ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Образовательная программа «Фундаментальная и компьютерная лингвистика»,
по направлению 45.03.03 Фундаментальная и прикладная лингвистика

Андрушко Тарас Иванович

Распознавание эмоций в аудиозаписях
Recognition of Emotions in Audio Recordings

Выпускная квалификационная работа студента 4-го курса

Академический руководитель
Доцент Школы лингвистики
Ю.А.Ландер

Научный руководитель
Приглашенный преподаватель
О.А. Сериков

Москва, 2023

1. Introduction

Emotions play a key role in the decisions we make, which is why they are so interesting from a research perspective. Knowing and understanding a person's emotions we can, for example, adapt to the conversation, detect a loss of trust on the part of the interlocutor and, if necessary, change the tone and direction of the dialogue. The importance of such things is well described in work "*The Role of Trust in Proactive Conversational Assistants" (Kraus M. Wagner N. Callejas Z. & Minker W. 2021)*.

Linguists can use speech emotion recognition to study how emotions are expressed across different languages and cultures. For example, they can examine how certain emotions are expressed differently in other languages or how cultural norms influence the expression of emotions through language.

Nowadays systems for emotion detection in oral speech are complicated and a bit inaccurate, and also extremely expensive to work with, as described in detail in "*The Conversational Interface: Talking to Smart Devices"(McTear M. Callejas Z. & Griol D. 2016).*

In spite of this, some models can even work in real-time, such as the one described by Anvarjon in his paper "*A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features"(Anvarjon, T., Mustaqeem, & Kwon, S. 2020)*.

In addition, in this area, an attempt has been made to recognize emotions in speech, not only with the help of audio features but also with the help of facial features. The paper "*A Proposal for Multimodal Emotion Recognition Using Aural*

Transformers and Action Units on RAVDESS Dataset" (Luna-Jiménez C. Kleinlein R. Griol D. Callejas Z. Montero J. M. & Fernández-Martínez F. 2021) used two models connected by late fusion strategy.

All of these scientific works somehow set themselves the goal of improving the accuracy of automatic emotion recognition algorithms, but at the same time, they did not go much deeper into what parameters are key. This paper will raise the question of which parameters and features are key and why.

However, this work is inspired by the work "*Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions*." *(Rodero E. 2011).* In this paper, the author answers the question of what is more important when expressing emotions, the pitch level or the intonation contour. As a result of this, the question arose, what is more, important when a machine learning algorithm recognizes the emotions: intonation contour or pitch level? In Rodero's work[1], the result is that intonation contour is more important than pitch level during expression and recognizing of the emotions.

In this light, that this particular paper is extremely important because it will give us an understanding of which features are most important in recognizing emotions not only in terms of the accuracy of the recognition model but also from the perspective of one of the most exciting parts of linguistics - phonetics. By the end of the work, we will be able to say with confidence how to improve existing algorithms for recognizing emotions.

The current work consists of several stages, which include:

- Literature review - a review of relevant literature on the topic.

[1] *Rodero, E. (2011). Intonation and emotion: influence of pitch levels and contour type on creating emotions. Journal of voice, 25(1), e25-e34.*

- Theory - explanation of the main concepts and concepts used in this work.
- Methodology - description of the methods by which experiments will be carried out in this work.
- Results
- Analysis of results
- Conclusions and the Prospects for subsequent research

The main hypothesis: intonation contour is more important than pitch level in emotion recognition tasks.

The main goal of this work is to check with machine learning methods whether the intonation contour is more important than the pitch level during emotion recognition tasks.

T**wo additional goals**: 1. Find out which pairs of emotions will be recognized incorrectly more often than others. 2. Find out which emotion is the most difficult emotion to recognize. 3. Prove that intonation contour and pitch level better work together during emotion recognition tasks. 4. ‘Sad’ is the most recognizable emotion.

2. Literature Review

In order to fully understand what this work is about and how the methodology will be structured, then first it is worth getting acquainted with the key concepts and notions that are extremely important in writing this work. To do this, a review of the literature will be given which will best explain these concepts.

2.1 Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions

Since the main hypothesis is built around the work of Rodero, E., I think it is worth giving a more detailed description of what happens there.

In this work, as mentioned earlier, the author sets himself the task of proving that intonation plays a key role in recognizing emotions. In addition, he compares which concept is key in recognizing emotions - pitch level or contour type.

To test his hypotheses, a researcher conducted an experimental study in which they assigned specific emotions (joy, anxiety, sadness, and calmness) to different pitch patterns, including pitch levels and contour types. After that, a group of one hundred people are shown patterns and recorded emotions and asked to rate emotions on a bipolar scale with opposite pairs.

The main achievement of this study is the conclusion that both pitch level and contour type affect the expression of emotions. Although the author noted the contribution of both parameters, he also clarified that the type of contour is more important.

In addition, the author also said that when perceiving emotions, respondents most often correctly recognized the emotion of sadness, followed by sadness, followed by joy, calmness, and anxiety.

2.2 *Machine learning basis*

The topic of machine learning, is not new, and already has many articles that describe it. One article that is often recommended as a base is *“A Few Useful Things to Know About Machine Learning” (Domingos P. 2012)*. It was published

in the Communications of the ACM journal in 2012 and has since become a popular reference in the machine learning field

Machine learning algorithms - are designed to learn from data, in order to make predictions or take actions based on patterns or relationships that are not explicitly programmed into the algorithm. This process typically involves training the algorithm on a large dataset and then using it to make predictions or decisions on new data.

2.3 *Contemporary Models and Intonation Contour*

If we are talking about working with ready-made models in the field of emotion recognition, then there are several factors to consider. Firstly, there is a model for each task, and if it is suitable for one thing, it is unlikely to work well with something else. Secondly, it is very important to understand on which dataset the models were trained since this can be critical when assessing quality.

This information can be found by reading the work *"Speech emotion recognition methods: A literature review" (Basharirad B. & Moradhaseli M. 2017)*.

In this paper, the authors review various machine-learning models that have been used to recognize emotions in speech. The most useful thing is that they discussed the advantages and disadvantages of each model and conducted a comparative analysis of their work on various datasets, thus identifying a universal model.

In addition, in their work, we can find a discussion of various features used for training (including spectral, prosodic, and voice quality features).

In our case, what is also important for us are audio features, the extraction process of which is described in *“Acoustic Emotion Recognition: A Review of the State of the Art” (Schuller B. W. 2012)*.

In his study, Björn W. Schulle reviews the various features that are present in the audio signal including spectral, prosodic, and voice quality features. They also discuss the different techniques for extracting these features, such as Mel Frequency Cepstral Coefficients (MFCCs), pitch tracking, and spectral flux.

In addition, in his work, he also discusses the complexity of such a task as recognizing emotions in speech and says that training requires large datasets and resources.

The work of Björn W. Schulle covers the issue of acoustic emotion detection, but not emotion from speech. The issue is covered in more detail in work *“Feature extraction algorithms to improve the speech emotion recognition rate” (Koduru A. Valiveti H. B. & Budati A. K. 2020)*.

The authors use MFCC as a basic feature. Mel-frequency cepstral coefficients (MFCC) are a type of feature commonly used in speech and audio signal processing. They are a representation of the short-term power spectrum of a sound signal, based on the principle that the human auditory system is more sensitive to changes in frequency at lower frequencies than at higher frequencies.

The final product of their work is a trained model on the RAVDESS dataset which can recognize 4 emotions: anger, sadness, joy, and neutral.

And although the results of the authors were satisfactory the Decision tree model gives 85%, they rather poorly explain the choice of parameters and the choice of features, besides the RAVDESS dataset contains 8 emotions, but only 4

were used in their work (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

As a consequence, I think it would be extremely useful to conduct a series of experiments with the features, in order to find out what effect each of them has when recognizing each of the emotions.

Apart from the mediocre explanation of the choice of features for machine learning, the authors of these papers do not seem to me to take into account an important feature of speech, namely the intonation contour.

The intonation contour is described for the Russian language in the book *“Sovremennyj russkij literaturnyj yazyk: fonetika, orfoe`piya, grafika i orfografiya” (Knyazev S. V. & Pozharitskaya S. K. 2009)*[Modern Russian literary language: phonetics, orthoepy, graphics, and orthography]. The intonational contour is, in turn, determined by the intonational construction. The intonation construction (IC) is the type of ratio of tone, timbre, intensity, and duration, capable of contrasting the incompatible in one context semantic differences of statements. In their work, they distinguish 7 types of intonation constructions, each of which differs in the previously mentioned parameters. It seems to me that these parameters highlighted in the work of Knyazev and Pozharitskaya can be applied in practice when teaching the model, which can improve the effectiveness of emotion recognition.

3. Theory

In order to fully understand this work, it is worth taking the time to describe the main concepts and terms that will be used in it. One of the most important and complex definitions in this work is emotion. It is worth looking at the concept of

emotion from different points of view and understanding what characterizes, for example, anger and sadness, and for this, you will have to figure out what is filling or composing emotions. In addition, other important definitions are intonation contour and pitch level, the components of which will be further used for machine learning.

3.1 Defining emotion

There are several popular theories about what emotion is.

One of them is the James-Lange theory (William James and Carl Lange), according to which emotion is the result of psychological reactions to external stimuli. In their theory, they also suggest that we experience emotions as a result of processes occurring in our body (For example, we feel fear when the heart beats faster).

Another popular theory about emotions is called Basic Emotions Theory by Paul Ekman. In his theory, he identifies six basic emotions (happiness, sadness, fear, anger, surprise, and disgust) which, in his opinion, are universally readable and reproducible through specific facial expressions.

And the most important and basic theory, at least for this work, is the Differential Emotions Theory proposed by Carroll Ellis Izard.

Carroll Ellis Izard, an American psychologist, and specialist in the field of emotions, in his book Human Emotions, defines emotion as follows:

Emotion - is a process of producing physiological, experiential, expressive, and behavioral responses that an individual produces in response to stimuli. In addition, he emphasizes the importance of the fact that emotion is not only a

reaction to a stimulus but also a subjective assessment of a situation to the result of an external stimulus.

According to Izard, emotions are characterized by three components:

1. Experienced or perceived in the psyche by the sensation of emotion
2. Processes occurring in the respiratory, digestive, nervous, endocrine, and other systems of the body.
3. Observed expressive complexes: gestures, changes in the face, changes in intonation, etc.

Of all the theories described above, the Differential Emotions Theory is the best representative for this work, because, unlike the other two, it gives us an understanding of what emotions are expressed through voice, which is key to this work.

Therefore, from the theory of Carroll Ellis Izard, I propose to derive the term emotions specifically for the current work, which will sound like this:

Emotion is a response to a stimulus expressed verbally by a human and marked with a change in the tone of the speaker.

But in addition to defining what emotion is, it is also worth understanding what emotions exist and how we define them.

3.2 Content of emotion

The definition of emotion is a complex and multifaceted process that includes the following parts: Cross-cultural studies, Emotion recognition studies, Acoustic analysis, and Expert consensus.

According to these parts, scientists first conduct a study of emotions among different cultures to identify universal emotional expressions, after which various experiments come into play in which people listen to and analyze various emotions, it is precisely in these experiments that patterns of emotions are revealed that are later confirmed by acoustic analysis. Last, but not least, is the emotion expert's judgment of what is what.

Attempts to describe emotions have been made in many scientific works, I propose to briefly consider some of them, since each is interesting in its own way.

The already mentioned Carroll Ellis Izard also worked on the classification of emotions and singled out 10 primary emotions such as joy, interest-excitement, surprise, sadness, anger, disgust, contempt, fear, shame, and guilt. According to Izard's theory, these primary emotions are distinct and serve adaptive functions. He argues that each primary emotion has a unique neural and physiological profile, expressive behaviors, and subjective experience. Izard also emphasized the role of facial expressions in the communication of emotions, suggesting that certain facial expressions are universally associated with specific emotions.

In *"Handbook of Emotion" (Lewis M. Haviland-Jones J. M. & Barrett L. F. (Eds.). 2010)* we can find information about how emotions have already been described by various researchers and the general approach to classifying emotions and studying them.

In general, if you look at the studies described in *“Handbook of Emotion”*, as well as the conclusions of the authors, then you can try to summarize the fact that at the moment there are 9 main emotions Happiness / Joy, Sadness, Anger, Fear, Surprise, Disgust, Contempt, Embarrassment, Love/Affection.

Each of these emotions has its own characteristic which is reflected in the voice of the subject.

The characteristics of these emotions are as follows:

1. **Happiness/Joy**: This emotion suggests a positive experience or enjoyment, most often expressed by raising the tone and energetically increasing speech patterns.
2. **Sadness**: Described by a feeling of loss, and disappointment and can be expressed by subdued and slow speech rate, lower pitch, decreased intensity, and, in addition, a monotonous voice.
3. **Anger**: includes sensations such as frustration, hostility, and irritability, most often expressed by raising the intensity of the voice, pitch, and speed of speech and also more forceful articulation.
4. **Fear**: Feelings of fear are apprehension, anxiety, or threat expressed through a confused delivery of the speech, which is also accelerated, but at the same time with frequent interruptions.
5. **Surprise**: It is the Emotion of something unexpected, expressed through a sudden change in volume, or speech rate. Often accompanied by sighs.
6. **Disgust**: It is a repulsive emotion in relation to something offensive or unpleasant expressed by an increased pitch, and voice quality, also the voice slows down and emphasizes certain points.
7. **Contempt**: Includes feelings of neglect and superiority expressed by lowering the pitch level narrowed vocal range.
8. **Embarrassment**: Feeling that arises from shame and is expressed through a hesitant and self-deprecating speech pattern, in addition, the intensity of the voice decreases and there are many pauses and delays in speech.

9. **Love/Affection:** A feeling of warmth and affection and care, which is expressed through a decrease in the speech rate, which at the same time becomes gentle and smooth.

However, in this work, a slightly different classification of emotions will be used, which contains 8 emotions (neutral, calm, happy, sad, angry, fearful, surprise, and disgust) 6 of them, in accordance with the work of Ekman P, Sorenson ER, Friesen WV. *Pan-cultural elements in facial displays of emotion* are considered basic or fundamental. This choice of emotions will be explained in more detail in Section 4 Data, which will also contain information about the datasets.

After the concept of emotion has been defined we can go directly to what emotion consists of.

In this paper, we are interested in the phonetic component of emotions. In general, there are 7 phonetic components of emotions, namely:

1. **Pitch**: Emotions affect the level of the pitch and its contour, for example, a high pitch is associated with the emotion of surprise, while a low pitch is associated with sadness.
2. **Intensity**: Emotions affect the volume of speech, for example, strong emotions increase the intensity of the voice.
3. **Speech Rate**: Emotions affect the change in the pace and speed of speech, for example, Increased pace with joy.
4. **Voice Quality**: Emotions affect the quality and character of the voice, for example, amplification with anger leads to the fact that the voice becomes harshly.

5. **Articulation and Phonation**: How our voice sounds will largely determine our emotions, for example, intense emotions lead to changes in articulation and the purity of speech.
6. **Prosody**: Emotions are critical in influencing prosodic features. Prosody refers to the melodic and rhythmic aspects of speech, including stress, intonation, and rhythm.
7. **Vocalizations**: Sighs, sobs, and other non-verbal vocalizations can give us insight into certain emotions and serve as indicators of them.

Of all the above elements, we are primarily interested in the first and sixth points, since they contain the very components of emotion, thanks to which we will test the main hypothesis of this work. These components are intonation and pitch, and more specifically, intonation contour and pitch level.

David Crystal in his works *"Relative and Absolute in intonation analysis" (Crystal D. 1971)* reveals the concepts of intonation and pitch and gives them the following definitions:

Pitch: refers to the "highness' and "lowness" of sound. Pitch defines itself as the frequency of the sound waves. In terms of linguistics, pitch corresponds to the fundamental frequency (F0) which is representing the rate of vocal chord vibration.

Intonation: basically it is patterns of pitch changings in speech. Intonation is responsible for indicating sentence types such as statements or questions, expressing emotion, or highlighting certain words or phrases.

Now, knowing the concepts of intonation and pitch, we can consider the basic technical concepts for this work, namely pitch level and intonation contour

3.3 Intonation contour

In Levis, J. M., & Wichmann, A. *"English intonation–Form and meaning, The handbook of English pronunciation" (Levis J. M. & Wichmann A. 2015)*, the authors describe different approaches to what intonation and intonation contours are, as well as what traditions exist.

There have been many different attempts to describe the melody of speech by intonation. For example, the first attempt looked like general notes for music, and the description of intonation was done on the staves. But the stave approach seems unfortunate because the voice does not have the same fixed duration as a note.

Another system of interpretation was proposed in 2006 in Wales, where major and minor dots are used to represent intonation, as in the example below.

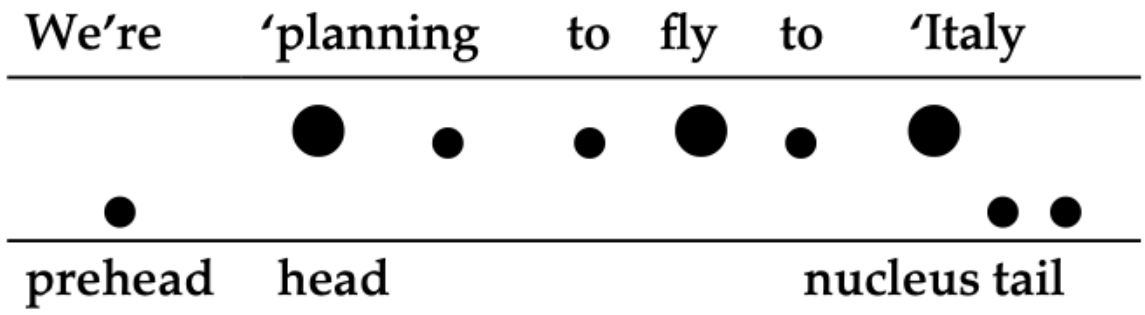

Figure 1. Intonation contour example('Levis, J. M., & Wichmann, A. English intonation–Form and meaning, The handbook of English pronunciation')

An important intonation feature is that not all the elements of melody have the same value. And during working with intonation we need to take into account that the pitch of stressed syllables tends to be more important than those associated with unstressed syllables.

According to the British system of analysis, there are groups of tones, and each of the groups contains at least one stressed syllable, which carries a change in pitch. If there is more than one stressed syllable, then the last one is called **‘nucleus’** and the change of tone is this case known as **‘nuclear tone’.** A phrase or group of tones may also contain additional stressed syllables, the first of which is called onset. The "head" of a phrase or group of tones refers to the stretch from the onset to the nucleus, while any unstressed syllables preceding it are considered **prehead** syllables. The tone group follows a structure of **[pre head, head, nucleus tail]**, with the nucleus being the only necessary component. This pattern is demonstrated in the provided illustration.

Two different traditions:

In the **British tradition**, nuclear tones are represented as contours and can be depicted iconically using simple keystrokes such as [**fall, rise /, fall-rise /, rise-fall/, level -**], which are placed before the syllable where the contour starts.

In **American approaches,** pitch contours are typically broken down into distinct levels or targets, and the resulting contour is considered the interpolation between these points. Put differently, a falling contour is the pitch movement between a high target and a lower target.

Here are some examples of different commonly recognized contours:

1. **Falling contour**: a falling contour demonstrates the completeness of the statement, usually characterized by a high pitch at the beginning and a gradual decrease towards the end of the phrase.

   Example: "I saw her yesterday."

2. **Rising contour**: a rising contour is often used in a yes/no question. Starts with a low pitch and then gradually rises.

   Example: "Did you see her?"

3. **Rising-falling contour**: it often occurs in wh-questions, indicating both the inquiry and a potential contrastive element. This type combines a rising and falling pitch movement.

