Speech Emotion Recognition

Alessandro Di Liberti
Dipartimento di Informatica, Università degli studi di Milano
alessandro.diliberti@gmail.com

***Abstract*—Speech Emotion Recognition (SER) plays a crucial role in human-computer interaction, yet accurately classifying emotions from speech remains a challenge, especially in real-world noisy conditions. This study presents a dual-method approach to SER: one based on handcrafted acoustic features processed with Random Forest and XGBoost classifiers, and another leveraging deep learning through log-Mel spectrograms and a 2D Convolutional Neural Network (CNN). The experiment evaluates model performance under two conditions—using a controlled dataset and incorporating an additional noisier dataset. Results indicate that while traditional machine learning models provide interpretability and stability, deep learning approaches exhibit better generalization in complex scenarios. Confusion matrices highlight specific emotion misclassification patterns, particularly when data variability increases. Performance metrics suggest that CNNs excel in handling diverse data distributions, whereas Random Forest maintains robustness in structured datasets. These findings offer insights into the trade-offs between feature-engineered models and deep learning techniques for SER applications.**

I. Introduction

Speech emotion recognition (SER) has emerged as a crucial field in human-computer interaction, with applications from mental health monitoring to customer service. Despite advances, accurately identifying emotional states from speech remains challenging, particularly in noisy real-world environments.

This project explores a dual-strategy framework for SER. The first approach extracts acoustic features to create a tabular dataset for Random Forest classification, leveraging its ensemble nature to capture complex patterns. The second transforms audio signals into log-Mel spectrograms as input for a 2D CNN, utilizing its ability to detect local patterns critical for emotion discrimination. This dual approach harnesses the interpretability of traditional machine learning alongside deep learning's capacity to learn complex representations directly from data.

A significant challenge in SER is dataset variability and audio sample collected in a non-controlled environment. To investigate this problem, and how it affects the training process, two training sets are created: one with ”clean” data from three well-structured datasets, and another incorporating an additional noisier dataset.

Confusion matrices are provided to evaluate classification performance across these conditions.

II. System overview

The framework has been structured into four distinct sections, as shown in Fig. 1, to facilitate a comprehensive evaluation of different approaches and datasets. This modular organization enables systematic testing and validation of the proposed speech emotion recognition methodologies, while maintaining clear separation between different experimental conditions.

A. Preprocessing

This module splits the dataset into train, test, and validation sets, with a strong focus on preventing data leakage. Duplicate instances were identified and removed to avoid the model learning test data features during training, reducing overfitting. Data augmentation was postponed to later phases to maintain strict separation between training and evaluation data.

B. Features Extraction

The feature extraction process is divided into two distinct methodologies, each designed to capture different aspects of the speech signal: traditional statistical feature extraction and spectrogram-based feature extraction.

The first approach involves extracting a set of handcrafted features from the raw audio signal. This method relies on well-established speech analysis techniques [1] [2], where the audio waveform is processed to derive meaningful characteristics. Here following the extracted features for each audio sample:

- Pitch: represents the perceived frequency of the voice, indicating the speaker's intonation and emotional state.
- Jitter: measures the frequency variation between consecutive speech cycles, reflecting the stability of the voice.
- Shimmer: assesses the amplitude variation between speech cycles, providing insights into voice quality and emotional expression.
- HNR: indicates the ratio of harmonics to noise in the voice signal, with lower values often associated with negative emotions.
- energy: represents the signal's power, correlating with the speaker's intensity and emotional emphasis.
- ZCR: counts the rate at which the signal changes sign, reflecting the signal's noisiness and emotional intensity.
- MFCC (1-13) [3]: captures the power spectrum of the speech signal, serving as a compact representation of the speech's spectral properties.

