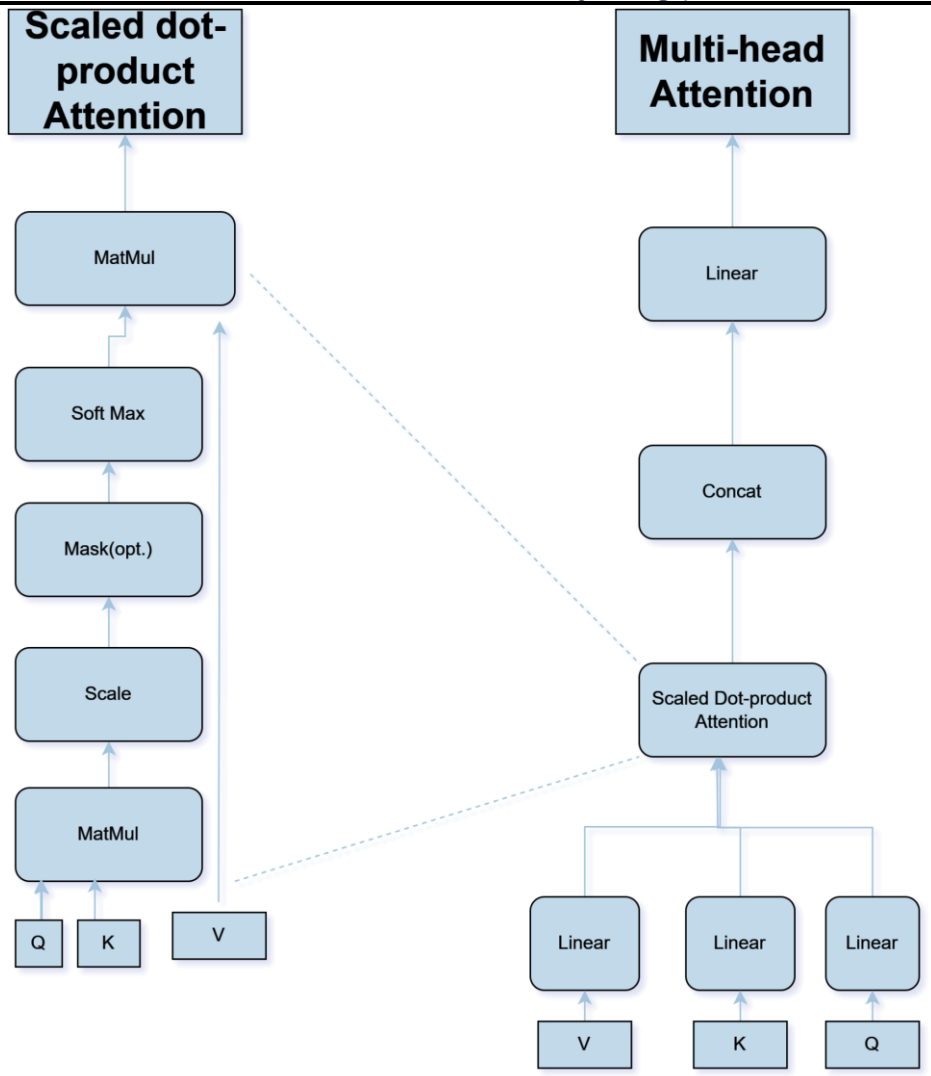

Figure 7. step-by-step breakdown of self attention

- **Scaled Dot-Product Attention:** The attention scores (weights) are scaled by the square root of the dimension of the key matrix. This helps prevent the gradients from becoming too small during backpropagation. The scaled scores are then used to compute the weighted sum of values.

- **Multi-Head Attention:** In practice, self-attention is performed in parallel multiple times, each time with different learned projections (queries, keys, and values). These parallel self-attentions are called "heads." The output of the self-attention mechanism from each head is concatenated and linearly projected to obtain the final output.