   Example: "Where did you go?"

4. **High-rising contour**: often associated with shock, surprise, and disbelief. A high pitch that only gets higher.

   Example: "You did what?"

5. **Low-rising contour**: used in impolite queries and when enumerating lists. The phrase in these types of contours starts with a low pitch and slowly rises.

   Example: "I need milk, bread, and eggs."

6. **Fall-rise contour**: it begins with a fall and then rises toward the end. This contour can convey a range of meanings, such as uncertainty, hesitation, or sarcasm.

   Example: "Oh, that's just great."

Modern technologies that were developing during the 20th century allow researchers all over the World to take a fresh look at the intonation question. Often researchers use 3 main visualizations to study intonation, namely: **the waveform spectrograms, and in particular the fundamental frequency trace**

Working with the results of an acoustic intonation analysis is a challenging task. On the one hand, it requires a basic understanding of acoustic phonetics and the capabilities of the software being used. On the other hand, it is important to acknowledge that our perception of language is not solely determined by sound waves, but also by our knowledge of language and the workings of our brain. In other words, while computer software can reveal the nature of the sounds we hear, it cannot demonstrate the linguistic interpretations that we make of them.

English intonation is one of several prosodic features that serve various linguistic purposes, such as:

1. Highlighting important words or syllables.
2. Conveying the connection between consecutive phrases, as determined by the pitch contour used.
3. Signaling the boundaries between phrases.

Generally speaking, **intonation contour** is a change in pitch behavior or variation in the fundamental frequencies of speech. The intonation contour is a demonstration of the melodic structure in the utterance and has a key role in conveying the meaning, mood, and emotion of the language.

3.4 Pitch level

In the field of phonetics, in addition to the intonation contour, the level of pitch is also actively studied. Pitch as a phonetic phenomenon is well described in the works of Kenneth L. Pike, in particular in "A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion".

Pike suggests that pitch is on par with other prosodic features such as rhythm and stress forms an integral part of language structure. He argued that the pitch has such linguistic properties as the expression of emotions.

Pike also suggested that a different pitch reflects different emotional moods in communication. For example, a high level of pitch is more often associated with excitement, enthusiasm, or happiness, while a low level may be associated with sadness, seriousness, or anger. He argues that changes in pitch level, along with other prosodic features like tempo and rhythm, contribute to the overall emotional impact of speech.

Thus, in phonetics, pitch is the ratio of the height and frequency of a sound wave, and the pitch, in particular, determines how the sound is perceived and whether it will be high or low for the listener.

3.5 Intonation contour vs Pitch level

Although pitch level and intonation contour are related concepts in the field of studying phonetics, there are still significant differences between them, the distinctive features of each of the concepts are listed below.

Pitch level:

- Pitch level refers to the perceived relative height or frequency of a sound wave, specifically the fundamental frequency (F0) of vocal fold vibration.

- Represents the overall pitch of the sound and whether it is perceived as high or low

- Pitch level is usually measured and analyzed in terms of frequency (Hertz)

- An inherent characteristic of any sound and can vary between different speech sounds, vowels, consonants or words

- The pitch level can contain information about stress, sentence boundaries, and emotional coloring.

Intonation contour:

- Changing the pitch pattern in a sentence, or phrase

- Shows the melodic structure of speech (rise and fall of the pitch over time).

- The intonation contour is described in terms of the movement of the pitch, its direction, and form and represents information about the communication functions of speech.

- This is the result of a combination of the pitch level with its fall, rise, or other movements.

- The intonation contour plays a key role in the designation of questions, and statements and contains information about emotions and pragmatic meanings.

Thus, the intonation contour is a continuum concept that contains information about the change in pitch throughout the entire utterance, while the pitch level contains information about the pitch or frequency of a single sound. At the same time, both concepts contain information about emotional color.

4. Data

In this work, as previously mentioned, 4 datasets taken from Kaggle were used. In this section, there will be a detailed description of each of them, in

particular, which methods were chosen for the selection of emotions, the content of the dataset, and the way it was collected.

4.1 General information

The final dataset for this work is a compilation of 4 datasets (TESS, RAVDESS, CREMA, SAVEE) that contain audio recordings of various phrases that reflect one of the eight emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised). The pandas table consisting of 3 columns (Labels, Path, Gender) stores information about the path to the audio file with the recording, the gender of the speaker, and the immediate emotion reflected in the audio recording.

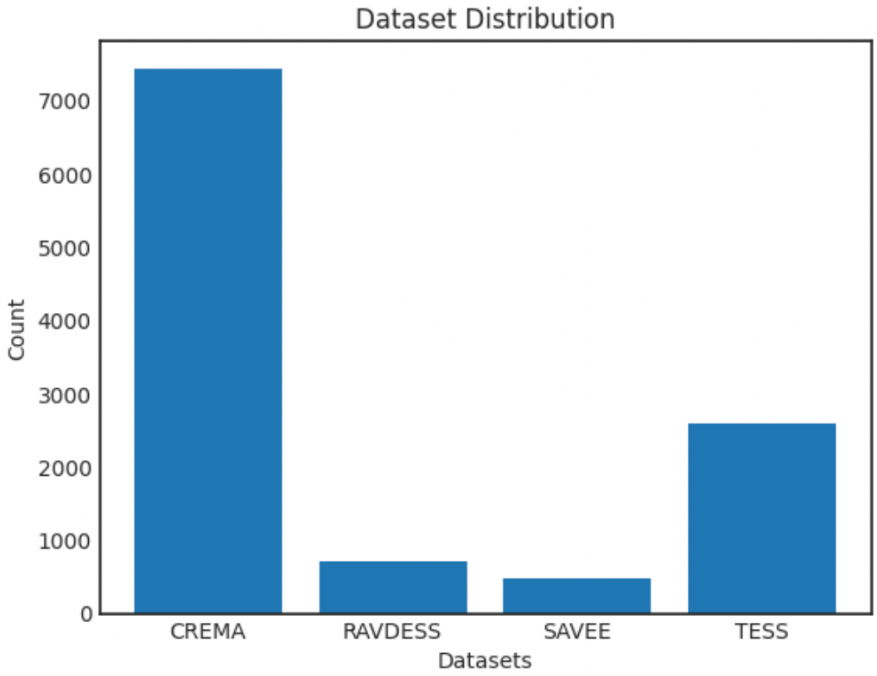

Figure 2. Dataset Distribution

You can see the distribution of audio recordings in datasets in Figure 2. The dataset with the largest amount of data is CREMA, TESS, Ravdess, and SAVEE.

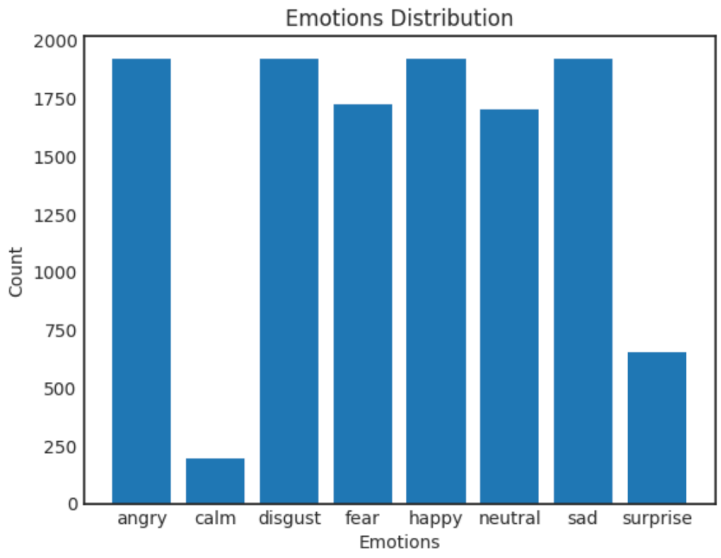

Figure 3. Emotion Distribution

In Figure 3, you can see how the emotions were distributed. In general, almost all emotions are equally distributed, but out of 8 emotions "calm" and 'neutral' are strongly knocked out of which there are about 200 and 600 pieces, it is important to take this into account when analyzing the results.

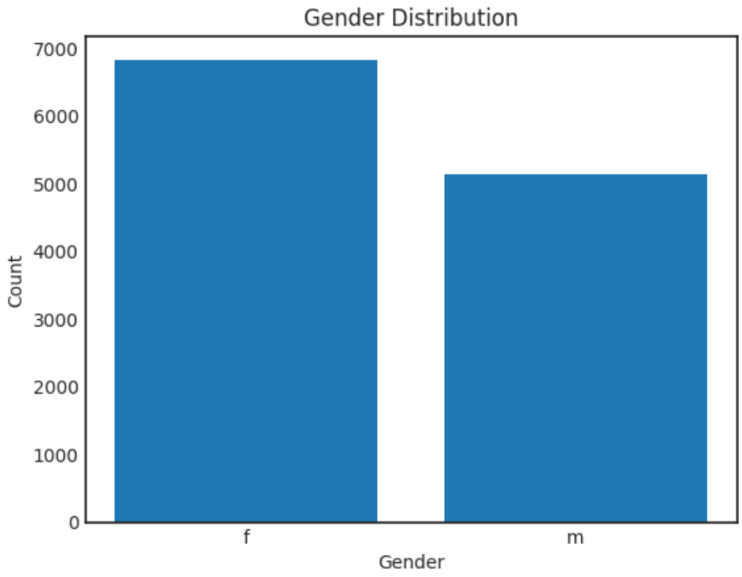

Figure 4. Gender Distribution

Figure 4 shows the distribution between male and female sexes in the dataset, in general, there are slightly more females, due to the fact that the TESS dataset consists exclusively of female audio recordings (for more details, see section 4.3 TESS)

With regards to each dataset, the most sensitive point is how the authors of these datasets defined emotions, and this issue should be discussed in more detail, since it is one of the key ones in this work.

4.2 RAVDESS

RAVDESS is a dataset that contains emotions in speech as well as in songs. The speech emotion dataset was recorded by 24 professional actors, all of whom have neutral North American accents. The dataset includes the previously named emotions neutral, calm, happy, sad, angry, fearful, surprise, and disgust.

Each of the 7356 entries was read at 2 levels and evaluated for emotion validity 10 times in order for the average listener to agree with the given emotion label. In total, there were 247 people from North America who were evaluators of the records and gave a label to the emotions. (However, in this work only a small available on Kaggle[2] piece of this dataset with about 700 audio recordings will be used.)

Of the eight emotions, two were baselines, namely calm and neutral, while the remaining six emotions are fundamental and universal in accordance with the work of Ekman P., Sorenson E.R, and Friesen W.V. Pan-cultural elements in facial displays of emotion. In addition, the authors of the dataset rely on already existing

[2] URL: [https://www.kaggle.com/datasets/barelydedicated/savee-database]

and recognized ones in the scientific world, datasets described in works 1[3], 2[4], 3[5], and 4[6].

4.3 TESS

Toronto emotional speech set (TESS) - Dataset of audio recordings containing various emotions collected by Eu Jin Lok. Unfortunately, unlike the Ravdess dataset, it does not have a publication or a detailed description of what methodology was used to create the dataset. Despite this, this dataset has significant advantages.

In addition to the fact that the emotions of this dataset coincide with the emotions of the previously described Ravdess(except ‘calm’), this dataset also contains only recordings of a female voice (2 professional actresses aged 26 and 64) in good quality. As the author notes, in datasets of this kind, there is often a preponderance in the number of records towards males, this dataset will help us balance the sample and even out the number of female and male speakers.

The records themselves are in **`.wav`** format and each of them are a record of the phrase "Say the word _" where there are various words in place of the dash.

---

[3] Mazurski EJ, Bond NW. A new series of slides depicting facial expressions of affect: a comparison with the pictures of facial affect series. Australian Journal of Psychology. 1993;45(1):41–7.
[4] Lundqvist D, Flykt A, Öhman A. The Karolinska directed emotional faces [Database of standardized facial images]: (Available from Psychology section, Department of Clinical Neuroscience, Karolinska Hospital, S-171 76 Stockholm, Sweden); 1998.
[5] Wang L, Markham R. The development of a series of photographs of Chinese facial expressions of emotion. Journal of Cross-Cultural Psychology. 1999;30(4):397–410.
[6] Kanade T, Cohn JF, Tian Y, editors. Comprehensive database for facial expression analysis. Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat No PR00580); 2000; Los Alamitos, CA: IEEE Computer Society Conference Publishing Services

4.4 SAVEE

Dataset includes 7 emotions and consists of 480 entries. The recordings were made by the actors, and the dataset itself was taken from the Kaggle[7] platform. The dataset was recorded by male speakers aged 27 to 31.

4.5 CREMA

CREMA-D dataset consists of more than 7000 audio recordings. The main advantage of this dataset is not only a large amount of data, but also a lot of different voices of actors, of which there were 91 in this dataset (48 men and 43 women aged 20 to 74). The emotions were determined by the listeners, they were asked to listen to the recording and choose one of 6 emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) after which the final dataset was formed with a label for each audio recording, which was based on the perception of the listeners.

5. Methods

The experiment itself looks like this:

- Choice of comparison parameters (in this case intonation contour and pitch level).
- Extracting features for machine learning from the selected parameters, in this way we get two sets of features.
- Model selection
- Training
- Analysis of results

[7] [URL:https://www.kaggle.com/]

5.1 Defining features

In order to extract features from the intonation contour and pitch level, we should understand what sound characteristics each of the concepts consists of.

As mentioned earlier, the intonation contour is a complex concept, but in general, it is changing the pitch pattern in a sentence, or phrase. One way or another, the intonation contour according to the work[8] of Bolinger, D. is formed by the following parts: pitch movement, pitch range, pitch direction, boundary tones, phrase-final lengthening, and prominence.

1. **Pitch movement**: The pitch of the voice rises and falls during the speech, creating a melodic pattern.
2. **Pitch range**: The range of pitch used in a sentence can vary from high to low, depending on the meaning and emphasis of the sentence.
3. **Pitch direction**: The direction of the pitch movement can be rising, falling, or level.
4. **Boundary tones**: Boundary tones signal the end of a sentence and can be rising, falling, or level.
5. **Phrase-final lengthening**: The length of the final syllable in a phrase is often lengthened, indicating the end of the phrase.
6. **Prominence**: Certain words or phrases in a sentence are emphasized, and the pitch contour will reflect this emphasis.

Pitch level is not such a complex concept as an intonation contour, moreover, it is not a continuum, so there are not so many technical components of it. In this paper, I propose to consider such parts of the pitch level as: average pitch

[8] Bolinger, D. (1982). Intonation and its parts. *Language*, 505-533.

level, pitch variability, pitch contour shape, and pitch range. (theese parts are also mentionet in the main reference of this work[9])

This features has the following explanation:

1. **Average pitch level**: The mean fundamental frequency (F0) of the speech signal over a segment of audio. Higher pitch levels may indicate excitement, happiness, or fear, while lower pitch levels may indicate sadness or anger.
2. **Pitch variability**: The variation of F0 values in the speech signal over time. Greater pitch variability may indicate more emotional expressiveness or variability in an emotional state.
3. **Pitch contour shape**: The shape of the pitch contour over time, which can reveal patterns of rising or falling pitch, such as a rising pitch contour for expressing surprise or a falling pitch contour for expressing disappointment.
4. **Pitch range**: The difference between the highest and lowest pitch levels in the speech signal. A wider pitch range may indicate greater emotional expressiveness or emphasis.

5.2 Feature extraction

After we have defined two sets of features, we need to extract their numerical values for further machine learning.

To extract the features, the python modules pydub, and NumPy were used.

Using the pydub module, each of the audio recordings was recognized by Python and converted to a mono recording with a frequency of 44.1 kHz, because. With this preprocessing, we reduce high-frequency noise that can be a hindrance

[9] Rodero, E. (2011). Intonation and emotion: influence of pitch levels and contour type on creating emotions. Journal of voice, 25(1), e25-e34.

(Also, many machine learning algorithms work with sound at a frequency of 44.1 kHz by default, because this frequency was previously the standard for CDs). In addition, by converting the sound to mono, we reduce its spatiality, which allows us to speed up the work with it.

After preprocessing, the features were already extracted using the Numpy module, which received numerical values from the pydub module, and then converted them and the final features were obtained.

5.3 Models

In this work, it was decided to use the Random Forest machine learning model, since it is well suited for the classification problem.

This model has obvious advantages for the current task, such as:

- The ability to measure which feature is more important
- Random Forest model is less prone to overfitting compared to individual decision trees
- Thanks to this model, it is possible to fix non-linear relationships between features and emotions, and this is very important because the task of recognizing emotions is a complex task of acoustics, linguistics and contextual cues

Disadvantages of random forest model also exist, for example, they are quite resource-intensive and require a large amount of memory and these models are sensitive to noisy and "dirty" data.

Despite this, memory problems are solved by banal competent sound preprocessing, and in our case, the data was already cleaned of noise and other interference.

One of the key things in working with machine learning models is the selection of hyperparameters. Often, it is the choice of hyperparameters that determines how well the model learns.

In this work, the following hyperparameters n_estimators, max_depth, max_features, min_samples_split, and min_samples_leaf are used. These hyperparameters are particularly suitable for the random forest model.

Each of the above parameters has the following meaning:

- **n_estimators**: how many individual decision trees will be built during training.
- **max_depth**: maximum depth of decision trees.
- **max_features**: the maximum number of features that are considered when splitting a node in each decision tree.
- **min_samples_split**: minimum number of samples required to split an internal node during the tree-building process.
- **min_samples_leaf**: minimum number of samples required to be at a leaf node.

The **best_params_** module built into the sklearn library[10] can help you choose the correct values for hyperparameters, so it will be used. This module works on the principle of enumeration, we simply pass into it a list of values that

[10] [URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html]

we have selected for a certain hyperparameter, and the module already chooses which of the values we proposed is best suited.

6. Results

After the model is trained, we can proceed to the analysis of the results. In total, more than 30 training sessions were carried out, each of which will allow us to draw a conclusion about the following things: which set of features is most suitable for recognizing emotions, what accuracy each of the models has on a certain set of features, and which emotions are most often confused.

All in all, there were 3 approaches:

1. The intonation contour approach
2. The pitch level approach
3. Approach with pitch level and intonation contour

It took more than 180 hours to train all the models, considering all the attempts.

Further, each of them will be considered and an answer will be given to the question of which turned out to be more effective.

6.1 Intonation contour approach

After training on the features of the intonation contour, the accuracy of the model turned out to be **56.16%**, which can only mean that these features do not give the random forest model much clarity about what emotion is in the audio recording.

The quality of work was also evaluated using other metrics their result you can see below:

- **F1 score**: 0.5546750731682341
- **Recall score**: 0.5616117706069219
- **Precision score** 0.5838748130785928

Using the built-in random forest module **feature_importances_**, it was possible to calculate which of the features were the most important. Below, in Figure 5, you can see the results for the first approach.

