Speech Emotion Recognition (SER)

Lim Sheng Xiang 1005005
Joel Tay 1005117
Lim Fuo En 1005125
Anthony Lim 1005264
Ankita Parashar 1005478

Agenda

Problem Formulation

The Current SER Landscape

SER Case Studies

Dataset(s) Used

Methodology

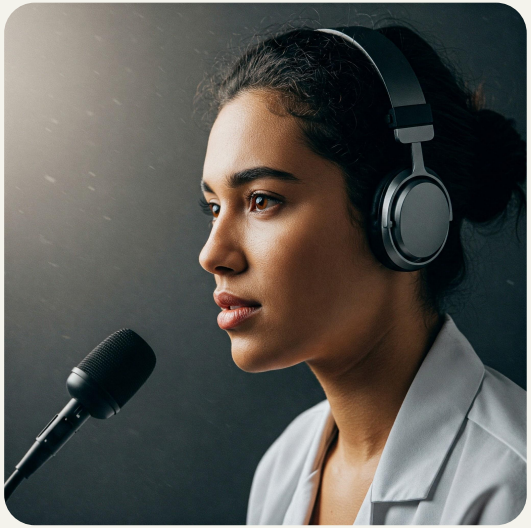
Problem Formulation

Background and Problem Formulation

Understanding emotions in speech is a fundamental yet challenging aspect of human communication.

Misinterpretation of emotional cues can lead to misunderstandings in critical fields such as law enforcement, customer service, and mental health support.

However, accurately identifying emotions from speech is **COMPLEX** due to **variations in tone**, **accent**, and **context**.



Solution

Speech Emotion Recognition (SER)

1

Identifies Emotional Cues

AI detects tone, pitch, and intensity to recognize core emotions like happiness, sadness, anger, and fear

2

Handles Speech Variability

Adapts to different accents, tones, and languages, improving accuracy across diverse speakers

3

Reduces misinterpretations

Provides objective emotional analysis, minimizing human bias and preventing misjudgments

4

Enhance Emotional Awareness

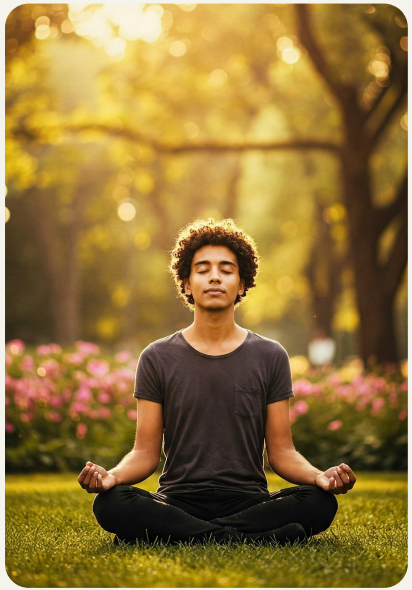
Helps individuals and professionals better understand emotional states for improved interactions

5

Supports Decision Making

Enables informed responses with better emotional insights of stakeholders

AI For Good - SER for Mental Health Monitoring



Why does Mental Health Matter?

- **1 in 8 people globally** live with a mental health condition
- Early detection **prevents escalation** and improves treatment outcomes

How SER Supports Mental Health

- Detects Emotional Distress
  - AI algorithms can analyze speech patterns for signs of depression and anxiety with high accuracy
- Enables Early Intervention
  - AI tracks emotional triggers and warning signs, allowing professionals to step in before conditions worsen
- Objective & Scalable Monitoring
  - Provides continuous, unbiased emotional assessment beyond self-reporting, offering deeper insights between therapy sessions

SER provides many benefits to the society, fitting into the theme of ‘AI for Good’

The Current SER Landscape

Traditional Approaches in SER

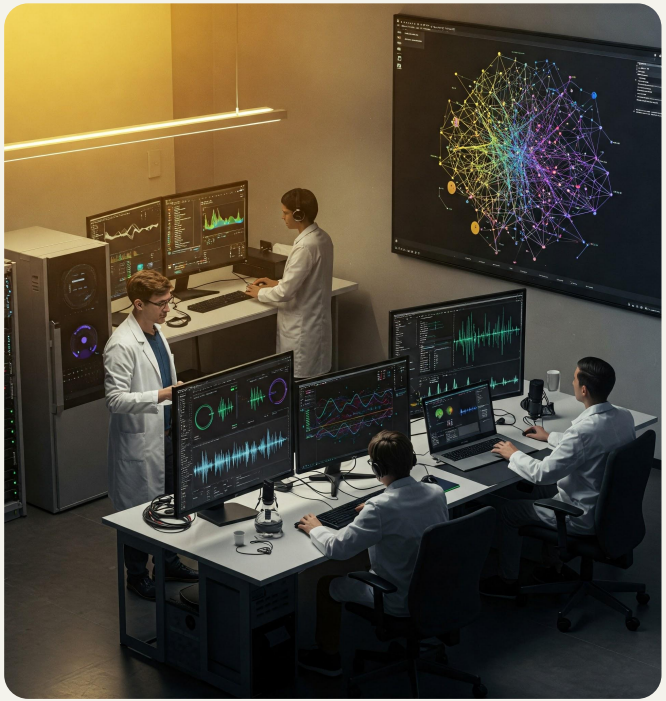
Acoustic **speech features** were extracted from raw speech signals:

- Spectral features, like MFCCs (Mel Frequency Cepstral Coefficients)
- Prosodic features, pitch, energy, and speech intensity

Classifiers used are:

- Support Vector Machines (SVMs)
  - Known to have good generalizability in noisy SER tasks
- Hidden Markov Models (HMMs)
- Decision Trees

Deep Learning Approaches in SER



Deep learning approaches are increasingly being used due to **enhanced recognition rates** and **robustness**.

