Multimodal Emotion Recognition

Andreas Kouridakis

Aris Tsilifonis

Introduction

According to the European Data Protection Supervisor, emotion recognition is the technology that analyzes signals such as facial expressions, voice tones, and body language in order to reveal information on one's emotional state.

Motivation:

- A deep interest in understanding how emotions work and how they can be recognized and interpreted by machines
- Health care: Depression Detection
- Customer services: Identifying customer emotions in a call center
- Online teaching: Help teachers identify students emotional state and tailor their teaching methods

Objectives:

- Evaluate the performance of ML models in different experiments
- Compare our best ML model with DL models

Approaches

- Mel-Frequency Cepstral Coefficients (MFCC) for speech feature extraction
- RNNs
- LSTM
- CNN
- MLP

Our approach

Machine Learning Models

- Multinomial Naive Bayes
- Random Forest
- XGBoost
- Logistic Regression

Experiments

- Speech only
- Text only
- Speech + Text

Metrics

- Confusion matrix
- Recall
- Precision
- F1 score
- Accuracy
- ROC

Different speech features (later..)

IEMOCAP Dataset

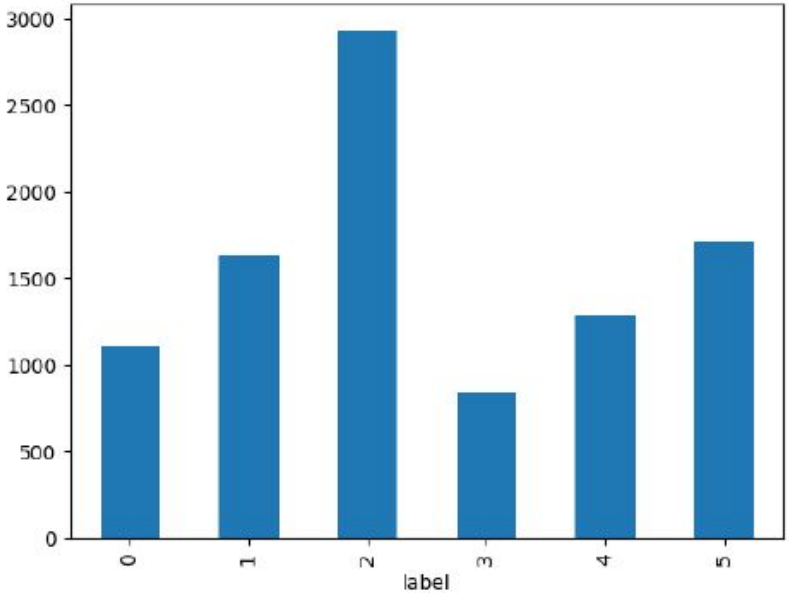
- 12 hours audio visual data + text transcriptions
- 5 sessions
- 5 different pairs (male + female)
- scripted dialogues + improvisational interactions
- 9 labels (emotions)

Emotions: anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state

Preprocessing of speech data

- Remove ‘other’ class
- Merged ‘happy’ & ‘excited’
- Merged ‘sad’ & ‘frustrated’
- Oversampling ‘fear’ & ‘surprise’

Final list of emotions:
0: angry
1: happy/excited
2: sad/frustrated
3:fear
4:surprise
5: neutral



