

Homework 3: Emotion Recognition

jl7199 Jinzi Luo

I confirm that I have not used any GPT-generated responses for any part of this assignment.

1. Feature analysis

I First create plot without normalization.

For the normalized versions of the plots, I applied per-speaker z-score normalization to both pitch and intensity. The goal is to remove speaker-specific differences in overall pitch range and loudness, so that we can better compare how emotions affect relative changes within each speaker.

Concretely, for each speaker we first extracted all frame-level pitch values (75–600 Hz, Praat autocorrelation) and all frame-level intensity values (minimum pitch 75 Hz, automatic time step) from all of their segments. I concatenated these arrays across all utterances of that speaker and computed a single global mean and standard deviation for pitch ($\mu_s^{\text{pitch}}, \sigma_s^{\text{pitch}}$) and for intensity ($\mu_s^{\text{int}}, \sigma_s^{\text{int}}$). Then, for every segment from speaker s , I normalized each frame x as

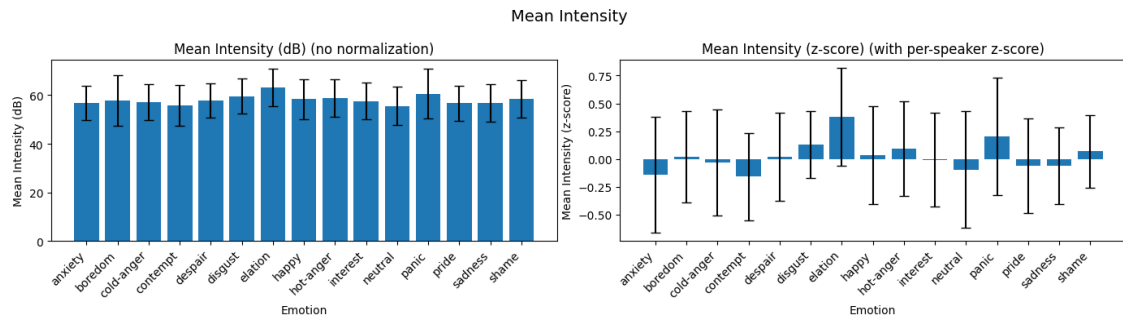
$$z = \frac{x - \mu_s}{\sigma_s}$$

After obtaining the frame-level z-scores, we recomputed the min, max, and mean of pitch and intensity for each segment from the normalized sequences. These six z-scored features (min/max/mean pitch z , min/max/mean intensity z) are used in the “with normalization” plots.

I chose per-speaker z-score normalization because it explicitly controls for large inter-speaker variability in absolute pitch and intensity (e.g., naturally higher voices vs. lower voices), while preserving the within-speaker relative deviations that are more likely to be driven by emotion. This makes the emotion-level comparisons across the corpus more interpretable than using raw Hz/dB values.

I got the 12 plots:





From the 12 plots (raw and normalized features):

1. High-arousal emotions systematically raise pitch, especially for panic.

In the raw plots of min/max/mean pitch, panic consistently has the highest values, with elation, happy, and hot-anger also clearly above most other emotions. After per-speaker z-score normalization, these emotions still have positive or near-zero pitch z-scores, while low-arousal emotions remain negative. This suggests that even after controlling for each speaker's baseline, high-arousal states are produced with systematically higher pitch.

2. Low-arousal and negative emotions show lower pitch relative to each speaker's baseline.

Emotions such as boredom, sadness, shame, and neutral tend to have lower min/max/mean pitch in the raw plots. In the normalized plots, their pitch z-scores are mostly negative, often among the lowest across all classes. I interpret this as evidence that speakers not only sound lower in absolute Hz, but also shift their pitch down relative to their own typical range when expressing these low-arousal or subdued emotions.

3. High-arousal emotions are also louder, especially in terms of maximum and mean intensity.

In the raw max and mean intensity plots, elation, happy, hot-anger, panic, and pride are clustered at the higher end of the dB scale, whereas boredom, sadness, and neutral tend to be softer. After normalization, these high-arousal emotions generally have positive intensity z-scores, while low-arousal emotions are closer to zero or negative. This pattern suggests a consistent coupling between emotional arousal and vocal loudness, not just pitch.

4. Panic and elation show large variability across speakers, while neutral and boredom are more stable.

The error bars for panic and elation are relatively large in many of the pitch and intensity plots, both before and after normalization. I take this to mean that speakers use a wider range of acoustic strategies to express these emotions (some sounding extremely high-pitched or loud, others less so). In contrast, neutral and boredom usually have smaller error bars, which suggests that these states are expressed in a more consistent and constrained acoustic space.

5. Normalization removes strong speaker effects but preserves the relative ordering of emotions.

In the raw pitch plots, the absolute differences between emotions are very large (on the order of 100–200 Hz), and these differences can easily be influenced by which speakers contribute more samples to a given emotion. After per-speaker z-score normalization, all emotions are centered around the zero baseline, and the dynamic range is compressed, but the ranking across emotions is largely preserved (panic and elation still high, boredom and sadness still low). I learn from this that a substantial part of the raw variation is due to speaker-specific baselines, and that z-score normalization helps me focus on emotion-driven deviations from each speaker's typical voice.

6. Normalization sharpens intensity-based differences that looked small in raw dB.

In the raw max/mean intensity plots, most emotions cluster in a narrow range around ~70–80 dB, so the differences are not visually striking. After per-speaker z-score normalization, I see a much clearer separation: elation, happy, hot-anger, panic, and partly pride shift to positive intensity z-scores, while boredom, sadness, and neutral move to negative values. This tells me that absolute dB values are heavily influenced by microphone and speaker loudness, and normalization is crucial to reveal emotion-driven loudness patterns.