End-to-end architectures are used to **learn features** directly from speech signals:

- Deep Neural Networks (DNNs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory (LSTMs)

Why Noisy SER Matters

Noise is a **significant challenge** in SER, hindering the performance of SER models.

Noisy SER is SER in **uncontrolled environments**, with noisy speech signals as inputs.

Our project is considered noisy SER, as our input speech signals are **not going to be recorded in controlled environments**.

Noisy SER Techniques

Preprocessing

- **Noise reduction** that reduces unwanted background noise from a speech signal
- **Voice Activity Detection (VAD)** that isolates speech portions from non-speech ones (i.e., silence, noise)
- Signal **normalisation**

Noise Robust Feature Extraction

- The process of **identifying and extracting crucial information** from speech signals to create noise robust data
- Features that are **robust to noise**:
  - Mel-Frequency Cepstral Coefficients (MFCCs)
  - Weighted Wavelet Packet Cepstral Coefficients (W-WPCCs)
  - Prosodic features like rhythm, intonation, stress patterns
- **Mel spectrograms** are commonly used for deep learning approaches like CNNs to extract meaningful features

Data Augmentation

- Artificial noise addition:
  - **Additive Gaussian White Noise (AGWN)**, pink noise, and environmental noise are often added to speech signals in datasets from controlled environments
- Helps to **improve robustness** of SER in different situations

Performance of Existing Noisy SER Models

Accuracy is the most commonly-used evaluation metric in noisy SER as a classification task.

Model	Test Dataset	Result (Accuracy)
SVM with RBF Kernel	• EMO_DB • IEMOCAP	• **EMO-DB: 85%** • IEMOCAP: 77%
DNN	• EMO_DB • IEMOCAP	• **EMO_DB: 76.77%** • IEMOCAP: 53.55%
CNN	• EMO_DB • RAVDESS	• EMO_DB: 76% • **RAVDESS: 82%**
Ensembled	• EMO_DB • TESS • CREMA-D	• EMO_DB: 95.42% • **TESS: 99.46%** • CREMA-D: 90.47%
CNN + LSTM	• EMO_DB • IEMOCAP • RAVDESS	• **EMO_DB: 99.76%** • IEMOCAP: 98.13% • RAVDESS: 99.50%

SER Case Studies

CASE 1: Schuller et al. (2011)

Aspect	Details
Acoustic Features	• Prosodic: Pitch, energy, duration • Spectral: MFCCs
Modeling Techniques	• Traditional: GMMs, HMMs • Advanced: SVMs, ANNs
Databases	• Berlin Emotional Speech Database • IEMOCAP
Challenges	• Cultural variability • Speaker-specific traits • Recording conditions

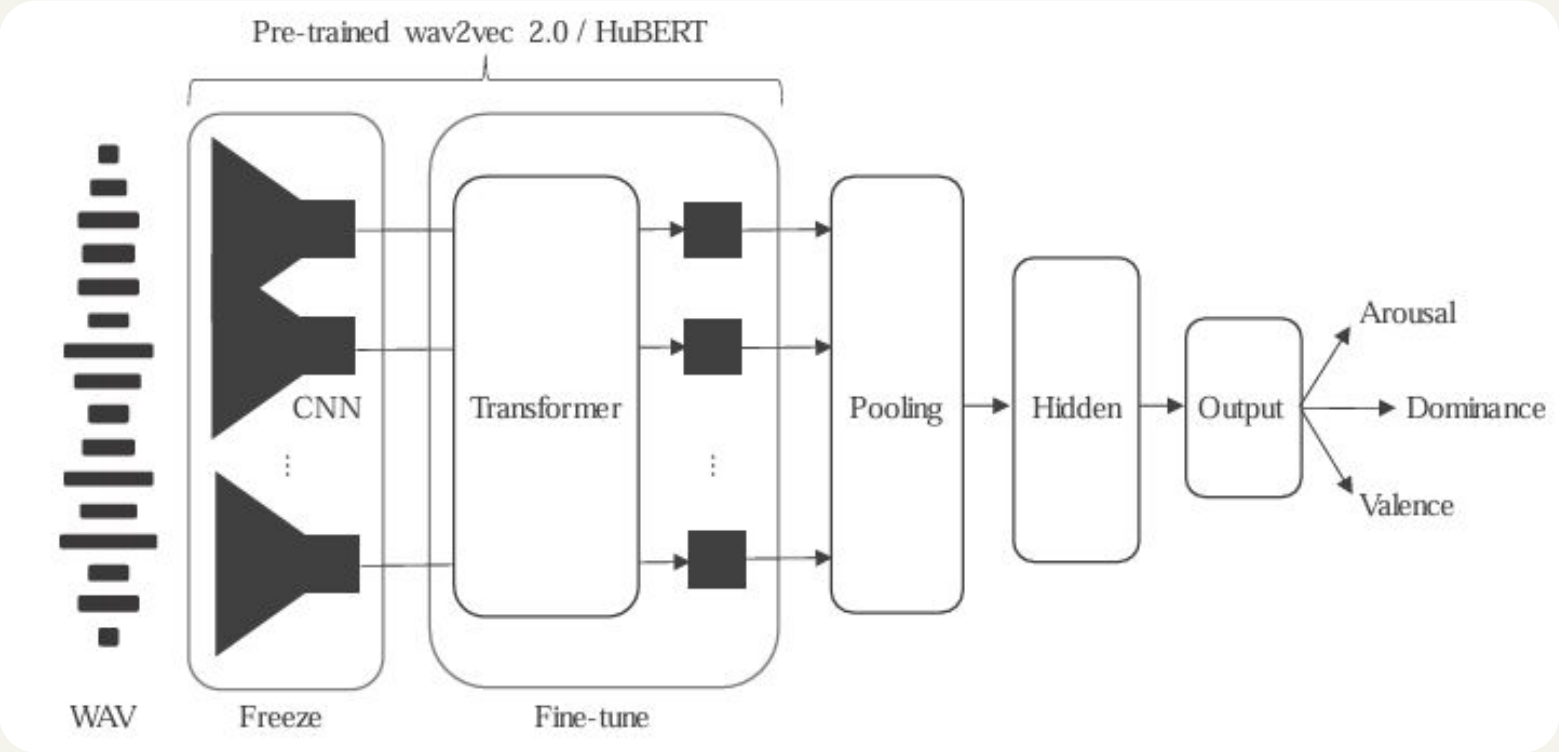
CASE 2: Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap

- Transformer-based models (wav2vec 2.0, HuBERT) for SER
- Focus on valence recognition
- Evaluates models on multiple datasets for robustness, fairness and generalisation
- Implicit extraction of linguistic cues by transformer models

Datasets Used

Datasets Used	Description
MSP-Podcast	62 hours of natural speech, labeled for valence, arousal, dominance. (Primary training dataset)
IEMOCAP	12 hours of acted/improvised speech (Cross-domain evaluation).
MOSI	YouTube movie reviews (Cross-corpus valence evaluation).

Software Architecture



Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). *Dawn of the transformer era in speech emotion recognition: Closing the valence gap*. arXiv preprint arXiv:2203.07378. https://arxiv.org/abs/2203.07378

Evaluation

Metric	Purpose
Concordance Correlation Coefficient (CCC)	Measures agreement between predicted & actual values.
Fairness Analysis	Evaluates if models perform equally well across different groups (e.g. gender).
Robustness Testing	Tests model performance under distortions like noise & speed changes.

Dataset(s) Used

Datasets

Source Datasets

- CREMA-D: **7442** samples
- MELD: **9989** samples
- MLEnd: **32654** samples
- RAVDESS: **1440** samples
- SAVEE: **480** samples
- TESS: **2800** samples
- ESD: **17500** samples
- JL Corpus: **2400** samples

Combine

Some original datasets include other information about the speaker, intensity of speech, and the text. For our project, we are only interested in the raw audio file.

Combined Dataset

Total: **74705** samples

Contains raw audio files in English with their corresponding emotion labels

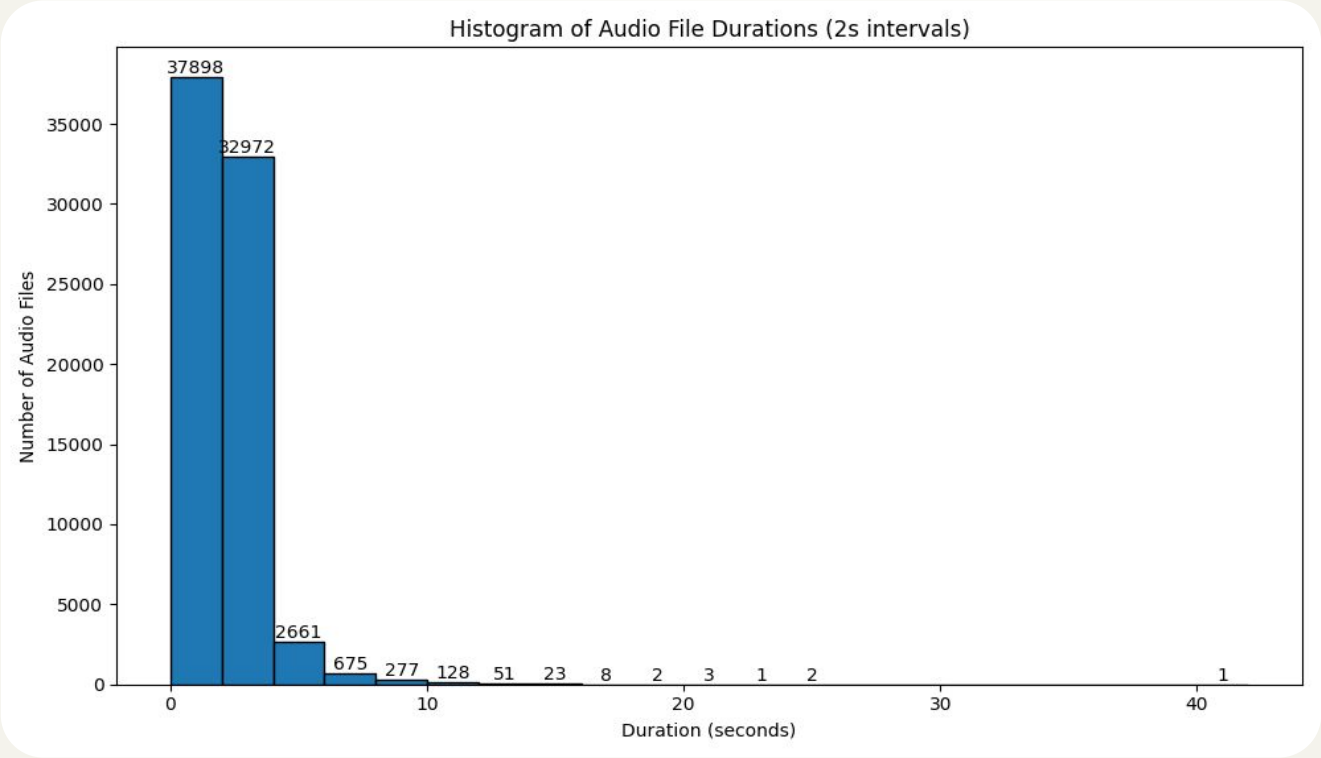
Removed corrupted audio
1 from MELD, 2 from TESS