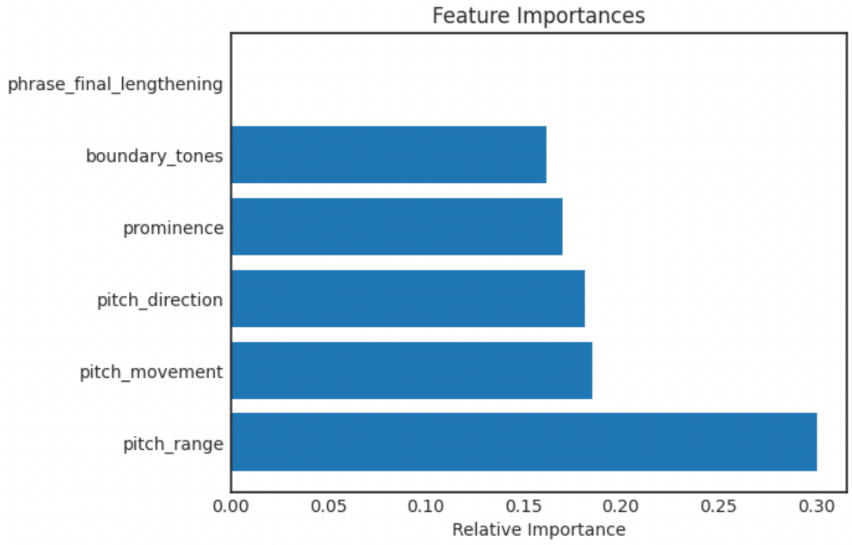

Figure 5. Feature importances for Intonation Contour approach (General Dataset)

Figure 5 clearly shows that the pitch_range feature is almost **1.5 - 2** times more important than all other features and has an importance factor of about **0.30**. While all other features, except phrase_final_lengthening, fluctuate between **0.16-0.18**.

As for phrase_final_lengthening, this feature does not have any significant contribution to the machine learning algorithm at all. Having understood the

technical aspect of the **parselmouth**[11] library, I managed to find out that the module that extracted this feature does not readily cope with short audio recordings, and in this work, the recordings were no longer than 5 seconds, although the creators of the library note that audio excerpts longer than 8 seconds are required for successful operation. Thus, the phrase_final_lengthening feature was assigned the default value of 1 and, accordingly, this did not affect anything, because all audio in this feature had a value equal to one. (I suggest further in the analysis to ignore this feature)

6.2 Pitch level approach

In the pitch level approach, give us a result of accuracy equal to 51,99%.

As you can see, the results differ, but not significantly, the overall accuracy has a difference of just over 4 percent, which is not a big difference from the previous result.

Below in Figure 6, features were also displayed and their importance was measured.

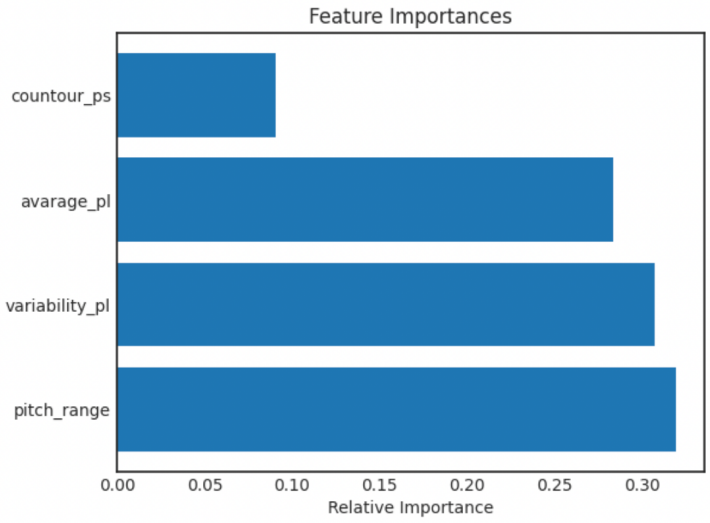

Figure 6. Feature importances for the Pitch Level approach (General Dataset)

[11] URL:[https://parselmouth.readthedocs.io/en/stable/installation.html]

As you can see from Figure 6, all features except the pitch shape(contour_ps) are almost equally important, which may indicate that the model relied on these features when solving: average pitch level, variability of pitch and pitch range.

6.3 Approach with pitch level and intonation contour

In the case of two approaches, the results, although on average, are much better. The accuracy of the model is **62.93%** which is already a great difference in comparison to the two previous approaches.

The accuracy is almost 7 percent more than in the experiment with intonation contour features and almost 11 percent more than in the pitch level approach.

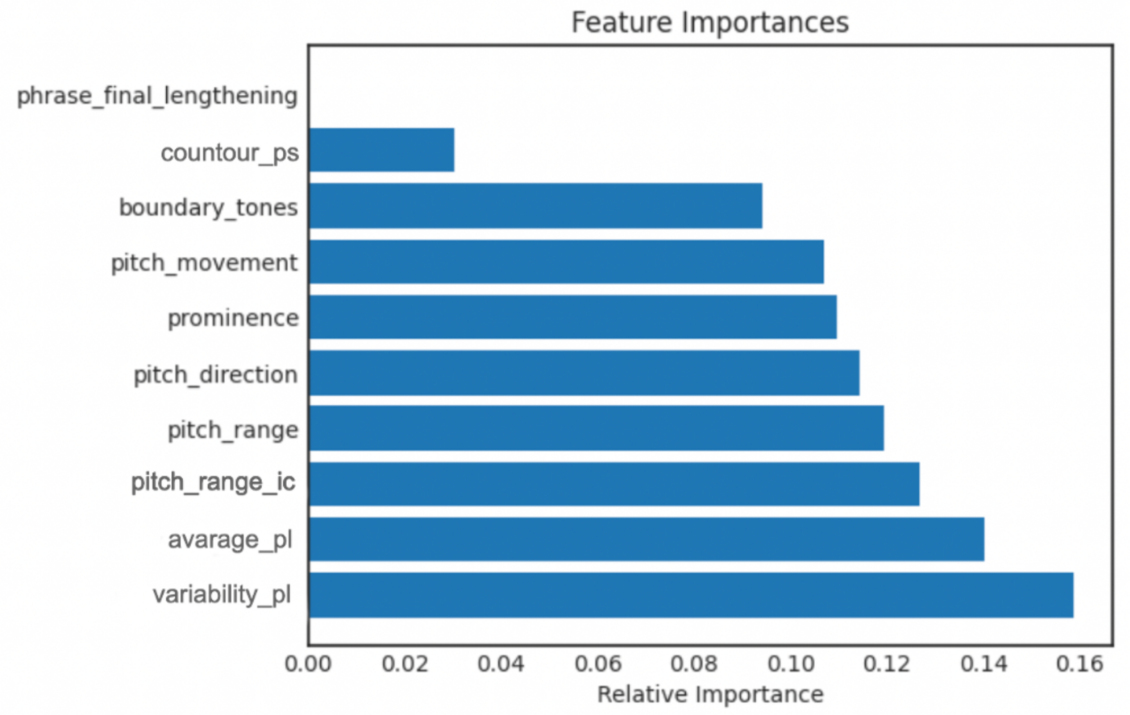

Figure 7. Feature importances for both approaches (General Dataset)

In Figure 7 above, you can also see how the importance ratios were distributed between features in a dual approach.

Average pitch level and variability of pitch are the two most important features of this approach. Then they are followed with approximately the same importance by pitch_range and pitch_range_ic, it is not surprising that the same characteristic has the same values in different approaches.

The remaining features scored approximately similar results, ranging from 0.10 to 0.12. Outsiders were phrase_final_lengthing(this result was expected) and countour_ps.

For greater representativeness, a confusion matrix was built for each of the experiments that can be seen in Figures 8,9,10.

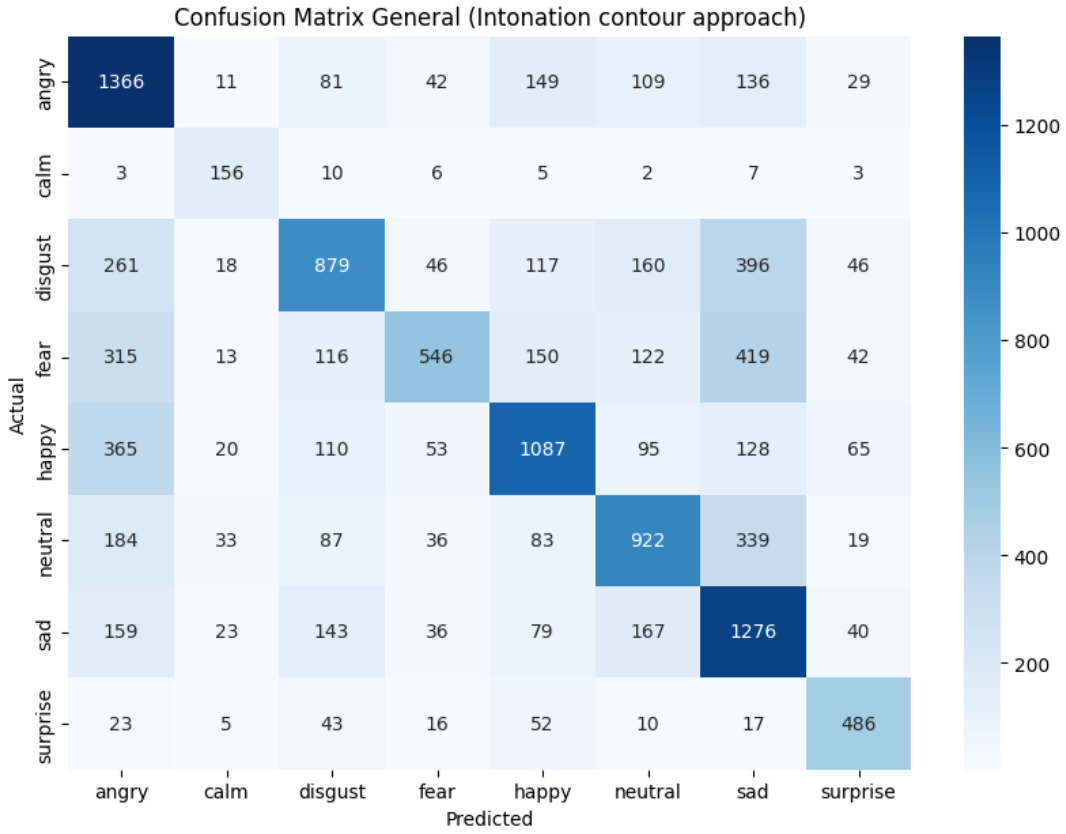

Figure 8. Confusion Matrix General (Intonation contour approach)

From Figure 8, it is obvious that the emotions of *anger*, *sadness,* and *happy* were determined best for the approach with the intonation contour, and the worst of all were *surprise* and *calm*.

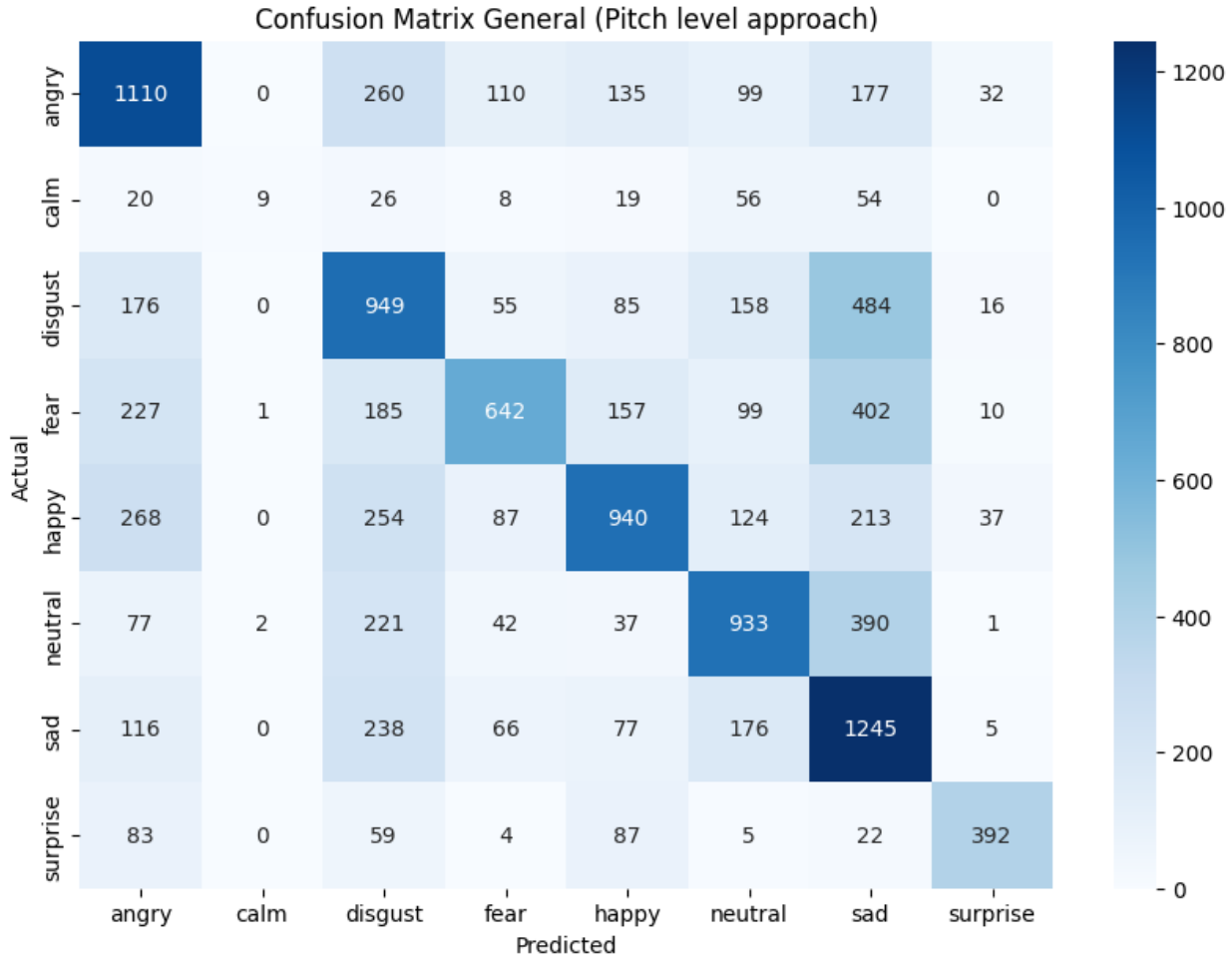

Figure 9. Confusion Matrix General (Pitch level approach)

In Figure 9, we can see that the pitch level approach did not do much worse overall and was slightly worse in *anger*, *surprise*, *happiness*, and *sadness*, but slightly better in *disgust*, *fear*, and *neutral* emotion. But at the same time, it is

important to note that this approach is a complete failure in relation to the emotion of *calmness*.

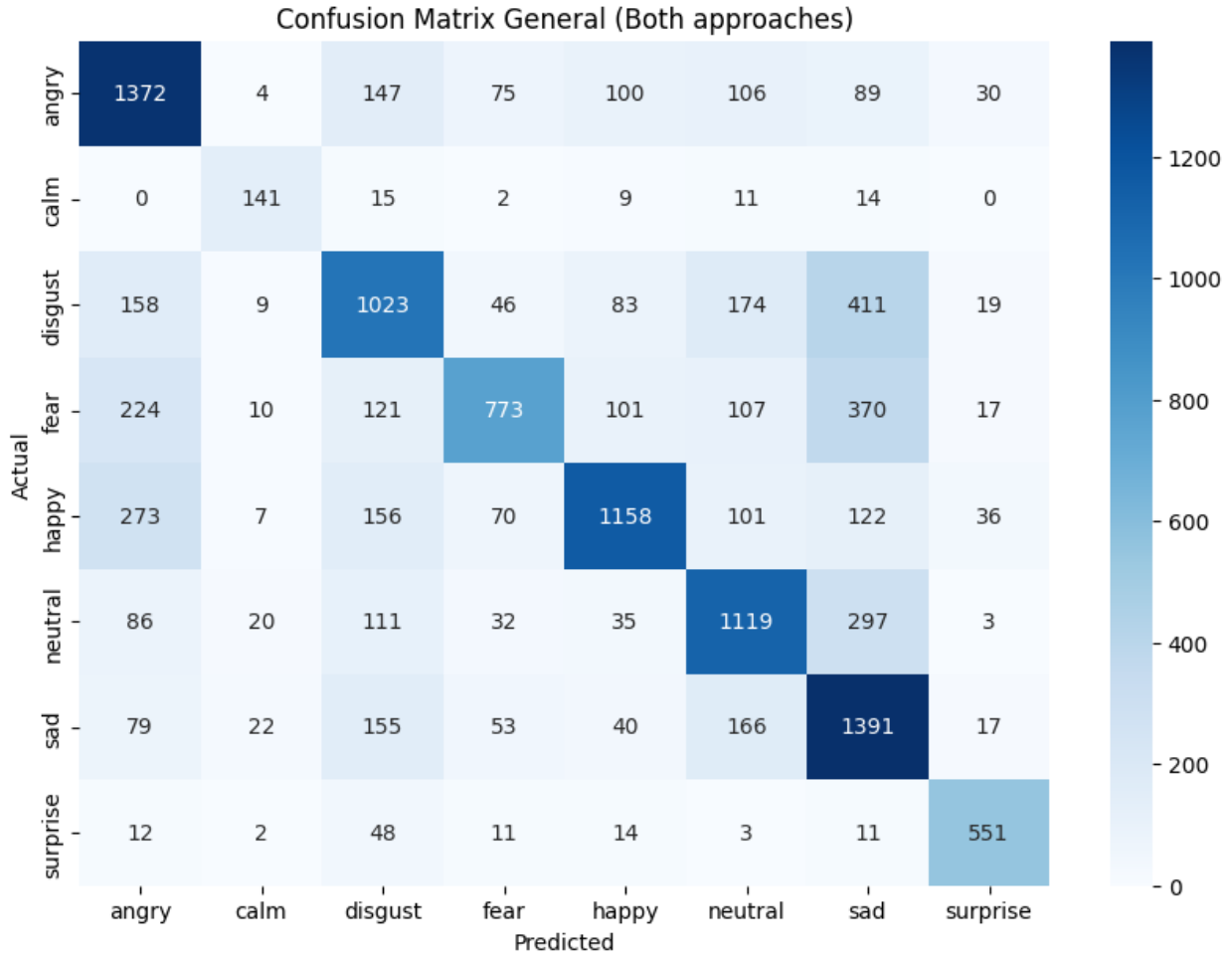

Figure 10. Confusion Matrix General (Both approaches)

In Figure 10, the first thing that catches your eye is a significant increase in the percentage of correct predictions regarding the emotion of calm compared to Figure 9, but it is still less than in Figure 8. This can only tell us that with a joint approach where we use intonation contour and pitch level features, pitch level features, to some extent, prevent intonation contour features from making correct predictions. For the rest of the emotions, we see only an increase in comparison with the approaches shown in Figures 8, 9. Similar results suggest that despite the

emotion of calmness, which was less correctly predicted, the idea of combining features was a good one.

In general, given the specifics of the dataset and the fact that it was not fully balanced in terms of emotions, I consider the result to be satisfactory. With nearly 63 percent accuracy in the two-feature approach, the most commonly confused emotions were:

```
Actual - Predicted

disgust - sad          411

fear - sad             370

neutral - sad          297

happy - angry          273

fear - angry           224
```

6.4 Separate datasets

In order to understand how different the datasets are and to draw final conclusions about the accuracy of the model, it was decided to conduct a similar experiment as above on each of the datasets separately.

