Speech Emotion Recognition (SER)

Model: Long Short-Term Memory (LSTM)

Problem Statement:

The aim of this project is to construct and employ an LSTM classification model for SER.

GOAL:

Accurately assess the emotional state of speakers in audio recordings.

HOW:

Using the librosa library standardize and extract features voice signals from a dataset created for SER modeling and then build, train and test the SER model.

DELIVERABLE:

MVP - A functioning LSTM model

What is LSTM

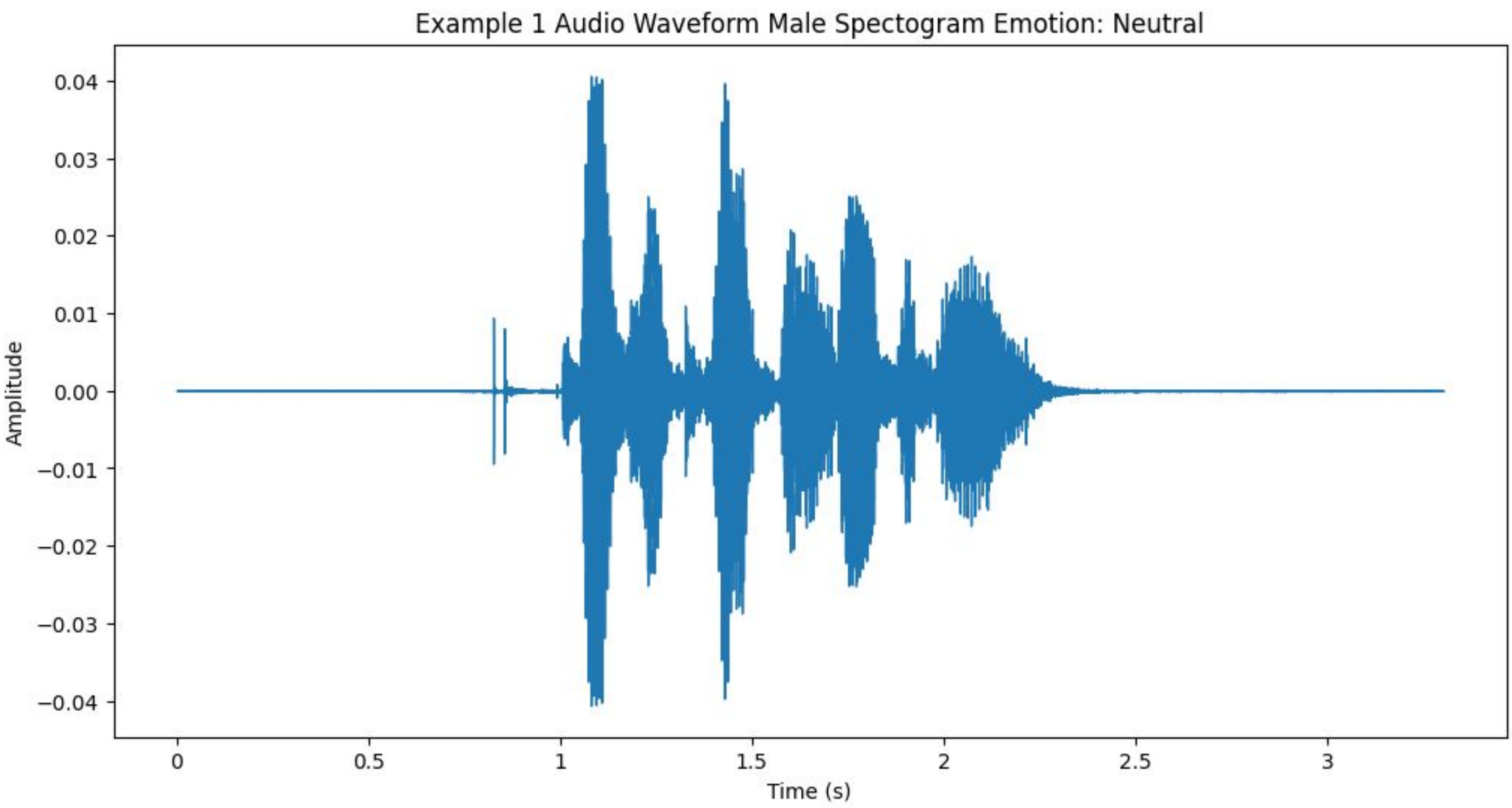
Data:

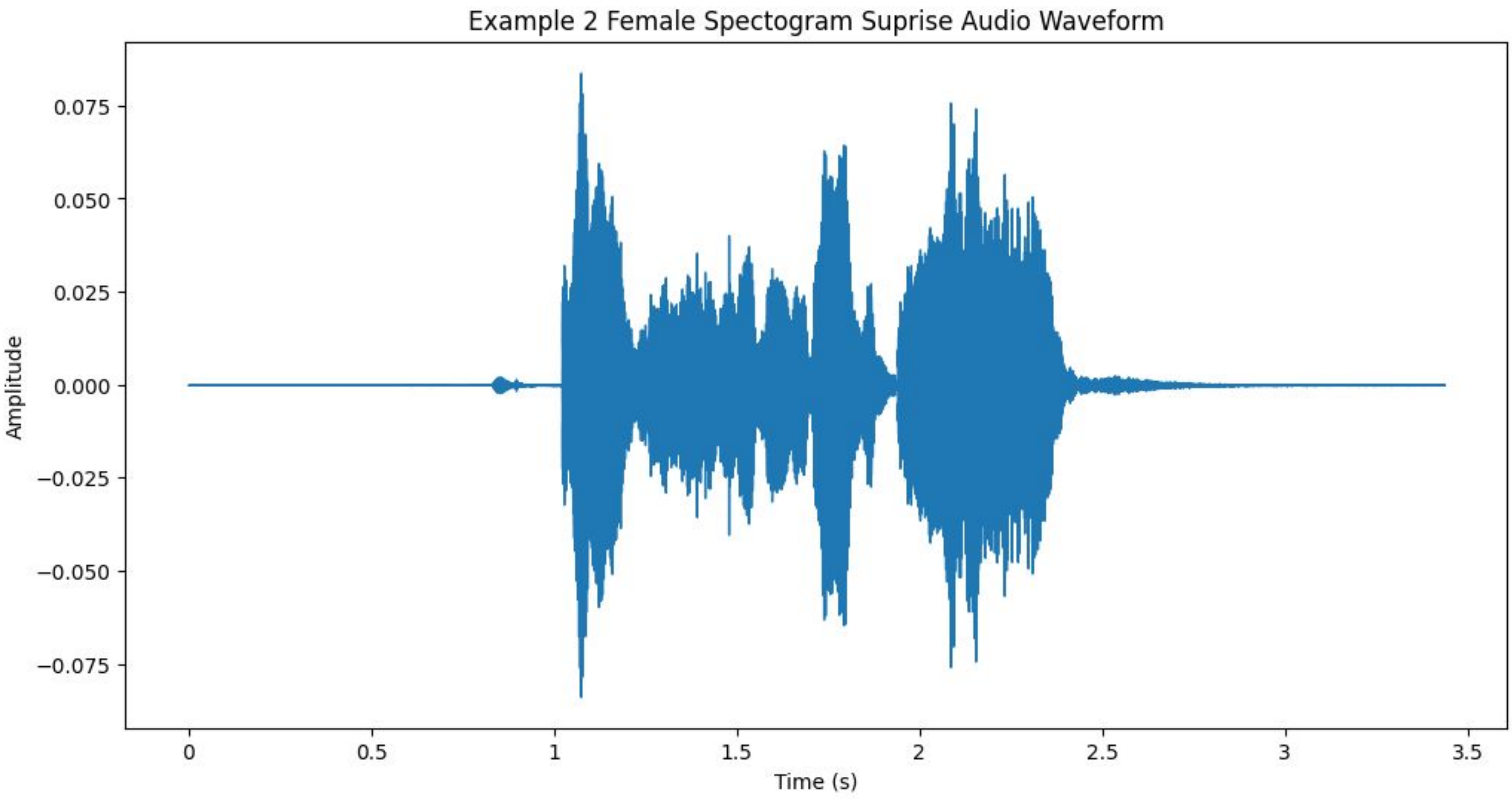
[Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)](https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio)