- **Arousal Transformer:** The Arousal Transformer extracts emotional features related to the arousal dimension. This dimension characterizes the intensity of an emotion, such as excitement or calmness. The input audio features, including energy, pitch, and intensity, are processed through multiple Transformer encoders. The positional encoding ensures that the model understands the sequential nature of audio data. The self-attention mechanism allows the model to emphasize relevant emotional cues while suppressing noise, thus capturing the varying intensity of emotions in speech.

- **Valence Transformer:** In contrast, the Valence Transformer focuses on the valence dimension, which signifies the positivity or negativity of an emotion. Spectral features like centroid, spread, crest, and energy capture spectral characteristics of the audio. These features are concatenated and processed through Transformer encoders. By identifying spectral variations, the Valence Transformer deciphers emotional shifts, such as moving from happiness to sadness or vice versa.

- **Final Transformer:** The outputs of the Arousal and Valence Transformers are concatenated and passed through another set of Transformer encoders. This final step combines the two emotional dimensions, yielding a holistic representation of emotional content in speech. The linear and softmax layers then map the encoded features to emotion classes.

Self-attention's ability to capture contextual relationships is what makes transformers powerful for tasks like speech emotion recognition. It allows the model to focus on relevant parts of the input at different layers, effectively learning to recognize patterns and relationships within the data. The multi-head attention mechanism further enhances the model's capacity to capture diverse and intricate relationships in the input sequence. This architecture's advantage lies in its ability to process sequences of variable length and capture complex dependencies, making it well-suited for tasks that require understanding emotional nuances in speech.

4. EXPERIMENTS

The analysis of acoustic features was performed on a database which is made by us. In this we tried multiple architectures for getting good results.

4.1 Emotion Database(E-DB)

E-DB is used in the emotional speech database in SED, and E-DB was made by us for this research, consisting of seven emotions, namely anger, happiness, neutral, sad, surprise, fear and disgust. The utterances in this database were created by multiple persons. We chose four emotions for our experiments. The main reasons for this was that these four emotions share similar acoustic features, and it is hard to discriminate these emotions in the valence dimensions. Moreover, these emotions are easily accessible in almost all other emotional speech databases. We randomly chose 666 audio for each selected emotion category to avoid an imbalanced data problem. Information about the databases,emotion categories and the number files in each emotion category are given in the table.

Emotion	E-DB
Anger	200
Happy	200
Neutral	112
Sadness	154

Figure 8. number of samples in each emotion category.

First, gathering audio data featuring spoken expressions of emotions involves sourcing recordings from various sources. Typically, these recordings encompass a spectrum of emotional states, including but not limited to happiness, sadness, anger, and fear. These audio samples are derived from sources that encompass scripted sentences designed to convey specific emotions, as well as unscripted, spontaneous conversations that naturally express a range of emotions. To guarantee the richness and representativeness of the dataset, it is imperative to procure these samples from a diverse array of individuals, spanning different demographics such as age, gender, and cultural backgrounds. After the data collection phase, the next step involves refining the dataset to focus solely on vocal content while eliminating any musical elements or background noise present in the audio recordings. Subsequently, the process of annotating this data is typically carried out by human annotators. The labeling can be done through categorical labels, where distinct emotional categories are assigned. To ensure the quality and consistency of these annotations, it's crucial to furnish annotators with clear guidelines for precise and uniform labeling. to facilitate this process, we developed a user interface (UI) script that streamlines the audio labeling task after listing the individual audio samples, allowing for efficient and accurate emotion categorization.