7. Some emotions look strong in raw dB but are “normal” after normalization.

For example, pride has relatively high raw max/mean intensity, but in the intensity z-score plots it is much closer to zero than elation or panic. This

suggests that the speakers who produce pride already tend to speak loudly in general, so pride is not as strong a within-speaker loudness increase as elation or panic. Normalization helps me distinguish this emotion uses loud speakers from this emotion actually makes individual speakers louder than usual.

2. Classification experiments

Extract features

I use two groups of acoustic-prosodic features in our experiments.

(1) openSMILE IS09 Emotion Challenge feature set (baseline).

As main acoustic feature representation, I extract the INTERSPEECH 2009 Emotion Challenge ("IS09") feature set using the openSMILE toolkit and the official IS09_emotion.conf configuration file. Concatenating all functionals over all LLDs yields a 384-dimensional fixed-length vector per utterance. This IS09 set is a widely used baseline for emotion recognition and provides a broad coverage of prosodic, spectral, and cepstral information.

(2) Additional prosodic features from Part 1 (Parselmouth).

In addition to the IS09 features, I also include a small set of manually designed prosodic features that I used in Part 1 for feature analysis. Using Parselmouth (a Python interface to Praat), I compute utterance-level summary statistics of pitch and intensity: minimum, maximum, and mean F0, as well as minimum, maximum, and mean intensity. These six values explicitly capture the global level and range of fundamental frequency and loudness for each utterance. We chose this subset of features because:

- they are directly interpretable and closely related to perceived arousal and valence (e.g., higher pitch and larger intensity range often correspond to high-arousal emotions such as panic or hot anger);
- our Part 1 analysis showed clear differences in these statistics across emotion categories;
- they add only six dimensions on top of IS09, providing complementary information without excessively increasing feature dimensionality.

All openSMILE (IS09) and Parselmouth features were z-score normalized per speaker (same procedure as in Part 1).

Train a multiclass classifier

Classifier and training setup

I treat emotion recognition as a 15-way multiclass classification problem. All models are implemented in scikit-learn. For evaluation, I use `sklearn.metrics.classification_report` to obtain per-class precision, recall, and F1 scores, as well as macro and weighted averages for each test speaker.

Main classifier: RBF SVM.

MY primary model is a support vector machine with a radial basis function (RBF) kernel, implemented via `sklearn.svm.SVC`. SVMs are well-suited for high-dimensional feature spaces with relatively limited training data, which matches our setting (hundreds of acoustic-prosodic features but only a few thousand utterances). I use the default multiclass handling in scikit-learn (one-vs-rest strategy) with the following hyperparameters:

- kernel: "rbf"
- C: 10
- gamma: "scale"
- random_state: 0

I chose a slightly larger C than the default ($C = 1$) to allow a bit more flexibility in fitting the data. The same SVM configuration is used for both Model A (IS09 only) and Model B (IS09 + Parselmouth features).

For IS09 only:

Test speaker = cc

Emotion	Precision	Recall	F1-score	Support
anxiety	0.000	0.000	0.000	10
boredom	0.074	0.133	0.095	15
cold-anger	0.067	0.067	0.067	15
contempt	0.308	0.364	0.333	22
despair	0.125	0.333	0.182	9
disgust	0.364	0.129	0.190	31
elation	0.167	0.250	0.200	16
happy	0.190	0.174	0.182	23

hot-anger	0.333	0.500	0.400	14
interest	0.200	0.235	0.216	17
neutral	0.667	0.222	0.333	18
panic	0.583	0.389	0.467	18
pride	0.167	0.043	0.069	23
sadness	0.308	0.308	0.308	13
shame	0.250	0.143	0.182	21

Overall metrics (speaker cc)

- Accuracy: 0.211 (n = 265)
- Macro avg – precision: 0.253, recall: 0.219, F1-score: 0.215
- Weighted avg – precision: 0.272, recall: 0.211, F1-score: 0.218

Speaker cc accuracy: 0.2113, weighted F1: 0.2180, n = 265

Test speaker = cl

Emotion	Precision	Recall	F1-score	Support
anxiety	0.226	0.333	0.269	21
boredom	0.400	0.552	0.464	29
cold-anger	0.476	0.370	0.417	27
contempt	0.289	0.440	0.349	25
despair	0.156	0.172	0.164	29
disgust	0.154	0.182	0.167	22
elation	0.273	0.333	0.300	27
happy	0.462	0.286	0.353	21
hot-anger	0.556	0.577	0.566	26
interest	0.346	0.346	0.346	26
neutral	0.000	0.000	0.000	17
panic	0.364	0.190	0.250	21
pride	0.294	0.208	0.244	24
sadness	0.174	0.148	0.160	27
shame	0.233	0.269	0.250	26

Overall metrics (speaker cl)

- Accuracy: 0.304 (n = 368)
- Macro avg – precision: 0.294, recall: 0.294, F1-score: 0.287
- Weighted avg – precision: 0.300, recall: 0.304, F1-score: 0.295

Speaker cl accuracy: 0.3043, weighted F1: 0.2953, n = 368

Test speaker = gg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.528	0.633	0.576	30
boredom	0.316	0.400	0.353	30
cold-anger	0.234	0.407	0.297	27
contempt	0.286	0.231	0.255	26
despair	0.226	0.250	0.237	28
disgust	0.474	0.176	0.257	51
elation	0.304	0.500	0.378	28
happy	0.231	0.400	0.293	30
hot-anger	0.562	0.409	0.474	22
interest	0.237	0.300	0.265	30
neutral	1.000	0.111	0.200	9
panic	0.583	0.519	0.549	27
pride	0.278	0.200	0.233	25
sadness	0.000	0.000	0.000	33
shame	0.346	0.375	0.360	24

