Task 1 (Feature Analysis):
Extract six features from each speech segment:
- The min, max, mean of pitch
- The min, max, mean of intensity

Since each speaker naturally has a different pitch range and other voice qualities like intensity, you need to normalize the features accordingly by speaker. There are at least two ways you may want to normalize. Please specify your method and provide a detailed description of how you calculated it and why you chose it.

Normalization Method: Z-score normalization

I employed Z-score normalization to standardize features using speaker-specific statistics. The process began by extracting raw pitch and intensity values from all audio files for each speaker, iterating through all .wav files while filtering out zeros and NaN values. For each speaker, I then concatenated all valid pitch values from all utterances across all emotions into a single vector, and similarly concatenated all valid intensity values into another vector. From these concatenated vectors, I calculated each speaker's overall mean (μ) and standard deviation (σ, with ddof = 0).

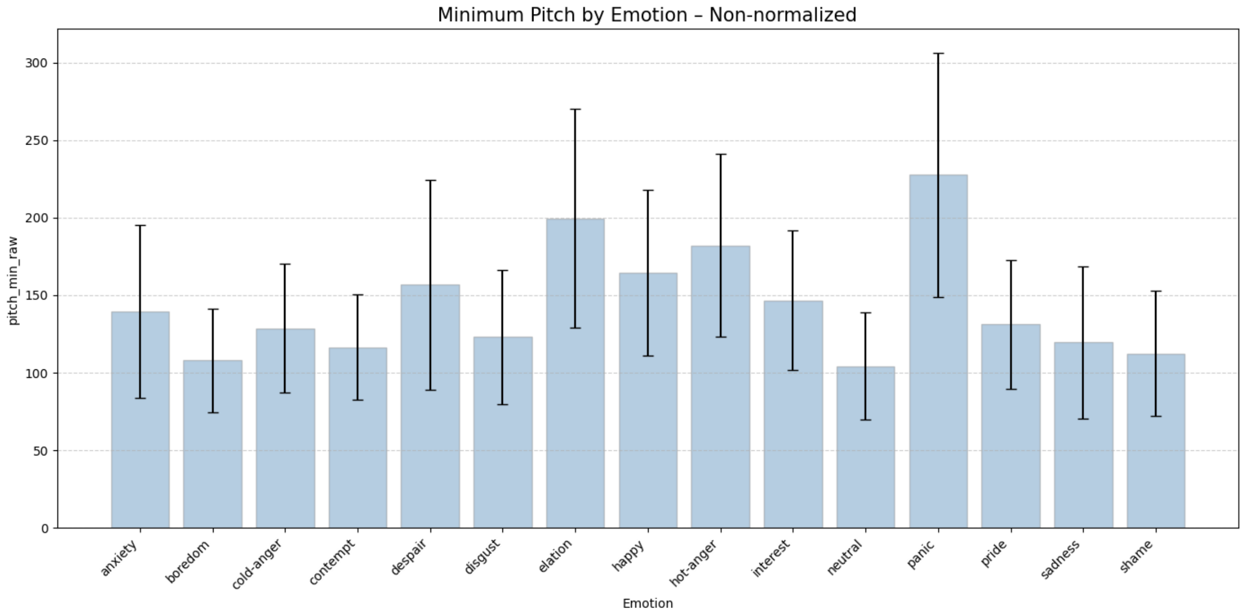
During feature extraction for individual speech segments, I applied the Z-score transformation formula ($Z = (x - \mu)/\sigma$), converting each raw sample (x) to a Z-score by subtracting the speaker's overall mean for that feature type and dividing by the corresponding standard deviation. This normalization recenters each speaker's distribution to have a mean of zero and unit variance. After normalization, I calculated the six required features (minimum, maximum, and mean values for both pitch and intensity) using these normalized values.

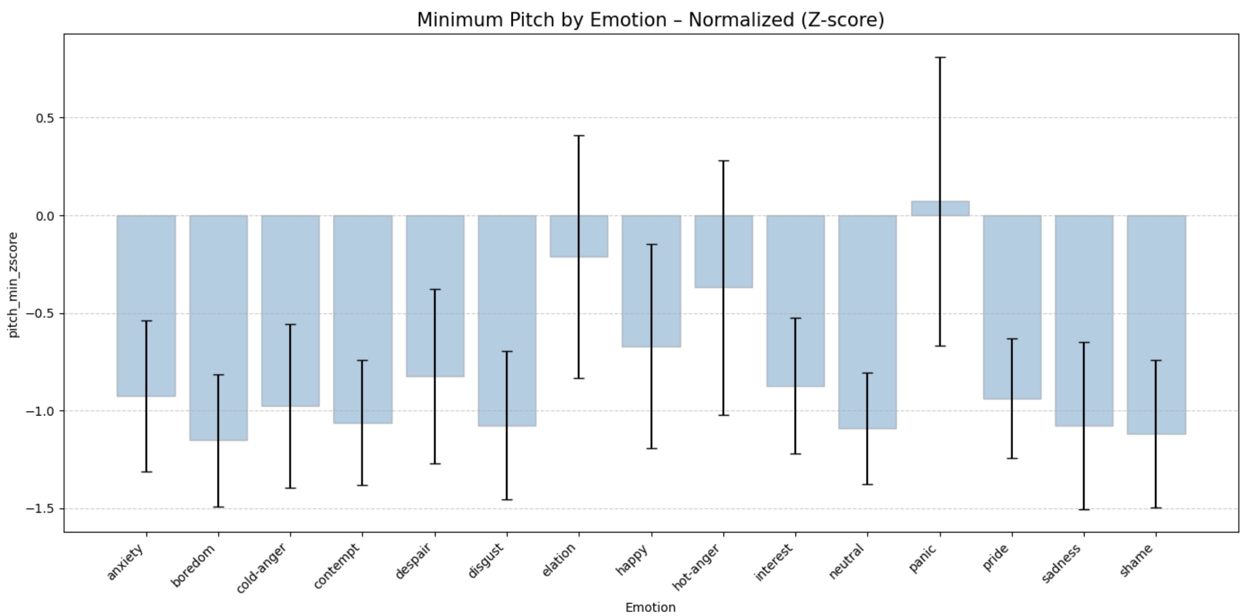
I selected Z-score normalization for its statistical robustness, as it relies solely on per-speaker means and standard deviations, making it effective even when acoustic features deviate from normal distributions. By recentering each speaker's pitch and intensity distributions at zero and rescaling them to unit variance, this method neutralizes variability stemming from physiological differences such as vocal tract length or vocal fold mass, enabling valid cross-speaker comparisons. Crucially, Z-score normalization preserves each speaker's internal prosodic relationships, ensuring that deviations from an individual's baseline remain interpretable. This approach effectively standardizes features across different speakers while maintaining the relative patterns within each speaker's emotional expressions.