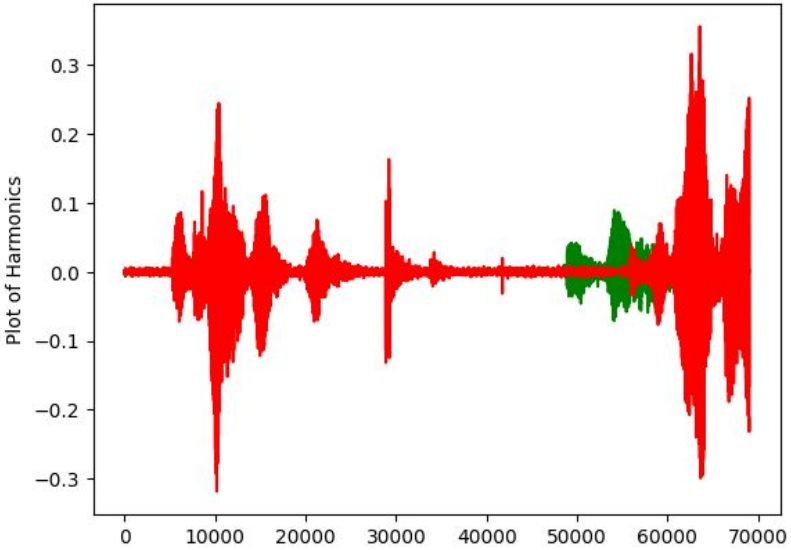
Feature extraction

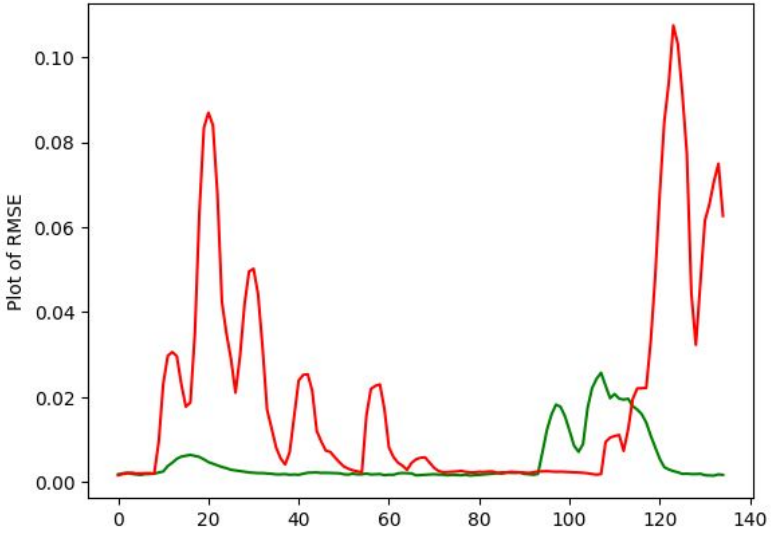
Width of the signal		Pitch		Energy		Silence	Harmonic
Mean	Std	Auto corr max	Auto corr std	RMSE mean	RMSE std	[1]	[2]

[1] Pr(y[n] < 0.4 * RMSE)

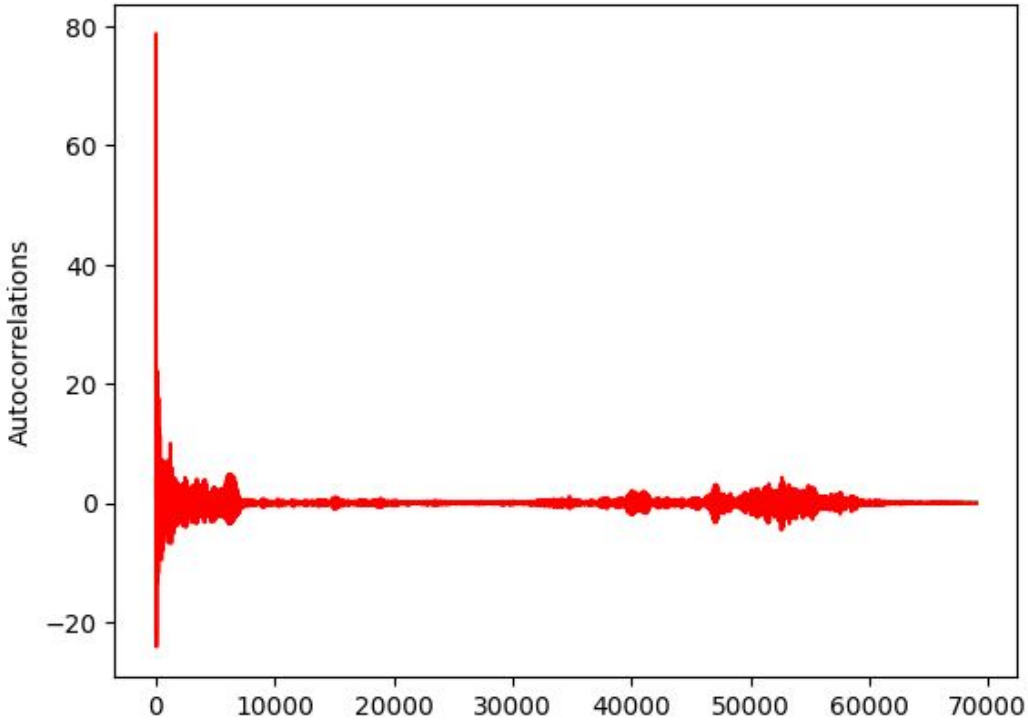
[2] libr.effects.hpss(y)