6.4.1 SAVEE experiments

The accuracy for the Intonation Contour approach on this dataset was almost 74%, in addition, other metrics below gave good results:

- **Accuracy**: 0.7395833333333334
- **F1 score**: 0.7304462536003785
- **Recall score**: 0.7395833333333334

- **Precision score**: 0.7711692177301129

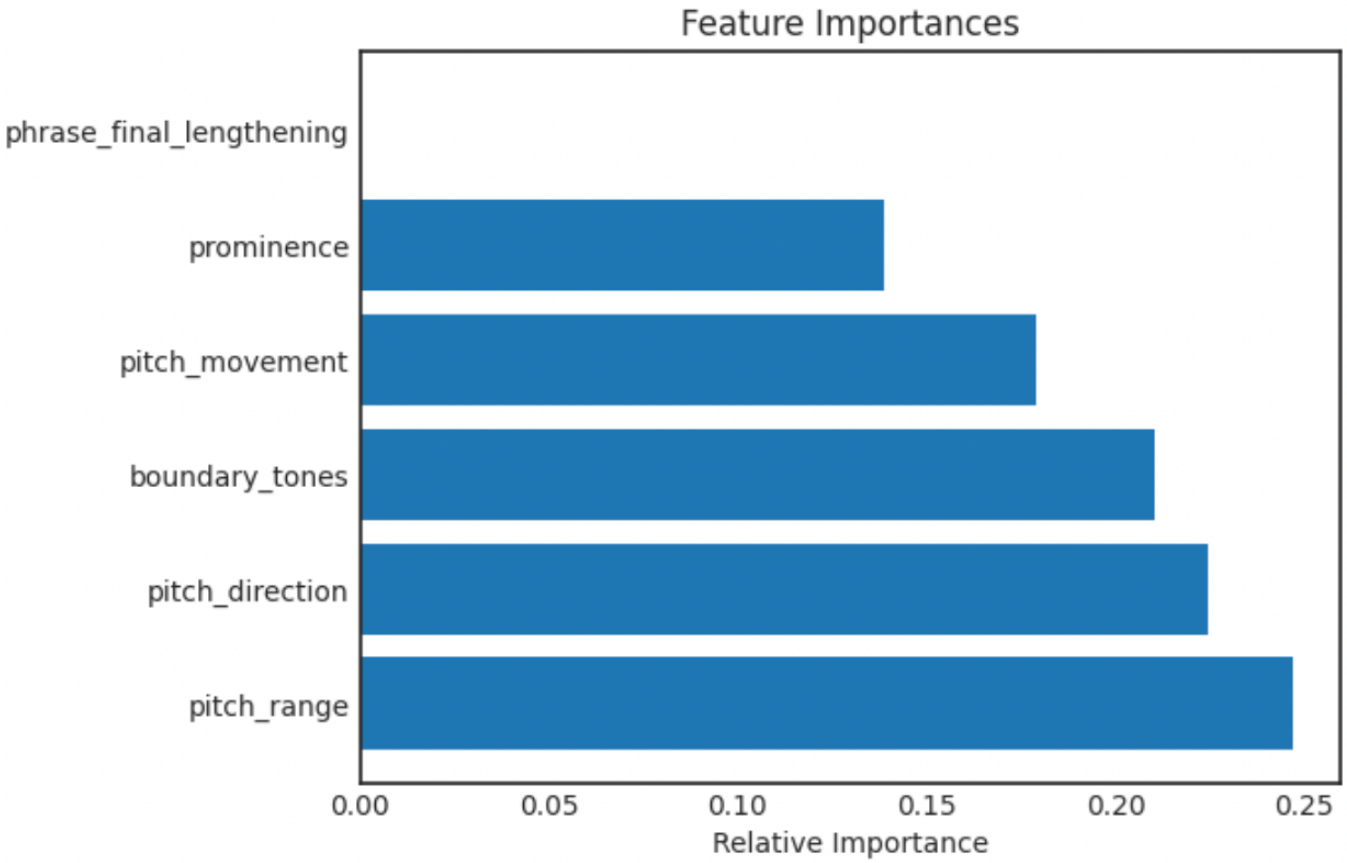

Figure 11. Feature Importances SAVEE (Intonation contour approach)

Figure 11 demonstrates that the most valuable feature in recognition for the SAVEE dataset when approached with an intonation contour is the pitch_range, then in approximately 0.25 steps they follow each other in the following order: pitch_direction, boundary_tones, pitch_movement, prominence.

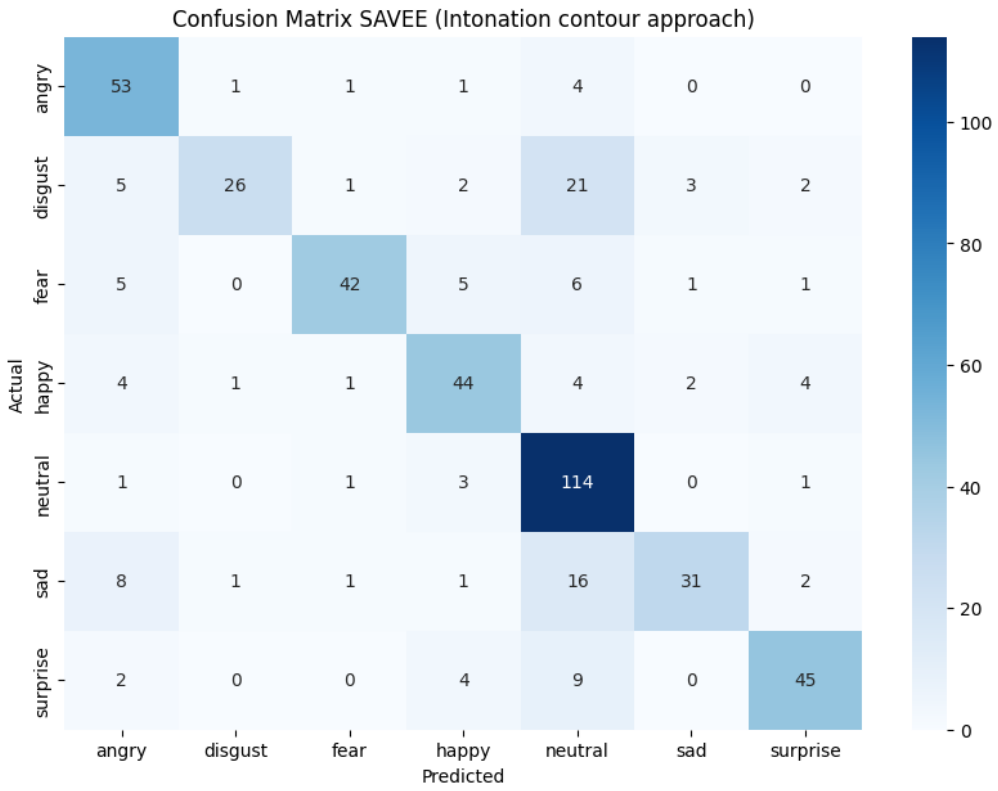

Figure 12. Confusion Matrix SAVEE (Intonation contour approach)

In Figure 12, we can see how the answers were distributed on this dataset. Best of all in this case, the model coped with a *neutral* emotion, it is not surprising, because it is the most often emotion in this dataset, and worst of all with the emotion of disgust, but in general I would not say that some kind of emotion is strongly knocked out.

The pitch level approach on this dataset turned out to be slightly worse than on the intonation contour and showed the following results in metrics:

- **Accuracy**: 0.6791666666666667
- **F1 score**: 0.6687005403138409
- **Recall score**: 0.6791666666666667
- **Precision score** 0.6855744600471767

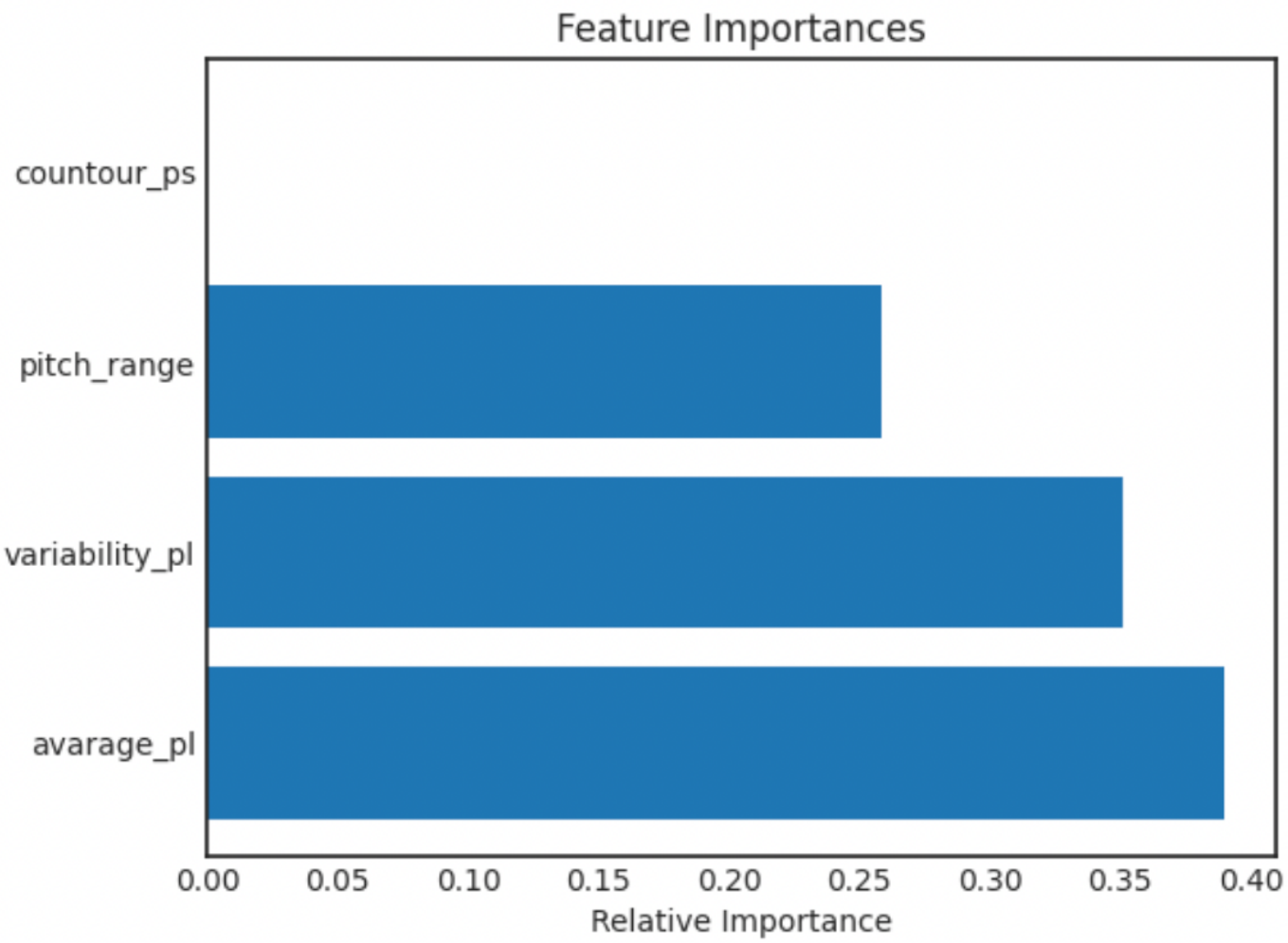

Figure 13. Feature Importances SAVEE (Pitch level approach)

Figure 13 shows us interesting results. For this dataset, the contour_ps feature does not matter at all and does not carry any importance, but at the same time, the value of the average pitch reaches an importance coefficient equal to almost 0.40.

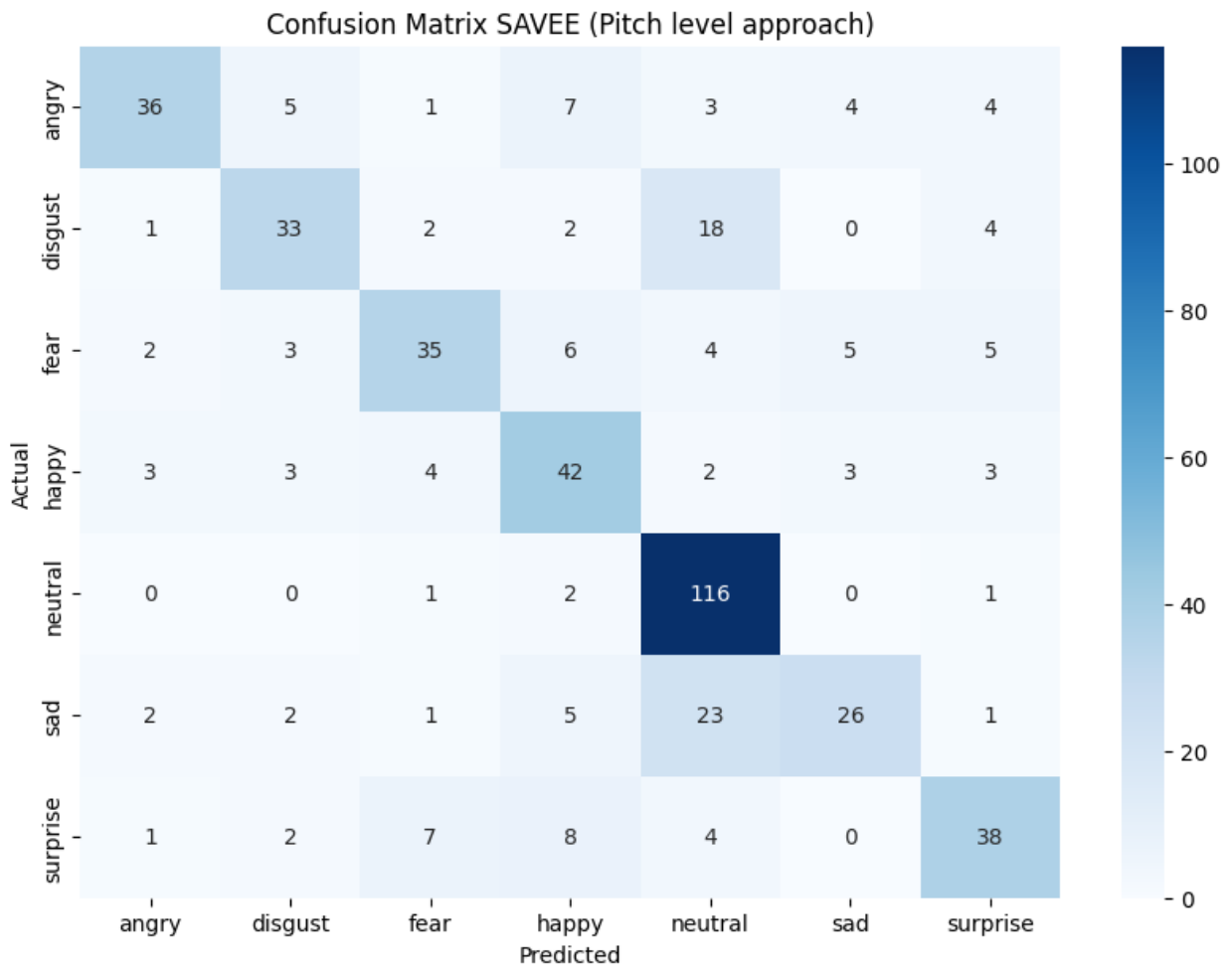

Figure 14. Confusion Matrix SAVEE (Pitch level approach)

In Figure 14, we can see how the answers were distributed for Pitch Level approach. Best of all in this case again was a *neutral* emotion, and worst of all with the emotion of sadness.

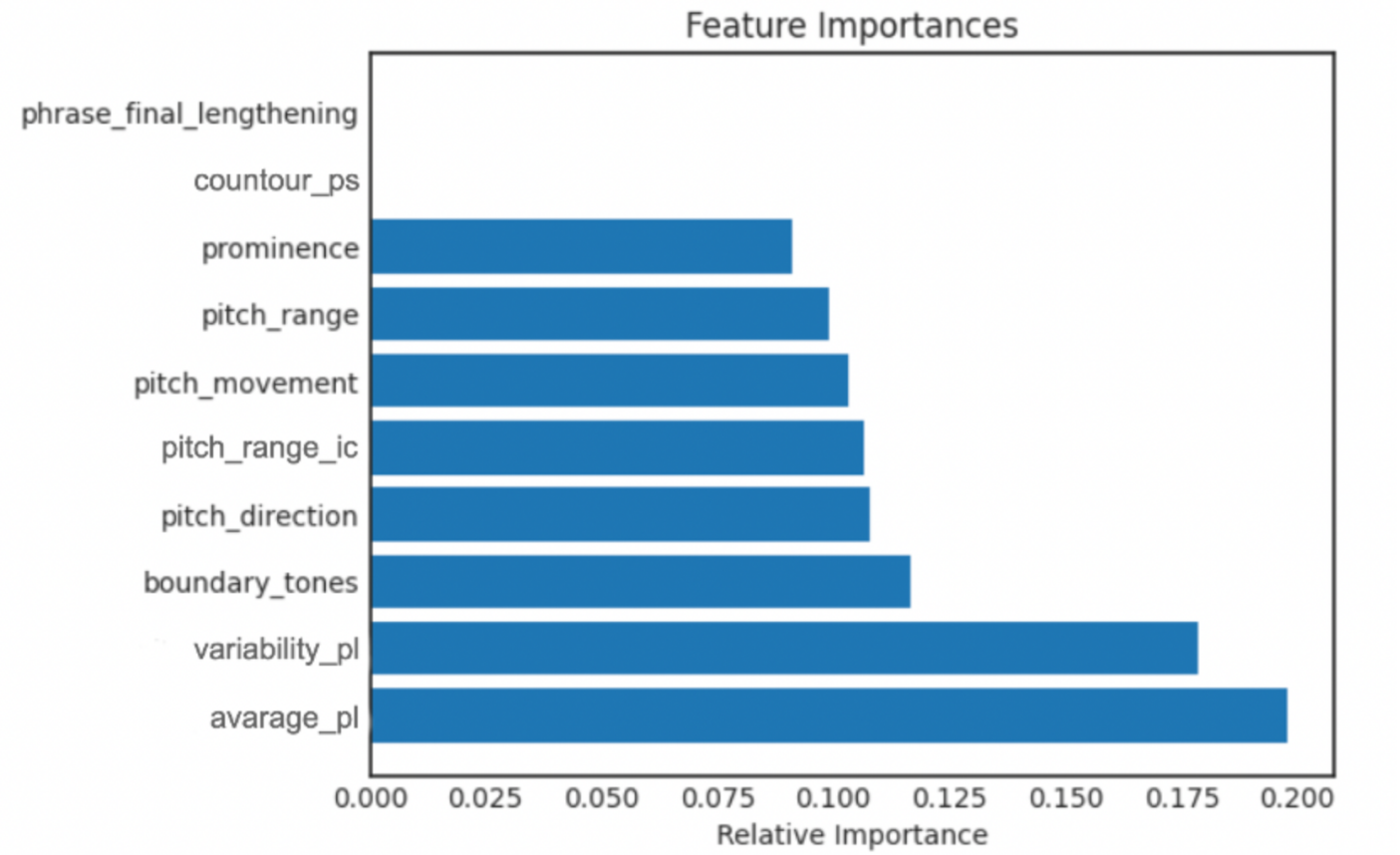

Figure 15. Feature Importances SAVEE (Both approaches)

In this case, in Figure 15, the pitch level features are more important than the intonation contour features, the shape of the pitch contour still has a value of zero.

The value of the metrics in this case:

- **Accuracy:** 0.86875
- **F1 score:** 0.8673051013629389
- **Recall score:** 0.86875
- **Precision score** 0.8748124247584237

Again obvious increase in evaluation metrics.