Overall metrics (speaker gg)

- Accuracy: 0.326 (n = 420)
- Macro avg – precision: 0.374, recall: 0.327, F1-score: 0.315
- Weighted avg – precision: 0.344, recall: 0.326, F1-score: 0.312

Speaker gg accuracy: 0.3262, weighted F1: 0.3121, n = 420

Test speaker = jg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.120	0.158	0.136	19
boredom	0.118	0.143	0.129	14
cold-anger	0.250	0.136	0.176	22
contempt	0.235	0.174	0.200	23
despair	0.053	0.048	0.050	21
disgust	0.296	0.348	0.320	23
elation	0.217	0.250	0.233	20
happy	0.118	0.100	0.108	20
hot-anger	0.238	0.278	0.256	18
interest	0.125	0.158	0.140	19
neutral	0.000	0.000	0.000	8
panic	0.125	0.071	0.091	14
pride	0.138	0.222	0.170	18
sadness	0.286	0.211	0.242	19
shame	0.000	0.000	0.000	15

Overall metrics (speaker jg)

- Accuracy: 0.165 (n = 273)
- Macro avg – precision: 0.155, recall: 0.153, F1-score: 0.150
- Weighted avg – precision: 0.168, recall: 0.165, F1-score: 0.162

Speaker jg accuracy: 0.1648, weighted F1: 0.1623, n = 273

Test speaker = mf

Emotion	Precision	Recall	F1-score	Support
anxiety	0.150	0.136	0.143	22
boredom	0.304	0.259	0.280	27
cold-anger	0.136	0.150	0.143	20
contempt	0.526	0.227	0.317	44
despair	0.080	0.125	0.098	16
disgust	0.028	1.000	0.054	1
elation	0.080	0.077	0.078	26

happy	0.105	0.087	0.095	23
hot-anger	0.455	0.476	0.465	21
interest	0.227	0.263	0.244	19
neutral	0.818	0.900	0.857	10
panic	0.643	0.750	0.692	12
pride	0.000	0.000	0.000	18
sadness	0.059	0.050	0.054	20
shame	0.389	0.350	0.368	20

Overall metrics (speaker mf)

- Accuracy: 0.237 (n = 299)
- Macro avg – precision: 0.267, recall: 0.323, F1-score: 0.259
- Weighted avg – precision: 0.274, recall: 0.237, F1-score: 0.244

Speaker mf accuracy: 0.2375, weighted F1: 0.2445, n = 299

Test speaker = mk

Emotion	Precision	Recall	F1-score	Support
anxiety	0.103	0.103	0.103	29
boredom	0.231	0.300	0.261	20
cold-anger	0.222	0.435	0.294	23
contempt	0.160	0.190	0.174	21
despair	0.261	0.113	0.158	53
disgust	0.087	0.095	0.091	21
elation	0.162	0.261	0.200	23
happy	0.179	0.167	0.173	42
hot-anger	0.208	0.227	0.217	22
interest	0.333	0.250	0.286	44
neutral	0.857	0.750	0.800	8
panic	0.389	0.333	0.359	21
pride	0.056	0.043	0.049	23
sadness	0.115	0.136	0.125	22

shame	0.292	0.280	0.286	25
-------	-------	-------	-------	----

Overall metrics (speaker mk)

- Accuracy: 0.212 (n = 397)
- Macro avg – precision: 0.244, recall: 0.246, F1-score: 0.238
- Weighted avg – precision: 0.223, recall: 0.212, F1-score: 0.209

Speaker mk accuracy: 0.2116, weighted F1: 0.2093, n = 397

Test speaker = mm

Emotion	Precision	Recall	F1-score	Support
anxiety	0.600	0.308	0.407	39
boredom	0.462	0.316	0.375	19
cold-anger	0.100	0.100	0.100	20
contempt	0.300	0.316	0.308	19
despair	0.310	0.500	0.383	18
disgust	0.240	0.261	0.250	23
elation	0.206	0.368	0.264	19
happy	0.438	0.778	0.560	18
hot-anger	0.462	0.375	0.414	16
interest	0.167	0.238	0.196	21
neutral	1.000	0.111	0.200	9
panic	0.571	0.143	0.229	28
pride	0.333	0.316	0.324	19
sadness	0.158	0.176	0.167	17
shame	0.429	0.529	0.474	17

Overall metrics (speaker mm)

- Accuracy: 0.318 (n = 302)
- Macro avg – precision: 0.385, recall: 0.322, F1-score: 0.310
- Weighted avg – precision: 0.381, recall: 0.318, F1-score: 0.313

Speaker mm accuracy: 0.3179, weighted F1: 0.3131, n = 302

Aggregated results for Model A: IS09 only

- Aggregated average accuracy: **0.2586**
- Aggregated average weighted F1: **0.2550**

For combined:

Test speaker = cc

Emotion	Precision	Recall	F1-score	Support
anxiety	0.000	0.000	0.000	10
boredom	0.074	0.133	0.095	15
cold-anger	0.091	0.067	0.077	15
contempt	0.357	0.455	0.400	22
despair	0.087	0.222	0.125	9
disgust	0.429	0.194	0.267	31
elation	0.192	0.312	0.238	16
happy	0.143	0.130	0.136	23
hot-anger	0.400	0.571	0.471	14
interest	0.278	0.294	0.286	17
neutral	0.500	0.167	0.250	18
panic	0.615	0.444	0.516	18
pride	0.000	0.000	0.000	23
sadness	0.231	0.231	0.231	13
shame	0.200	0.143	0.167	21