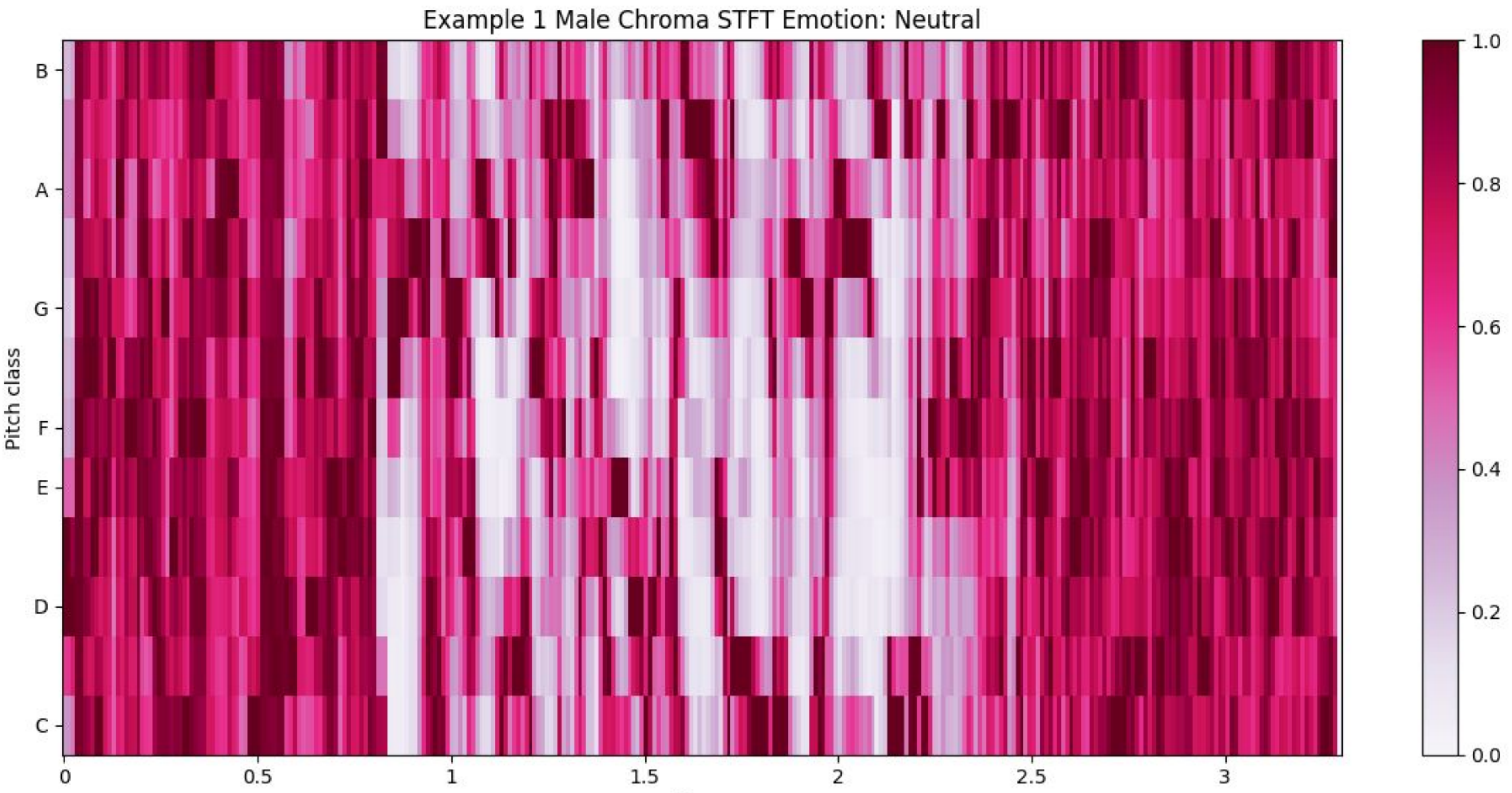
[Toronto emotional speech set (TESS)](https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess)

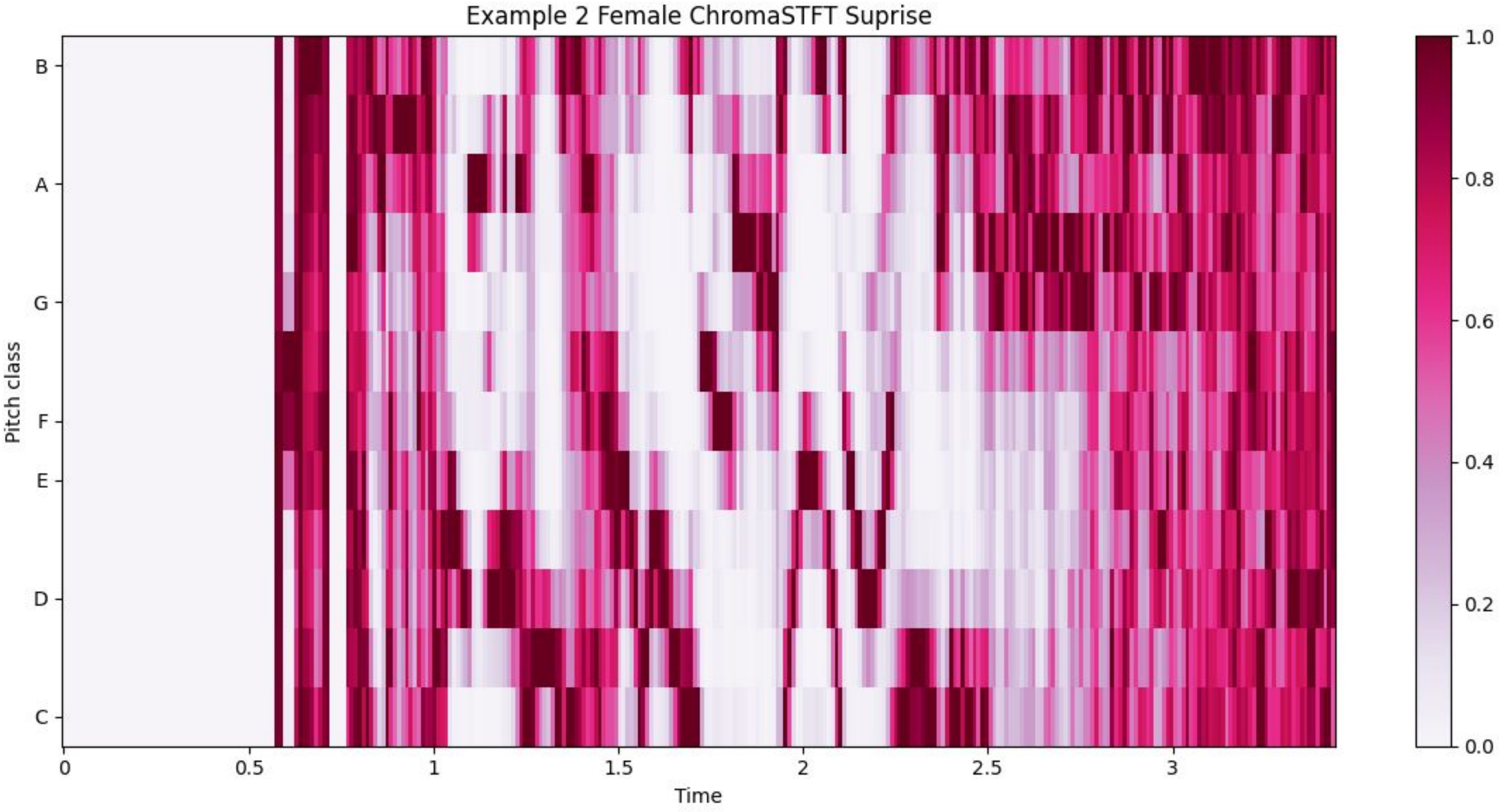
[Surrey Audio-Visual Expressed Emotion (SAVEE)](https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee)

[Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)](https://www.kaggle.com/datasets/ejlok1/cremad)

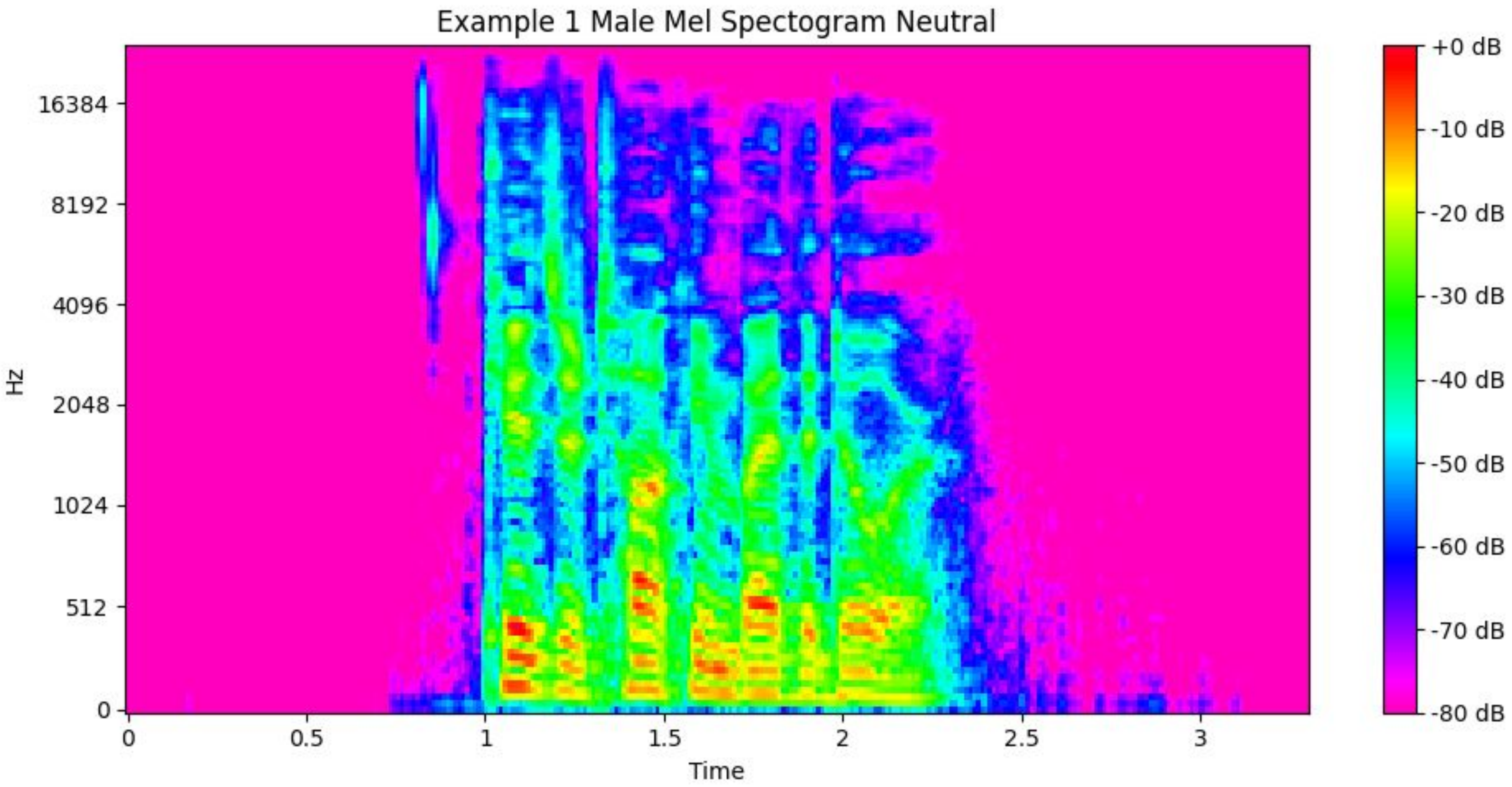

Example 1 Audio Waveform Male Spectogram Emotion: Neutral
Amplitude
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
0
0.5
1
1.5
2
2.5
3
Time (s)