We extracted the many features and most used acoustic features in the SED field, which include Energy, pitch, intensity, spectral features, using open-source framework librosa. Furthermore, it's important to note that we can derive acoustic characteristics from various input representations, including the original audio signal and derived forms like the Short-Term Fourier Transform (STFT) amplitude, power, temporal energy envelope, and harmonic components. In our research, we took local acoustic features from these diverse input representations and conducted experiments to ascertain which one was most appropriate for our specific task. Following these experiments, it was determined that the STFT amplitude and harmonic input representations were the most suitable for our purposes. The selection was based on the criterion of classification accuracy, with these representations demonstrating the best performance. For more detailed information regarding the extraction of timbre features, please refer to [source reference]. It's worth noting that our timbre features consisted of 4 spectral attributes, all derived from the acoustic features. We use a Arousal and Valence dimensional approach for selecting features. Arousal and valence are two fundamental dimensions used to describe the emotional content of speech. We select energy, pitch, intensity as an arousal feature and Spectral centroid, Spectral spread, Spectral crest, Spectral energy as a valence features as an input. Arousal represents the level of emotional intensity, ranging from calm or low arousal to highly excited or high arousal. Valence, on the other hand, refers to the emotional polarity, spanning from negative emotions such as sadness and anger to positive emotions like happiness and excitement. These dimensions provide a structured way to quantify and analyze the emotional nuances within speech, enabling more comprehensive sentiment and emotion analysis in various applications, including speech emotion recognition and affective computing.

4.2 Experimental Setup

Our model's architecture presents an innovative blend of the Transformer framework and a novel two-dimensional feature representation. This distinctive combination is harnessed to decode the intricate emotional cues embedded within speech signals. We introduce three distinct architectural paradigms, each meticulously designed to unravel and express the nuanced emotions conveyed through speech. These architectures not only distinguish emotions but also provide insights into their composition.

4.2.1 Singular Transformer Paradigm:

The first architectural design, the Singular Transformer Paradigm, lays the foundation for our exploration. It commences by enhancing each of the seven audio features with positional encoding. This additional context is crucial for the model to understand the sequence of features over time, an essential aspect for capturing emotions' dynamic nature.As these encoded features progress through the Transformer's encoder, a pivotal mechanism called multi-head attention comes into play. Here, each feature is split into two distinct self-attention layers, a unique approach that scrutinizes different aspects of the feature's emotional content. These layers collaborate to emphasize relevant dimensions, thus heightening the model's sensitivity to subtle emotional cues.The iterative process of Transformer encoder layers refines these representations. Multi-head attention and feedforward neural networks work in harmony, modifying feature representations. These iterations culminate in a linear layer that further transforms the features into a suitable format for emotion classification. Finally, the softmax activation function brings forth emotion predictions, expressed as probabilities.