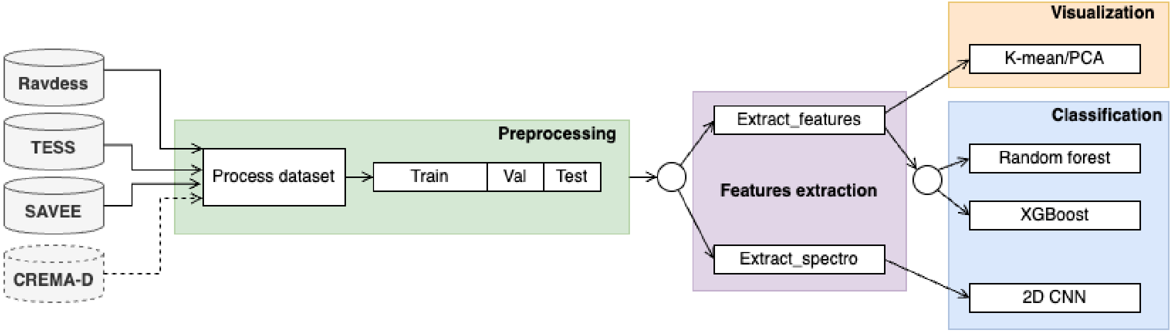

Fig. 1: Block diagram of the method

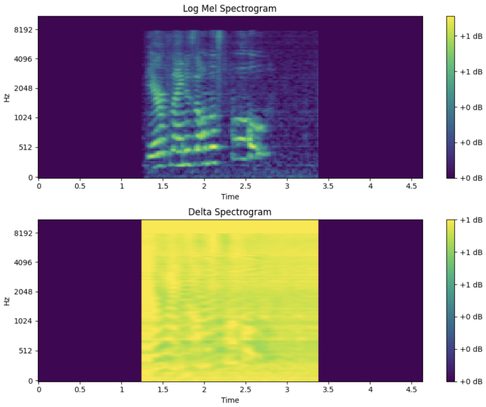

Fig. 2: Logmel and delta spectrograms of an individual audio sample

These features are then aggregated into a structured dataset suitable for traditional machine learning models. The implementation carefully ensures that feature computation remains consistent across all dataset samples, avoiding potential biases introduced by varying recording conditions. The extracted features serve as input to the Random Forest classifier, providing a robust and interpretable way to distinguish emotions from speech.

The second approach transforms raw audio signals into time-frequency representations using log-Mel spectrograms. The idea behind this approach [4], is to use the great ability of deep learning models to generalize and discover hidden patterns. Log-Mel spectrograms are preferred over regular spectrograms because they better mimic human auditory perception; also, a second channel was added: a first-order delta features providing temporal information by capturing the rate of change in speech, which helps recognize subtle variations in emotion. This stack of two images, Fig. 2, is then fed into a 2D CNN.

C. Visualization

Traditionally extracted features were used in this module to create a visualization of latent space. In order to observe this, 3 PCAs were used, Fig 3. It can be seen that some classes are distinct even at reduced dimensionality, while other blobs are mixed together. This may indicate that selected features are actually characterizing some classes. However, solving this problem through an unsupervised algorithm does not produce sufficient results: accuracy is 0.31 and f1 score is 0.32.

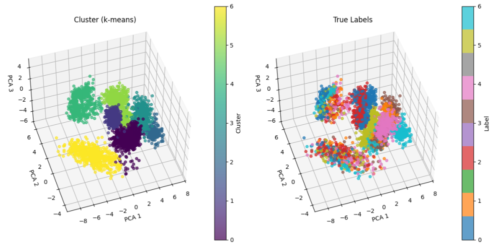

Fig. 3: Comparison of K-means clustering vs ground truth, 3 PCA

D. Classification

The task is a multiclass classification of 6 different emotions plus one neutral class (total of 7 classes). Random Forest and XGBoost classifiers were implemented for the tabular data approach, with Random Forest providing robust handling of high-dimensional features and XGBoost offering potential performance improvements through its gradient boosting mechanism. Both models underwent hyperparameter optimization through grid search with cross-validation. The CNN architecture processes log-mel spectrograms as 2D matrices, treating the emotion recognition task as an image classification problem. The network comprises multiple convolutional layers with max-pooling operations, designed to capture both temporal and frequency patterns characteristic of emotional expressions in speech signals.

III. Experimental Setup

A. Dataset

The initial experimental phase utilized three widely-recognized emotional speech datasets.

RAVDESS [5] contains speech samples from 24 professional actors (12 female, 12 male) expressing eight emotions in two different statements. The eighth class of this set (calm) is ignored and all the audio samples beloging to this class are discarded; this decision was guided by the fact that all the other datasets have only 7 classes, except CREMA which has 6. The dataset is structured so that each sentence is repeated twice by the same author. To avoid data leakage, the second repetition has been removed.

TESS [6] consists of recordings from two female actors speaking 200 target words in seven different emotions.

