Multimodal Emotion Detection With an Emphasis on Speech Modal

Hasti Khajeh

September 2023

Project Advisor: Prof. B. Nasersharif
Project Examiner: Prof. M. Vali

Outline

- Speech Emotion Recognition
- Machine Learning
- Experiment & Results
- Conclusion
- References

Speech Emotion Recognition

Speech Emotion Recognition & Applications

A key source of emotional information is the spoken expression, which may be part of the interaction between the human and the machine.

Speech emotion recognition (SER) has various applications.

- Automatic translation systems
- Anger detection for audio portals or call centers

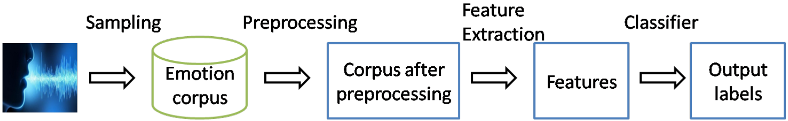
Challenges in Speech Emotion Recognition

- Lack of large and high-quality datasets
- Diverse emotions
- Cultural and linguistic differences
- Uncertainty in labeling and annotation

Speech Emotion Recognition Process

Speech Emotion Recognition, like other speech processing systems, primarily consists of 3 stages:

- Pre-processing of the speech signal
- Extraction of speech features
- Classifier training & Output speech emotion labels



Feature Extraction

The features used in speech emotion analysis are typically divided into two main categories:

- Time domain
- Frequency domain

Each of these categories extracts specific features from speech data for the purpose of emotion detection and interpretation.

Time Domain Features

- Magnitude

$$M_s = \frac{1}{N}\sum_{n=0}^{N-1}|S[n]|$$

- Energy

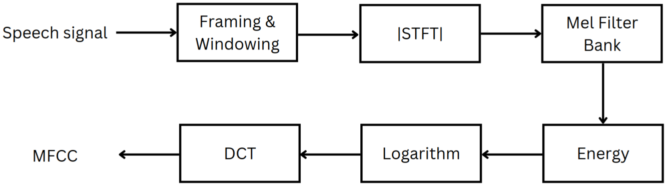
$$E_s = \frac{1}{N}\sum_{n=0}^{N-1}S^2[n]$$

- Energy Entropy
- Zero Crossing Rate

$$ZCR = \frac{1}{N}\sum_{n=1}^{N-1}\frac{|sgn[S(n)] - sgn[S(n-1)]|}{2}$$

Frequency Domain Features

- Pitch Frequency
- Formant Frequencies
- Spectrum
- Mel Frequency Cepstral Coefficient (MFCC)



Frequency Domain Features (Cont'd)

- Chromagram
- Spectral Centroid

$$\mathit{spectral\ centroid} = \frac{\sum_{k=0}^{N-1} f[k]|X[k]|}{\sum_{k=0}^{N-1} |X[k]|}$$

- Spectral Entropy

$$H(p) = -\sum_{k=0}^{N-1} p_k \log_2 p_k$$

$$p_k = \frac{|X[k]|}{\sum_{k=0}^{N-1}|X[k]|}$$

Frequency Domain Features (Cont’d)

- Spectral Rolloff
- Spectral Flux

$$spectral\ flux = \left\| \frac{|X_{\hat{n}}|}{\sum_{k=0}^{N-1}|X_{\hat{n}}[k]|} - \frac{|X_{\hat{n}+1}|}{\sum_{k=0}^{N-1}|X_{\hat{n}+1}[k]|} \right\|$$

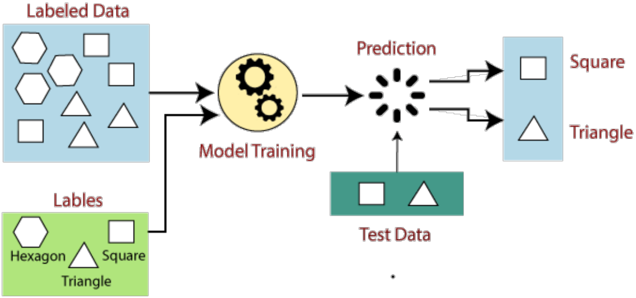
- Spectral Spread

$$spectral\ spread = \sqrt{\frac{\sum_{k=0}^{N-1}(f_k - SC)^2|X[k]|}{\sum_{k=0}^{N-1}|X[k]|}}$$

Machine Learning

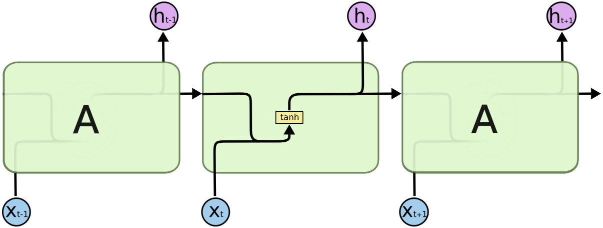
Supervised Learning

The training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly.