Dataset without corrupt audio

Total: **74702** samples

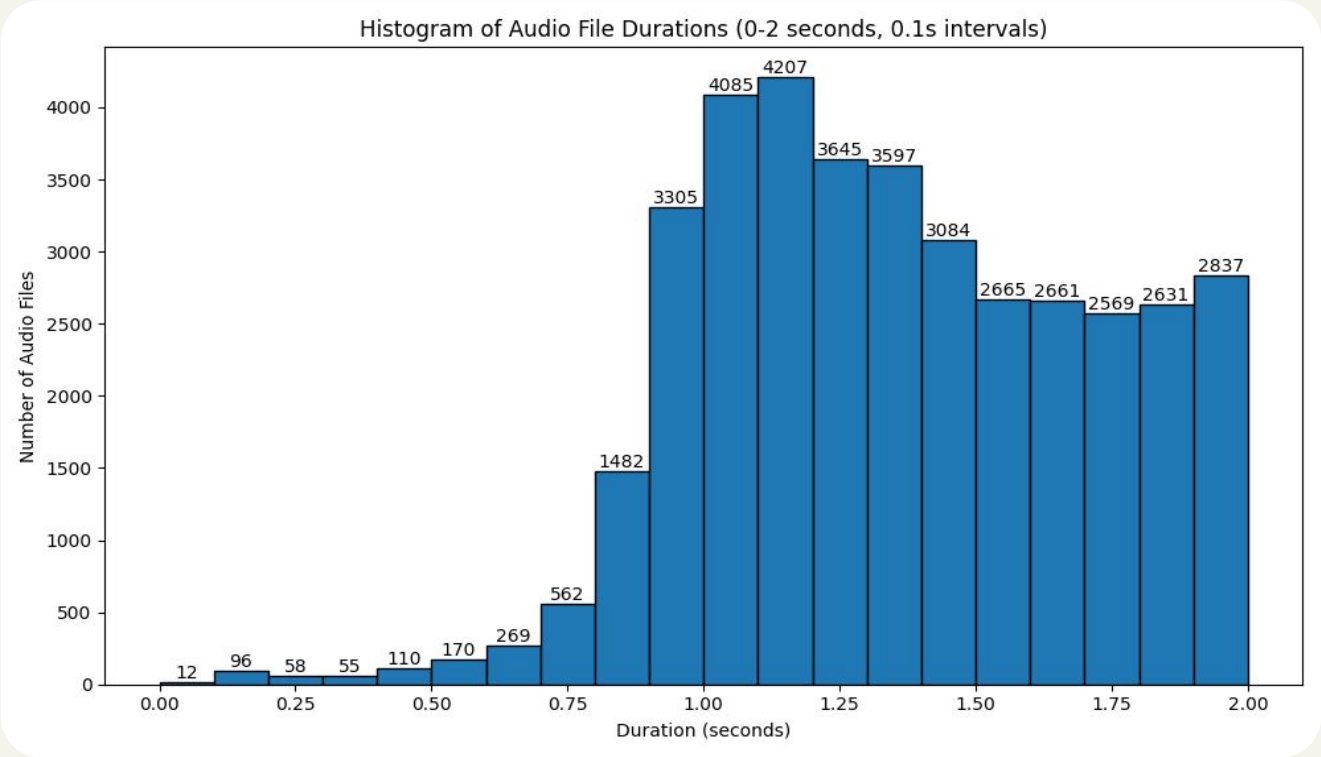
EDA - Distribution of raw audio duration

Long tail distribution plot reveals some outliers, they have very long audio duration