Example 2 Female Spectogram Suprise Audio Waveform
Amplitude
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0
0.5
1
1.5
2
2.5
3
3.5
Time (s)

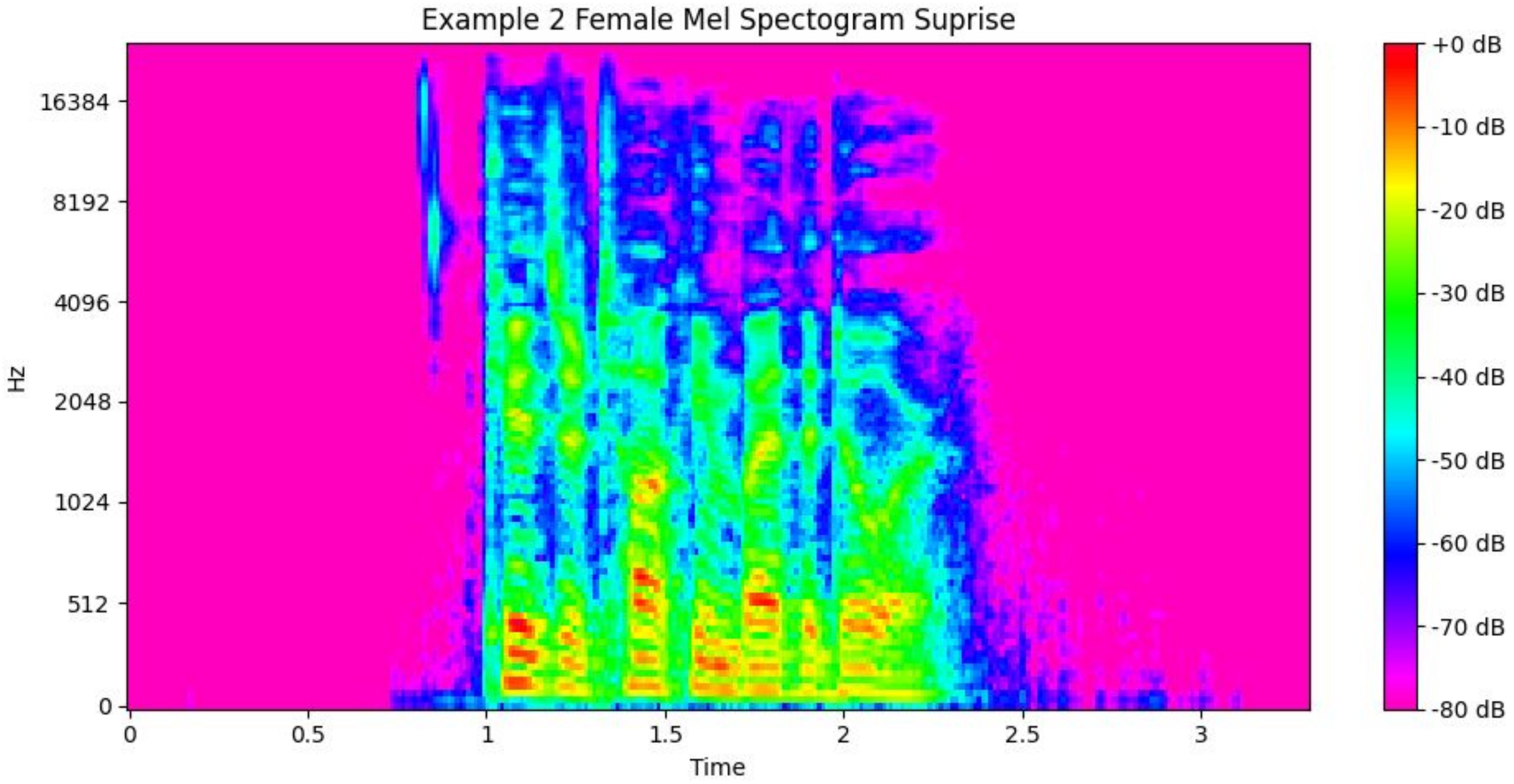

Example 1 Male Chroma STFT Emotion: Neutral
Pitch class
B
A
G
F
E
D
C
0
0.5
1
1.5
2
2.5
3
1.0
0.8
0.6
0.4
0.2
0.0