Overall metrics (speaker cc)

- Accuracy: 0.223 (n = 265)
- Macro avg – precision: 0.240, recall: 0.224, F1-score: 0.217
- Weighted avg – precision: 0.258, recall: 0.223, F1-score: 0.224

Speaker cc accuracy: 0.2226, weighted F1: 0.2244, n = 265

Test speaker = cl

Emotion	Precision	Recall	F1-score	Support
anxiety	0.250	0.381	0.302	21
boredom	0.381	0.552	0.451	29

cold-anger	0.526	0.370	0.435	27
contempt	0.289	0.440	0.349	25
despair	0.182	0.207	0.194	29
disgust	0.217	0.227	0.222	22
elation	0.265	0.333	0.295	27
happy	0.429	0.286	0.343	21
hot-anger	0.552	0.615	0.582	26
interest	0.364	0.308	0.333	26
neutral	0.000	0.000	0.000	17
panic	0.333	0.190	0.242	21
pride	0.286	0.167	0.211	24
sadness	0.125	0.111	0.118	27
shame	0.156	0.192	0.172	26

Overall metrics (speaker cl)

- Accuracy: 0.302 (n = 368)
- Macro avg – precision: 0.290, recall: 0.292, F1-score: 0.283
- Weighted avg – precision: 0.296, recall: 0.302, F1-score: 0.291

Speaker cl accuracy: 0.3016, weighted F1: 0.2912, n = 368

Test speaker = gg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.613	0.633	0.623	30
boredom	0.333	0.400	0.364	30
cold-anger	0.205	0.333	0.254	27
contempt	0.235	0.308	0.267	26
despair	0.233	0.250	0.241	28
disgust	0.611	0.216	0.319	51
elation	0.348	0.571	0.432	28
happy	0.234	0.367	0.286	30
hot-anger	0.625	0.455	0.526	22
interest	0.225	0.300	0.257	30

neutral	1.000	0.111	0.200	9
panic	0.591	0.481	0.531	27
pride	0.300	0.240	0.267	25
sadness	0.000	0.000	0.000	33
shame	0.346	0.375	0.360	24

Overall metrics (speaker gg)

- Accuracy: 0.336 (n = 420)
- Macro avg – precision: 0.393, recall: 0.336, F1-score: 0.328
- Weighted avg – precision: 0.371, recall: 0.336, F1-score: 0.328

Speaker gg accuracy: 0.3357, weighted F1: 0.3281, n = 420

Test speaker = jg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.103	0.158	0.125	19
boredom	0.200	0.214	0.207	14
cold-anger	0.133	0.091	0.108	22
contempt	0.316	0.261	0.286	23
despair	0.050	0.048	0.049	21
disgust	0.345	0.435	0.385	23
elation	0.263	0.250	0.256	20
happy	0.111	0.100	0.105	20
hot-anger	0.333	0.389	0.359	18
interest	0.160	0.211	0.182	19
neutral	0.000	0.000	0.000	8
panic	0.143	0.071	0.095	14
pride	0.179	0.278	0.217	18
sadness	0.286	0.211	0.242	19
shame	0.000	0.000	0.000	15

Overall metrics (speaker jg)

- Accuracy: 0.194 (n = 273)

- Macro avg – precision: 0.175, recall: 0.181, F1-score: 0.174
- Weighted avg – precision: 0.187, recall: 0.194, F1-score: 0.187

Speaker jg accuracy: 0.1941, weighted F1: 0.1872, n = 273

Test speaker = mf

Emotion	Precision	Recall	F1-score	Support
anxiety	0.263	0.227	0.244	22
boredom	0.261	0.222	0.240	27
cold-anger	0.115	0.150	0.130	20
contempt	0.474	0.205	0.286	44
despair	0.115	0.188	0.143	16
disgust	0.028	1.000	0.054	1
elation	0.111	0.115	0.113	26
happy	0.100	0.087	0.093	23
hot-anger	0.556	0.476	0.513	21
interest	0.263	0.263	0.263	19
neutral	0.818	0.900	0.857	10
panic	0.615	0.667	0.640	12
pride	0.000	0.000	0.000	18
sadness	0.125	0.100	0.111	20
shame	0.368	0.350	0.359	20

Overall metrics (speaker mf)

- Accuracy: 0.244 (n = 299)
- Macro avg – precision: 0.281, recall: 0.330, F1-score: 0.270
- Weighted avg – precision: 0.285, recall: 0.244, F1-score: 0.254

Speaker mf accuracy: 0.2441, weighted F1: 0.2537, n = 299

Test speaker = mk

Emotion	Precision	Recall	F1-score	Support
anxiety	0.107	0.103	0.105	29
boredom	0.233	0.350	0.280	20

cold-anger	0.263	0.435	0.328	23
contempt	0.148	0.190	0.167	21
despair	0.300	0.113	0.164	53
disgust	0.087	0.095	0.091	21
elation	0.200	0.304	0.241	23
happy	0.167	0.167	0.167	42
hot-anger	0.240	0.273	0.255	22
interest	0.303	0.227	0.260	44
neutral	0.857	0.750	0.800	8
panic	0.368	0.333	0.350	21
pride	0.056	0.043	0.049	23
sadness	0.138	0.182	0.157	22
shame	0.348	0.320	0.333	25

Overall metrics (speaker mk)

- Accuracy: 0.222 (n = 397)
- Macro avg – precision: 0.254, recall: 0.259, F1-score: 0.250
- Weighted avg – precision: 0.233, recall: 0.222, F1-score: 0.218