Recurrent Neural Network

Recurrent Neural Networks (RNNs) handle sequences by considering both current and past inputs, impacting present output through memory of prior inputs.

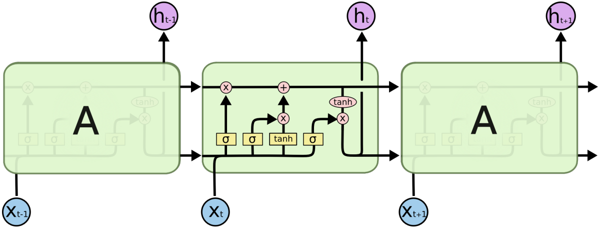


- The 'Vanishing gradient' problem

Long Short-Term Memory

Long Short-Term Memory (LSTM) is a specialized RNN. An LSTM has three of these gates (forget, input, output), to protect and control the cell state.

- Input gate: manages the information to store from new input.
- Forget gate: decides what to discard from the previous state.
- Output gate: controls the information to be used as output.



Experiment & Results

Dataset

- Utilized IEMOCAP dataset in this thesis.
- Dataset comprises 12 hours of audio, video, and English text data from 17 actors, divided into 5 sessions.
- Dialogues segmented into shorter sections lasting 3 to 15 seconds each.
- Evaluation involved 3 or 4 assessors who labeled the sections with various emotions: neutral state, happiness, sadness, anger, surprise, fear, disgust, Frustration, Excited, and Other.

Dataset (Cont’d)

- Note: Combined labels of sadness, anger, and neutral, as well as merged happiness and Excited labels for this thesis. Only **spontaneous** data was used. **5-fold** method used for evaluation based on sessions present in database.

happiness	sadness	anger	neutral
473	304	144	549

Pre-Processing

- Speech Features:
    - Segments of 0.2 seconds, step size 0.1 seconds, and 16 kHz sampling rate.
    - Each segment: 30 ms frame length, 75% overlap.
    - Total of 34 features:
        - 13 MFCCs
        - 13 chromagram-based features
        - 8 Time Spectral Features
- Frame Processing:
    - Each sample considers 100 frames.
    - If frames > 100, excess frames are removed.
    - If frames < 100, zero-featured frames added.
    - Ensures a consistent 100 frames for each sample.
- Input to Model:
    - Results in a matrix of shape (100, 34).

Speech Models Architecture

Abbreviated Model Name	Model Name	Layer	Number of Neurons	Activation Function
M1	MLP	Dense	256	ReLU
		Dense	128	ReLU
		Dense	4	Softmax
M2	LSTM	LSTM	128	Tanh
		LSTM	64	Tanh
		Dense	64	ReLU
		Dense	4	Softmax
M3	BLSTM	BLSTM	128	Tanh
		Dense	256	ReLU
		Dense	4	Softmax

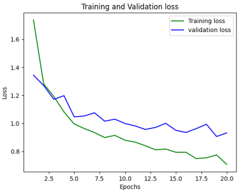
Model Compilation and Training Setup

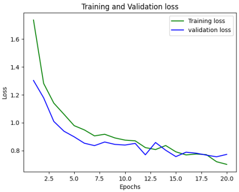
- Model Compilation:
  - Optimizer: Adam
  - Learning Rate: 3e-4
  - Loss Function: Categorical Cross entropy
  - Metrics: Accuracy
- Training Setup:
  - Data Split: Train-Validation (90% Train, 10% Validation)
  - Batch Size: 64
  - Epochs: 20

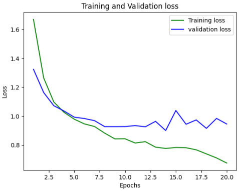
Speech Models Result

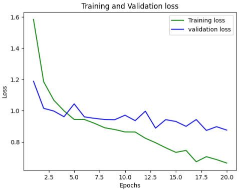
Abbreviated Model Name	session1	session2	session3	session4	session5	Average
M1	59.4%	62.2%	52.7%	59.2%	49.3%	56.6%
M2	58.5%	54.5%	48.3%	38.6%	52.2%	50.4%
M3	58.5%	62.9%	55.9%	63.1%	53.1%	58.7%