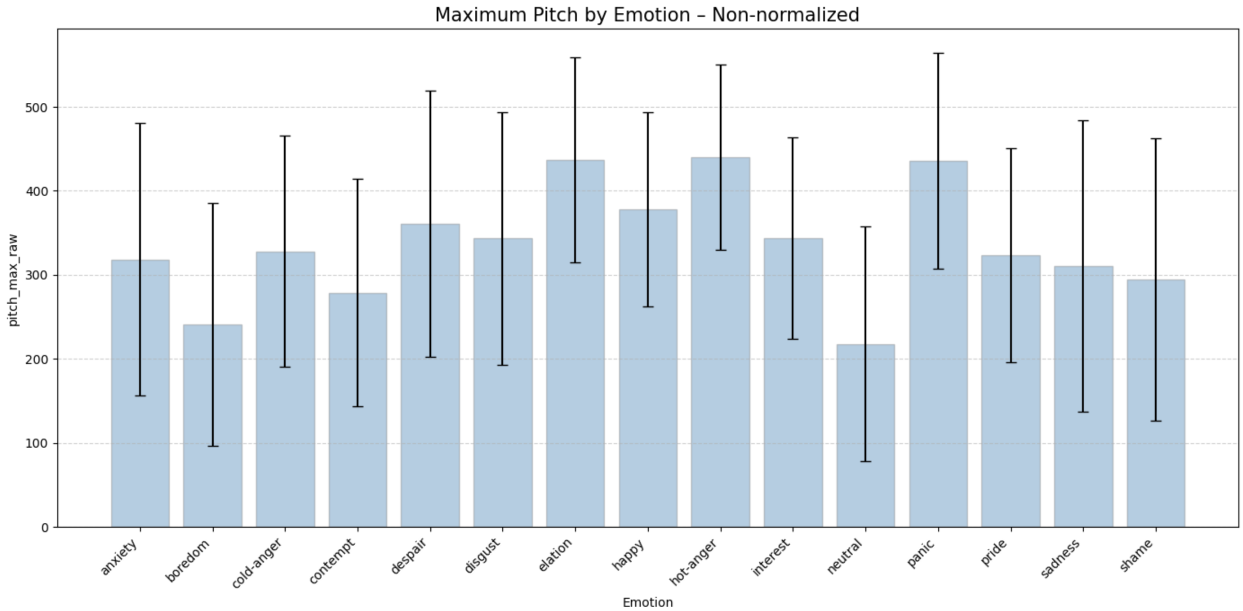
Compared to neutral baseline normalization, the Z-score approach offers several advantages. It derives μ and σ from each speaker's entire corpus of utterances across all emotions, providing more stable statistics than neutral baseline normalization, which typically relies on only eight to ten neutral clips per speaker. This larger sample size reduces sampling noise and yields more reliable normalized features. Additionally, whereas neutral baseline normalization can fail when neutral data is insufficient, Z-score normalization remains robust regardless of class imbalance, making it particularly suitable for datasets with uneven emotional category distributions.

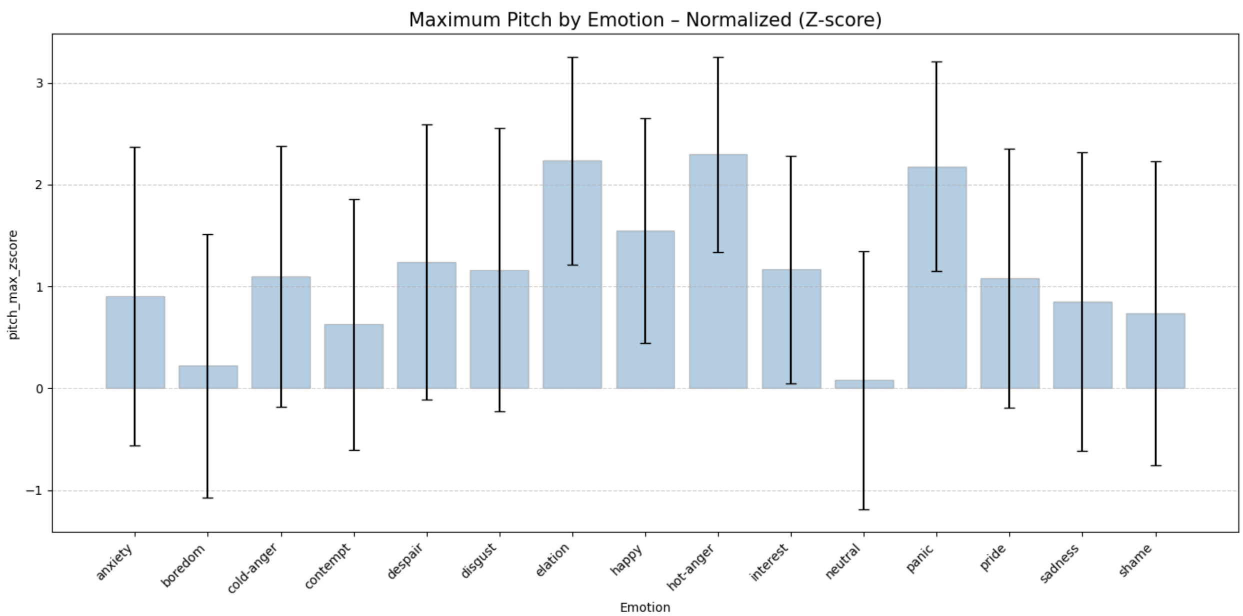
You need to turn in plots of the mean and standard deviation of each feature for all of the 15 emotion classes. Please also specify for each plot whether it was created a) without normalization; b) with normalization (tell us what normalization method you used, how you calculated it, and why you chose this method). Specifically, create 2 plots for each feature, one without normalization, and one with normalization.