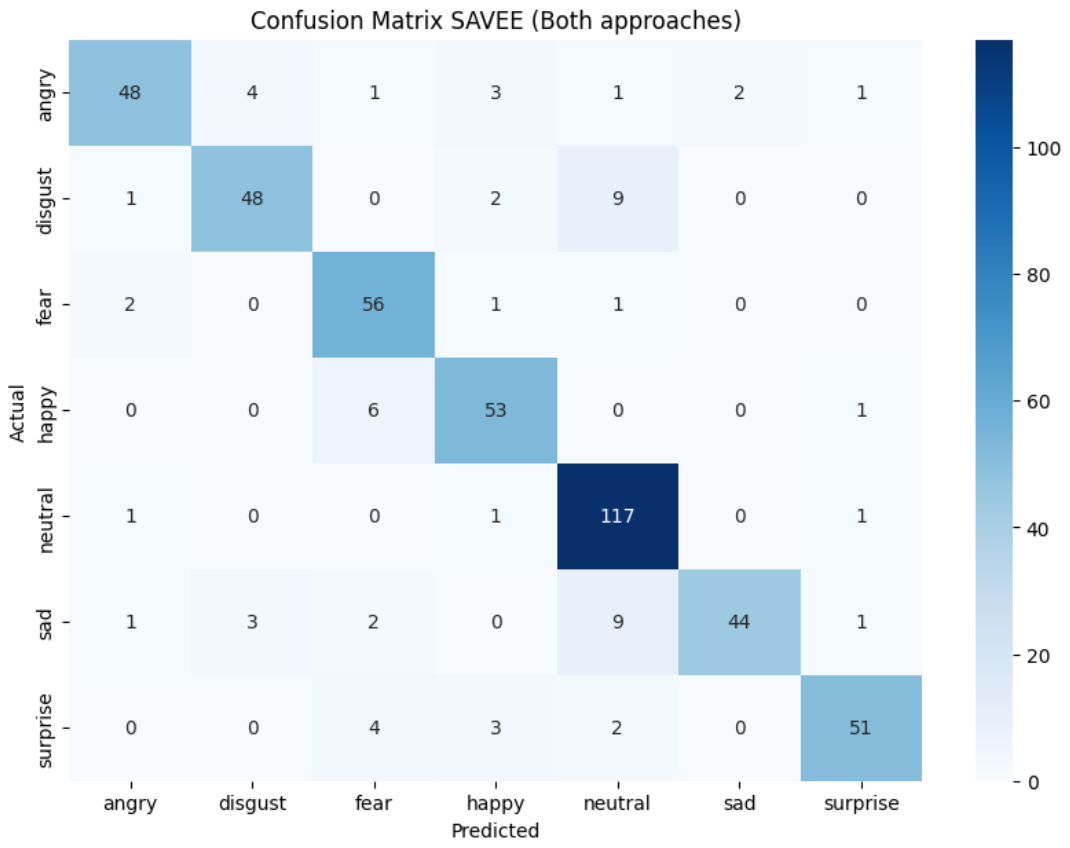

Figure 16. Confusion Matrix SAVEE (Botch approaches)

In Figure 16, we can see how the model worked out with two approaches.

Five of the most confusing emotions in this case:

```
Actual - Predicted

disgust - neutral      9

sad - neutral          9

happy - fear           6

surprise - fear        4

angry - disgust        4
```

6.4.2 RAVDESS experiments

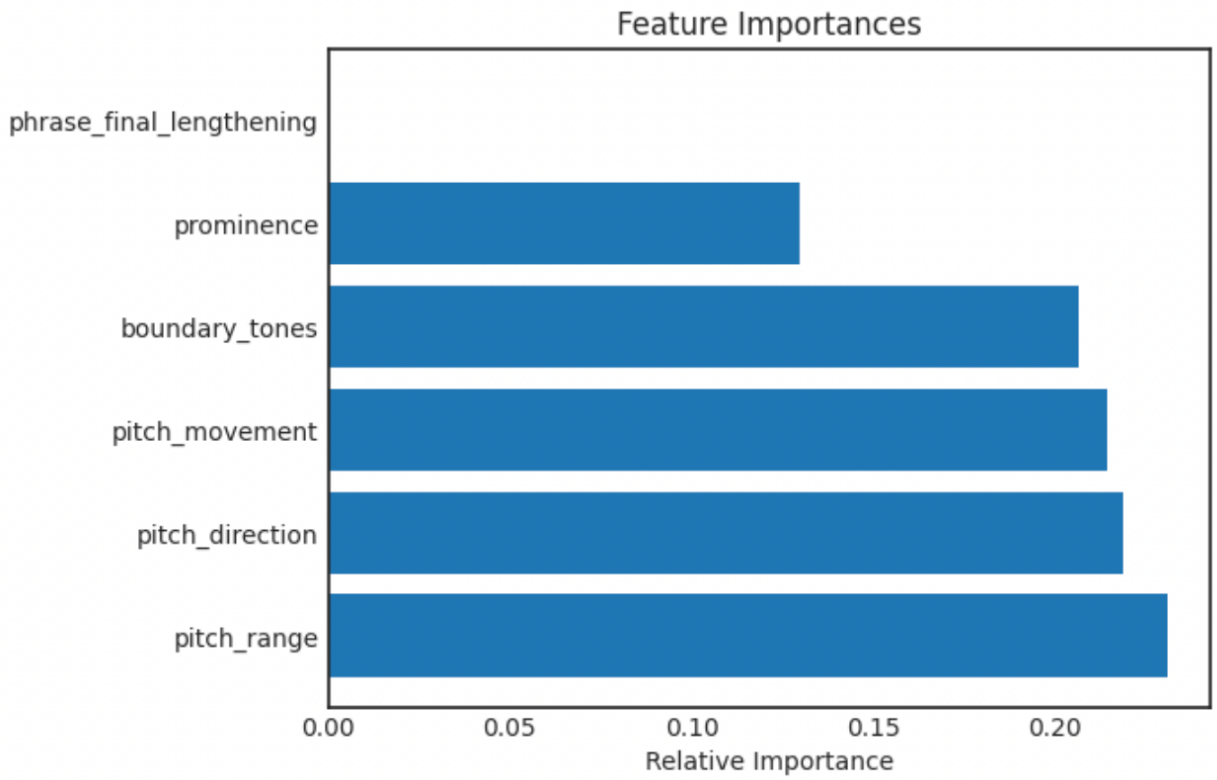

Figure 17. Feature Importances RAVDESS (Intonation contour approach)

The results shown in Figure 17 are very similar to those shown in Figure 11. This may be due to the fact that the sizes of these datasets are approximately the same.

Metric values:

- **Accuracy**: 0.8708333333333333
- **F1 score**: 0.8707628583023144
- **Recall score**: 0.8708333333333333
- **Precision score** 0.873322429106841

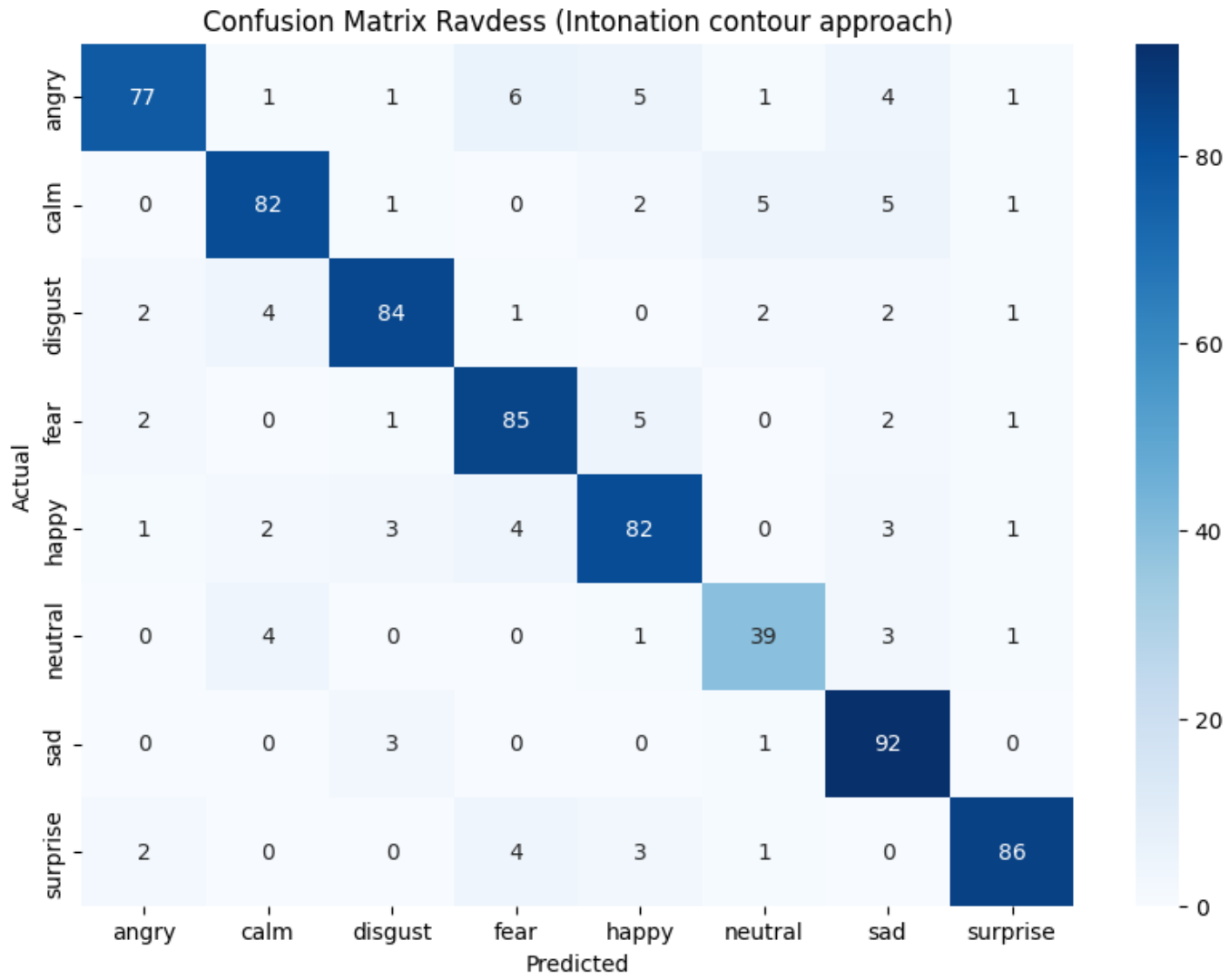

Figure 18. Confusion Matrix RAVDESS (Pitch level approach)

Again we can see great performance from random forest with intonation contour features in Figure 18.

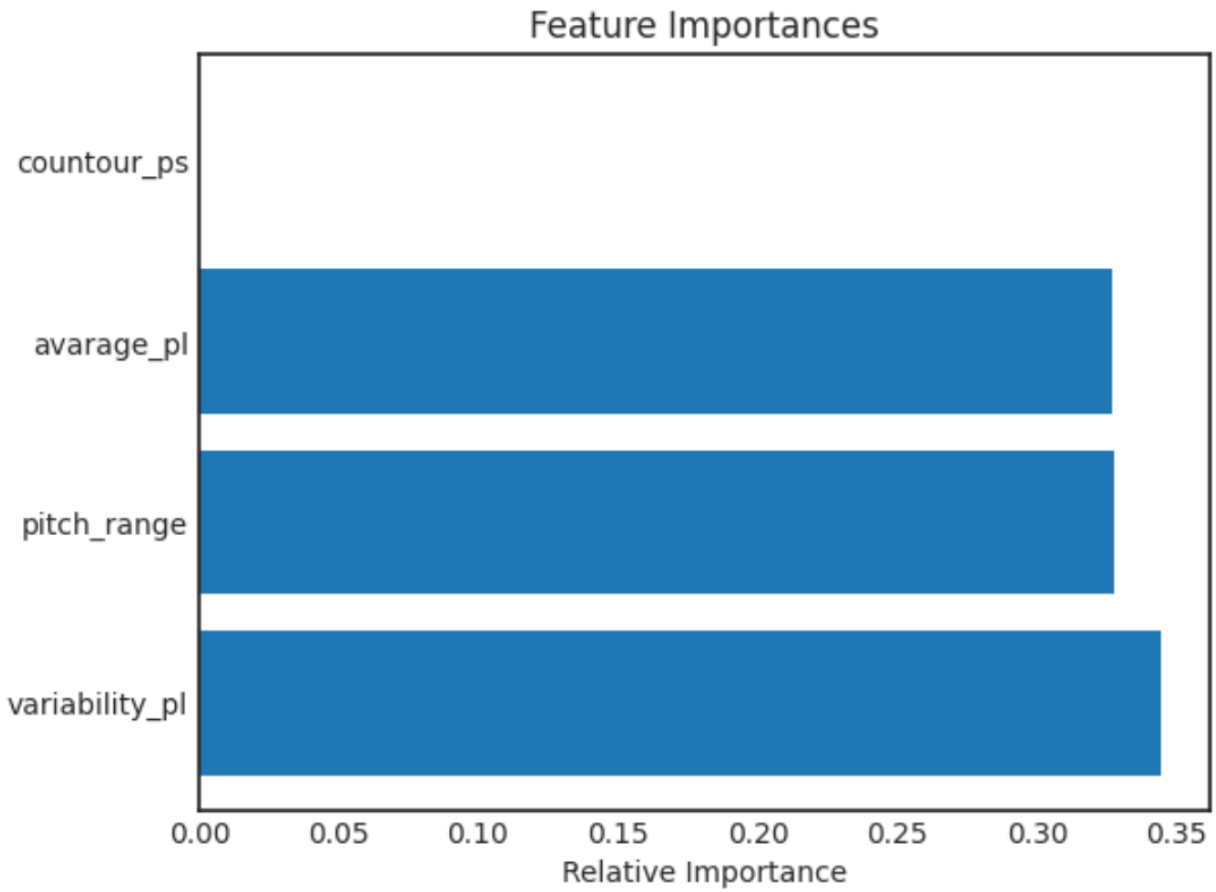

Figure 19. Feature Importances RAVDESS (Pitch level approach)

In this Figure 19, we see that compared to the previous dataset, in this case, almost all features have an equal importance coefficient, except for the pitch shape, which is still equal to zero.

Metrics:

- Accuracy: 0.8569444444444444
- F1 score: 0.8559434631569088
- Recall score: 0.8569444444444444
- Precision score 0.8595549612303806

In this case metrics a little bit higher than in the previous case.

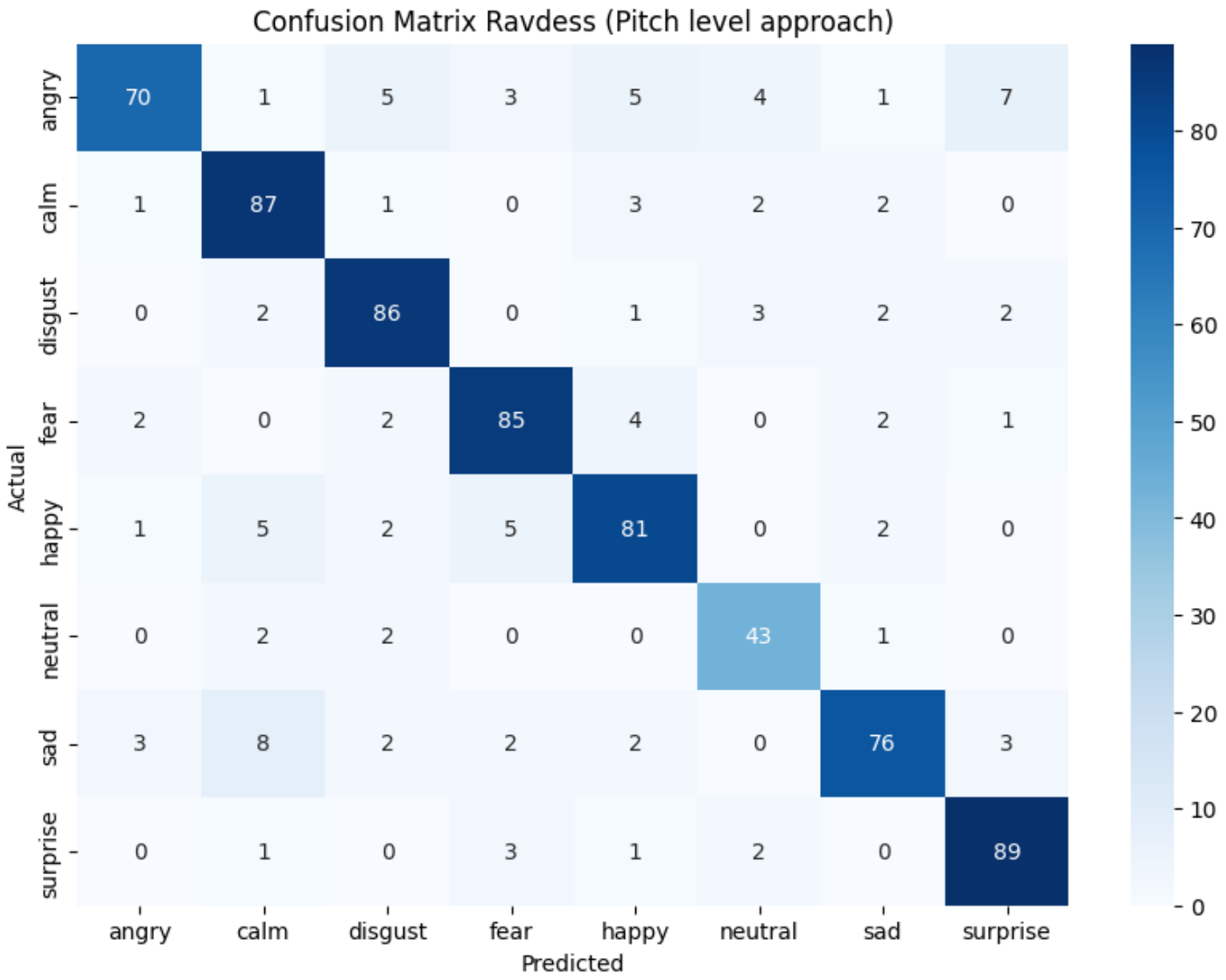

Figure 20. Confusion Matrix RAVDESS (Pitch level approach)

Another great performance from the model is in Figure 20.

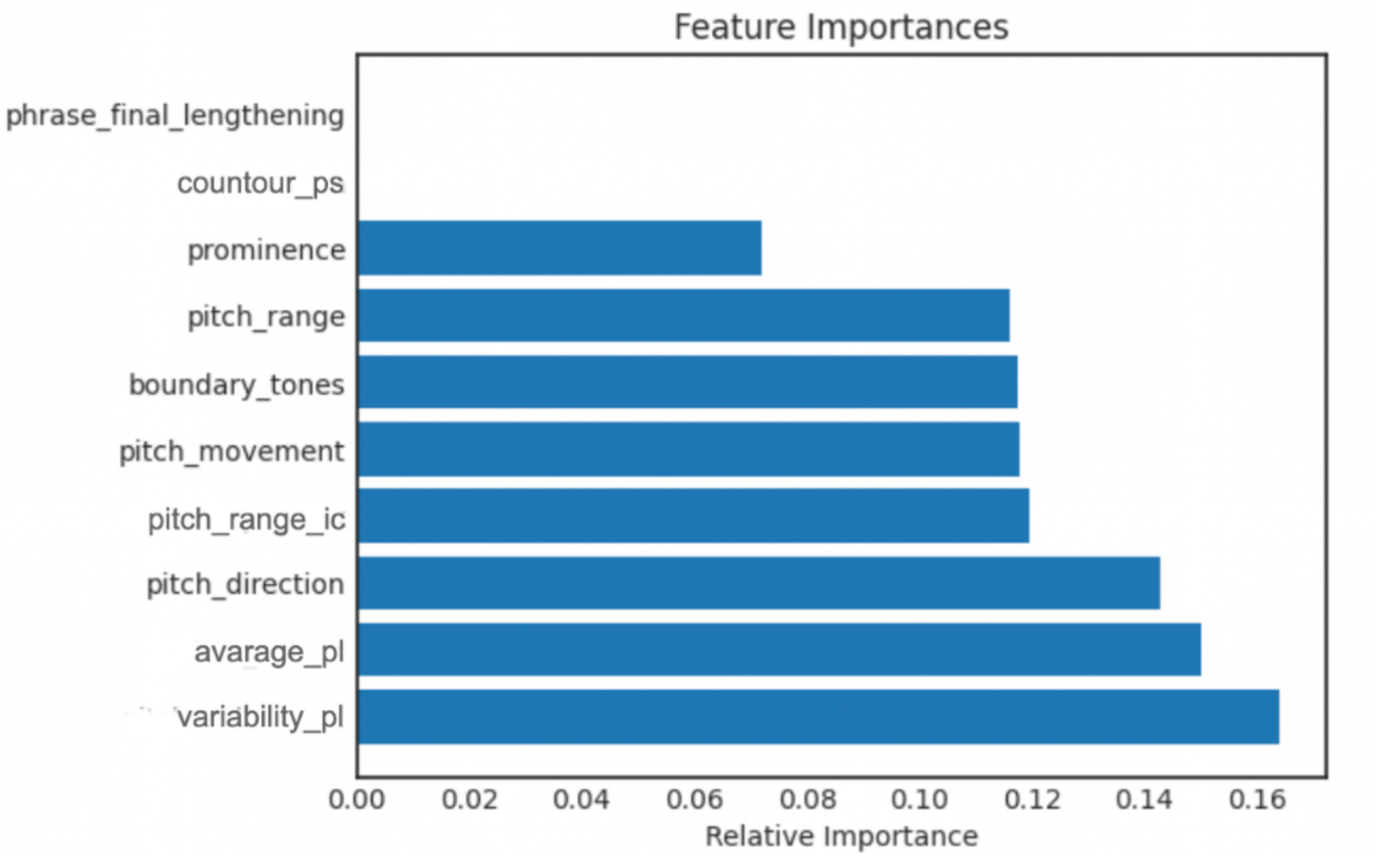

Figure 21. Feature Importances RAVDESS (Both approaches)

In Figure 21, we can see 3 distinct feature leaders in the 2-feature approach, followed by 4 features with nearly equal values and an outsider in the form of prominence.