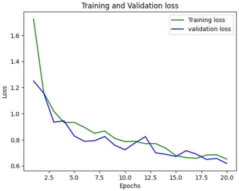
M3 Model Loss



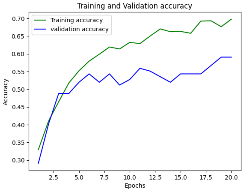


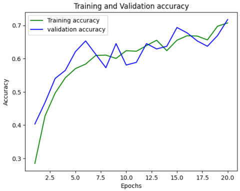


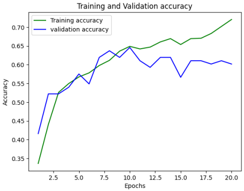


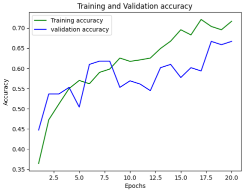


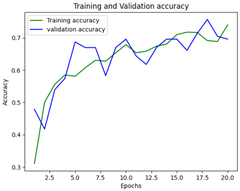
M3 Model Accuracy



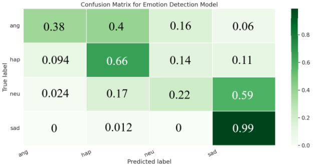


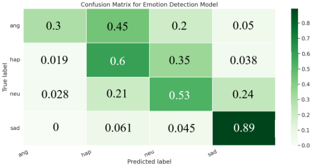


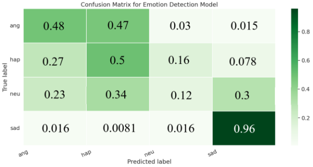


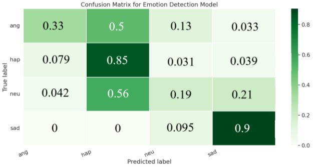


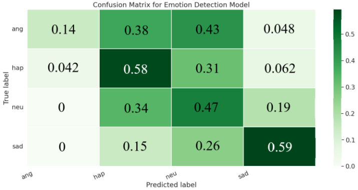
M3 Model Confusion Matrix



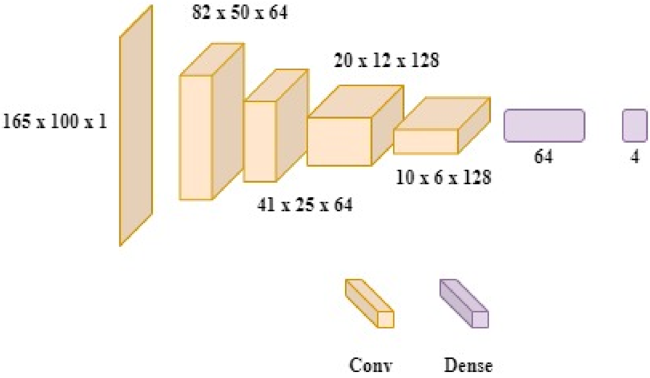




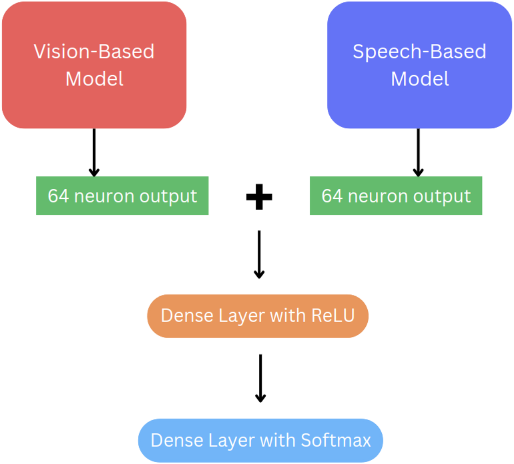




Vision-Based Model Architecture



Multi-modal Model Architecture



Model Compilation and Training Setup

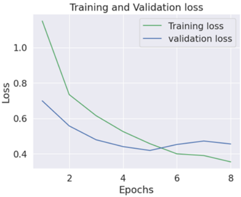
- Model Compilation:
  - Optimizer: Adam
  - Learning Rate: 3e-4
  - Loss Function: Categorical Cross entropy
  - Metrics: Accuracy
- Training Setup:
  - Data Split: Train-Validation (90% Train, 10% Validation)
  - Batch Size: 64
  - Epochs: 8

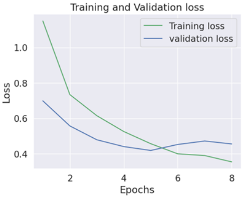
Multi-modal Model Result

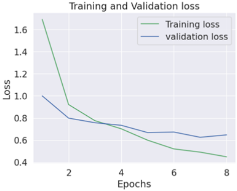
session1	session2	session3	session4	session5	Average
71.55%	70.23%	68.38%	71.12%	50.57%	66.37%

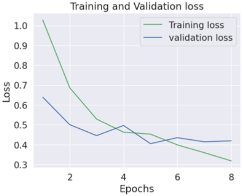
In terms of accuracy the multi-modal model is better by a margin of almost 8%.