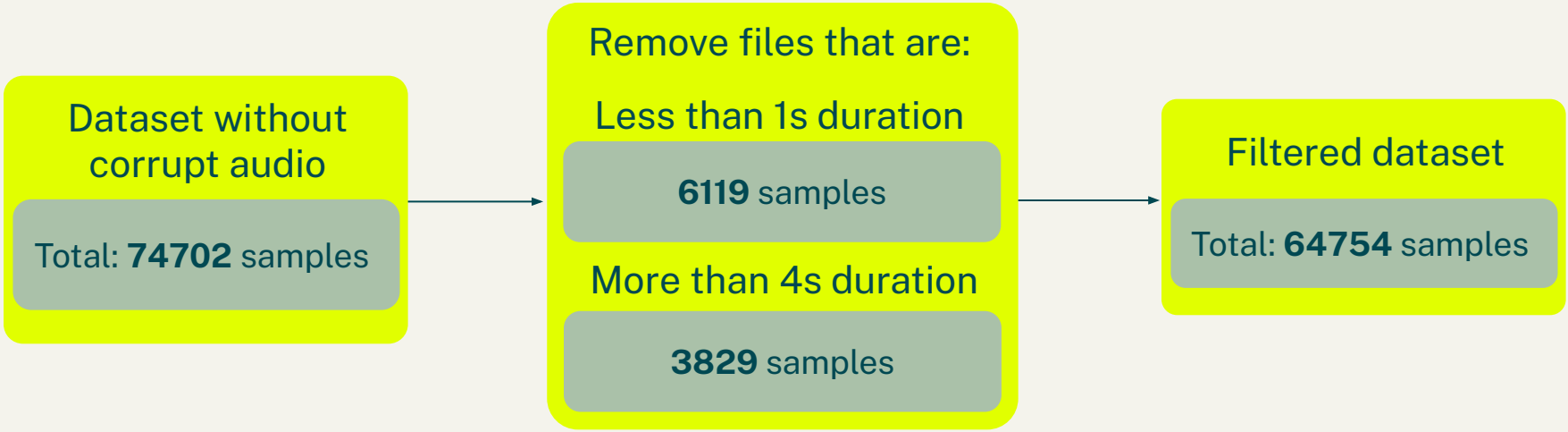
EDA - Distribution of raw audio duration

Distribution plot for shorter durations reveals some outliers, they have very short audio duration



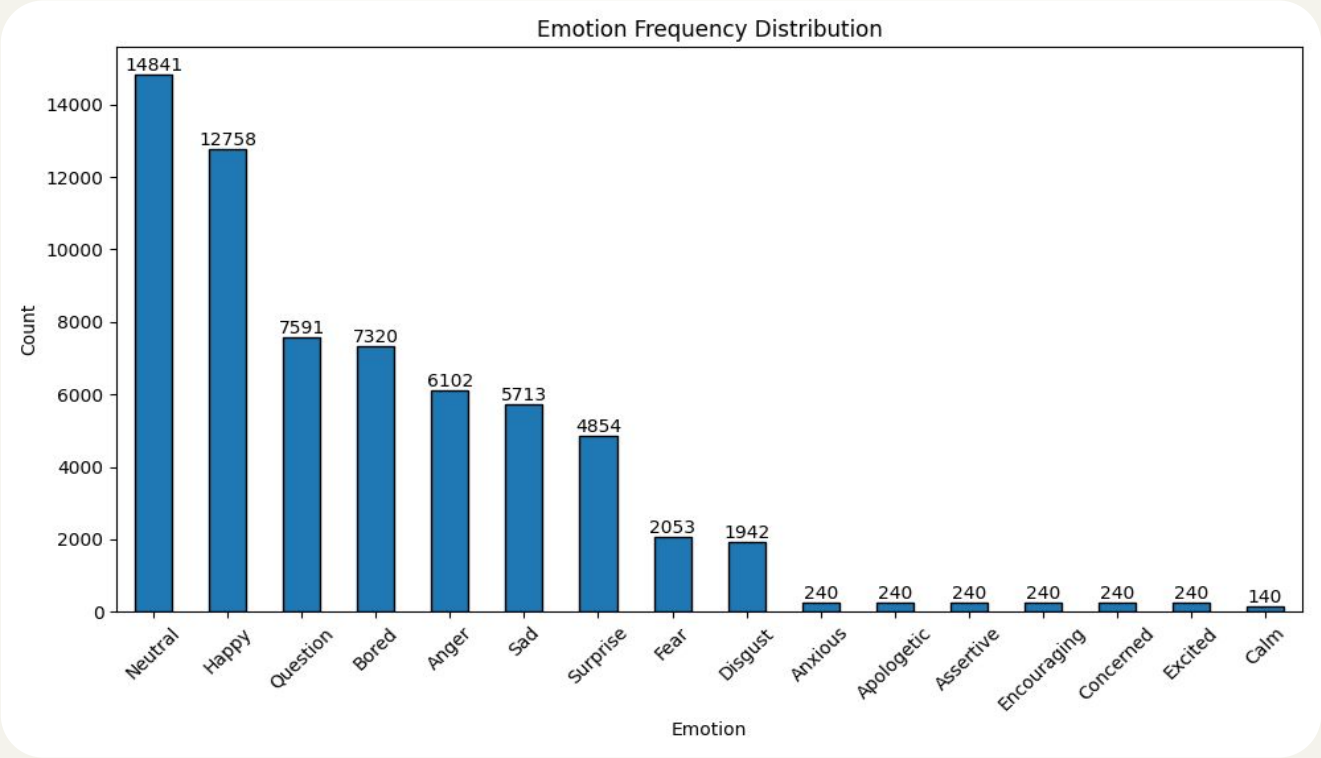
Removal of outliers

Based on the previous two distributions, we filtered out the audio samples which have a very short duration (< 1s) or a very long duration (> 4s) since these might affect model training.