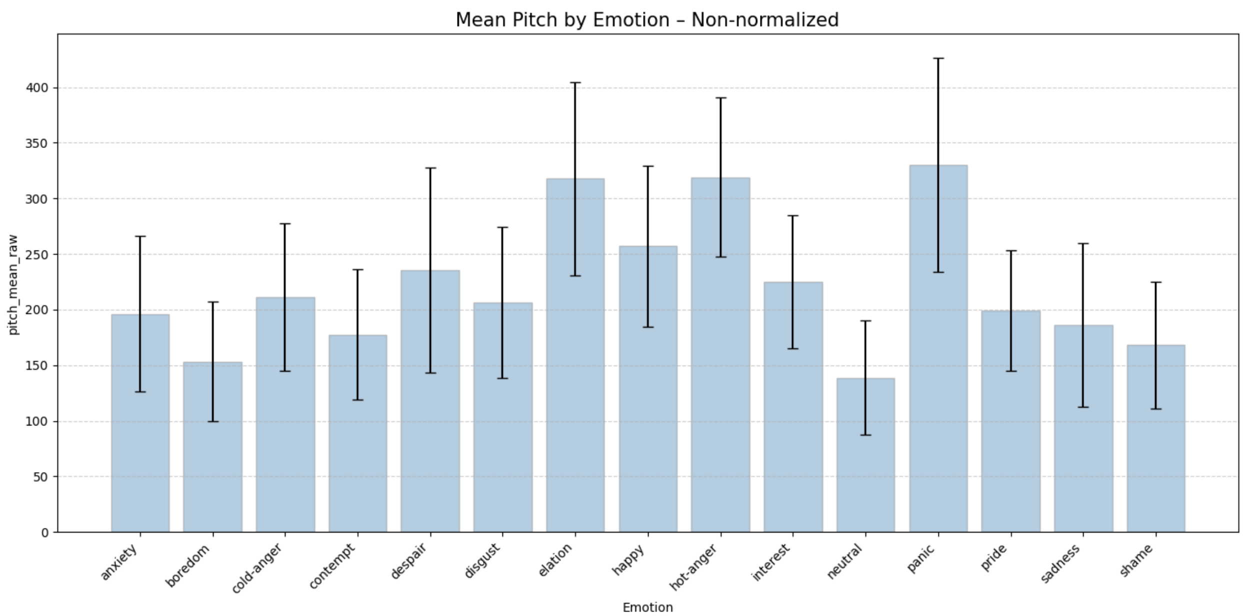
Plots:

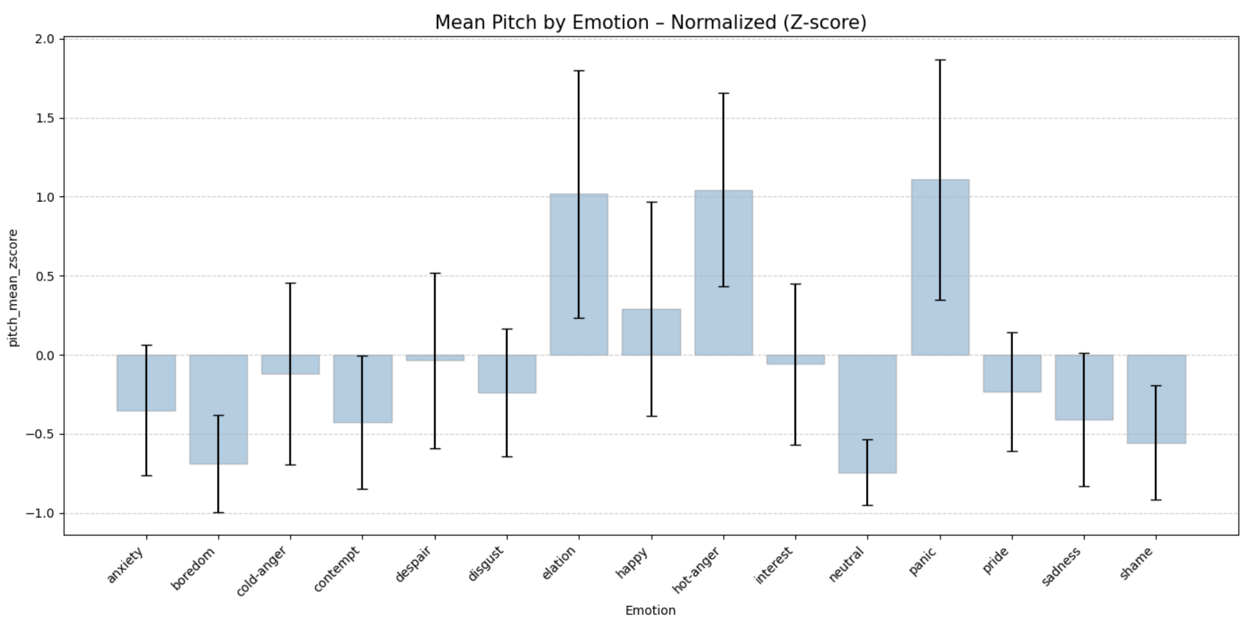

Minimum Pitch by Emotion – Non-normalized
pitch_min_raw
0
50
100
150
200
250
300
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion

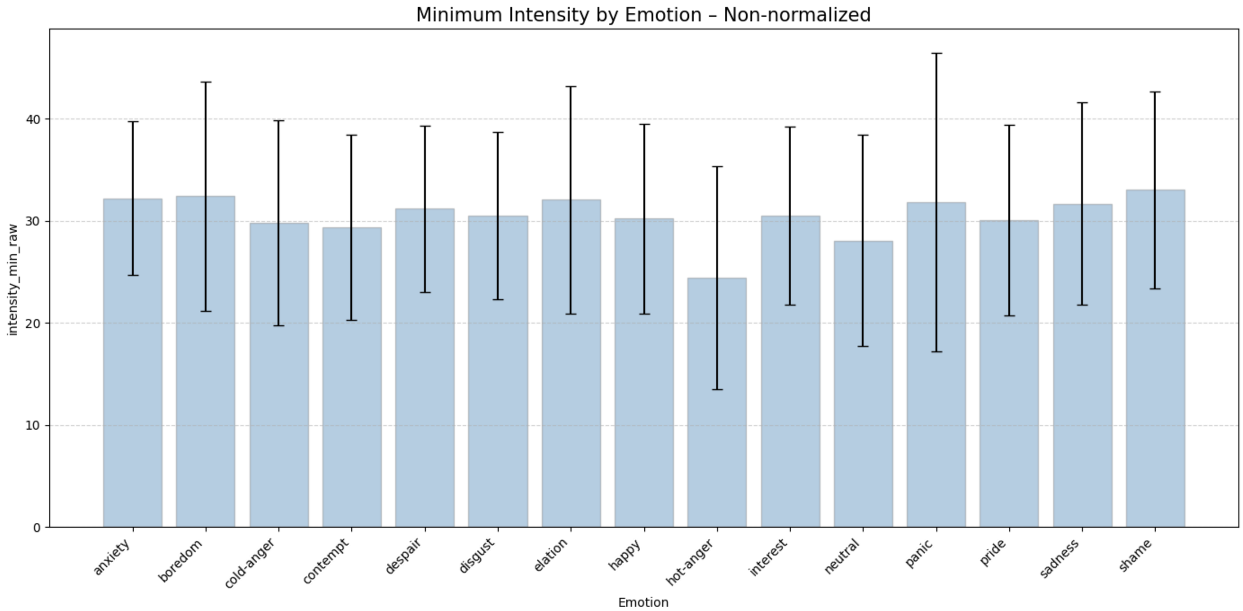

Minimum Pitch by Emotion – Normalized (Z-score)
pitch_min_zscore
0.5
0.0
−0.5
−1.0
−1.5
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion

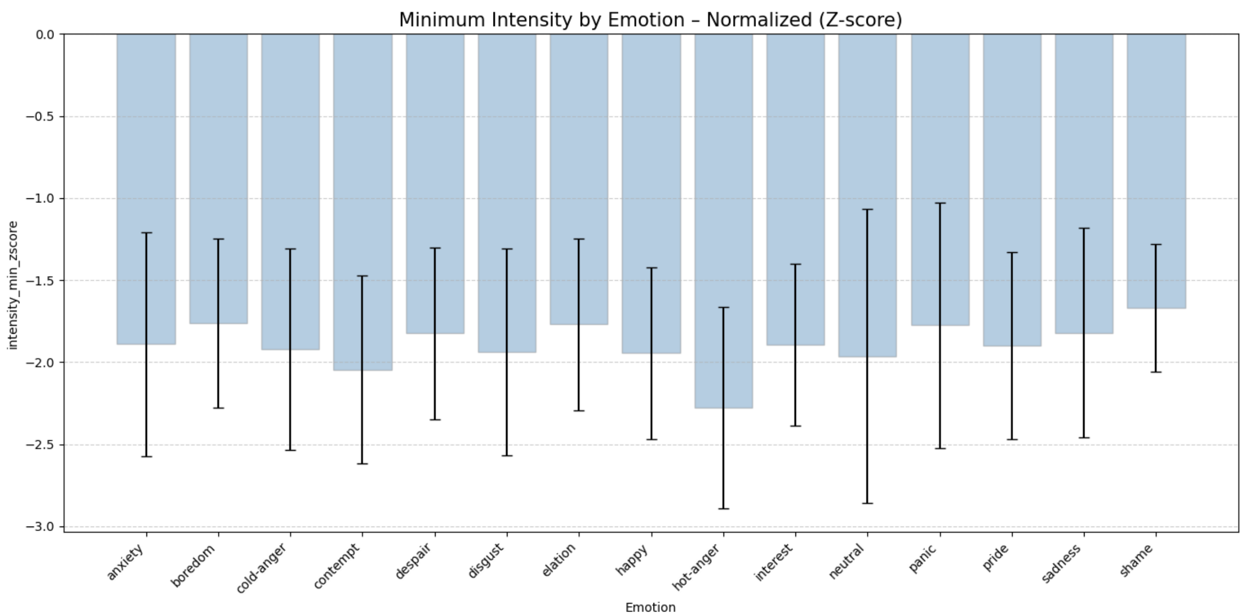

Maximum Pitch by Emotion – Non-normalized
pitch_max_raw
0
100
200
300
400
500
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion

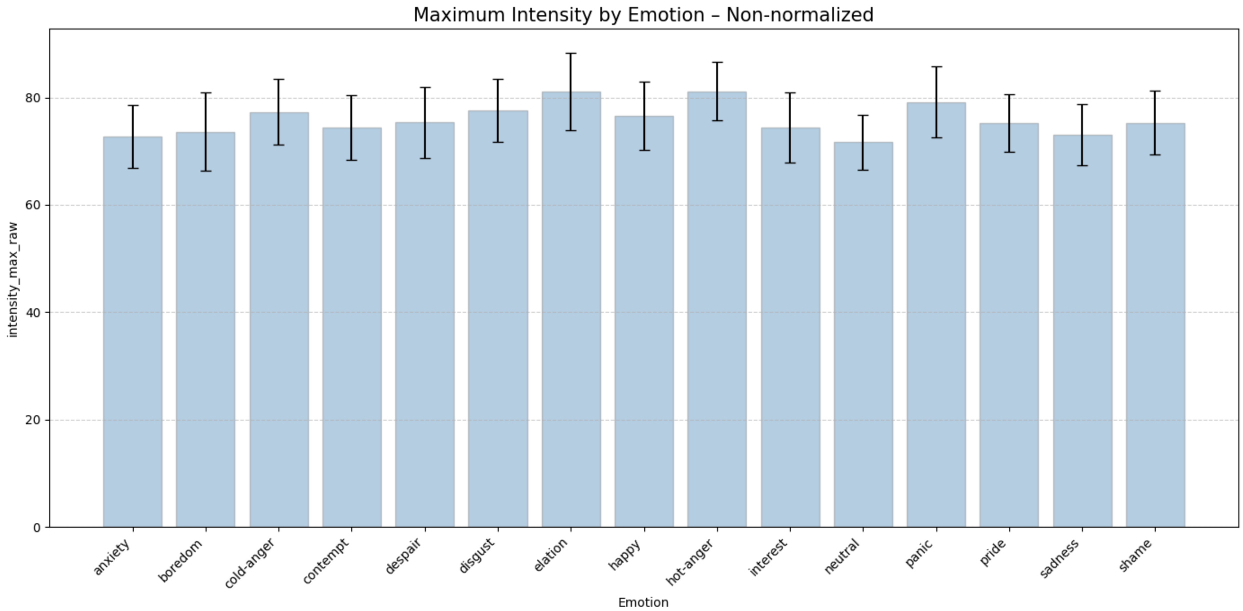

Maximum Pitch by Emotion – Normalized (Z-score)
pitch_max_zscore
−1
0
1
2
3
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion


Mean Pitch by Emotion – Non-normalized
pitch_mean_raw
0
50
100
150
200
250
300
350
400
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion

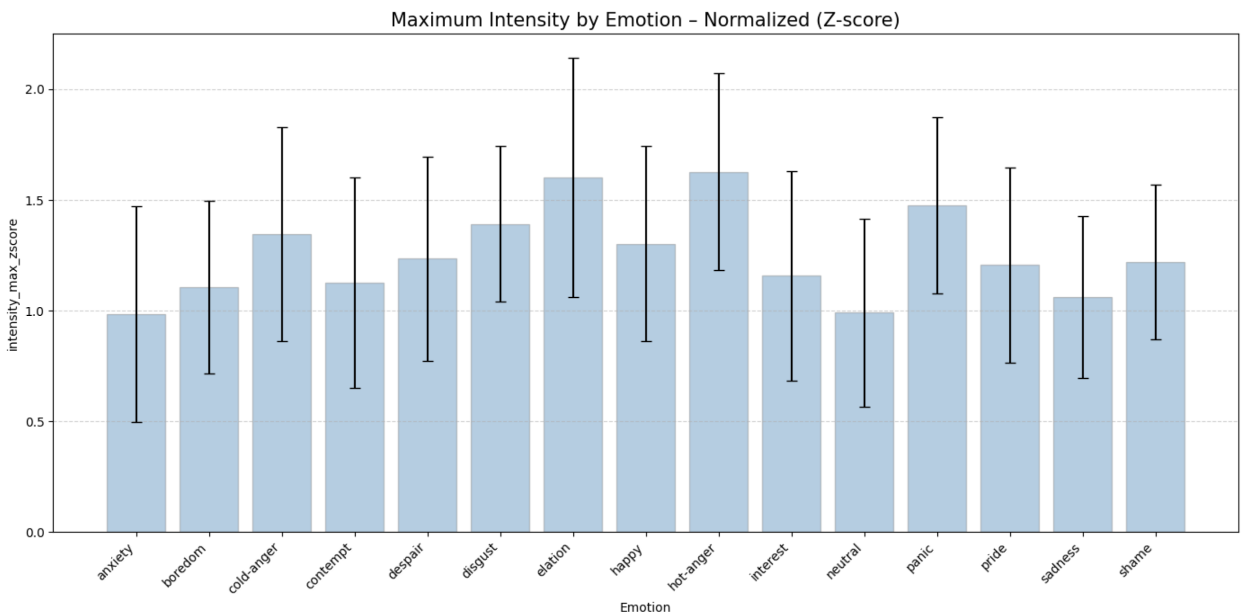

Mean Pitch by Emotion – Normalized (Z-score)
pitch_mean_zscore
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion


Minimum Intensity by Emotion – Non-normalized
intensity_min_raw
0
10
20
30
40
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion

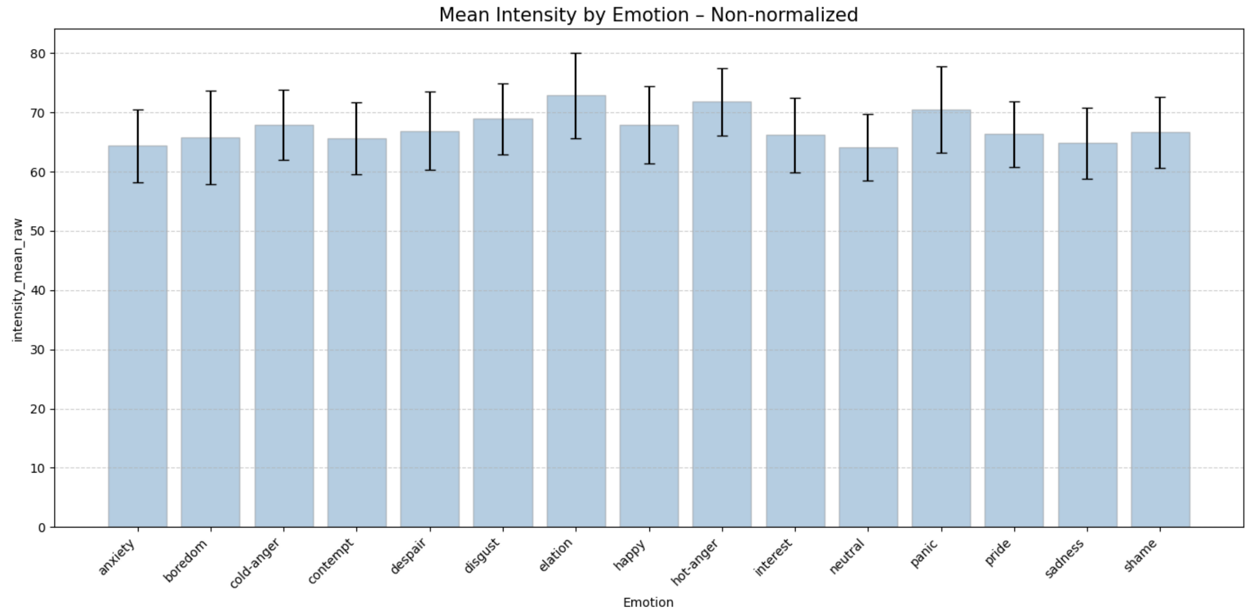

Minimum Intensity by Emotion – Normalized (Z-score)
intensity_min_zscore
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion

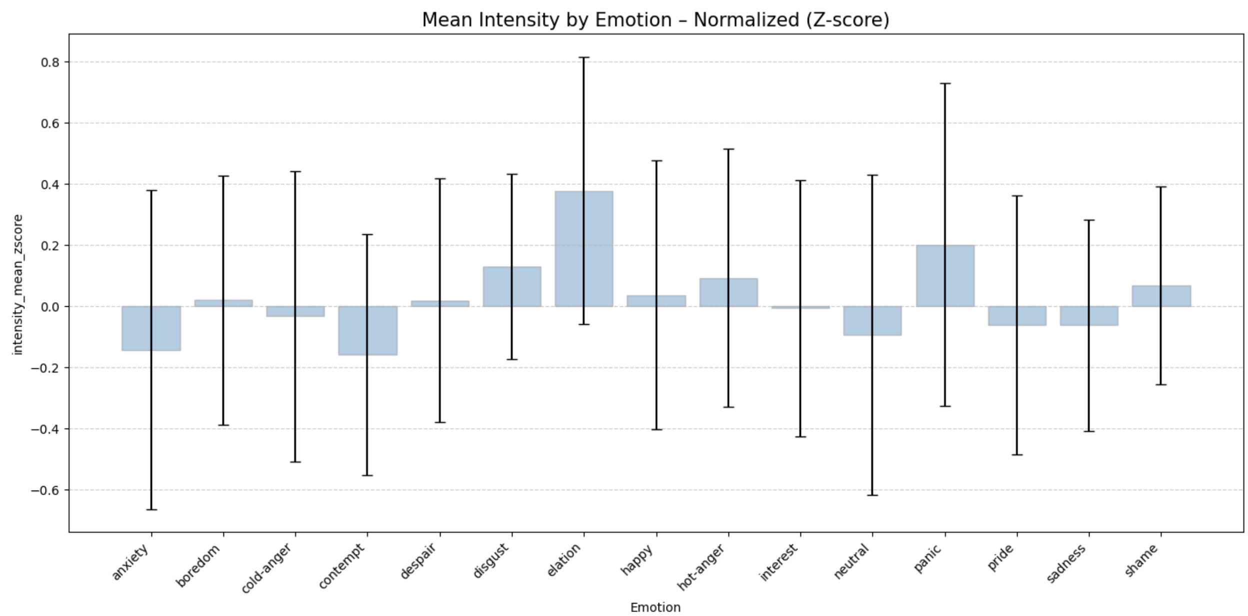


Maximum Intensity by Emotion – Non-normalized
intensity_max_raw
0
20
40
60
80
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion