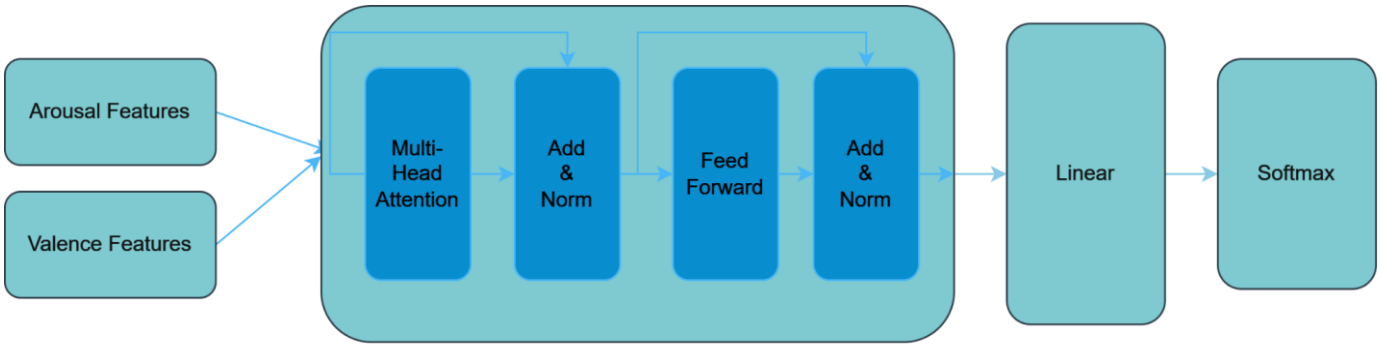

Figure 9. Singular Transformer model

4.2.2 Arousal and Valence-Centric Transformers:

The second architectural configuration centers around the Arousal and Valence-Centric Transformers, two models focusing on distinct emotional dimensions. Out of the seven audio features, three contribute to arousal and four to valence. This deliberate division creates the groundwork for two specialized Transformers, each catering to the exploration of a specific emotion dimension.These Transformers echo the multi-head attention's concept, employing self-attention layers to decipher unique aspects of arousal and valence. Their outputs encapsulate the essence of their respective dimensions. When these outputs converge, a harmonious blend emerges, mirroring the complex interplay of emotions in human expression. This fusion is then guided through a linear layer, refining it for the subsequent emotion prediction step. The softmax activation provides the final emotional landscape in the form of probability distributions.

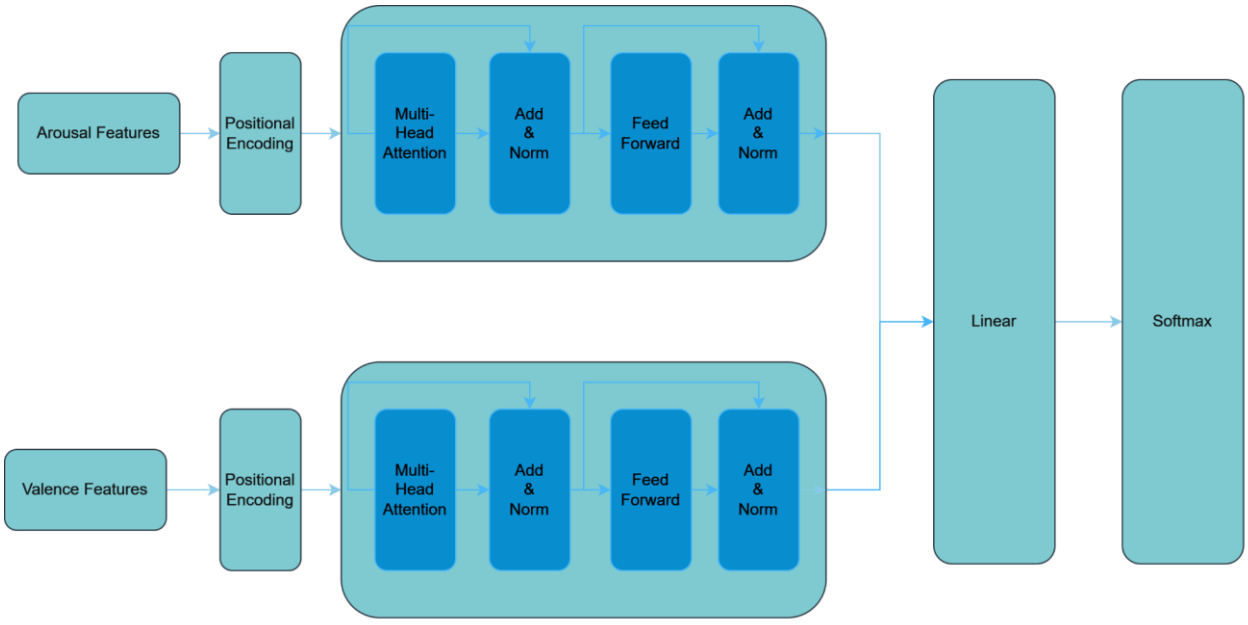

Figure 10. Arousal and Valence-Centric Transformers model

4.2.3 Holistic Fusion of Features:

Our third architectural approach, the Holistic Fusion of Features, represents the culmination of our architectural exploration. It goes beyond the segregation of arousal and valence dimensions, uniting their outputs into a single representation. This unified representation encapsulates the amalgamation of emotional dimensions, creating a holistic understanding of the emotional landscape.This amalgamated representation embarks on a fresh journey through a dedicated Transformer encoder. Here, the feature interactions, previously distinct, intertwine and enrich one another. Multi-head attention layers continue to unravel nuances, culminating in a linear layer that molds the representation for the final step. As the softmax activation envelopes the output, it imparts a life-like quality to the model's emotional predictions.