Audio analysis





Audio analysis



Text preprocessing

- **Extract transcriptions** from the transcription folder inside each session's dialog folder.
- **Extract labels** from the EmoEvaluation folder of IEMOCAP dataset
- Normalized text by **lowercasing**
- **Remove special characters** except for (. , ?, !). Are you here? -> Are you here ?
- Remove **stopwords** and **lemmatization**
- **Oversampling** of the more underrepresented classes(fear and surprise)
- **Merged classes** of excited and happy, frustrated and sad
- **Discard examples** as “other” [angry, happy, sad, fear, surprise, neutral]
- Create **tf-idf vectors** of text data (unigrams, bigrams)
- Split initial dataset to **train and test** (80% train and 20% test) to feed the classifiers

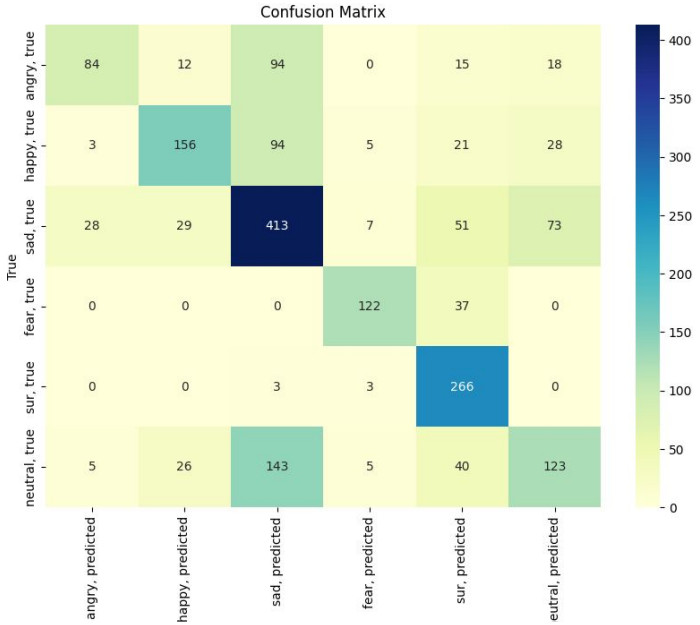
Combined Dataset(Audio and Text)

- Simple Fusion to combine the two modalities
- Concatenate audio and text feature columns
  - Each TF-IDF vector has the size of the TF-IDF vocabulary for every sample
  - The resulting arrays contain **both types of features side by side** for each sample
- Columns of dataframe are more representative for each sample since there is more information about the data. The new sample is more expressive than the initial one

Dataframe	Audio Features	Text Features
sample 1	<k1,l1, m1,…>	< x1, y1, z1. >

Results(Text)

Random Forest



VII. Results

Text (only) - accuracy	
Random Forest	61%
XGBoost	54%
Logistic Regression	61%
Multinomial NB	58%