Metrics values:

- Accuracy: 0.8736111111111111
- F1 score: 0.8725816969463731
- Recall score: 0.8736111111111111
- Precision score 0.8763191384759507

An almost imperceptible increase in accuracy when combining features.

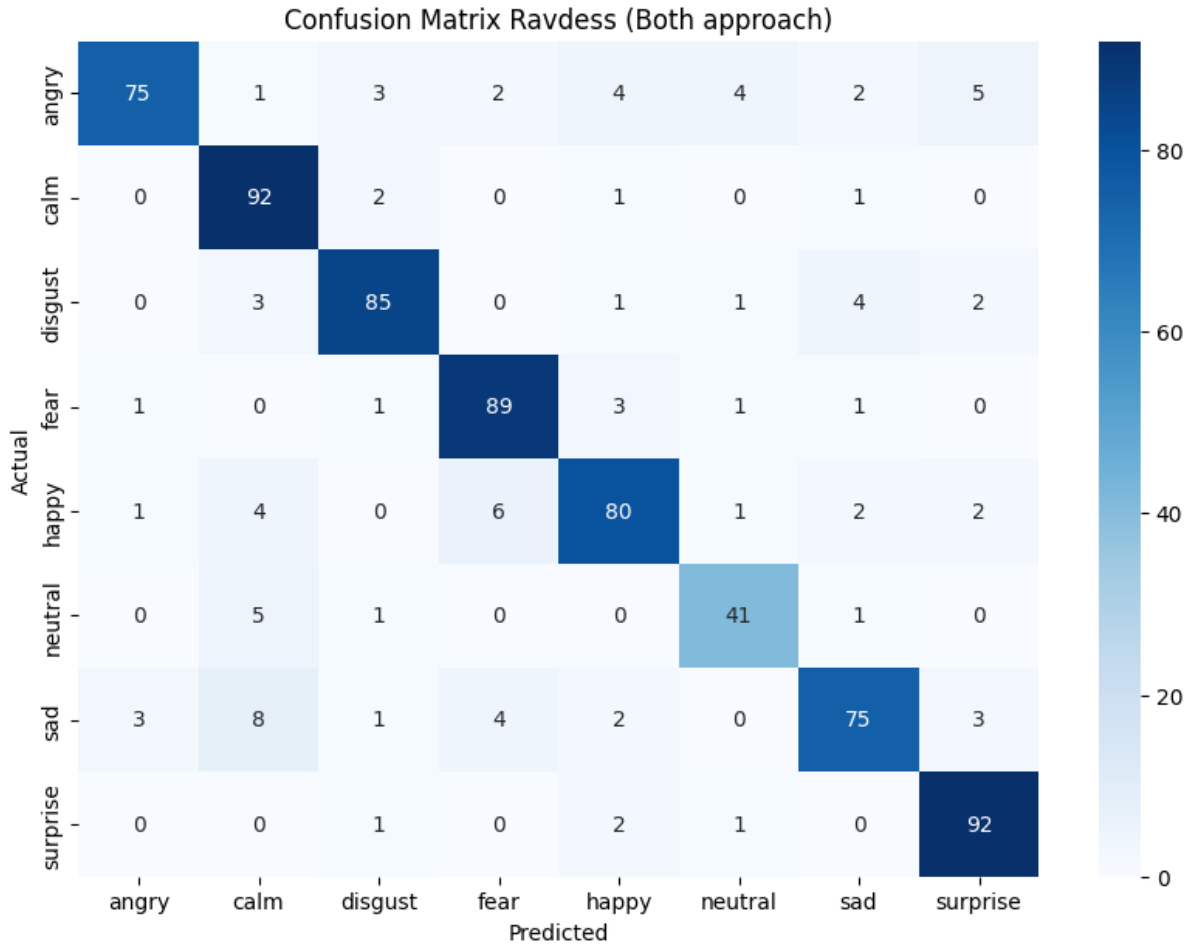

Figure 22. Confusion Matrix RAVDESS (Both approaches)

Figure 22 perfectly describes itself as a great performance.

Most confusing emotions in this case:

```
Actual - Predicted

sad - calm              8

happy - fear            6

angry - surprise        5
```

```
neutral - calm          5

disgust - sad           4
```

6.4.3 TESS experiments

The metrics in this dataset are pleasantly surprised at all stages of the experiment, here are the results of the intonation contour approach for TESS dataset:

- **Accuracy**: 0.936923076923077
- **F1 score**: 0.9370290894439729
- **Recall score**: 0.936923076923077
- **Precision score** 0.9373988366601921

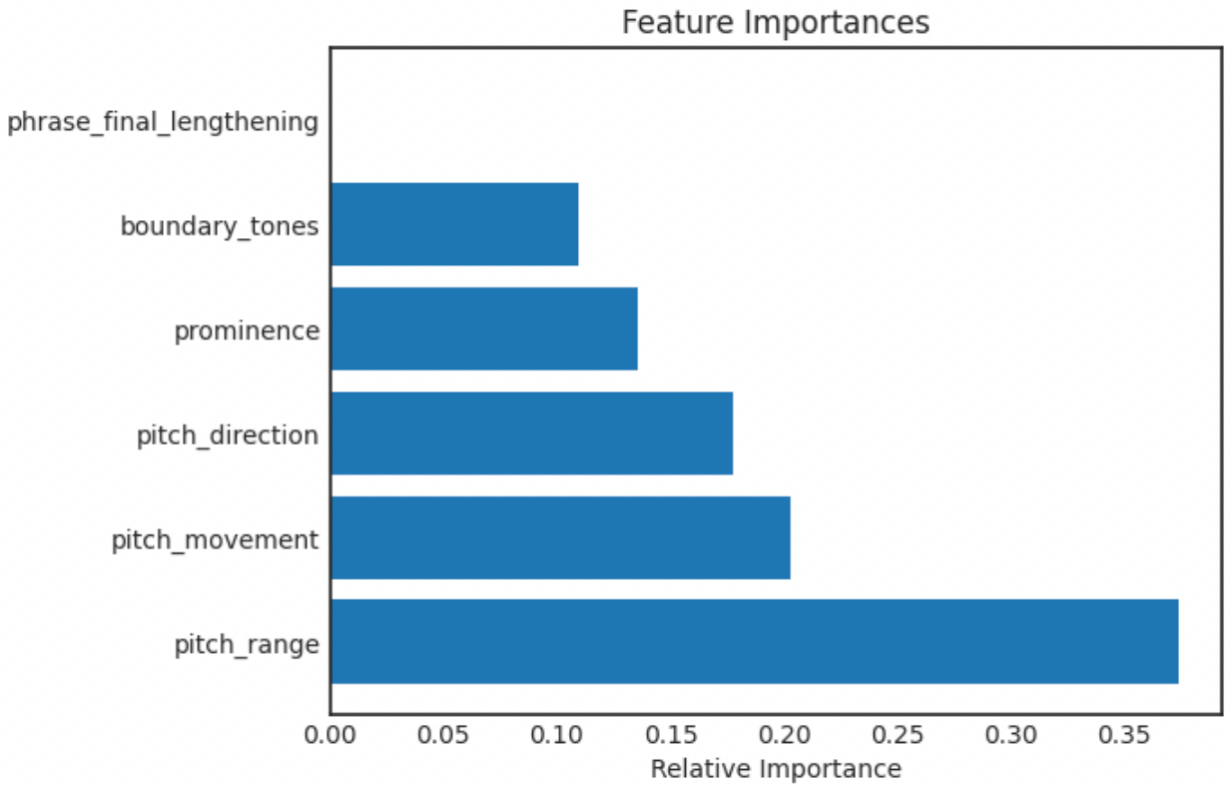

Figure 23. Feature Importances TESS (Intonation contour approach)

In Figure 23, you can also immediately notice the difference in comparison with the two previous datasets, the pitch_range feature with a value close to 0.37 is clearly highlighted here.

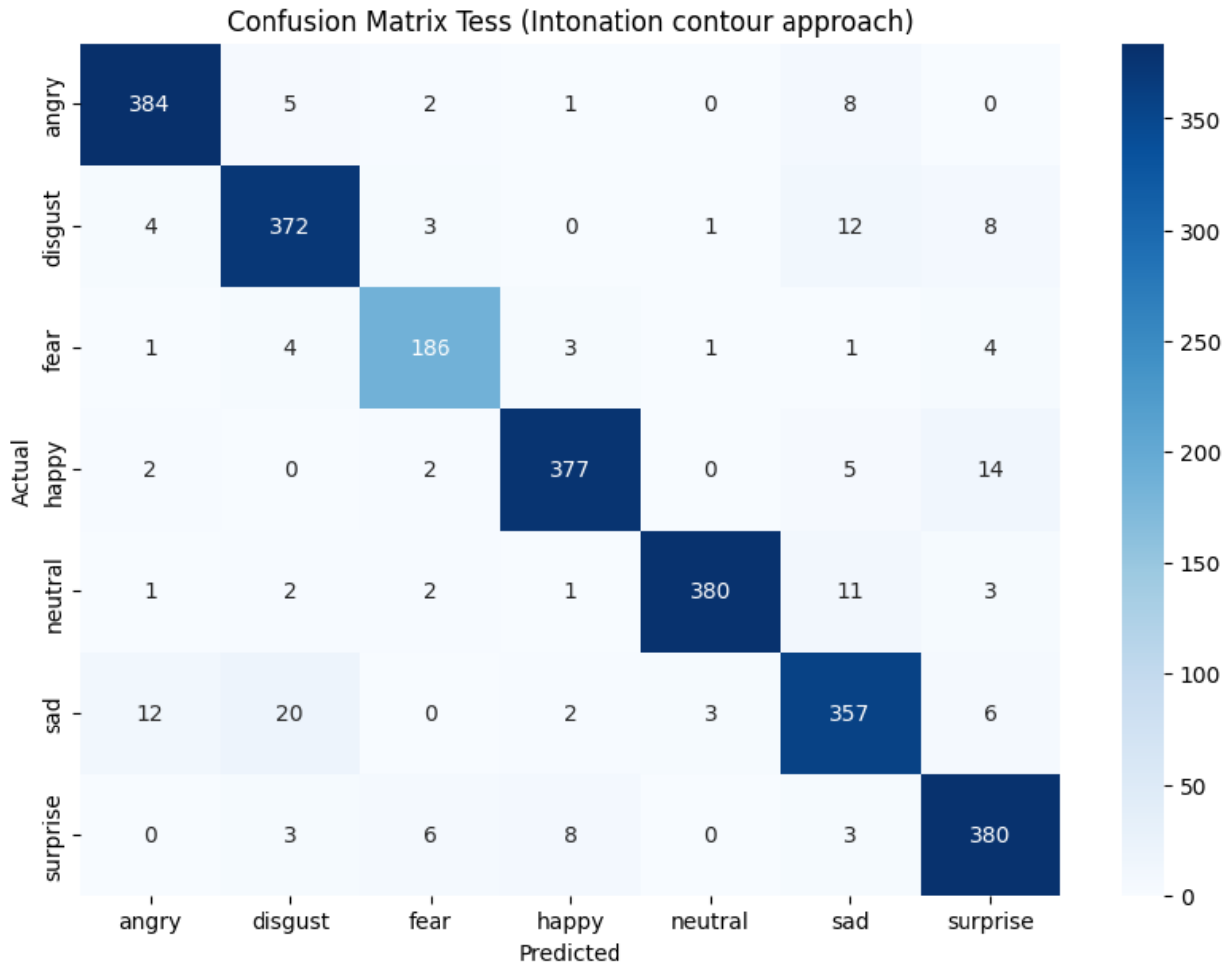

Figure 24. Confusion Matrix TESS (Intonation contour)

The results in Figure 24 are incredibly good, we can see that the model coped with the task of emotion recognition almost flawlessly, making only occasional errors. Considering that this dataset is quite large, these results can be considered representative.

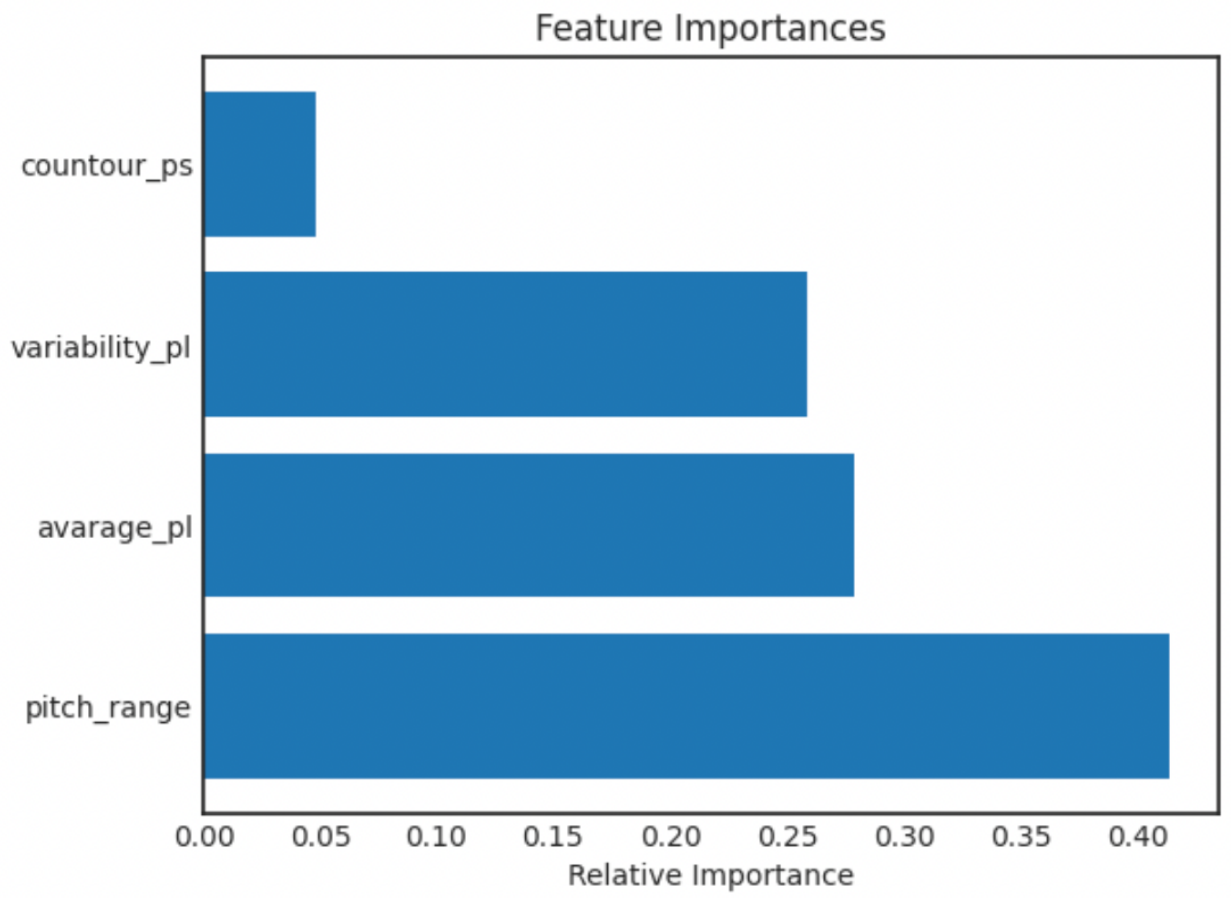

Figure 25. Feature Importances TESS (Pitch level approach)

Figure 25, again obvious differences from the two previous datasets, although the features are in their places, their coefficients have changed significantly and if the pitch range has the highest value, then the pitch variability and its average value are almost equal. And finally, the pitch contour appears as a significant feature, although it has little value compared to the rest.

Also metrics:

- **Accuracy**: 0.9496153846153846
- **F1 score**: 0.9496107024829924
- **Recall score**: 0.9496153846153846

- **Precision score** 0.9497736447649258

Again, there is a difference in metrics when compared with the intonation contour approach, but it is quite small, I would even say within the margin of error.

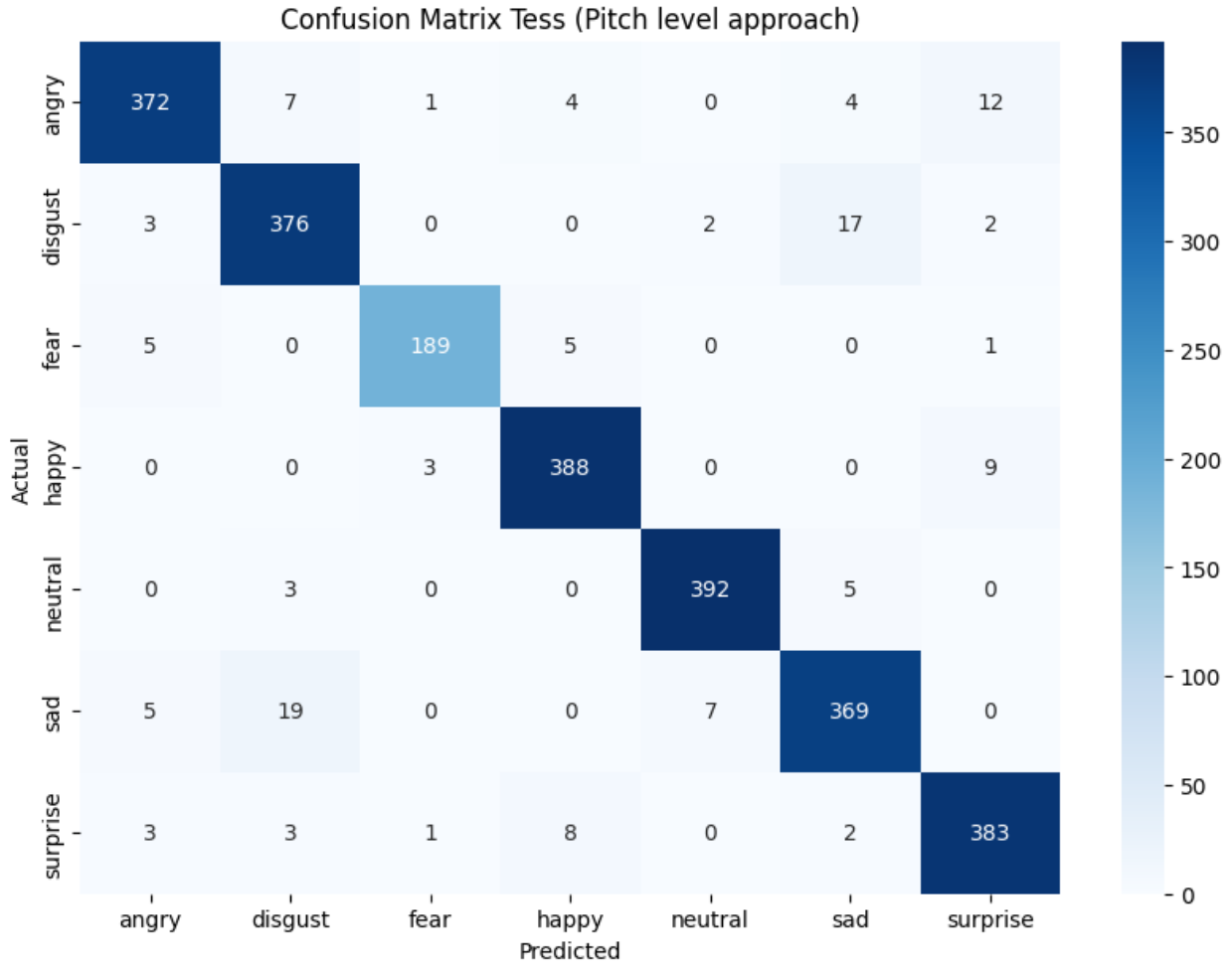

Figure 26. Confusion Matrix TESS (Pitch level approach)

As in Figure 24, Figure 25 has extremely good results.