In summary, our model architectures blend the power of Transformers with the subtleties of two-dimensional features, providing a nuanced framework for speech emotion recognition. Through these architectural designs, we delve into the intricacies of multi-head attention, the specialization of Transformers, and the orchestration of holistic emotional fusion. By skillfully integrating these elements, we unravel a pathway to decode emotions hidden within the audio's narrative.

5. Conclusion

In our research, we've explored the intriguing connection between the amount and variety of data used and the effectiveness of our speech emotion recognition models. While we concentrated on a modest collection of 666 emotion-infused speech samples, the potential for enhancing our models becomes evident when considering the expansion of our dataset. By increasing the number of samples and embracing a broader range of emotions like fear, disgust, and surprise in addition to neutral, happy, sad, and angry, we open doors to improved model performance. This expansion is about more than just numbers; it's about giving our models a deeper understanding of the intricate emotions expressed through speech.

In our exploration to enhance the accuracy of speech emotion recognition, we engaged in a systematic evaluation of different Transformer architectures. The key objective was to unravel the intricate relationship between model complexity and its ability to accurately classify emotions in speech.

Beginning with the 1 Transformer architecture, we witnessed a training accuracy of 80%. This signifies the model's adeptness at learning from the training dataset, comprehending the underlying emotional nuances. Moreover, this prowess extended beyond the training data, as evidenced by both validation and test accuracies of 74% and 72%. The precision and recall remained at 74% and 72% respectively. The model seems to exhibit a decent level of performance when confronted with new inputs, suggesting that it is somewhat responsive to the intricacy of its design. We can observe the balance between complexity and adaptability in this context, as the inclusion of the extra Transformer component might have resulted in the model becoming too specialized for the training data, which could lead to overfitting.

Transformer ONE

	precision	recall	f1-score	support
0	0.60	0.78	0.68	132
1	0.75	0.60	0.67	132
2	0.84	0.85	0.85	121
3	0.77	0.64	0.70	85
accuracy			0.72	470
macro avg	0.74	0.72	0.72	470
weighted avg	0.74	0.72	0.72	470

Transformer TWO

	precision	recall	f1-score	support
0	0.58	0.81	0.68	132
1	0.80	0.77	0.78	132
2	0.80	0.63	0.70	117
3	0.82	0.60	0.69	84
accuracy			0.71	465
macro avg	0.75	0.70	0.71	465
weighted avg	0.74	0.71	0.72	465

Transformer THREE

	precision	recall	f1-score	support
0	0.79	0.86	0.83	132
1	0.89	0.64	0.74	132
2	0.91	0.97	0.94	119
3	0.77	0.93	0.84	90
accuracy			0.84	473
macro avg	0.84	0.85	0.84	473
weighted avg	0.85	0.84	0.83	473

Figure 11. Classification report.

Based on the comparison of precision, recall, and true percentage scores for the three transformer-based models with different numbers of transformers (one, two, and three), we can evaluate their performance as follows:

- **One Transformer Model:**
  - **Precision:** Gradually improves with increasing threshold values, indicating that as you increase the threshold, the model's positive predictions become more precise.
  - **Recall:** Similar to precision, it also improves with increasing thresholds, indicating that as the threshold increases, the model correctly identifies more true positives.
  - **True Percentage:** Initially high, meaning a significant portion of the data has a confidence score above 60. However, it decreases as the threshold increases, suggesting that with stricter thresholds, less data is considered positive.