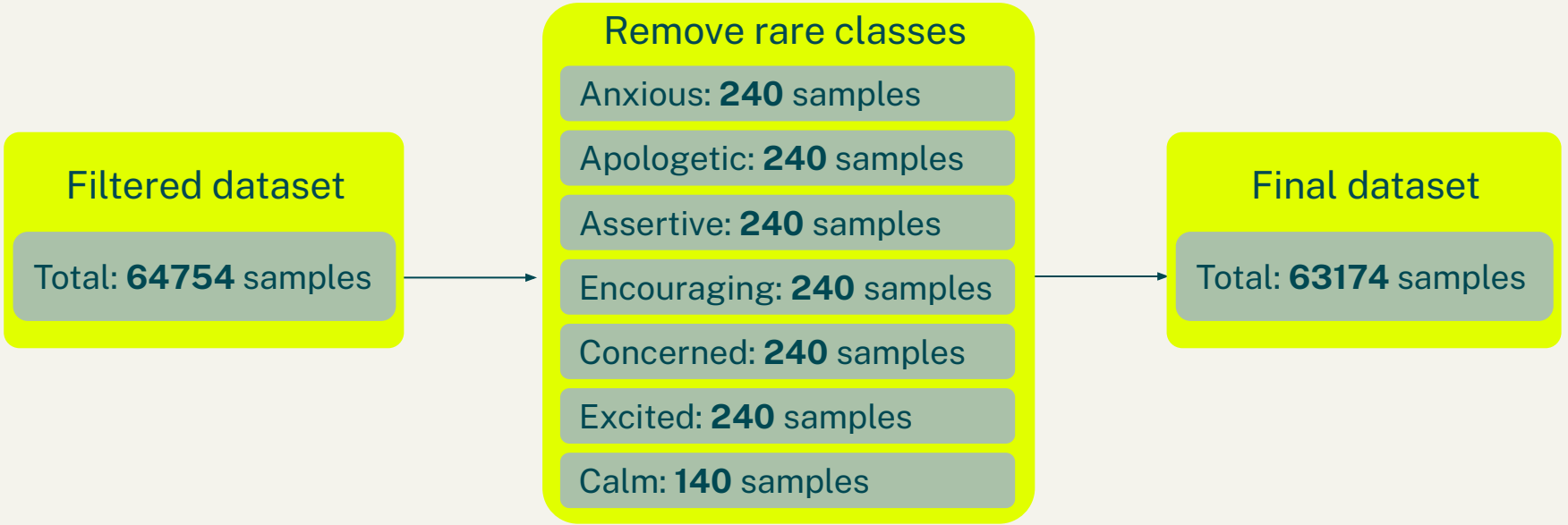
EDA - Emotion label distribution

After removing outliers, there are **16** emotion labels present total in the filtered dataset