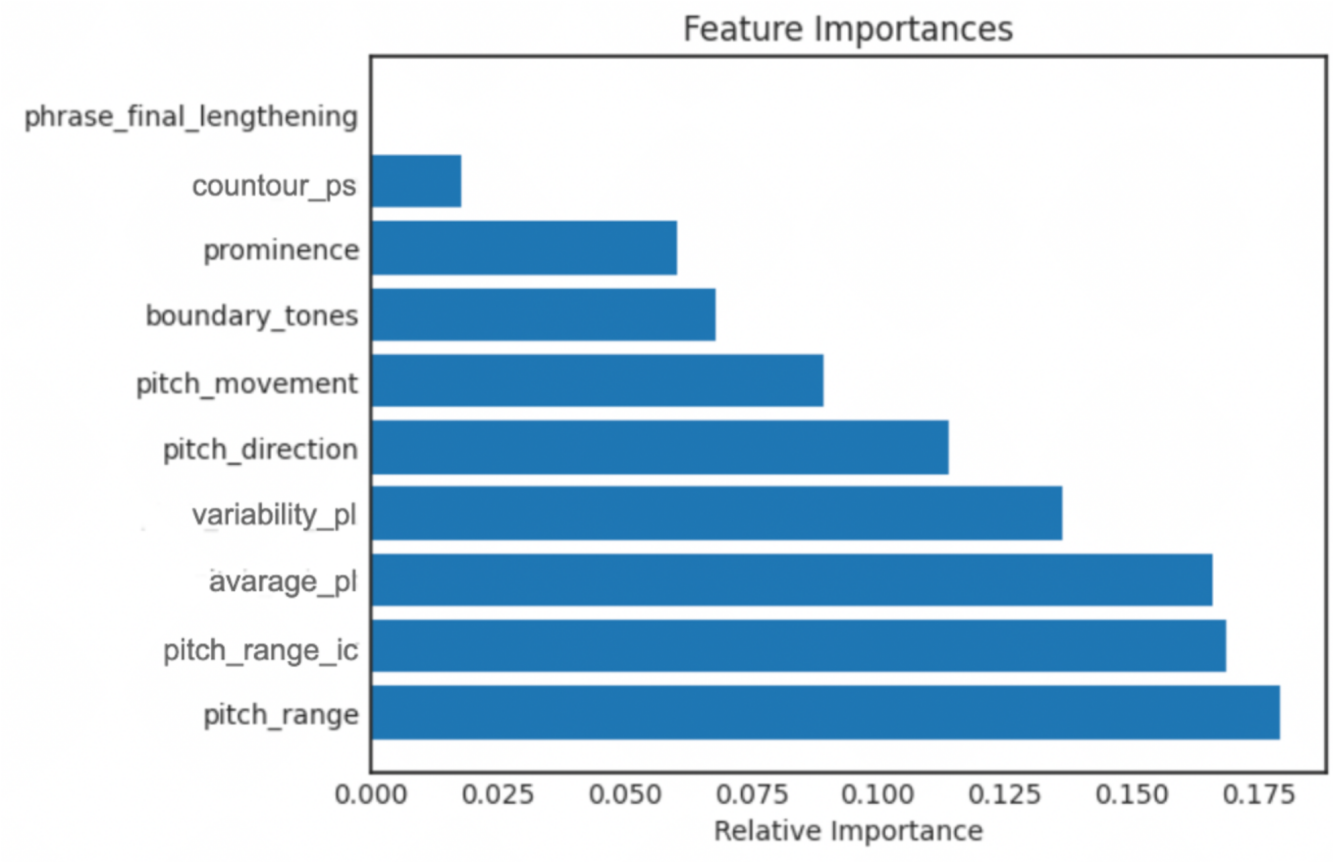

Figure 27. Feature Importances TESS (Both approaches)

Figure 27 clearly shows that out of the 4 most significant features for our task, 3 of them belong to the pitch level features set, which indicates that for the TESS dataset, the pitch level features are more important than the features of the intonation contour.

Also metrics:

- **Accuracy**: 0.9807692307692307
- **F1 score**: 0.9807763594212275
- **Recall score**: 0.9807692307692307
- **Precision score** 0.9808207761766206

Looking ahead, on this dataset, but I think it is already obvious, we got the best results in terms of prediction accuracy.

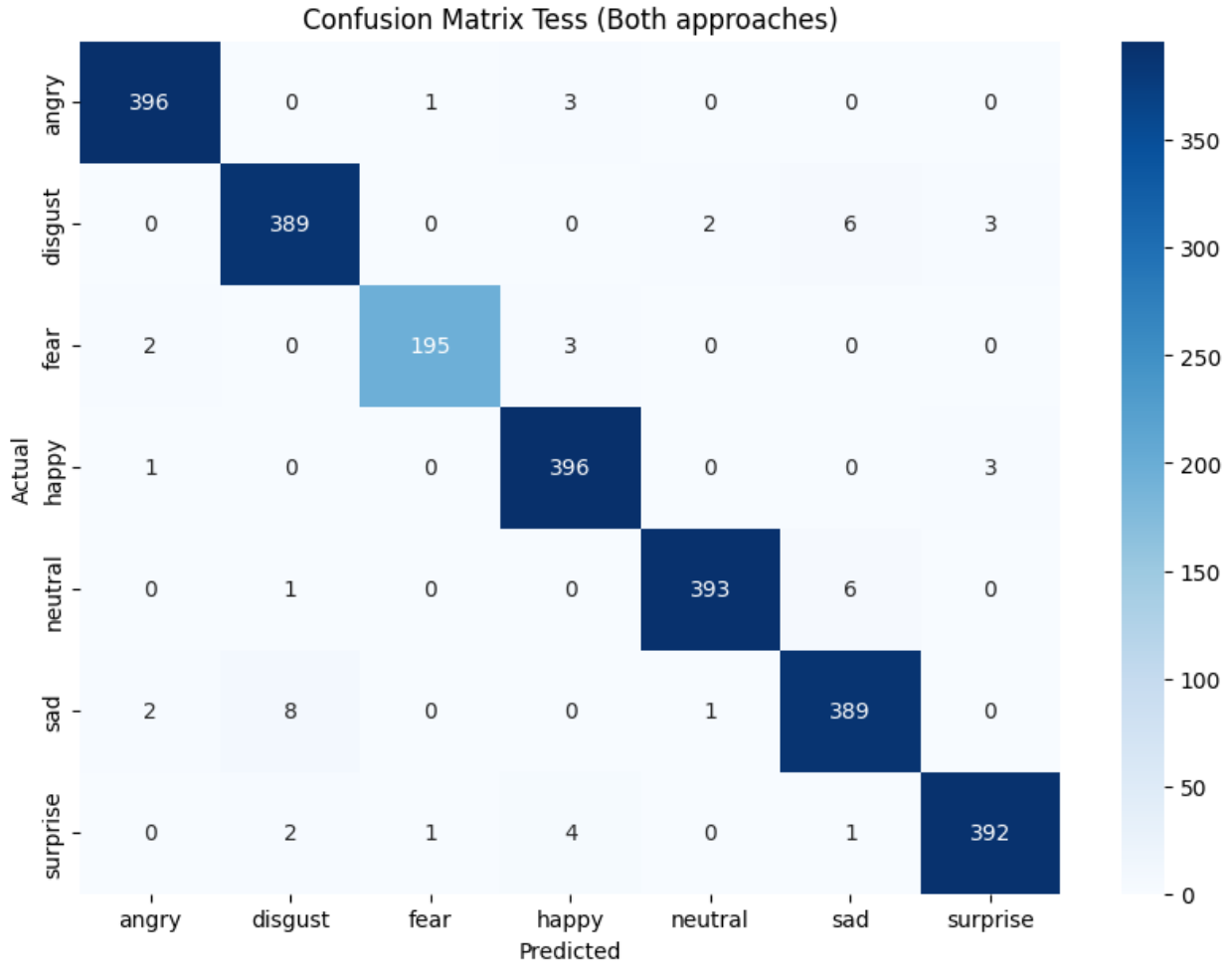

Figure 28. Confusion Matrix TESS (Both approaches)

In Figure 28, we can see that the combined set of features gives an even more accurate result.

Most confused emotions:

```
Actual - Predicted

sad - disgust          8
```

```
disgust - sad          6

neutral - sad          6

surprise - happy       4

angry - happy          3
```

6.4.4 CREMA experiments

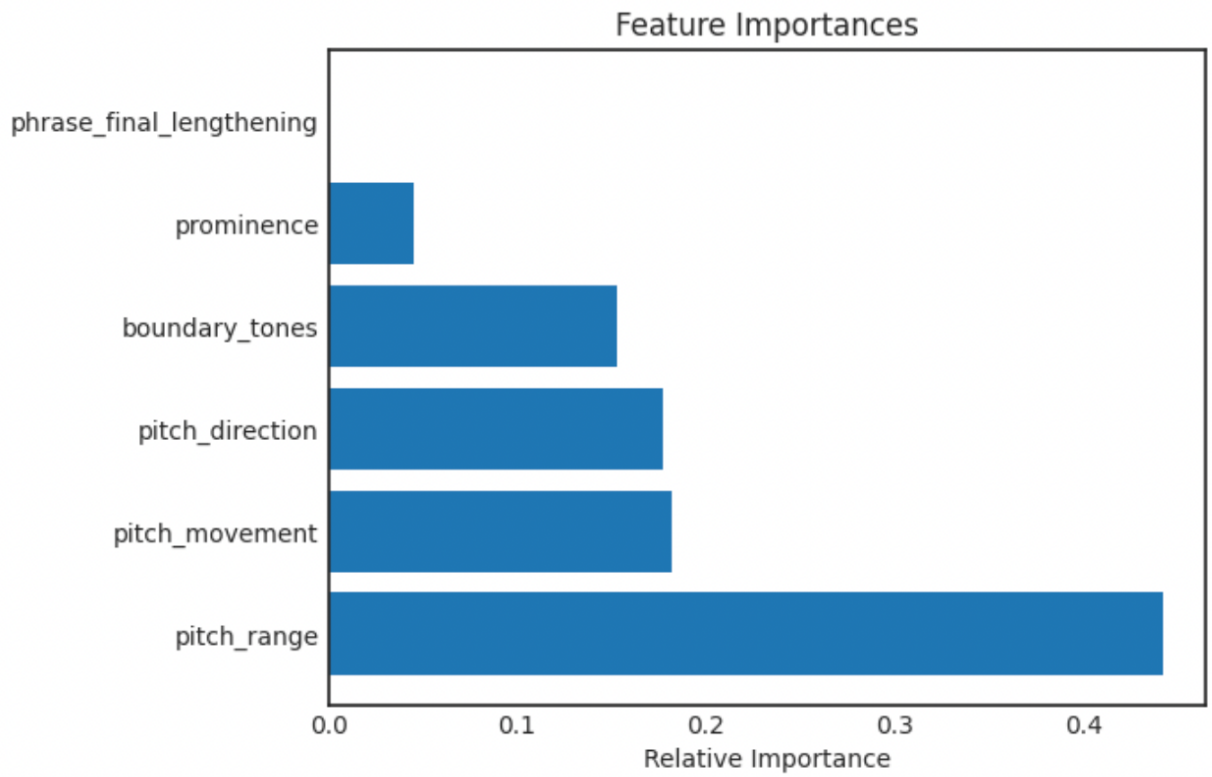

Figure 29. Feature Importances CREMA (Intonation contour approach)

Compared to Figure 23, in Figure 29 we can see that the importance of the pitch range has become even greater, but the importance of prominence has decreased significantly, while the importance of boundary_tones has increased significantly.

Metrics for CREMA dataset (Pitch level approach):

- **Accuracy**: 0.3586401504971782
- **F1 score**: 0.3213474117605809
- **Recall score**: 0.3586401504971782
- **Precision score** 0.3718902842361408

Among all the results so far, this is the worst.

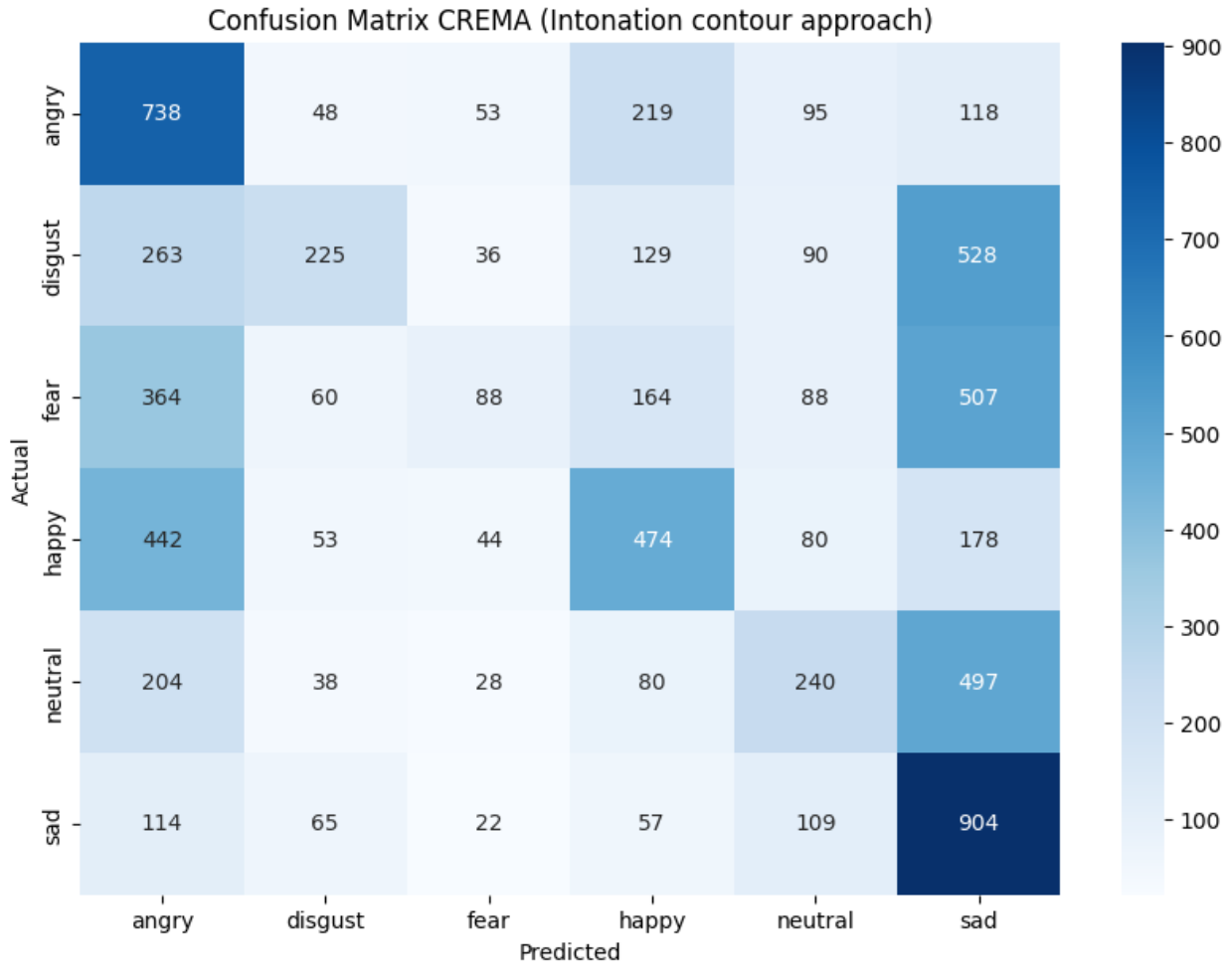

Figure 30. Confusion Matrix CREMA(Intonation contour approach)

Figure 30 perfectly shows us that the model coped extremely poorly with the task of all emotions, anger and sadness were acceptable predicted. Most likely, this result can be explained by the presence of a large number of actors who were involved in the creation of the dataset, and as a result, the feature of the pitch range increased as well. corny more different voices and it was important for the model to understand what ranges exist, but this did not help either.

Most confusing emotions:

```
Actual - Predicted

disgust - sad        528

fear - sad           507

neutral - sad        497

happy - angry        442

fear - angry         364
```

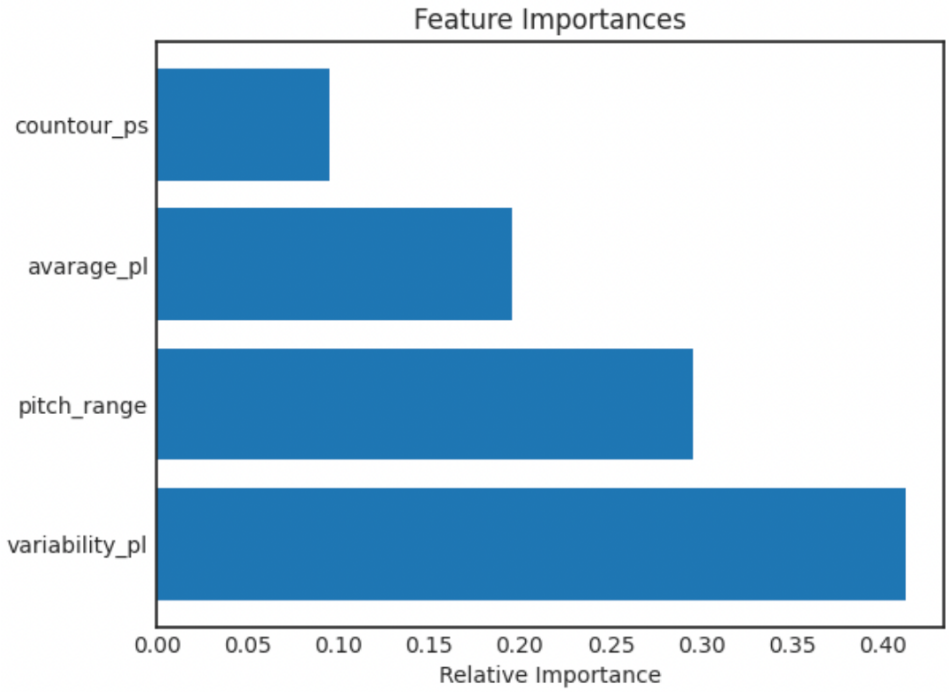

Figure 31. Feature Importances CREMA (Pitch level approach)

In the approach with the pitch level, the pitch variability became the main feature that helped to distinguish emotions from each other, here, as in the previous dataset, unlike the first two, there is a contour feature, but also slightly.