- **Two Transformers Model:**
  - **Precision:** Similar to the one transformer model, it gradually improves with increasing threshold values.
  - **Recall:** Also shows improvement with higher thresholds.
  - **True Percentage:** Like precision and recall, it follows a similar pattern but tends to be slightly lower than the one transformer model, indicating a slightly stricter model.

- **Three Transformers Model:**
  - **Precision:** Starts at a high value and remains consistently high across different thresholds, indicating a high degree of confidence in positive predictions.
  - **Recall:** Maintains a relatively high value across thresholds, suggesting that the model consistently identifies true positives.
  - **True Percentage:** Remains high even with stricter thresholds, indicating that a significant portion of data consistently has high confidence scores.

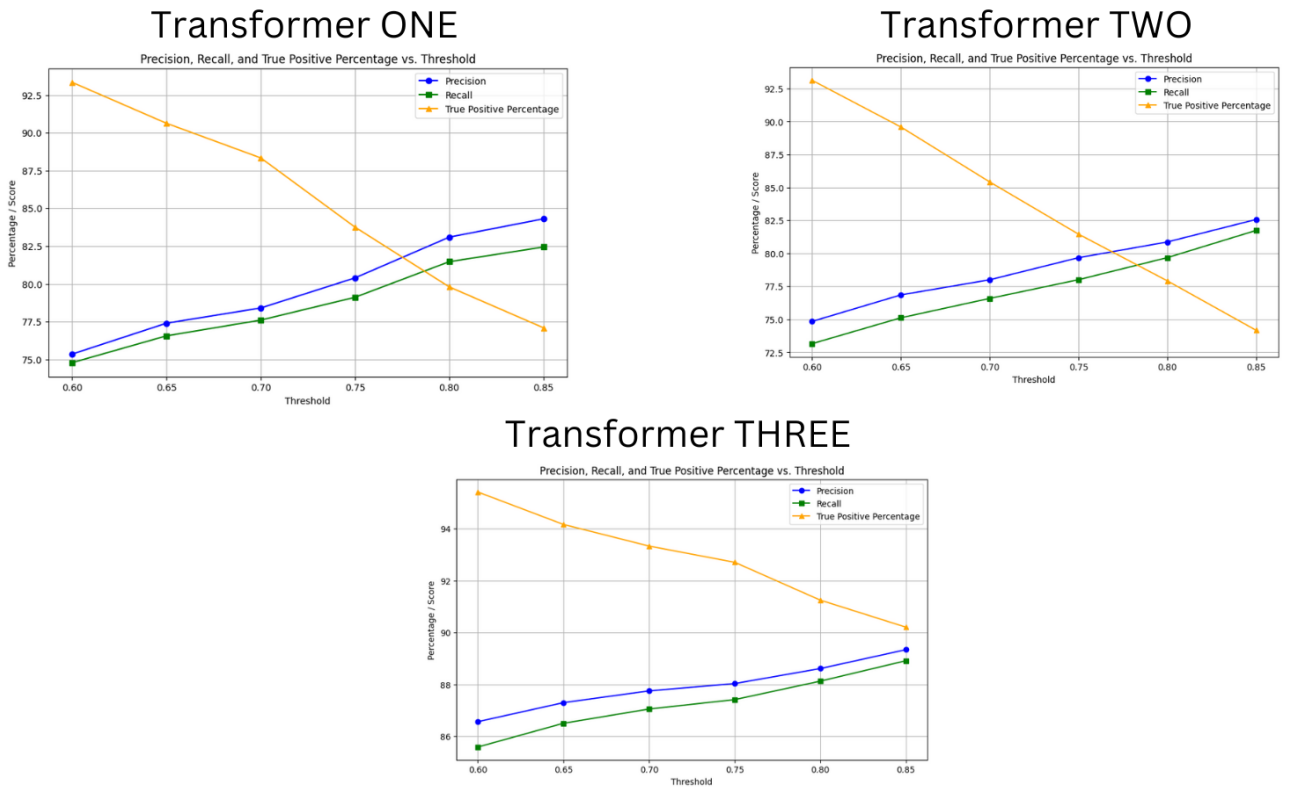

Figure 12. Precision, Recall and True Positive vs Threshold

Evaluation, The three transformers model consistently outperforms the one and two transformers models in terms of precision, recall, and true percentage. It achieves the highest precision and recall scores, indicating a better balance between positive predictions and true positive identifications. The true percentage score remains consistently high, even with stricter thresholds, suggesting that it maintains high confidence in its predictions across different confidence score levels.

