Speech Emotion Recognition (SER) project

Vaggelis Lamprou

Presentation

Crema-D dataset

- 91 speakers : 48 males, 43 females
- 4 classes of interest : happy, neutral, sad, angry
- 4414 sample points
- 136-dim vector per speech sample
- Each speaker contributes about 14 happy, sad, angry instances and 12 neutral (few exemptions exist)

- Speaker independent training/validation/test sets
- training set : 66 speakers ($\backsim$ 72%)
- validation set : 16 speakers ($\backsim$ 18%)
- test set : 9 speakers ($\backsim$ 10%)

Dataset visualization : PCA

Transform points according to the first 3 principal components.

Figure 2.1

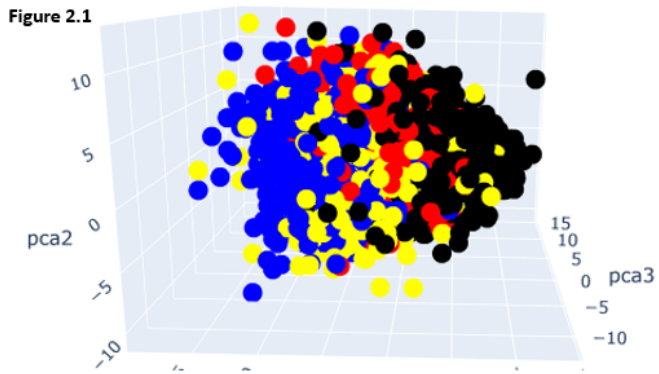


Figure 2.2

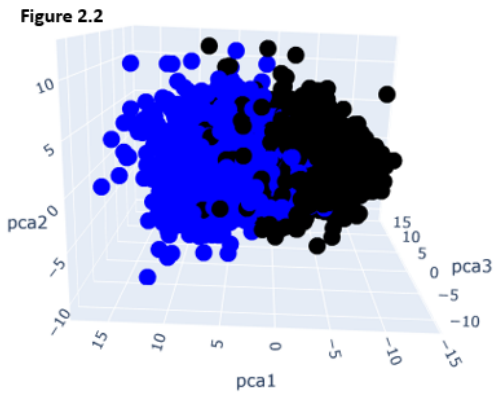


Figure 2.3

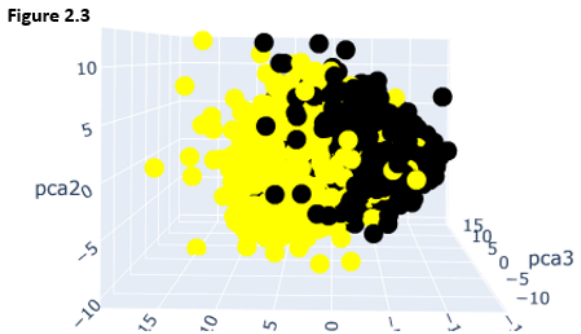
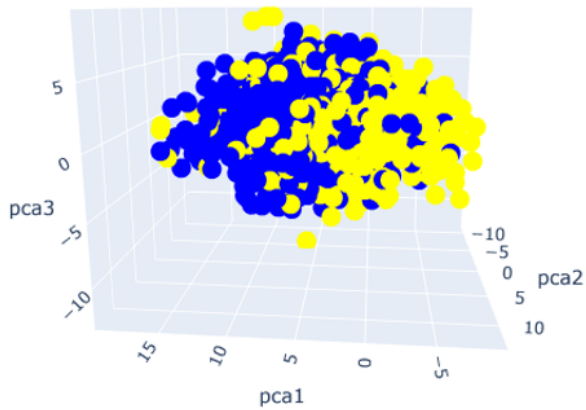
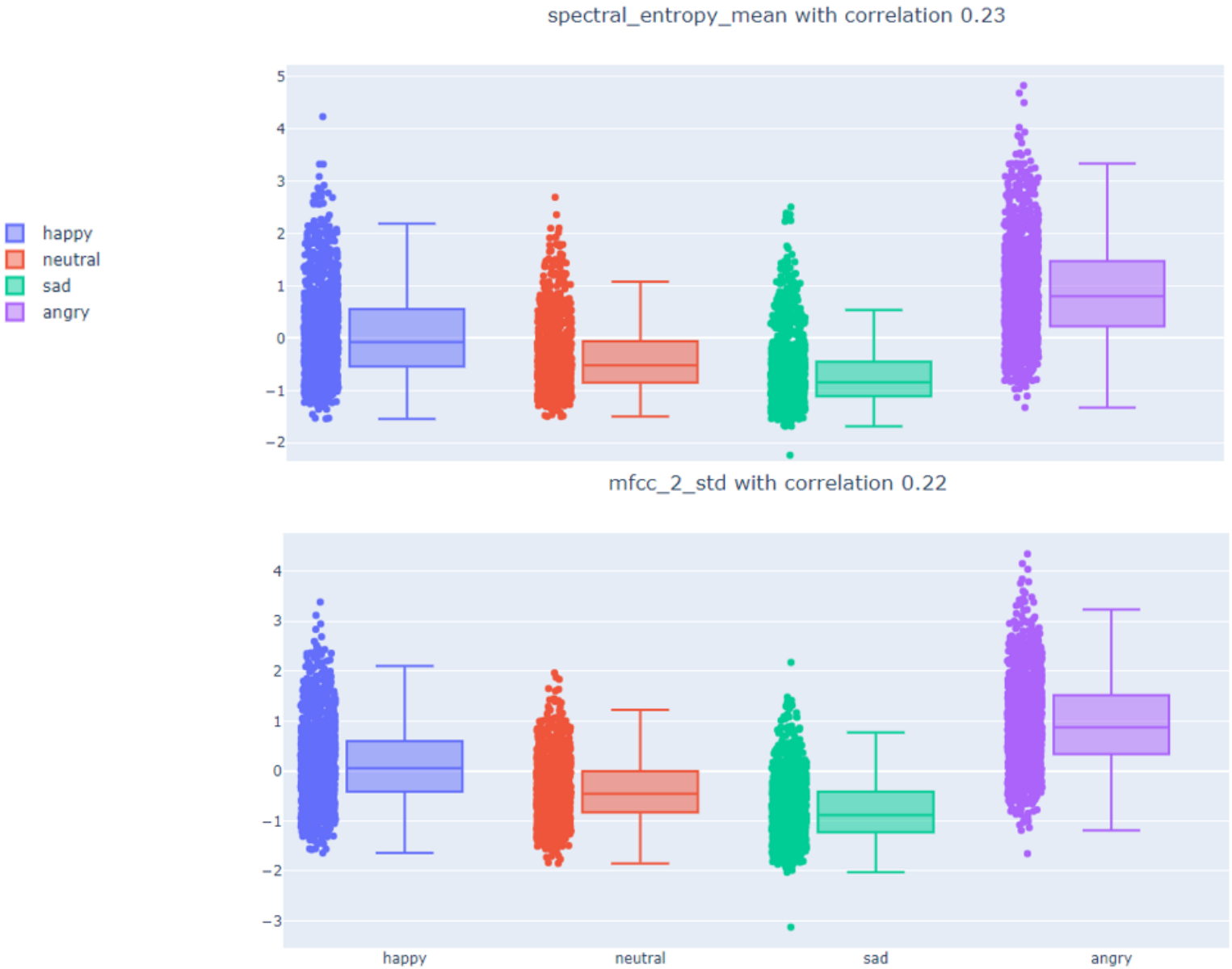