Maximum Intensity by Emotion – Normalized (Z-score)
intensity_max_zscore
0.0
0.5
1.0
1.5
2.0
anxiety
boredom
cold-anger
contempt
despair
disgust
elation
happy
hot-anger
interest
neutral
panic
pride
sadness
shame
Emotion





Interesting Observations:

1. Examining the mean pitch and mean intensity plots, boredom and neutral emotions exhibit the lowest values in both raw and normalized versions. After Z-score normalization, these emotions remain approximately 0.7 standard deviations below the mean, reflecting their characteristically low physiological arousal states. The remarkably small standard deviation whiskers for boredom in the pitch plots indicate monotone-like delivery.
2. Although sadness and despair share low mean pitch values, their acoustic patterns diverge notably. Sadness maintains consistently low intensity while showing slightly elevated maximum pitch (+0.85σ) compared to neutral or boredom. In contrast, despair exhibits a much wider pitch range, evidenced by significantly larger standard deviation whiskers in the pitch plots. This pattern suggests that despair involves occasional pitch excursions, creating intonation patterns absent in ordinary sadness.
3. The maximum and minimum intensity plots reveal that hot anger, cold anger, and panic display the largest intensity variance, with standard deviation whiskers spanning more than 2σ. This wide range

captures the dynamic bursts characteristic of these emotions. Notably, contempt shows a distinctive pattern with moderate variance but consistently negative mean values across intensity measures, suggesting that speakers reduce their volume when expressing this emotion, a feature that differentiates it from other negative emotions.

4. The maximum pitch plots with Z-score normalization clearly demonstrate that high-arousal emotions (panic, hot anger, and elation) exhibit significantly elevated maximum pitch values, approximately 2.2 standard deviations above the speaker mean. These three emotions stand apart from other emotional classes with values exceeding 2.0 on the y-axis, suggesting that intense emotions push speakers toward higher frequency ceilings.

5. While the maximum pitch gap between elation and hot anger narrows after normalization, their intensity difference actually widens, with hot anger showing greater loudness. Elation demonstrates dramatic increases in pitch metrics (mean +1σ) but only modest rises in intensity (+0.38σ). These plots reveal trade-offs between pitch and intensity across emotions, suggesting that these acoustic features track different dimensions of emotional expression and that combining both provides valuable complementary information for classification systems.

Task 2 (Classification Experiments):

Extract a set of acoustic-prosodic features using the openSMILE toolkit. Normalize your extracted features (as in Part 1.Feature Analysis) and use leave-one-speaker-out cross-validation to predict the emotion classes. Leave-one-speaker-out cross-validation means, for each speaker S, train on all other six other speakers combined and test on S. Report the classification results (screenshot or copy-paste sklearn classification reports) for all 7 experiments (leave one speaker out as the test set in each experiment). Also, compute and report aggregated average accuracy and weighted F1 scores over all the experiments and emotions.

```
report for speaker: cc
              precision    recall  f1-score   support

     anxiety      0.071     0.200     0.105        10
     boredom      0.061     0.133     0.083        15
  cold-anger      0.071     0.067     0.069        15
    contempt      0.400     0.455     0.426        22
     despair      0.091     0.222     0.129         9
     disgust      0.400     0.258     0.314        31
     elation      0.250     0.438     0.318        16
       happy      0.286     0.261     0.273        23
   hot-anger      0.435     0.714     0.541        14
    interest      0.083     0.059     0.069        17
     neutral      0.286     0.111     0.160        18
       panic      0.455     0.278     0.345        18
       pride      0.444     0.174     0.250        23
     sadness      0.333     0.154     0.211        13
       shame      0.500     0.143     0.222        21

    accuracy                          0.245       265
   macro avg      0.278     0.244     0.234       265
weighted avg      0.306     0.245     0.249       265
```

report for speaker: cl

	precision	recall	f1-score	support
anxiety	0.194	0.333	0.246	21
boredom	0.298	0.483	0.368	29
cold-anger	0.458	0.407	0.431	27
contempt	0.357	0.400	0.377	25
despair	0.219	0.241	0.230	29
disgust	0.133	0.091	0.108	22
elation	0.200	0.222	0.211	27
happy	0.280	0.333	0.304	21
hot-anger	0.520	0.500	0.510	26
interest	0.346	0.346	0.346	26
neutral	0.000	0.000	0.000	17
panic	0.267	0.190	0.222	21
pride	0.308	0.167	0.216	24
sadness	0.111	0.074	0.089	27
shame	0.212	0.269	0.237	26
accuracy			0.280	368
macro avg	0.260	0.271	0.260	368
weighted avg	0.268	0.280	0.268	368

report for speaker: gg

	precision	recall	f1-score	support
anxiety	0.370	0.567	0.447	30
boredom	0.282	0.367	0.319	30
cold-anger	0.333	0.519	0.406	27
contempt	0.304	0.269	0.286	26
despair	0.000	0.000	0.000	28
disgust	0.611	0.216	0.319	51
elation	0.333	0.571	0.421	28
happy	0.259	0.500	0.341	30
hot-anger	0.769	0.455	0.571	22
interest	0.179	0.167	0.172	30
neutral	0.000	0.000	0.000	9
panic	0.550	0.407	0.468	27
pride	0.200	0.200	0.200	25
sadness	0.000	0.000	0.000	33
shame	0.200	0.208	0.204	24
accuracy			0.302	420
macro avg	0.293	0.296	0.277	420
weighted avg	0.313	0.302	0.286	420