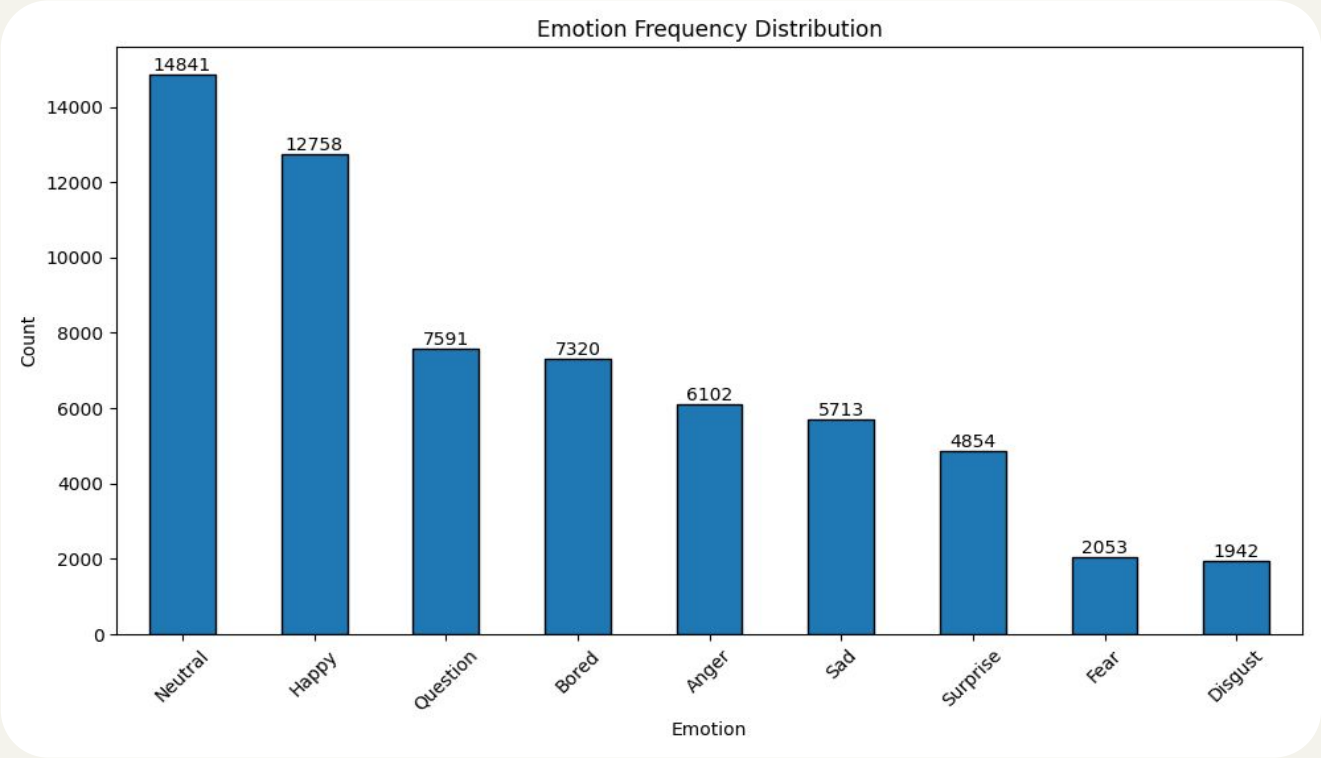
Removal of rare classes

Based on the emotion label distribution, there are very little data for some classes of emotions. This is a huge class imbalance and may affect the performance of the models we train.



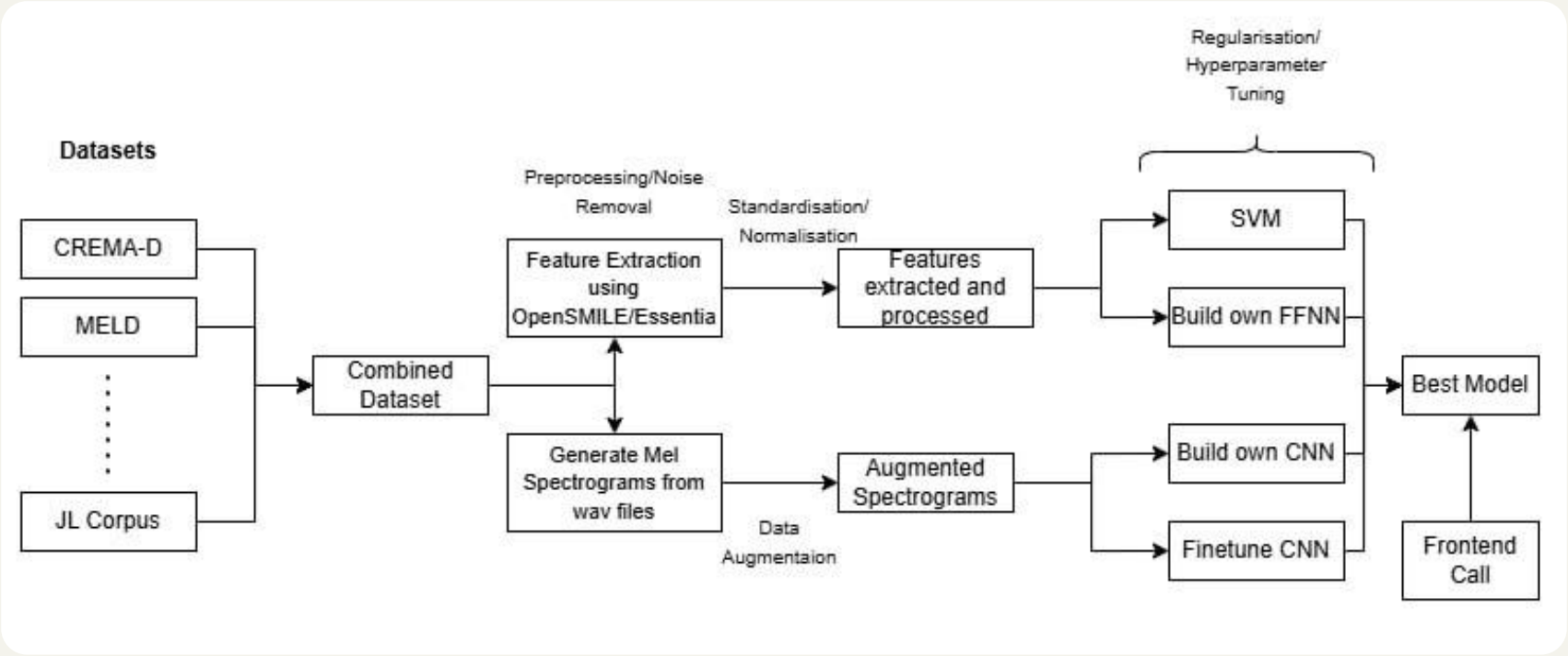
Final dataset distribution

There is still class imbalance, although less. We will employ **stratified sampling** is done to ensure each class is proportionally represented in the train and test set in an 80-20 split.