Speaker mk accuracy: 0.2217, weighted F1: 0.2181, n = 397

Test speaker = mm

Emotion	Precision	Recall	F1-score	Support
anxiety	0.524	0.282	0.367	39
boredom	0.462	0.316	0.375	19
cold-anger	0.125	0.100	0.111	20
contempt	0.417	0.526	0.465	19
despair	0.320	0.444	0.372	18
disgust	0.250	0.261	0.255	23
elation	0.229	0.421	0.296	19
happy	0.452	0.778	0.571	18
hot-anger	0.615	0.500	0.552	16

interest	0.172	0.238	0.200	21
neutral	1.000	0.111	0.200	9
panic	0.571	0.143	0.229	28
pride	0.368	0.368	0.368	19
sadness	0.238	0.294	0.263	17
shame	0.391	0.529	0.450	17

Overall metrics (speaker mm)

- Accuracy: 0.344 (n = 302)
- Macro avg – precision: 0.409, recall: 0.354, F1-score: 0.338
- Weighted avg – precision: 0.397, recall: 0.344, F1-score: 0.336

Speaker mm accuracy: 0.3444, weighted F1: 0.3355, n = 302

Aggregated results for Model B: IS09 + Parselmouth prosodic features

- Aggregated average accuracy: **0.2707**
- Aggregated average weighted F1: **0.2665**

Additional experiment: small neural network (MLP).

To explore a neural network–based classifier, I also conducted a secondary experiment with a small feed-forward multilayer perceptron using `sklearn.neural_network.MLPClassifier`. The network has two hidden layers with 256 and 64 units, ReLU activations, and uses the Adam optimizer with adaptive learning rate:

- `hidden_layer_sizes`: (256, 64)
- `activation`: "relu"
- `alpha` (L2 regularization): 1e-4
- `batch_size`: 64
- `learning_rate`: "adaptive"
- `max_iter`: 100
- `early_stopping`: True
- `random_state`: 0

For IS09 only:

Test speaker = cc

Emotion	Precision	Recall	F1-score	Support
---------	-----------	--------	----------	---------

anxiety	0.074	0.200	0.108	10
boredom	0.100	0.200	0.133	15
cold-anger	0.267	0.267	0.267	15
contempt	0.226	0.318	0.264	22
despair	0.158	0.333	0.214	9
disgust	0.455	0.161	0.238	31
elation	0.227	0.312	0.263	16
happy	0.067	0.043	0.053	23
hot-anger	0.316	0.429	0.364	14
interest	0.214	0.176	0.194	17
neutral	0.375	0.167	0.231	18
panic	0.318	0.389	0.350	18
pride	0.100	0.043	0.061	23
sadness	0.133	0.154	0.143	13
shame	0.143	0.048	0.071	21

Overall metrics (speaker cc)

- Accuracy: 0.200 (n = 265)
- Macro avg – precision: 0.211, recall: 0.216, F1-score: 0.197
- Weighted avg – precision: 0.224, recall: 0.200, F1-score: 0.193

Speaker cc accuracy: 0.2000, weighted F1: 0.1932, n = 265

Test speaker = cl

Emotion	Precision	Recall	F1-score	Support
anxiety	0.200	0.238	0.217	21
boredom	0.318	0.483	0.384	29
cold-anger	0.462	0.222	0.300	27
contempt	0.243	0.360	0.290	25
despair	0.208	0.172	0.189	29
disgust	0.125	0.091	0.105	22
elation	0.324	0.407	0.361	27

happy	0.208	0.238	0.222	21
hot-anger	0.379	0.423	0.400	26
interest	0.281	0.346	0.310	26
neutral	0.000	0.000	0.000	17
panic	0.250	0.143	0.182	21
pride	0.095	0.083	0.089	24
sadness	0.182	0.148	0.163	27
shame	0.091	0.115	0.102	26

Overall metrics (speaker cl)

- Accuracy: 0.242 (n = 368)
- Macro avg – precision: 0.224, recall: 0.231, F1-score: 0.221
- Weighted avg – precision: 0.233, recall: 0.242, F1-score: 0.230

Speaker cl accuracy: 0.2418, weighted F1: 0.2302, n = 368

Test speaker = gg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.194	0.233	0.212	30
boredom	0.100	0.067	0.080	30
cold-anger	0.259	0.259	0.259	27
contempt	0.379	0.423	0.400	26
despair	0.150	0.214	0.176	28
disgust	0.524	0.216	0.306	51
elation	0.429	0.536	0.476	28
happy	0.161	0.167	0.164	30
hot-anger	0.320	0.364	0.340	22
interest	0.170	0.267	0.208	30
neutral	0.000	0.000	0.000	9
panic	0.565	0.481	0.520	27
pride	0.125	0.160	0.140	25
sadness	0.000	0.000	0.000	33
shame	0.275	0.458	0.344	24

Overall metrics (speaker gg)

- Accuracy: 0.257 (n = 420)
- Macro avg – precision: 0.243, recall: 0.256, F1-score: 0.242
- Weighted avg – precision: 0.263, recall: 0.257, F1-score: 0.249

Speaker gg accuracy: 0.2571, weighted F1: 0.2487, n = 420

Test speaker = jg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.179	0.263	0.213	19
boredom	0.200	0.214	0.207	14
cold-anger	0.143	0.136	0.140	22
contempt	0.200	0.130	0.158	23
despair	0.042	0.048	0.044	21
disgust	0.214	0.261	0.235	23
elation	0.375	0.150	0.214	20
happy	0.000	0.000	0.000	20
hot-anger	0.259	0.389	0.311	18
interest	0.091	0.105	0.098	19
neutral	0.000	0.000	0.000	8
panic	0.182	0.143	0.160	14
pride	0.179	0.278	0.217	18
sadness	0.200	0.211	0.205	19
shame	0.154	0.133	0.143	15