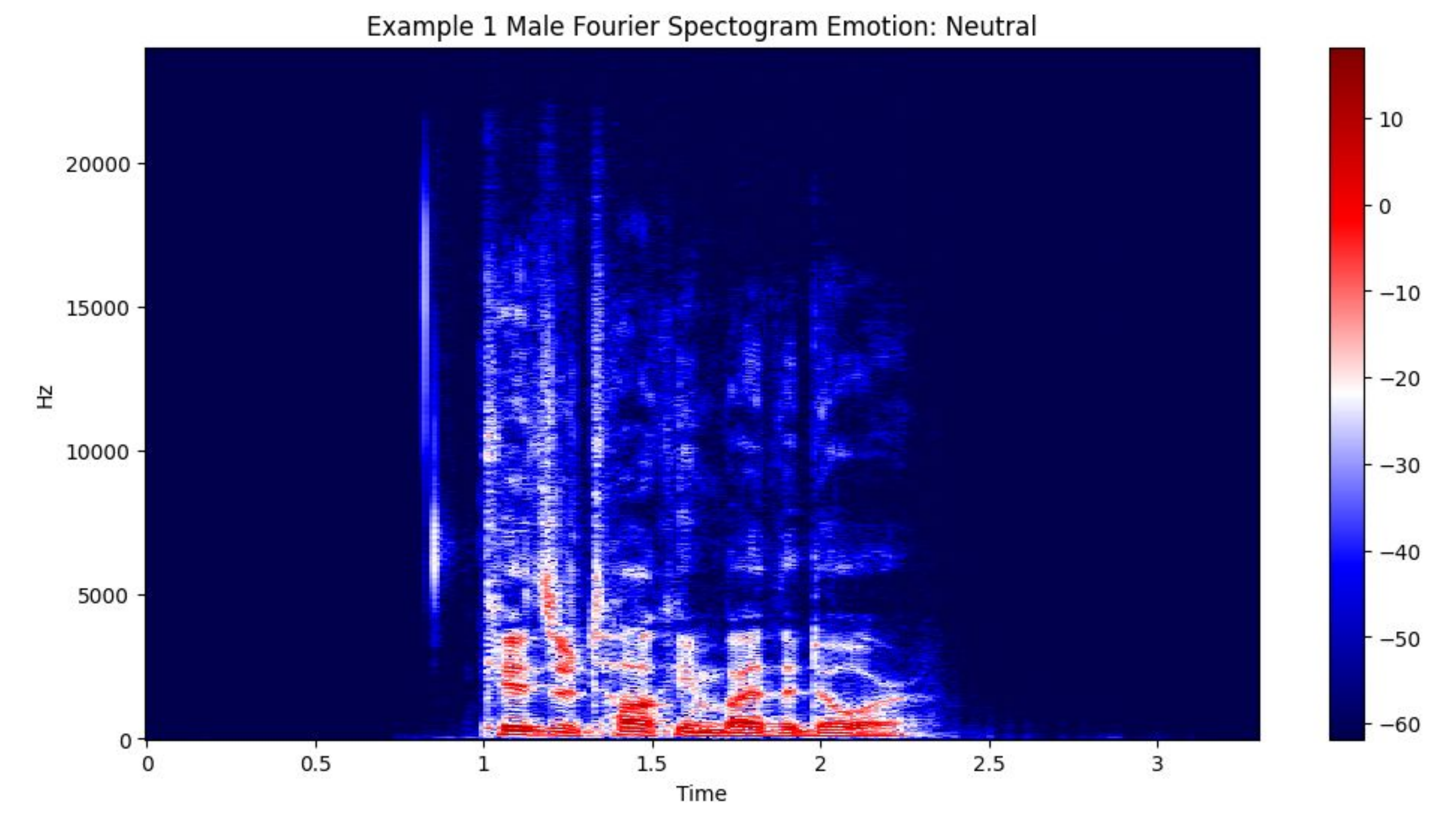
Example 2 Female ChromaSTFT Suprise
Pitch class
B
A
G
F
E
D
C
Time
0
0.5
1
1.5
2
2.5
3
1.0
0.8
0.6
0.4
0.2
0.0