SAVEE [7] provides emotional expressions from four male speakers, covering seven emotions across 15 phonetically-balanced statements.

For the second phase of experimentation, CREMA-D [8] was incorporated into the training process. This dataset was specifically selected for its more naturalistic recording conditions, featuring 91 actors (48 female, 43 male) expressing six emotions with varying intensities. The dataset's inherently noisier nature, resulting from less controlled recording conditions, makes it particularly valuable for evaluating model generalization to real-world scenarios.

For each scenario the dataset was splitted with a ratio 0.8 of training set and 0.1 both for validation set and test set. Resulting in subsets made by number of elements indicated in TABLE I.

	without CREMA-D	**with CREMA-D**
Total	3904	11346
Train	3073	8934
Test	440	1277
Validation	391	1135

TABLE I: Number of element after the split for both the scenarios

B. Feature extraction

For the tabular representation, a set of statistical features was extracted from each audio sample to capture the acoustic properties relevant to emotion recognition. The feature set includes the mean and standard deviation of pitch, energy, zero-crossing rate (ZCR), and the first 13 Mel-frequency cepstral coefficients (MFCCs). Mean and standard deviation were selected as they effectively summarize the central tendency and variation of these parameters across the entire utterance, providing a compact representation of the overall acoustic characteristics while reducing dimensionality compared to frame-level features. There is no data augmentation process since it emerges that no significant change was introduced by augmenting the data with noise injection, pitching, stretching nor shifting.

C. Log-mel spectrogram

The proposed methodology employs a dual-channel image representation derived from audio samples to enable effective emotion recognition using a 2D CNN architecture. The processing pipeline begins with the extraction of log-mel and delta spectrograms implemented through librosa library.

Following extraction, both spectrograms undergo min-max normalization to scale values between 0 and 1, ensuring consistent feature ranges while preserving relative intensity patterns. Each normalized spectrogram is then either trimmed or zero-padded to a uniform length of 200 frames, creating standardized inputs for the CNN model while maintaining temporal relationships within the original signal. Also in this extraction, no data augmentation was done.

D. Building and training random forest

Two distinct ensemble-based implementations were developed for the classification task: one using Random Forest from scikit-learn [9] and another using XGBoost [10]. XGBoost was selected as a complementary approach to Random Forest due to its gradient boosting methodology, which offers potential performance improvements through sequential error correction, advanced regularization capabilities, and more efficient handling of complex non-linear relationships in the emotional speech features. Both approaches incorporated rigorous hyperparameter tuning to optimize model performance.

For the Random Forest implementation, hyperparameter optimization was performed over four key parameters:

- maximum depth,
- minumum samples per leaf,
- minimum samples per split,
- number of estimator

These parameters control the complexity of individual decision trees and the overall ensemble size.

The XGBoost implementation underwent similar optimization, but with a more extensive hyperparameter search space including:

- number of estimator,
- maximum depth,
- learning rate,
- subsample ratio,
- columns sampling by tree,
- minimum child weight

This broader parameter exploration reflects XGBoost's more complex boosting mechanism and regularization options.

Both implementations employed 5-fold cross-validation during the hyperparameter optimization process to ensure robust evaluation across different data subsets. While accuracy served as the primary optimization metric for both models, the XGBoost implementation additionally utilized multiclass logarithmic loss (mlogloss) as its objective function during training. This choice of loss function is particularly well-suited for multiclass classification problems, providing more nuanced gradient information during the boosting process.

E. Building and training 2D CNN

The 2D CNN was implemented using Keras to process the dual-channel spectrogram inputs. The network architecture followed the structure specified in the provided table, Table II, designed to extract relevant features from both log-mel and delta spectrograms. Training was conducted for 15 epochs with a batch size of 32, using sparse categorical cross-entropy loss and the Adam optimizer.