Methodology

Overview


Datasets
CREMA-D
MELD
JL Corpus
Combined Dataset
Preprocessing/Noise Removal
Feature Extraction using OpenSMILE/Essentia
Standardisation/ Normalisation
Features extracted and processed
Generate Mel Spectrograms from wav files
Data Augmentaion
Augmented Spectrograms
Regularisation/ Hyperparameter Tuning
SVM
Build own FFNN
Build own CNN
Finetune CNN
Best Model
Frontend Call

Training Procedures

Dataset:
- Mixture of datasets
- 63174 samples
- Train-test split (80-20) through stratified sampling

Loss function
- Binary cross-entropy loss since this is classification task

Hyperparameters (Model-dependent, but include):
- Learning rate, optimizers
- Number of hidden layers and their dimensions (e.g. for a feedforward NN)
- Dropout rate (if applicable)
- Epochs, batch size
- Data preprocessing: number of features

Data Preprocessing

Feature extraction with openSMILE:

- In particular, need to assess what features to extract, and if we require dimensionality reduction via LDA for example
- Normalisation

Wav files to spectrograms

- Inputs to the neural networks
- Potentially explore data augmentation of spectrograms (e.g. time warping to shorten/lengthen inputs/addition of Gaussian White noise)

Proposed Architecture

We aim to **branch out and explore** a few architectures before converging into our best one.

Classical Machine Learning

- SVM
  - Useful for classification in general
  - Past works have shown some 85% accuracy
- Feedforward NN
  - Past works have shown about 75-85% accuracy (measured on a variety of datasets, and different ones than ours)

Deep Learning

- CNN
  - Past works have shown about 80-90% accuracy (on EmoDB, variation due to different emotions)

Fine-tuning pre-trained models

- Commonly used model: Wav2Vec2
- Some existing work: 81.82% accuracy on the RAVDESS dataset

Evaluation Metrics

Accuracy

- Accuracy is the most commonly used evaluation criterion in most studies

Other metrics

- Precision & F1 scores
  - Since we are dealing with imbalanced classes, might be worth to look into metrics other than accuracy

Thank You!

Questions?

References

1. Neuroscience News. (2019, May 6). *AI detecting depression via speech.* Retrieved from https://www.mood-me.com/how-emotion-detection-ai-is-revolutionizing-mental-healthcare/
2. ITRex Blog. (2024, July 24). *AI spotting early signs and reducing stigma in mental health.* Retrieved from https://www.mood-me.com/how-emotion-detection-ai-is-revolutionizing-mental-healthcare/
3. Vlisides-Henry, R. D., Gao, M., Thomas, L., Kaliush, P. R., Conradt, E., & Crowell, S. E. (2021). Digital phenotyping of emotion dysregulation across lifespan transitions to better understand psychopathology risk. *Frontiers in Psychiatry, 12*, Article 618442. https://doi.org/10.3389/fpsyt.2021.618442
4. George, Swapna Mol, and P. Muhamed Ilyas. “A Review on Speech Emotion Recognition: A Survey, Recent Advances, Challenges, and the Influence of Noise.” Neurocomputing, vol. 568, Feb. 2024, p. 127015. ScienceDirect, https://doi.org/10.1016/j.neucom.2023.127015.
5. Hashem, Ahlam, et al. “Speech Emotion Recognition Approaches: A Systematic Review.” Speech Communication, vol. 154, Oct. 2023, p. 102974. ScienceDirect, https://doi.org/10.1016/j.specom.2023.102974.
6. Schuller, B., & Batliner, A. (2011). *The automatic recognition of emotions in speech*. In *Emotion-Oriented Systems* (pp. 147–156). Springer.
7. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). *Dawn of the transformer era in speech emotion recognition: Closing the valence gap*. arXiv preprint arXiv:2203.07378. https://arxiv.org/abs/2203.07378

Dataset References

1. Standing-O. (n.d.). *Combined dataset for speech emotion recognition* [GitHub repository]. GitHub. Retrieved from https://github.com/standing-o/Combined_Dataset_for_Speech_Emotion_Recognition
2. Cheyney Computer Science. (n.d.). *CREMA-D dataset*. [GitHub repository]. Retrieved from https://github.com/CheyneyComputerScience/CREMA-D
3. S. Poria, D. Hazarika, N. Majumder, G. Naik, R. Mihalcea, E. Cambria. *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation* (2018)
4. Chen, S.Y., Hsu, C.C., Kuo, C.C. and Ku, L.W. *EmotionLines: An Emotion Corpus of Multi-Party Conversations. arXiv preprint arXiv:1802.08379* (2018)
5. Jesús Requena Carrión, and Nikesh Bajaj. (2022). *MLEnd Spoken Numerals* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3650468
6. "*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*" by Livingstone & Russo is licensed under CC BY-NA-SC 4.0
7. Lok, E. J. (n.d.). *Surrey Audio-Visual Expressed Emotion (SAVEE) dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
8. Lok, E. J. (n.d.). *Toronto emotional speech set (TESS) dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess
9. Kun Zhou, Berrak Sisman, Rui Liu and Haizhou Li, "*Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset*" ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
10. Jesin James, Li Tian, Catherine Watson, "*An Open Source Emotional Speech Corpus for Human Robot Interaction Applications*", in Proc. Interspeech, 2018