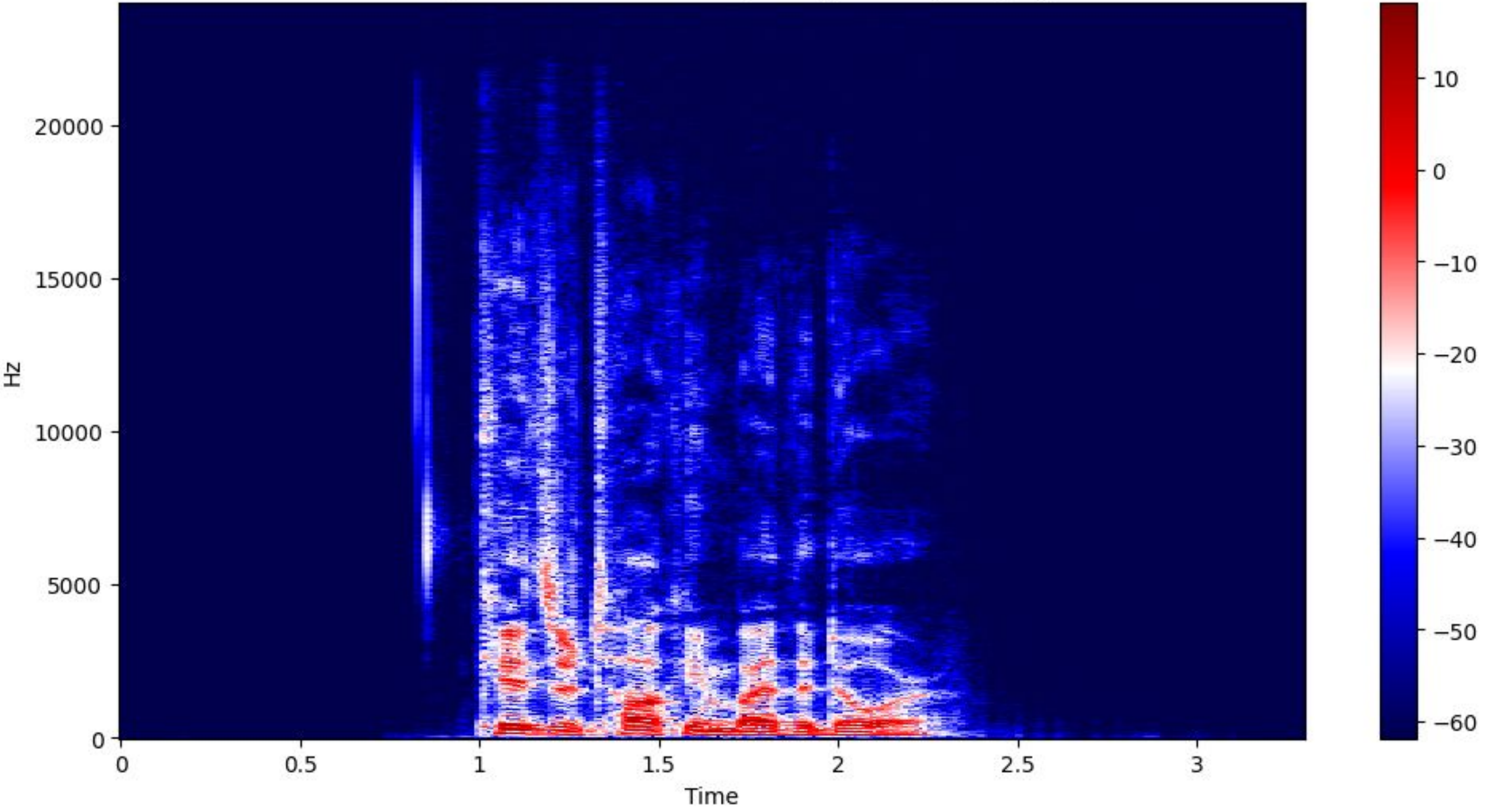
Example 1 Male Mel Spectogram Neutral
Hz
16384
8192
4096
2048
1024
512
0
0
0.5
1
1.5
2
2.5
3
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