Metrics in this case:

- **Accuracy**: 0.35192152647137864
- **F1 score**: 0.3263978169384115
- **Recall score**: 0.35192152647137864
- **Precision score** 0.3790971247329874

The metrics are hardly different from the previous approach.

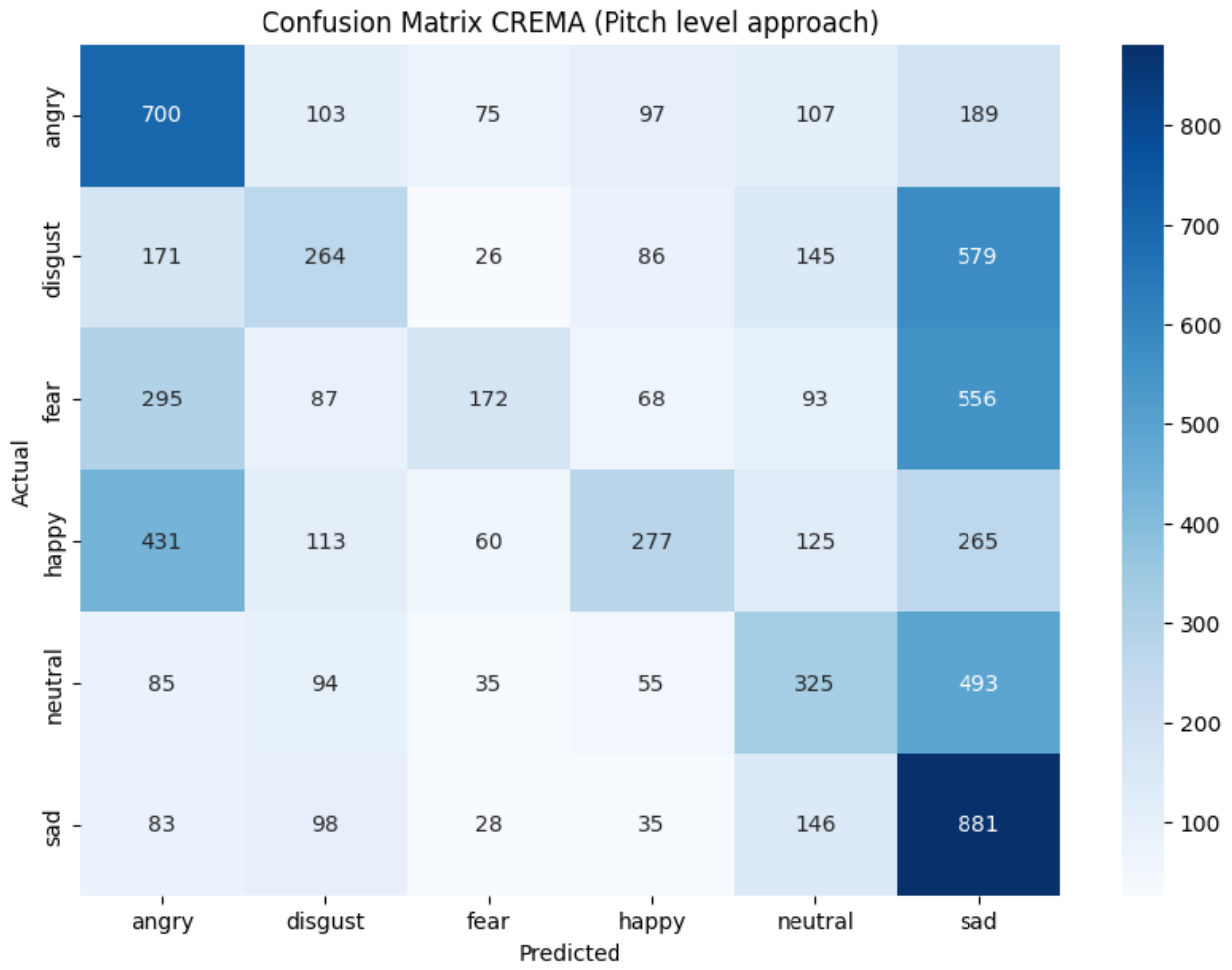

Figure 32. Confusion Matrix CREMA(Pitch level approach)

As expected, the model also performed lousy as in the previous approach, the only thing is that in this case, fear is better predicted and happiness is much worse.

List of the most confusing emotions:

```
Actual - Predicted

disgust - sad        579

fear - sad           556

neutral - sad        493
```

```
happy - angry          431

fear - angry           295
```

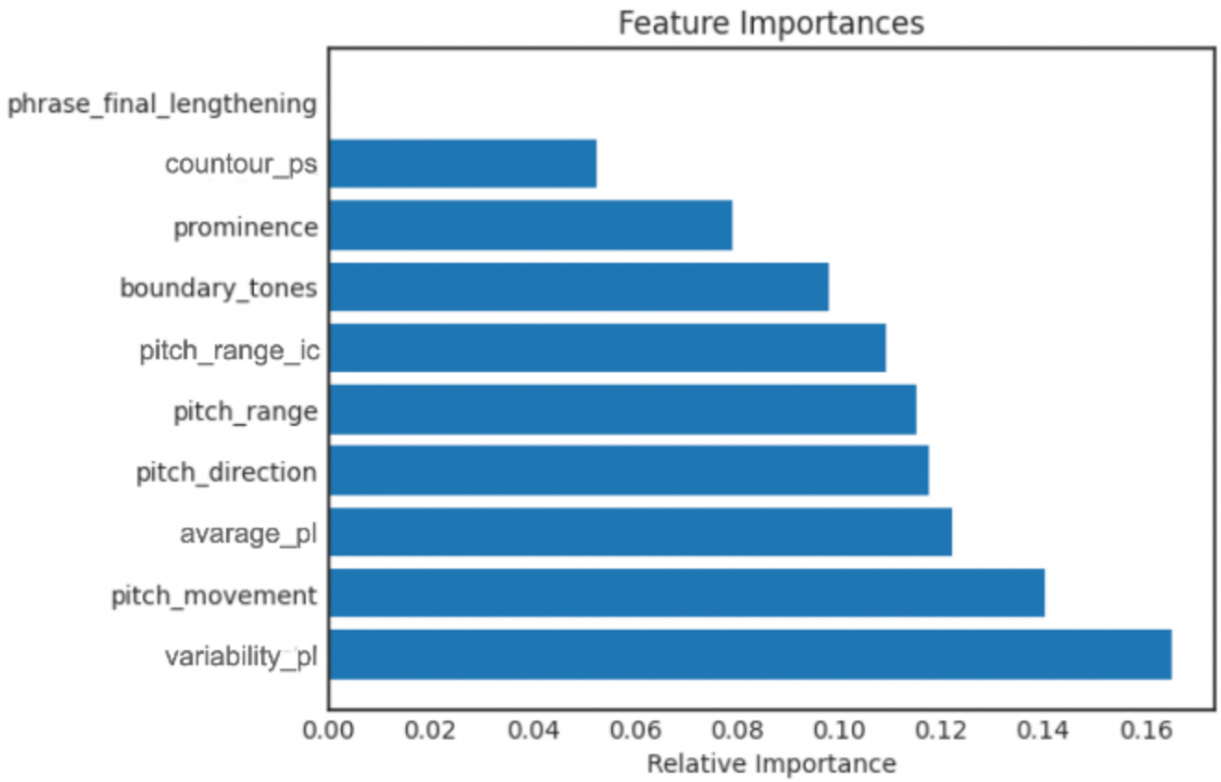

Figure 33. Feature Importances CREMA (Both approaches)

In Figure 33, we again see that 3 of the top four features in importance are pitch level features, which again tells us that pitch level is more important in the task of emotion recognition than intonation contour.

Metrics for this approach:

- **Accuracy**: 0.6561408223595807
- **F1 score**: 0.6547114203616595
- **Recall score**: 0.6561408223595807
- **Precision score** 0.6878202499023715

The metrics show a significant increase in accuracy, almost twice, which indicates that the two feature datasets work better in unison.

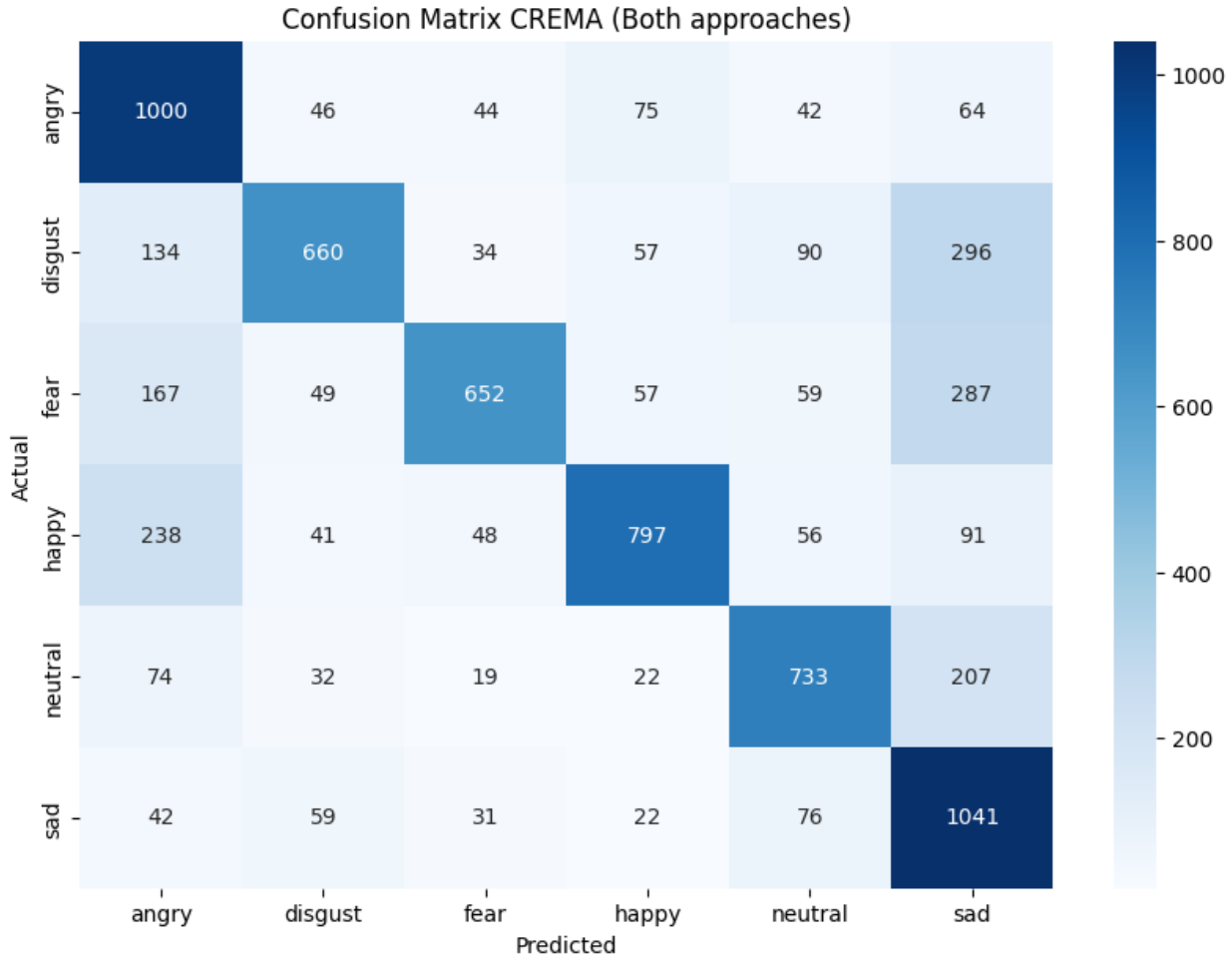

Figure 34. Confusion Matrix CREMA (Both approaches)

Figure 34 clearly shows that the increase in the number of correct predictions, compared to the two previous approaches, is observed across all emotions in the approach with two sets of features.

```
Actual - Predicted

disgust - sad        296
```

```
fear - sad          287

happy - angry       238

neutral - sad       207

fear - angry        167
```

6.5 Summary of results

	SAVEE	**RAVDESS**	**TESS**	**CREMA**
Intonation contour	73,95%	87,07%	93,69%	35,86%
Pitch level	67,91%	85,69%	94,96%	35,19%
Both approaches	86,87%	87,36%	98,07%	65,61

Table 1. Overall accuracy across all datasets.

So, summing up the general results, we can look at Table number 1, which contains the accuracy data for each approach for each dataset. The table clearly suggests that an approach with two sets of features, namely intonation contour and pitch level features, always gives a significantly better result than separate approaches. These results allow us to conclude that **both sets of features are important in recognizing emotions in audio recordings.**

The TESS and CREMA dataset deserves special attention since they showed the best and worst results in this work. I believe that such a successful result with the TESS dataset is due to the fact that it was technically much simpler than CREMA. Firstly, TESS was recorded by only two female actresses(and recordings were differ from each other only by 1 word), while CREMA has voice recordings of 91 speakers, which greatly complicates the recognition task for the model, since there is more variance in the parameters. Secondly, TESS, unlike CREMA, was

recorded by professional actresses, and it seems to me that the actors express emotions more expressively, professionally and correctly, and most likely from this we can conclude that **it is easier for emotion recognition models to understand the recordings of professional actors** than a handful of students.

Based on the obtained data and conducted by the experimenters, the main hypothesis of this study, that the intonation contour is more important in the task of sound recognition than pitch level, was tested. **This hypothesis was not confirmed**, and as a result, during the experiments, data were obtained that the **pitch level is more important** than the intonation contour in the tasks of recognizing emotions. This conclusion was made on the basis of **Figures 7, 15, 21, 27, and 33** where we clearly saw that the **pitch level features have a higher importance coefficient** compared to the intonation contour features.

On average, as in the work of Rodero E., it turned out that the **most often correctly predicted emotion is *sad***, in addition, another often correctly predicted emotion was ***angry***, this also follows from Figures 8, 9, 10, 18, 24, 28, 30, 32, 34.

On average, the models had the **most difficulty predicting the emotions of *disgust* and *fear***; this follows from Figures 10, 30, 32, 34.

The most **important features** in emotion recognition task, according to this work, are the ***pitch range*** and ***average pitch*** features, which follows from Figures 5, 6, 7, 11, 23, 25, 27, 29.

7. Conclusions

Despite the fact that the main hypothesis of this work was not confirmed, this does not negate the fact that in the course of this study, important conclusions were made regarding the recognition of emotions in oral speech. In addition to testing the main hypothesis and fulfilling the main goal of the work, all secondary goals were also met and answers were given to the questions posed. In the future, the results of this work can be used in research on emotion recognition. Thanks to this work, it can now be reliably said that intonation contour and pitch level are important components in emotion recognition and work best in combination rather than separately.

You can find all the materials related to this work in Appendix 1, in addition, in Appendix 2, 3, 4, and 5 there are links for downloading datasets for testing ready-made models that are the result of this work.

8. References

1. Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors, 20(18), 5212.

2. Basharirad B. & Moradhaseli M. (2017 October). Speech emotion recognition methods: A literature review. In AIP Conference Proceedings (Vol. 1891 No. 1 p. 020105). AIP Publishing LLC.

3. Bolinger D. (1982). Intonation and its parts. Language 505-533.

4. Cao H. Cooper D. G. Keutmann M. K. Gur R. C. Nenkova A. & Verma R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing 5(4) 377-390.

5. Crystal D. (1971). Relative and absolute in intonation analysis. Journal of the International Phonetic Association 1(1) 17-28.

6. Domingos P. (2012). A few useful things to know about machine learning. Communications of the ACM 55(10) 78-87.

7. Ekman P Sorenson ER Friesen WV. Pan-cultural elements in facial displays of emotion. Science. 1969;164(3875):86–8. pmid:5773719

8. Ekman P. (1992). An argument for basic emotions. Cognition & Emotion 6(3-4) 169-200. doi:10.1080/02699939208411068

9. H. M. Fayek M. Lech and L. Cavedon "Towards real-time Speech Emotion Recognition using deep neural networks" 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS) Cairns QLD Australia 2015 pp.

10. Han S. Leng F. & Jin Z. (2021 May). Speech emotion recognition with a ResNet-CNN-Transformer parallel neural network. In 2021 International Conference on Communications Information System and Computer Engineering (CISCE) (pp. 803-807). IEEE.

11. Izard C. E. (2013). Human emotions. Springer Science & Business Media.

12. James W. & Lange C. G. (1922). The emotions. In C. Murchison (Ed.) A history of psychology in autobiography (Vol. 1 pp. 370-396). Clark University Press.

13. Knyazev S. V. & Pozharitskaya S. K. (2009). Sovremennyj russkij literaturnyj yazyk: fonetika orfoe`piya grafika i orfografiya [Modern Russian literary language: phonetics orthoepy graphics and orthography]. Textbook for universities 166-179

14. Koduru A. Valiveti H. B. & Budati A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology 23(1) 45-55.

15. Kraus M. Wagner N. Callejas Z. & Minker W. (2021). The role of trust in proactive conversational assistants. IEEE Access 9.

16. Levis J. M. & Wichmann A. (2015). English intonation–Form and meaning. The handbook of English pronunciation 139-155.

17. Lewis M. Haviland-Jones J. M. & Barrett L. F. (Eds.). (2010). Handbook of emotions. Guilford Press.

18. Livingstone SR Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

19. Luna-Jiménez C. Kleinlein R. Griol D. Callejas Z. Montero J. M. & Fernández-Martínez F. (2021). A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset. Applied Sciences 12(1) 327.

20. McTear M. Callejas Z. & Griol D. (2016). The conversational interface: Talking to smart devices: Springer International publishing.

21. Pell M. D. Paulmann S. Dara C. Alasseri A. & Kotz S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. Journal of Phonetics 37(4) 417-435.

22. Pike K. L. (1945). The Intonation of American English.

23. Pike K. L. (1964). Tone Languages: A Technique for Determining the Number and Type of Pitch Contrasts in a Language with Studies in Tonemic Substitution and Fusion.

24. Rodero E. (2011). Intonation and emotion: influence of pitch levels and contour type on creating emotions. Journal of voice 25(1) e25-e34.

25. S. Haq and P. J. B. Jackson "Machine Audition: Principles Algorithms and Systems" Hershey PA 2010 pp. 398-423.

26. Scherer K. R. (2005). What are emotions? And how can they be measured? Social science information 44(4) 695-729.

27. Schuller B. W. (2012). Acoustic Emotion Recognition: A Review of the State of the Art.

28. Tuncer T. Dogan S. & Acharya U. R. (2021). Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowledge-Based Systems 211 106547.

9. Appendix

Appendix 1. Repository with the source code models and dataframes with results [URL: https://github.com/letitself/bachelor_thesis]

Appendix 2. RAVDESS dataset [URL: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio]

Appendix 3. TESS dataset [URL: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess]

Appendix 4. CREMA dataset [URL: https://www.kaggle.com/datasets/ejlok1/cremad]

Appendix 5. SAVEE dataset[URL: https://www.kaggle.com/datasets/barelydedicated/savee-database]