Transitioning to the 2 Transformer configuration, we noted a training accuracy of 84%. While marginally reduced, this accuracy still indicates the model's effective learning from the training dataset. However, the validation and test accuracies, settling at 70% and 71% respectively, the precision and recall remained at 74% and 71% respectively, pose a question regarding generalization ability. The model appears to perform moderately well on novel inputs, indicating a degree of sensitivity to architecture complexity. The trade-off between complexity and generalization is evident here, as the additional Transformer potentially led to overfitting on the training data**.**

Classifier Architecture	Training Accuracy	Validation Accuracy	Test Accuracy	Precision	Recall
One Transformer	80%	74%	72%	74%	72%
Two Transformer	84%	70%	71%	74%	71%
Three Transformer	90%	86%	84%	85%	84%

Figure 13. Accuracy rate in all experiments

Lastly, the 3 Transformer architecture revealed intriguing insights. With a training accuracy of 90%, the model exhibited a commendable understanding of the training dataset's emotional characteristics. However, similar to the previous architectures, the validation and test accuracies remained at 86% and 84%, respectively. the precision and recall remained at 85% and 84%, respectively. This suggests that the model's capacity to generalize its learning to new, unseen instances is robust. Here, the architecture's complexity seemingly hindered its generalization prowess. The heightened intricacies might have caused the model to capture more noise than actual emotional patterns, resulting in limited performance on unseen data. Collectively, these findings underscore the importance of striking a balance between model complexity and generalization.

So, the Transformer 3 architecture stands out as a compelling choice, boasting impressive training accuracy while demonstrating consistent and robust performance across validation and test datasets thanks to the implementation of cross-validation. This indicates that the architecture's inherent design, coupled with its ability to distill pertinent emotional cues, is pivotal in achieving accurate speech emotion recognition. Additionally, the use of regularization techniques has played a crucial role in preventing overfitting and enhancing generalization. As we navigate the evolving landscape of emotion recognition, these insights pave the way for future advancements. Fine-tuning model architectures, with careful consideration of regularization and optimizing techniques, to strike the right balance between complexity and generalization is imperative for achieving superior performance in real-world scenarios. Ultimately, our research advocates for a thoughtful consideration of architectural configurations and regularization methods to unravel the intricate tapestry of emotions embedded within human speech.

In conclusion, Based on the provided performance metrics and your evaluation criteria, the three transformers model appears to be the best-performing model among the three. It demonstrates better precision, recall, and true percentage scores, making it a more reliable choice for speech emotion recognition.

Our research has thus not only contributed to advancing the field of speech emotion recognition but has also provided valuable insights into optimizing model confidence thresholds for improved real-world applicability.

REFERENCES

[1] Discriminating Emotions in the Valence Dimension from Speech Using Timbre Features Anvarjon Tursunov 1, Soonil Kwon 1,* and Hee-Suk Pang 2, Appl. Sci. 2019, 9, 2470; doi:10.3390/app9122470

[2] Perceptual audio features for emotion detection Mehmet Cenk Sezgin, Bilge Gunsel* and Gunes Karabulut Kurt, Sezgin et al. EURASIP Journal on Audio, Speech, and Music Processing 2012, 2012:16

[3] Jay Alammar Visualizing machine learning one concept at a time. @JayAlammar on Twitter The Illustrated Transformer.