Overall metrics (speaker jg)

- Accuracy: 0.168 (n = 273)
- Macro avg – precision: 0.161, recall: 0.164, F1-score: 0.156
- Weighted avg – precision: 0.167, recall: 0.168, F1-score: 0.161

Speaker jg accuracy: 0.1685, weighted F1: 0.1609, n = 273

Test speaker = mf

Emotion	Precision	Recall	F1-score	Support
anxiety	0.259	0.318	0.286	22
boredom	0.250	0.185	0.213	27
cold-anger	0.083	0.100	0.091	20
contempt	0.476	0.227	0.308	44
despair	0.136	0.188	0.158	16
disgust	0.029	1.000	0.057	1
elation	0.118	0.077	0.093	26
happy	0.000	0.000	0.000	23
hot-anger	0.476	0.476	0.476	21
interest	0.318	0.368	0.341	19
neutral	0.500	0.700	0.583	10
panic	0.556	0.833	0.667	12
pride	0.214	0.167	0.188	18
sadness	0.071	0.050	0.059	20
shame	0.467	0.350	0.400	20

Overall metrics (speaker mf)

- Accuracy: 0.251 (n = 299)
- Macro avg – precision: 0.264, recall: 0.336, F1-score: 0.261
- Weighted avg – precision: 0.277, recall: 0.251, F1-score: 0.252

Speaker mf accuracy: 0.2508, weighted F1: 0.2517, n = 299

Test speaker = mk

Emotion	Precision	Recall	F1-score	Support
anxiety	0.091	0.103	0.097	29
boredom	0.053	0.050	0.051	20
cold-anger	0.167	0.304	0.215	23
contempt	0.100	0.143	0.118	21
despair	0.200	0.075	0.110	53
disgust	0.125	0.095	0.108	21
elation	0.182	0.261	0.214	23

happy	0.394	0.310	0.347	42
hot-anger	0.074	0.091	0.082	22
interest	0.344	0.250	0.289	44
neutral	0.429	0.375	0.400	8
panic	0.188	0.143	0.162	21
pride	0.069	0.087	0.077	23
sadness	0.097	0.136	0.113	22
shame	0.138	0.160	0.148	25

Overall metrics (speaker mk)

- Accuracy: 0.169 (n = 397)
- Macro avg – precision: 0.177, recall: 0.172, F1-score: 0.169
- Weighted avg – precision: 0.189, recall: 0.169, F1-score: 0.171

Speaker mk accuracy: 0.1688, weighted F1: 0.1711, n = 397

Test speaker = mm

Emotion	Precision	Recall	F1-score	Support
anxiety	0.500	0.231	0.316	39
boredom	0.429	0.158	0.231	19
cold-anger	0.033	0.050	0.040	20
contempt	0.250	0.263	0.256	19
despair	0.226	0.389	0.286	18
disgust	0.261	0.261	0.261	23
elation	0.263	0.263	0.263	19
happy	0.306	0.611	0.407	18
hot-anger	0.429	0.188	0.261	16
interest	0.263	0.238	0.250	21
neutral	0.667	0.667	0.667	9
panic	0.684	0.464	0.553	28
pride	0.154	0.211	0.178	19
sadness	0.091	0.118	0.103	17

shame	0.438	0.412	0.424	17
-------	-------	-------	-------	----

Overall metrics (speaker mm)

- Accuracy: 0.288 (n = 302)
- Macro avg – precision: 0.333, recall: 0.301, F1-score: 0.300
- Weighted avg – precision: 0.341, recall: 0.288, F1-score: 0.295

Speaker mm accuracy: 0.2881, weighted F1: 0.2950, n = 302

Aggregated results for Model A2: IS09 only

- Aggregated average accuracy: **0.2259**
- Aggregated average weighted F1: **0.2223**

For Combined:

Test speaker = cc

Emotion	Precision	Recall	F1-score	Support
anxiety	0.000	0.000	0.000	10
boredom	0.071	0.133	0.093	15
cold-anger	0.000	0.000	0.000	15
contempt	0.382	0.591	0.464	22
despair	0.133	0.222	0.167	9
disgust	0.417	0.161	0.233	31
elation	0.118	0.125	0.121	16
happy	0.105	0.087	0.095	23
hot-anger	0.333	0.571	0.421	14
interest	0.182	0.235	0.205	17
neutral	0.000	0.000	0.000	18
panic	0.562	0.500	0.529	18
pride	0.000	0.000	0.000	23
sadness	0.158	0.231	0.188	13
shame	0.333	0.238	0.278	21

Overall metrics (speaker cc)

- Accuracy: 0.208 (n = 265)
- Macro avg – precision: 0.186, recall: 0.206, F1-score: 0.186
- Weighted avg – precision: 0.207, recall: 0.208, F1-score: 0.195

Speaker cc accuracy: 0.2075, weighted F1: 0.1948, n = 265

Test speaker = cl

Emotion	Precision	Recall	F1-score	Support
anxiety	0.179	0.238	0.204	21
boredom	0.327	0.621	0.429	29
cold-anger	0.238	0.185	0.208	27
contempt	0.346	0.360	0.353	25
despair	0.156	0.172	0.164	29
disgust	0.133	0.091	0.108	22
elation	0.480	0.444	0.462	27
happy	0.133	0.095	0.111	21
hot-anger	0.367	0.423	0.393	26
interest	0.314	0.423	0.361	26
neutral	0.667	0.118	0.200	17
panic	0.308	0.190	0.235	21
pride	0.136	0.125	0.130	24
sadness	0.263	0.185	0.217	27
shame	0.034	0.038	0.036	26