Figure 2.4



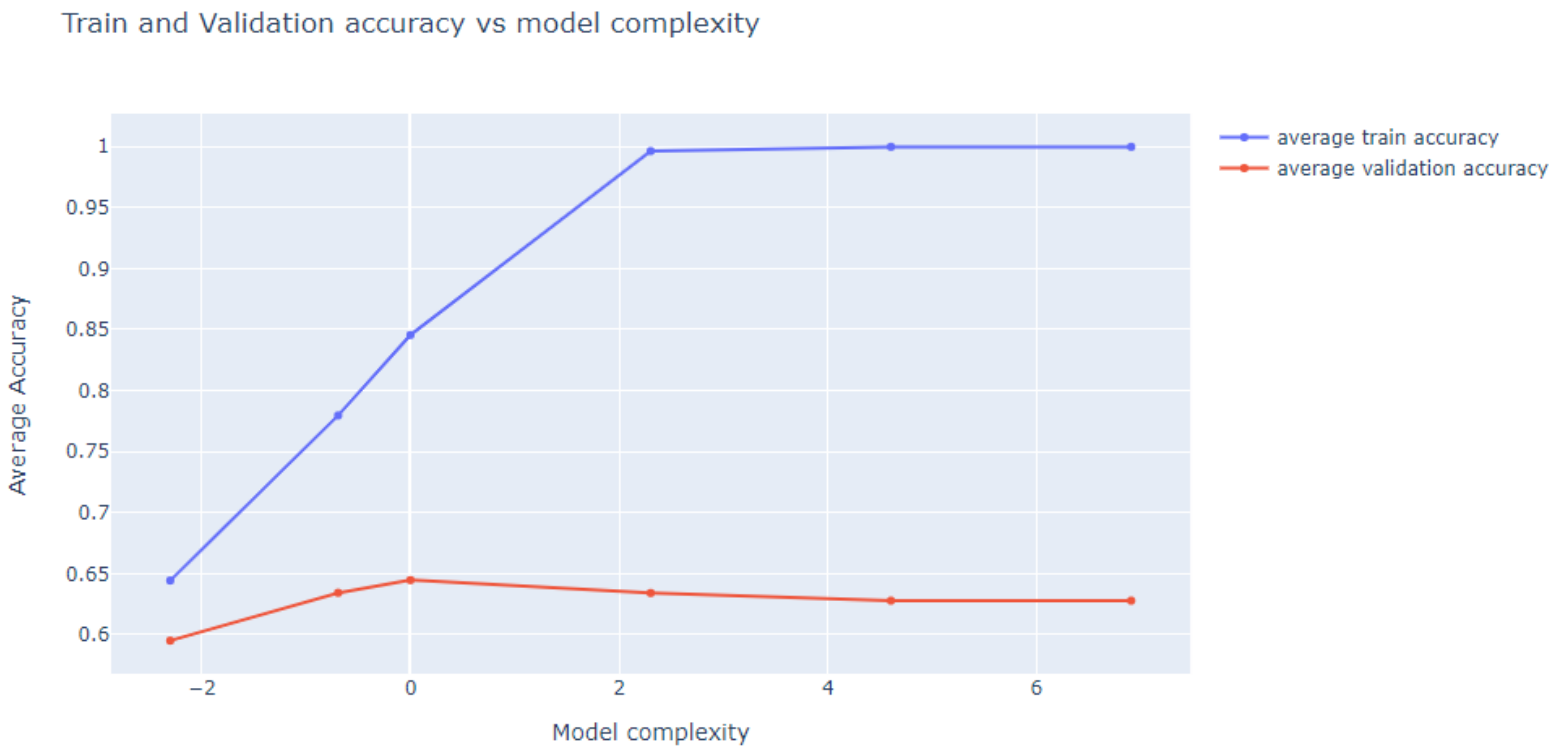
Values Distribution : Boxplots

Examine correlations between features and target.



The optimal model

- After 20 experiments (train/validation splits), the optimal model is a SVM with $C = 1$, average validation accuracy 0.645 and average F1-score 0.641 (the selection criterion).
- Trained on 136-features (many correlation levels examined)



Testing

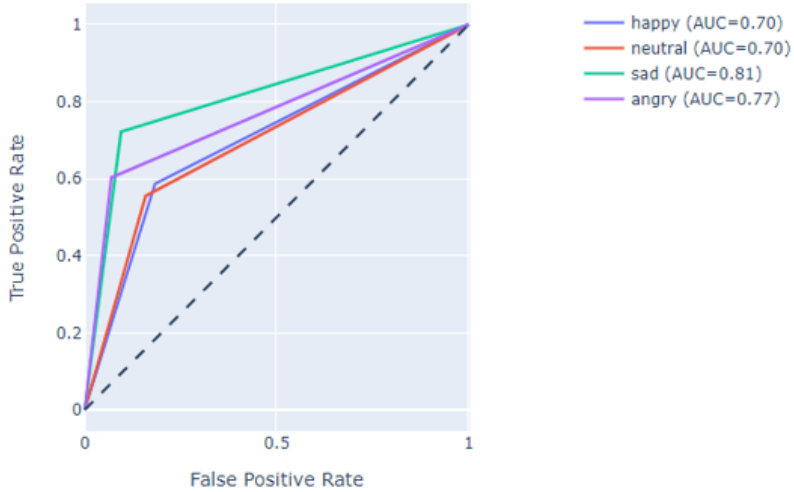
- Confusion matrix

	happy	neutral	sad	angry
happy	74	25	11	16
neutral	21	60	19	8
sad	13	21	91	1
angry	32	14	4	76

Classification report

	precision	recall	f1-score	support
happy	0.53	0.59	0.56	126
neutral	0.50	0.56	0.53	108
sad	0.73	0.72	0.73	126
angry	0.75	0.60	0.67	126
accuracy			0.62	486
macro avg	0.63	0.62	0.62	486
weighted avg	0.63	0.62	0.62	486

- ROC curves



PR curves