report for speaker: jg

```
              precision    recall  f1-score   support

     anxiety      0.120     0.158     0.136        19
     boredom      0.176     0.214     0.194        14
  cold-anger      0.182     0.182     0.182        22
    contempt      0.227     0.217     0.222        23
     despair      0.250     0.095     0.138        21
     disgust      0.292     0.304     0.298        23
     elation      0.133     0.100     0.114        20
       happy      0.042     0.050     0.045        20
   hot-anger      0.333     0.333     0.333        18
    interest      0.381     0.421     0.400        19
     neutral      0.000     0.000     0.000         8
       panic      0.333     0.286     0.308        14
       pride      0.000     0.000     0.000        18
     sadness      0.273     0.316     0.293        19
       shame      0.077     0.067     0.071        15

    accuracy                          0.190       273
   macro avg      0.188     0.183     0.182       273
weighted avg      0.197     0.190     0.190       273


report for speaker: mf
              precision    recall  f1-score   support

     anxiety      0.393     0.500     0.440        22
     boredom      0.312     0.185     0.233        27
  cold-anger      0.103     0.150     0.122        20
    contempt      0.581     0.409     0.480        44
     despair      0.269     0.438     0.333        16
     disgust      0.053     1.000     0.100         1
     elation      0.043     0.038     0.041        26
       happy      0.143     0.087     0.108        23
   hot-anger      0.556     0.476     0.513        21
    interest      0.067     0.053     0.059        19
     neutral      0.500     0.700     0.583        10
       panic      0.471     0.667     0.552        12
       pride      0.062     0.056     0.059        18
     sadness      0.143     0.100     0.118        20
       shame      0.368     0.350     0.359        20

    accuracy                          0.281       299
   macro avg      0.271     0.347     0.273       299
weighted avg      0.296     0.281     0.279       299


report for speaker: mk
              precision    recall  f1-score   support

     anxiety      0.048     0.069     0.056        29
```

```
    boredom      0.160     0.200     0.178        20
  cold-anger     0.154     0.261     0.194        23
   contempt     0.176     0.286     0.218        21
    despair     0.333     0.151     0.208        53
    disgust     0.045     0.048     0.047        21
    elation     0.211     0.348     0.262        23
      happy     0.297     0.262     0.278        42
   hot-anger     0.312     0.227     0.263        22
   interest     0.357     0.227     0.278        44
    neutral     1.000     0.500     0.667         8
      panic     0.417     0.476     0.444        21
      pride     0.059     0.043     0.050        23
    sadness     0.087     0.091     0.089        22
      shame     0.208     0.200     0.204        25

   accuracy                         0.209       397
   macro avg     0.258     0.226     0.229       397
weighted avg     0.241     0.209     0.214       397


report for speaker: mm
              precision    recall  f1-score   support

    anxiety     0.481     0.333     0.394        39
    boredom      0.444     0.421     0.432        19
  cold-anger     0.222     0.200     0.211        20
   contempt     0.421     0.421     0.421        19
    despair     0.147     0.278     0.192        18
    disgust     0.200     0.130     0.158        23
    elation     0.077     0.105     0.089        19
      happy     0.316     0.667     0.429        18
   hot-anger     0.692     0.562     0.621        16
   interest     0.233     0.333     0.275        21
    neutral     0.500     0.111     0.182         9
      panic     0.500     0.143     0.222        28
      pride     0.353     0.316     0.333        19
    sadness     0.188     0.176     0.182        17
      shame     0.333     0.412     0.368        17

   accuracy                         0.305       302
   macro avg     0.341     0.307     0.301       302
weighted avg     0.345     0.305     0.303       302


aggregated results:
accuracy = 0.261
weighted F1 = 0.257

kept 128 / 382 features.
```

In the data preprocessing step, I dropped the columns 'frameTime', 'F0_sma_min', and 'F0_sma_minPos'. The 'frameTime' and 'F0_sma_min' columns contained zeros across all 2,324 samples, while 'F0_sma_minPos' exhibited zeros in 2,316 instances. Including these features would be computationally inefficient, increasing dimensionality without contributing discriminative information for emotional classification. Moreover, these features would cause NaN propagations during Z-score normalization, as division by zero standard deviation in the Z-score formula produces undefined results. Therefore, I removed these features to ensure numerical stability and computational efficiency.

```
columns containing 0:

frameTime                  2324
F0_sma_min                 2324
F0_sma_minPos              2316
```

Regarding the classifier, you can use either a traditional machine learning model, such as random forest and SVM, or a neural network model. Report the type and structure of the model you use. Please avoid excessive tuning of the hyperparameters of the classifier you use, since you want to avoid overfitting the dataset.

Model Used: Random forest feature selector + RBF-SVM

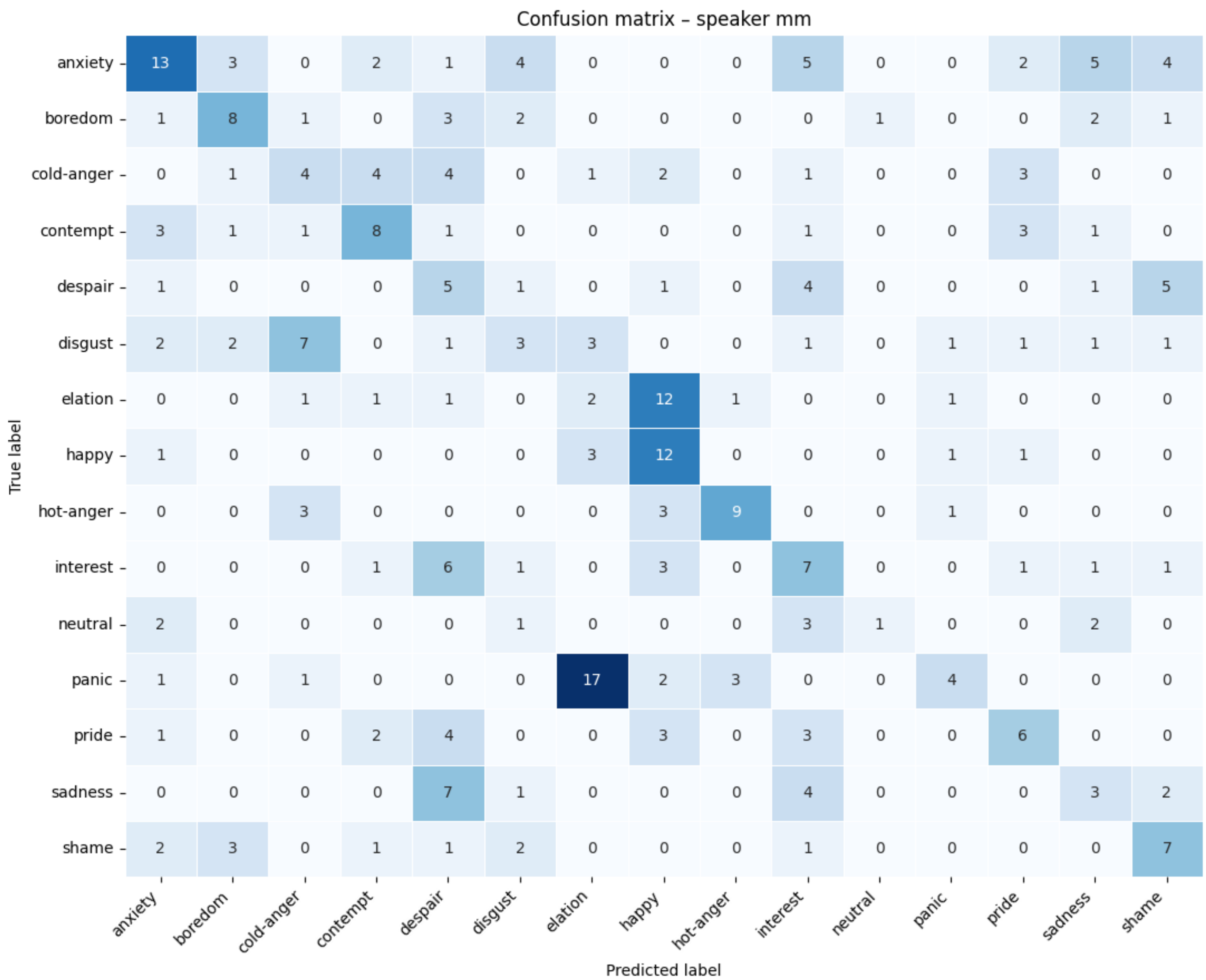
I employed a Random Forest feature selector with 100 estimators to identify the most informative acoustic features while maintaining balanced class representation. This selector reduced dimensionality from 382 to 128 features by retaining only those exceeding the mean importance threshold, effectively removing redundant information while preserving discriminative power across all 15 emotion classes. The Random Forest's ensemble nature proves particularly valuable for speech emotion features as it captures complex interactions between acoustic parameters that simple univariate methods might overlook.