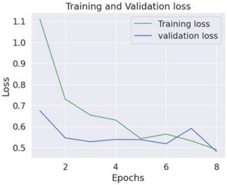
Multi-modal Model Loss



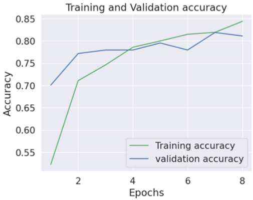


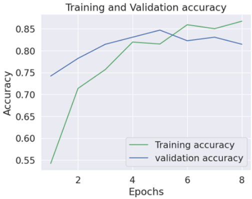


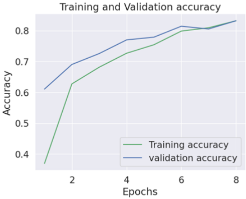


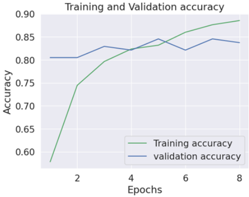


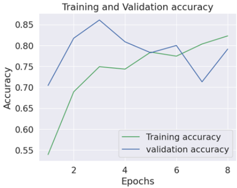
Multi-modal Model Accuracy



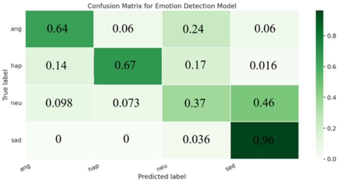


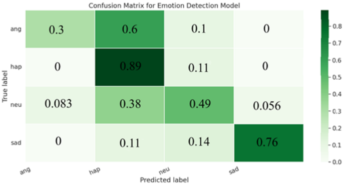


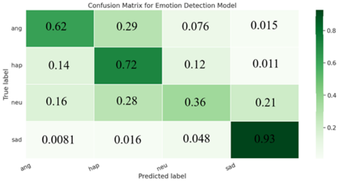


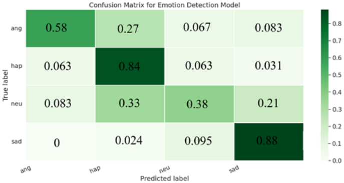


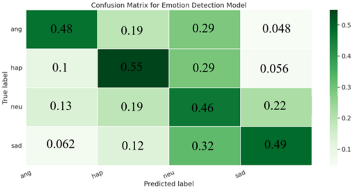
Multi-modal Model Confusion Matrix











Conclusion

Conclusion

- The results demonstrated that the bimodal model achieved a higher accuracy in emotion detection by leveraging data from both modalities.
- The speech model achieved an average accuracy of 58.7% in emotion detection, while the bimodal model achieved an accuracy of 66.37%.
- Improved detection of specific emotions like anger and neutral.

References

References

- Lieskovská, Eva, et al. ”A review on speech emotion recognition using deep learning and attention mechanism.” Electronics 10.10 (2021): 1163.
- Alías, Francesc, Joan Claudi Socoró, and Xavier Sevillano. ”A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds.” Applied Sciences 6.5 (2016): 143.
- Ewert, Sebastian. ”Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features.” Proc. ISMIR. 2011.
- Ayodele, Taiwo Oladipupo. ”Types of machine learning algorithms.” New advances in machine learning 3 (2010): 19-48.

References (Cont’d)

- Sherstinsky, Alex. ”Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.” Physica D: Nonlinear Phenomena 404 (2020): 132306.
- Hochreiter, Sepp, and Jürgen Schmidhuber. ”Long short-term memory.” Neural computation 9.8 (1997): 1735-1780.
- Busso, Carlos, et al. ”IEMOCAP: Interactive emotional dyadic motion capture database.” Language resources and evaluation 42 (2008): 335-359.
- Tripathi, Samarth, and Homayoon Beigi. ”Multi-modal emotion recognition on IEMOCAP with neural networks.” arXiv preprint arXiv:1804.05788 (2018).

Thank You!

Any Questions ?

hastikhajeh@email.kntu.ac.ir