Layer	Filt/Units	Kernel	Params
Conv2D	8	(3,3)	ReLU, Input: (128,200)
Batch Norm.	-	-	-
Dropout	-	-	Rate = 0.2
MaxPool2D	-	(2,2)	-
Conv2D	16	(3,3)	ReLU
Dropout	-	-	Rate = 0.2
MaxPool2D	-	(2,2)	-
Conv2D	32	(3,3)	ReLU
Dropout	-	-	Rate = 0.2
MaxPool2D	-	(2,2)	-
Flatten	-	-	-
Dense	512	-	ReLU
Batch Norm.	-	-	-
Dropout	-	-	Rate = 0.4
Dense	7	-	Softmax

TABLE II: Architecture of the 2D convolutional neural networks

F. Addition of noisier dataset

The second experimental phase incorporated the CREMA-D dataset alongside the original three datasets (RAVDESS, TESS, and SAVEE). The modular structure of the experimental setup facilitated the seamless integration of this additional dataset without requiring significant modifications to the processing pipeline or model architectures.

The new dataset was processed following the same preprocessing steps applied to the original dataset, maintaining consistency across the experimental framework.

It's important to highlight that the integration of CREMA-D produced an unbalanced dataset in the class ”Surprised”, hence in the evaluation process the data related to this class are not so meaningful.

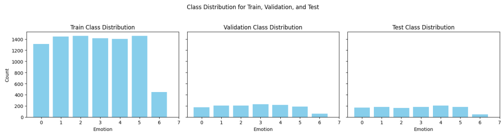

Fig. 4: Distribution of classes with CREMA-D integration

IV. Results

The experimental results demonstrate that all three classification approaches perform effectively in controlled settings with the initial datasets. However, notable performance variations emerge when introducing the more diverse CREMA-D dataset.

A. Random Forest

Random Forest maintains consistent recognition patterns across emotions in the controlled setting. When CREMA-D is incorporated, it shows increased confusion between Neutral-Sad and Neutral-Disgust pairs, while maintaining strong performance on Anger recognition.

Random forest (without CREMA-D)

	Net	Hap	Sad	Ang	Fea	Dis	Sur
Net	63	0	2	0	0	0	0
Hap	0	46	2	3	3	0	1
Sad	2	1	42	1	4	1	0
Ang	1	4	0	49	1	1	1
Fea	0	2	1	1	48	0	0
Dis	4	1	1	0	1	54	3
Sur	0	1	2	0	0	2	42

Random forest (with CREMA-D)

	Net	Hap	Sad	Ang	Fea	Dis	Sur
Net	122	7	18	0	4	20	0
Hap	17	110	3	22	6	19	2
Sad	23	10	111	2	8	11	0
Ang	6	17	1	142	4	10	2
Fea	13	28	33	19	98	14	0
Dis	24	14	25	14	2	101	3
Sur	2	2	1	0	2	3	40

TABLE III: Unnormalized confusion matrices on test set of random forest classifier

B. XGBoost

XGBoost exhibits similar confusion patterns to Random Forest but achieves slightly improved recognition of Fear and Disgust emotions. The incorporation of CREMA-D particularly affects its ability to distinguish between Fear-Happy and Disgust-Neutral pairs.

XGBoost (without CREMA-D)

	Net	Hap	Sad	Ang	Fea	Dis	Sur
Net	62	0	2	0	0	0	1
Hap	0	49	1	2	2	1	0
Sad	2	2	41	1	5	0	0
Ang	1	4	0	48	0	1	3
Fea	0	2	1	2	47	0	0
Dis	3	2	1	1	1	53	3
Sur	0	0	2	0	0	2	43

XGBoost (with CREMA-D)

	Net	Hap	Sad	Ang	Fea	Dis	Sur
Net	121	4	15	0	8	23	0
Hap	12	114	4	16	10	21	2
Sad	17	9	109	2	14	14	0
Ang	7	16	1	141	5	10	2
Fea	14	24	30	14	105	18	0
Dis	18	16	18	12	8	110	1
Sur	4	2	0	0	1	3	40

TABLE IV: Unnormalized confusion matrices on test set of XGBoost classifier

C. 2D CNN

The 2D CNN approach demonstrates strong performance on the controlled datasets with minimal confusion. However, it displays more pronounced sensitivity to the introduction of CREMA-D data, with increased confusion between acoustically similar emotions, particularly Fear-Disgust and Happy-Neutral pairs.