For classification, I implemented an SVM with RBF kernel (C=10, gamma='scale') to create complex non-linear decision boundaries capable of separating overlapping emotion categories. The RBF kernel transforms the feature space to enable nuanced separation between acoustically similar emotions where linear approaches would fail. The C=10 parameter provides sufficient flexibility for the model to adapt to intricate emotion patterns while avoiding overfitting. To address the inherent class imbalance in emotional speech data, I applied class weights inversely proportional to class frequencies, ensuring that less frequent emotions receive appropriate attention during training.

Task 3: Error Analysis

Analyze the errors made by your best performing leave-one-speaker-out experiment, i.e. the best results you got for one of the 7 speakers. What do you observe from the results you got for this speaker overall? And, in more detailed observations, which class(es) were easiest to predict? Why do you think they were easy? Which were the most difficult? Why do you think they were difficult? Based on this analysis, what ideas do you have to further improve your classifier?

Best Speaker: mm



per–class metrics (speaker mm)

	precision	recall	f1	support	top 5 confusions:
anxiety	0.481	0.333	0.394	39	interest, sadness, disgust, shame, boredom
boredom	0.444	0.421	0.432	19	despair, sadness, disgust, cold-anger, shame
cold-anger	0.222	0.200	0.211	20	contempt, despair, pride, happy, interest
contempt	0.421	0.421	0.421	19	anxiety, pride, cold-anger, despair, sadness
despair	0.147	0.278	0.192	18	shame, interest, anxiety, disgust, happy
disgust	0.200	0.130	0.158	23	cold-anger, elation, boredom, anxiety, sadness
elation	0.077	0.105	0.089	19	happy, panic, contempt, despair, hot-anger
happy	0.316	0.667	0.429	18	elation, pride, panic, anxiety
hot-anger	0.692	0.562	0.621	16	cold-anger, happy, panic
interest	0.233	0.333	0.275	21	despair, happy, pride, disgust, sadness
neutral	0.500	0.111	0.182	9	interest, sadness, anxiety, disgust
panic	0.500	0.143	0.222	28	elation, hot-anger, happy, anxiety, cold-anger
pride	0.353	0.316	0.333	19	despair, interest, happy, contempt, anxiety
sadness	0.188	0.176	0.182	17	despair, interest, shame, disgust
shame	0.333	0.412	0.368	17	boredom, disgust, anxiety, despair, interest

```
Overall accuracy   : 0.304635761589404
Overall weighted F1: 0.303074163743047
```

My error analysis of the best-performing speaker (mm) reveals moderate performance despite being the highest-scoring fold, with an overall accuracy of approximately 30.5% and weighted F1-score of 30.3% across the 15 emotion classes. This suggests that the model struggles to effectively classify emotions even for the optimal speaker.

The confusion matrix and per-class metrics reveal substantial performance variations across emotions. Hot anger (F1=0.621), happy (F1=0.429), and anxiety (F1=0.394) achieved the highest classification success. These emotions likely benefited from distinctive acoustic signatures that the selected features captured effectively. Hot anger typically exhibits high intensity and wide pitch range, while anxiety often presents with increased speech rate and vocal jitter patterns. Additionally, these classes had reasonable support sizes (16-39 samples), providing adequate training examples for model learning.

Conversely, elation (F1=0.089), disgust (F1=0.158), and despair (F1=0.192) proved most challenging to classify. The confusion matrix shows elation samples were predominantly misclassified as happy and panic, indicating the model's difficulty in differentiating between high-arousal emotions with different valence characteristics. Neutral speech was consistently confused with interest, sadness, and anxiety (all low-intensity emotional states) suggesting inadequate capture of neutrality's acoustic signature. Similarly, psychologically proximate emotions (disgust, contempt, and shame) were frequently confused with one another, all achieving F1 scores below 0.3.

These classification difficulties stem from several factors. First, the IS09 feature set has inherent limitations in distinguishing acoustically similar emotions, particularly subtle differences between emotions like disgust and contempt. The Random Forest selector can only operate on available features, potentially missing crucial acoustic markers absent from the original set. Second, there is a class imbalance problem where the neutral emotion has only 9 examples for this speaker which is insufficient for effective learning despite class weighting. This explains the scattered predictions for underrepresented classes in the confusion matrix. Third, speaker mm appears to express emotions atypically, producing softer elation and higher-pitched sadness than expected, making it difficult for the model to match these utterances to standard acoustic patterns.

To improve classifier performance, I would implement several enhancements based on observed error patterns. First, I would expand beyond the IS09 feature set by incorporating spectral and articulatory features that better capture emotional nuances, particularly voice quality parameters (jitter, shimmer, and harmonic-to-noise ratio) effective at differentiating similar emotions like contempt and disgust. Second, I would implement hierarchical classification, where an initial classifier determines arousal level (high/medium/low) before a second classifier identifies the specific emotion within that category, reducing competition between acoustically distant emotions. Third, I would address class imbalance more aggressively through SMOTE for minority classes and asymmetric loss functions that heavily penalize minority class misclassifications. Fourth, I would employ ensemble methods combining predictions from multiple base classifiers (SVM, Random Forest, and Gradient Boosting), which typically outperform single classifiers in emotion recognition tasks. Finally, I would incorporate speaker adaptation techniques, using limited labeled data from the target speaker to fine-tune the model, potentially addressing the speaker-specific realization patterns observed with speaker mm. These modifications would likely improve both overall performance and specifically address poor recall rates for frequently confused emotion categories.