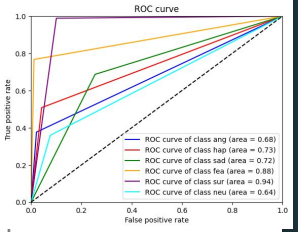
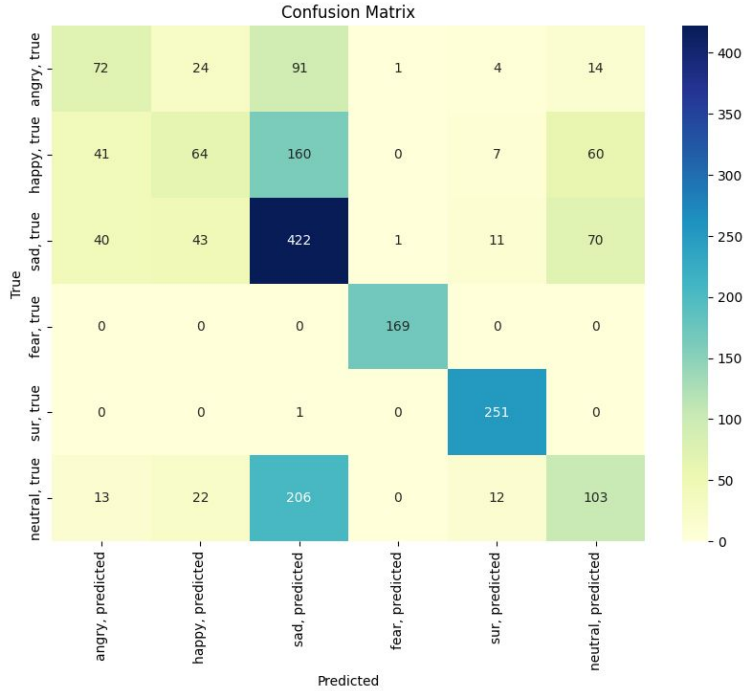


Table Random forest (text)

	precision	recall	f1	support
0	0.7	0.38	0.49	223
1	0.7	0.51	0.59	307
2	0.56	0.69	0.61	601
3	0.86	0.77	0.81	159
4	0.62	0.99	0.76	269
5	0.51	0.36	0.42	342

Results(Audio)

Random Forest



Speech (only) - accuracy	
Random Forest	56%
XGBoost	55%
Logistic Regression	32%
Multinomial NB	31%

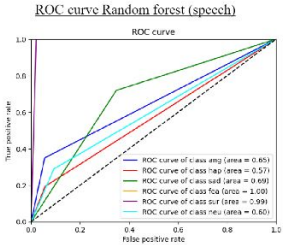


Table Random forest (speech)

	precision	recall	f1	support
0	0.43	0.35	0.39	206
1	0.42	0.19	0.26	332
2	0.48	0.72	0.58	587
3	0.99	1	0.99	169
4	0.88	1	0.94	251
5	0.42	0.29	0.34	356

Multimodal(Audio and Text)

Random Forest

Table Random forest (text+speech)

	precision	recall	f1	support
0	0.55	0.41	0.47	214
1	0.69	0.41	0.51	337
2	0.52	0.82	0.63	572
3	0.96	1	0.98	195
4	0.93	1	0.97	257
5	0.55	0.27	0.36	326

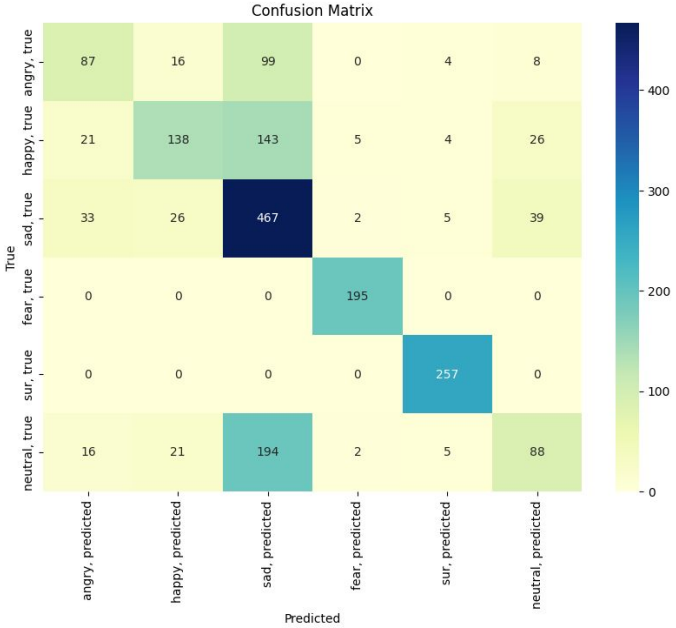
Comparison of our Best ML model to others