Overall metrics (speaker cl)

- Accuracy: 0.258 (n = 368)
- Macro avg – precision: 0.272, recall: 0.247, F1-score: 0.241
- Weighted avg – precision: 0.267, recall: 0.258, F1-score: 0.247

Speaker cl accuracy: 0.2582, weighted F1: 0.2472, n = 368

Test speaker = gg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.472	0.567	0.515	30

boredom	0.200	0.133	0.160	30
cold-anger	0.148	0.148	0.148	27
contempt	0.312	0.385	0.345	26
despair	0.135	0.179	0.154	28
disgust	0.529	0.176	0.265	51
elation	0.464	0.464	0.464	28
happy	0.222	0.267	0.242	30
hot-anger	0.409	0.409	0.409	22
interest	0.151	0.267	0.193	30
neutral	0.333	0.111	0.167	9
panic	0.379	0.407	0.393	27
pride	0.171	0.240	0.200	25
sadness	0.000	0.000	0.000	33
shame	0.308	0.500	0.381	24

Overall metrics (speaker gg)

- Accuracy: 0.279 (n = 420)
- Macro avg – precision: 0.282, recall: 0.284, F1-score: 0.269
- Weighted avg – precision: 0.289, recall: 0.279, F1-score: 0.267

Speaker gg accuracy: 0.2786, weighted F1: 0.2675, n = 420

Test speaker = jg

Emotion	Precision	Recall	F1-score	Support
anxiety	0.226	0.368	0.280	19
boredom	0.200	0.214	0.207	14
cold-anger	0.190	0.182	0.186	22
contempt	0.160	0.174	0.167	23
despair	0.100	0.095	0.098	21
disgust	0.208	0.217	0.213	23
elation	0.500	0.250	0.333	20
happy	0.133	0.100	0.114	20
hot-anger	0.407	0.611	0.489	18

interest	0.125	0.158	0.140	19
neutral	0.333	0.250	0.286	8
panic	0.250	0.071	0.111	14
pride	0.087	0.111	0.098	18
sadness	0.231	0.158	0.188	19
shame	0.067	0.067	0.067	15

Overall metrics (speaker jg)

- Accuracy: 0.201 (n = 273)
- Macro avg – precision: 0.215, recall: 0.202, F1-score: 0.198
- Weighted avg – precision: 0.210, recall: 0.201, F1-score: 0.197

Speaker jg accuracy: 0.2015, weighted F1: 0.1965, n = 273

Test speaker = mf

Emotion	Precision	Recall	F1-score	Support
anxiety	0.250	0.273	0.261	22
boredom	0.227	0.185	0.204	27
cold-anger	0.061	0.100	0.075	20
contempt	0.368	0.159	0.222	44
despair	0.129	0.250	0.170	16
disgust	0.000	0.000	0.000	1
elation	0.211	0.154	0.178	26
happy	0.077	0.043	0.056	23
hot-anger	0.571	0.571	0.571	21
interest	0.250	0.211	0.229	19
neutral	0.667	1.000	0.800	10
panic	0.529	0.750	0.621	12
pride	0.227	0.278	0.250	18
sadness	0.000	0.000	0.000	20
shame	0.286	0.200	0.235	20

Overall metrics (speaker mf)

- Accuracy: 0.244 (n = 299)
- Macro avg – precision: 0.257, recall: 0.278, F1-score: 0.258
- Weighted avg – precision: 0.261, recall: 0.244, F1-score: 0.241

Speaker mf accuracy: 0.2441, weighted F1: 0.2413, n = 299

Test speaker = mk

Emotion	Precision	Recall	F1-score	Support
anxiety	0.091	0.069	0.078	29
boredom	0.032	0.050	0.039	20
cold-anger	0.206	0.304	0.246	23
contempt	0.130	0.143	0.136	21
despair	0.235	0.075	0.114	53
disgust	0.042	0.048	0.044	21
elation	0.333	0.391	0.360	23
happy	0.206	0.167	0.184	42
hot-anger	0.375	0.409	0.391	22
interest	0.278	0.227	0.250	44
neutral	0.462	0.750	0.571	8
panic	0.269	0.333	0.298	21
pride	0.045	0.043	0.044	23
sadness	0.148	0.182	0.163	22
shame	0.243	0.360	0.290	25

Overall metrics (speaker mk)

- Accuracy: 0.202 (n = 397)
- Macro avg – precision: 0.206, recall: 0.237, F1-score: 0.214
- Weighted avg – precision: 0.203, recall: 0.202, F1-score: 0.194

Speaker mk accuracy: 0.2015, weighted F1: 0.1937, n = 397

Test speaker = mm

Emotion	Precision	Recall	F1-score	Support
anxiety	0.200	0.103	0.136	39

boredom	0.312	0.263	0.286	19
cold-anger	0.043	0.050	0.047	20
contempt	0.208	0.263	0.233	19
despair	0.296	0.444	0.356	18
disgust	0.353	0.261	0.300	23
elation	0.222	0.211	0.216	19
happy	0.303	0.556	0.392	18
hot-anger	0.364	0.250	0.296	16
interest	0.174	0.190	0.182	21
neutral	0.250	0.111	0.154	9
panic	0.611	0.393	0.478	28
pride	0.179	0.263	0.213	19
sadness	0.136	0.176	0.154	17
shame	0.333	0.353	0.343	17