2D CNN (without CREMA-D)							
	Net	Hap	Sad	Ang	Fea	Dis	Sur
Net	58	0	0	0	0	0	7
Hap	0	44	0	4	1	1	5
Sad	1	0	42	1	3	2	2
Ang	0	0	0	47	0	5	5
Fea	0	1	2	0	44	2	3
Dis	2	0	1	2	0	58	1
Sur	0	0	0	0	1	1	45

2D CNN (without CREMA-D)							
	Net	Hap	Sad	Ang	Fea	Dis	Sur
Net	99	11	7	3	16	35	0
Hap	5	102	2	14	26	29	1
Sad	9	6	78	2	35	34	1
Ang	1	11	3	120	16	31	0
Fea	4	8	14	4	137	36	2
Dis	5	11	14	9	20	124	0
Sur	2	1	0	1	2	7	37

TABLE V: Unnormalized confusion matrices on test set of 2D CNN classifier

D. Overall performances

Looking at the table Tab. VI, we can draw some conclusions about this experiment: the higher complexity of XGBoost compared to a simpler random forest does not bring any real improvement, this might change by tuning more intensively on the parameters. In a more controlled setting (without CREMA-D) the performance of the random forest and logmel spectrogram approaches are very similar, showing robustness even with a reduced dataset. It is rather interesting to note that the CNN approach seems to show better results in a more general setting. Another possible observation is that some classes are better classified by one classifier, rather than another, indicating that the features selected to train the classifier are more characterizing for that class.

V. Conclusion

The study demonstrates that different approaches to Speech Emotion Recognition yield varying strengths depending on dataset conditions. Traditional classifiers like Random Forest and XGBoost perform well in controlled environments but struggle with more diverse data. The 2D CNN, while initially comparable, exhibits superior generalization when trained on noisier datasets. The analysis suggests that deep learning techniques are more resilient to variability in emotional speech data, although feature-engineered models retain advantages in interpretability. Future work could explore hybrid approaches combining both methodologies to improve robustness and explainability in real-world applications.

References

[1] Mohammad Mahdi Rezapour Mashhadi and Kofi Osei-Bonsu. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLOS ONE*, 18(11):1–13, 11 2023.

[2] Nilu Singh, Prof. Raees Khan, and Raj Shree Pandey. Mfcc and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54:9–13, 09 2012.

[3] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[4] Azamat Mukhamediya, Siamac Fazli, and Amin Zollanvari. On the effect of log-mel spectrogram parameter tuning for deep learning-based speech emotion recognition. *IEEE Access*, 11:61950–61957, 2023.

[5] Steven R. Livingstone and Frank A. Russo. 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):1–35, 05 2018.

[6] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020.

[7] Philip Jackson and Sana ul haq. Surrey audio-visual expressed emotion (savee) database, 04 2011.

[8] Houwei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390, 10 2014.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, and Jiaming Yuan. *xgboost: Extreme Gradient Boosting*, 2024. R package version 1.7.8.1.

	RF (no CREMA-D)			**RF (CREMA-D)**			**XGboost (no CREMA-D)**			**XGBoost (CREMA-D)**			**2D CNN (no CREMA-D)**			**2D CNN (CREMA-D)**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Neutral	.90	**.97**	**.93**	.59	**.71**	.65	.91	.95	**.93**	.63	**.71**	.66	**.95**	.89	.92	**.79**	.58	**.67**
Happy	.84	.84	.84	.59	.61	.60	.83	**.89**	.86	.62	**.64**	**.63**	**.98**	.80	**.88**	**.68**	.57	.62
Sad	.84	**.82**	.83	.58	**.67**	.62	.85	.80	.83	.62	.66	**.64**	**.93**	**.82**	**.87**	**.66**	.47	.55
Angry	**.91**	**.86**	**.88**	.71	**.78**	.75	.89	.84	.86	.76	.77	**.77**	.87	.82	.85	**.78**	.66	.72
Fearful	**.94**	**.92**	.88	**.79**	.49	**.60**	.85	.90	.88	.70	.51	.59	.90	.85	**.87**	.54	**.67**	**.60**
Disgust	**.93**	.84	**.89**	**.57**	.55	.56	**.93**	.83	.88	.55	.60	**.58**	.84	**.91**	.87	.42	**.68**	.52
Surprised	**.89**	.89	**.89**	.85	**.80**	.82	.86	.91	**.89**	.89	**.80**	**.84**	.66	**.96**	.78	**.90**	.74	.81
Average	.89	.87	.87	.66	.65	.65	.87	.87	.87	.68	.67	.67	.87	.86	.86	.68	.61	.64

TABLE VI: Precision, recall and f1 score of all classes for each scenario