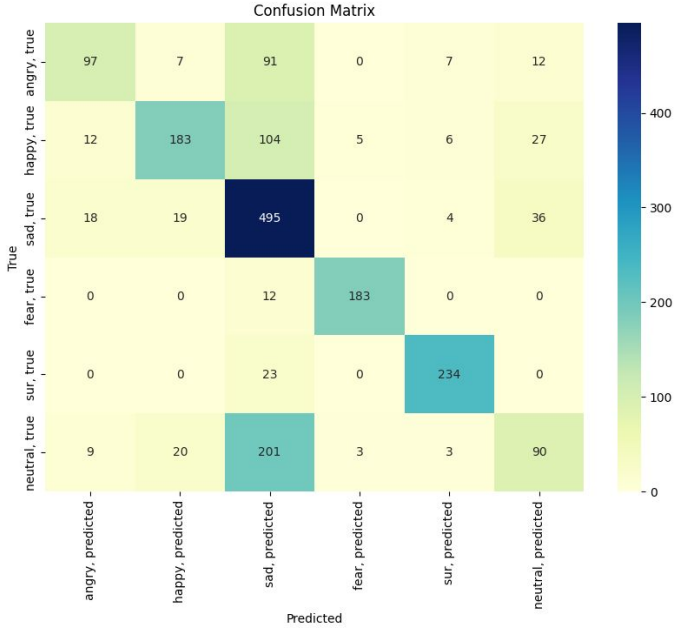
Model	**Accuracy**
Random Forest	65%
Approach [8]	68% – 71%
Approach [9]	71%
Approach [10]	68%

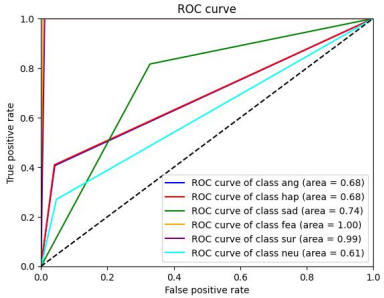
Note: Soft voting achieved 68% accuracy, outperforming random forest model by **3%**

Multimodal(Audio and Text)

Random Forest and Soft Voting







Less Ambiguities!

Results

Key points

- Random Forest was the most efficient individual model
- Soft voting improved scores only on the multimodal experiment
- Simple machine learning models achieve competitive results compared to complex neural networks
- Misclassification: Classifiers confused angry, happy and neutral sentiments with sad emotion
- Confusion matrix of the multimodal experiment: soft voting improves random forest on sentiments by correctly classifying more happy and sad emotions (diagonal)
- Roc curves: Fear and surprise sentiment lines are closer to the upper left corner
- Simple ML models are competitive!

Comparison with other approaches

Model	Accuracy
Random Forest	65%
Dual RNN	68-71%
LSTM + CNN + MLP	71%
LSTM + DenseNet + LSTM	68%

Conclusion: This comparison indicates that our Random Forest model's performance is close to these deep learning models, highlighting its effectiveness despite the simplicity of the approach.

Future work

- Split data into scripted + improvised segments
- Leave one session out cross validation
- Utilize spectograms for speech feature extraction
- Leverage advances word embeddings for text such as GloVe or BERT

- Gender Specific features and the use of Deep Learning models

  "A Deep Learning Method Using Gender-Specific Features for Emotion Recognition"
  L i, Y., Zhang, Y.T., Ng, G.W., Leau, Y.B., & Yan, H. (2023)