Overall metrics (speaker mm)

- Accuracy: 0.255 (n = 302)
- Macro avg – precision: 0.266, recall: 0.259, F1-score: 0.252
- Weighted avg – precision: 0.271, recall: 0.255, F1-score: 0.253

Speaker mm accuracy: 0.2550, weighted F1: 0.2528, n = 302

Aggregated results for Model B2: IS09 + Parselmouth prosodic features

- Aggregated average accuracy: **0.2375**
- Aggregated average weighted F1: **0.2298**

Then I got the acc:

Model	Features	Classifier	Aggregated Accuracy	Aggregated Weighted F1
A	IS09 (openSMILE)	RBF SVM	0.2586	0.2550
B	IS09 + Parselmouth prosodic (6 dims)	RBF SVM	0.2707	0.2665
A2	IS09 (openSMILE)	MLP (NN)	0.2259	0.2223
B2	IS09 + Parselmouth prosodic (6 dims)	MLP (NN)	0.2375	0.2298

We can see that the best is B, which is SVM with IS09 + Parselmouth prosodic (6 dims) features.

3. Error analysis

The best result I got is: speaker mm used as the test set in our SVM Model B (IS09 + Parselmouth features).

Test speaker = mm

Emotion	Precision	Recall	F1-score	Support
anxiety	0.524	0.282	0.367	39
boredom	0.462	0.316	0.375	19
cold-anger	0.125	0.100	0.111	20
contempt	0.417	0.526	0.465	19
despair	0.320	0.444	0.372	18
disgust	0.250	0.261	0.255	23
elation	0.229	0.421	0.296	19
happy	0.452	0.778	0.571	18
hot-anger	0.615	0.500	0.552	16
interest	0.172	0.238	0.200	21
neutral	1.000	0.111	0.200	9
panic	0.571	0.143	0.229	28
pride	0.368	0.368	0.368	19
sadness	0.238	0.294	0.263	17
shame	0.391	0.529	0.450	17

Overall metrics (speaker mm)

- Accuracy: 0.344 (n = 302)
- Macro avg – precision: 0.409, recall: 0.354, F1-score: 0.338
- Weighted avg – precision: 0.397, recall: 0.344, F1-score: 0.336

Speaker mm accuracy: 0.3444, weighted F1: 0.3355, n = 302

The model achieves an overall accuracy of 34.44% and a weighted F1 of 33.55%, which is above the random 15 baseline.

Overall pattern.

The per-class scores for speaker mm show that the model is much better at recognizing some emotions than others. Macro-average F1 (0.338) is lower than the weighted F1 (0.3355), indicating that performance on minority or difficult classes is worse than on the more frequent/easier ones. The confusion is especially pronounced among negative emotions and low-arousal states.

Easiest classes.

For this speaker, several emotions are relatively easy to predict:

- **happy**: precision 0.452, recall 0.778, F1 0.571
- **hot-anger**: 0.615 / 0.500 / 0.552
- **contempt**: 0.417 / 0.526 / 0.465

These emotions are often associated with strong, distinctive prosodic cues. For example, happy typically has higher mean pitch and intensity and a wider F0 range, while hot-anger tends to have high energy and sharp pitch excursions. Our feature set explicitly includes global pitch and intensity statistics from Parselmouth, as well as many energy- and F0-related descriptors from IS09. Thus, high-arousal or clearly marked emotional states are easier to separate in the acoustic-prosodic space. The high recall for happy suggests that, at least for this speaker, the classifier reliably detects strongly positive high-arousal speech.

Most difficult classes.

Other emotions remain much harder to recognize for speaker mm:

- **cold-anger**: precision 0.125, recall 0.100, F1 0.111
- **neutral**: precision 1.000, recall 0.111, F1 0.200
- **interest**: F1 \approx 0.200
- **panic**: F1 \approx 0.229

These classes are difficult for different reasons. Cold-anger has a prosodic profile closer to controlled or low-arousal negative emotions and likely overlaps with neutral, sadness, or mild anger; global statistics alone may not capture the subtle differences between these states. The neutral class shows an extreme precision–recall imbalance: the classifier almost never predicts “neutral”, but when it does, it

is correct. This suggests that the model is overly conservative about assigning the neutral label and tends to map neutral utterances to nearby low-arousal emotions such as boredom or sadness.

Similarly, interest and panic may share acoustic patterns with other high-arousal emotions (e.g., happy, elation, anxiety, hot-anger), making them hard to separate in the current feature space.

Ideas for improving the classifier.

This error pattern suggests several directions for future improvement:

1. Incorporate lexical features: Many distinctions that are hard to make acoustically (e.g., interest vs. happiness, cold-anger vs. neutral) may be clearer from the verbal content. Adding textual features from ASR transcripts (TF-IDF, or sentence embeddings) could help disambiguate emotions that share similar prosody but differ in semantics.
2. Model temporal dynamics of prosody: Our current features are mostly global statistics over the whole utterance. Using sequence models on frame-level pitch and energy, or richer openSMILE functionals, may better capture the shape of prosodic contours.
3. More expressive acoustic representations. Replacing or augmenting IS09 with deep speech embeddings (e.g., wav2vec2 or Whisper encoder features) could capture more fine-grained spectral and voice-quality cues that help distinguish low-arousal and subtle negative states such as sadness and cold-anger. I did not explore this direction here, but maybe it is a promising avenue.

Overall, the best speaker-specific fold (speaker mm) shows that my current acoustic-prosodic feature set is particularly effective for clearly marked high-arousal emotions but struggles with subtle or low-arousal emotions. Combining lexical information and temporal modeling would likely further improve performance on the most difficult classes.