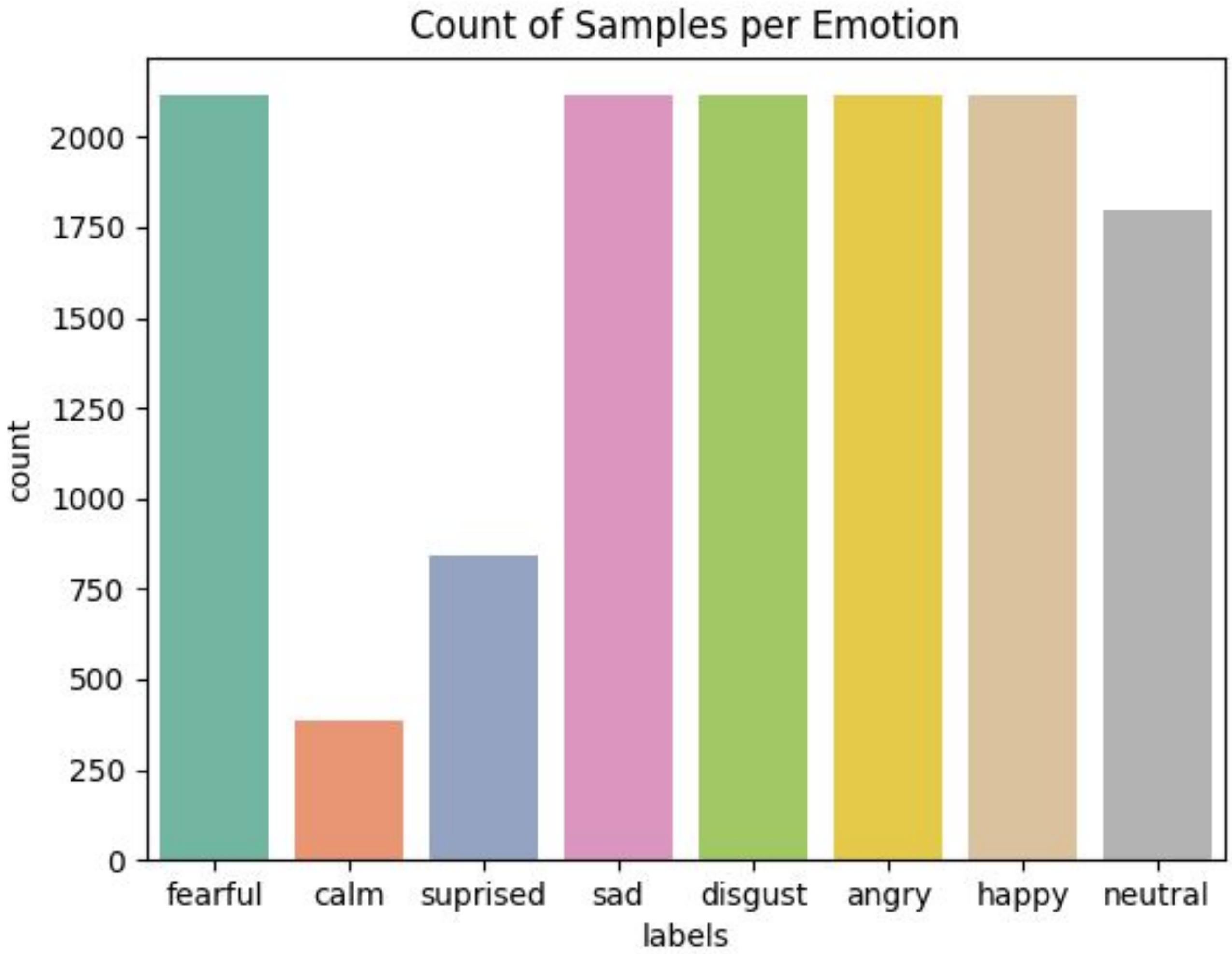


Example 2 Female Mel Spectogram Suprise
Hz
16384
8192
4096
2048
1024
512
0
Time
0
0.5
1
1.5
2
2.5
3
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

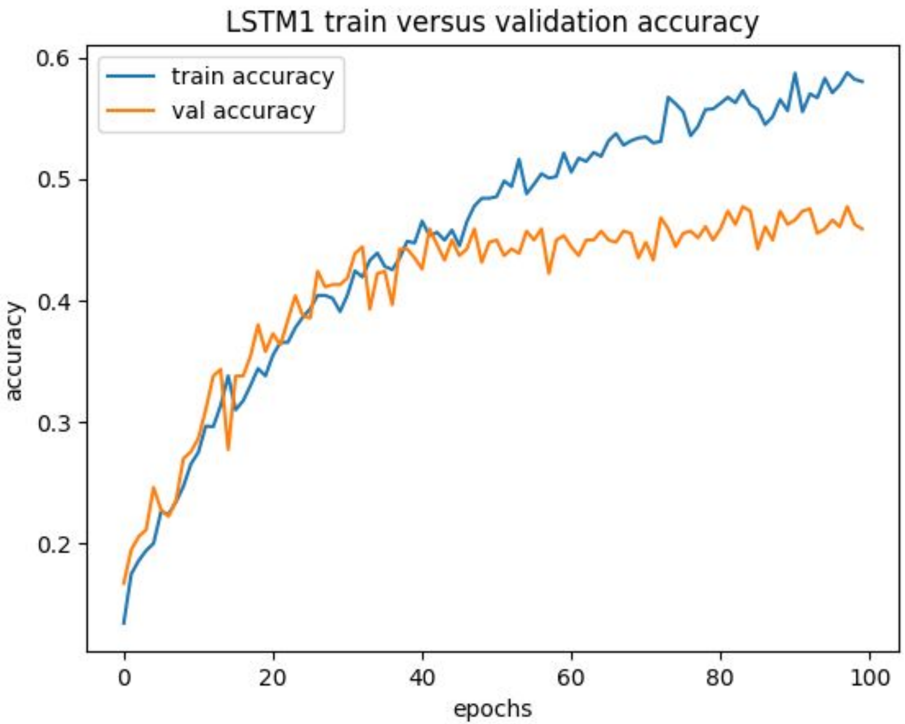

Example 1 Male Fourier Spectogram Emotion: Neutral
Hz
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3
Time
10
0
−10
−20
−30
−40
−50
−60

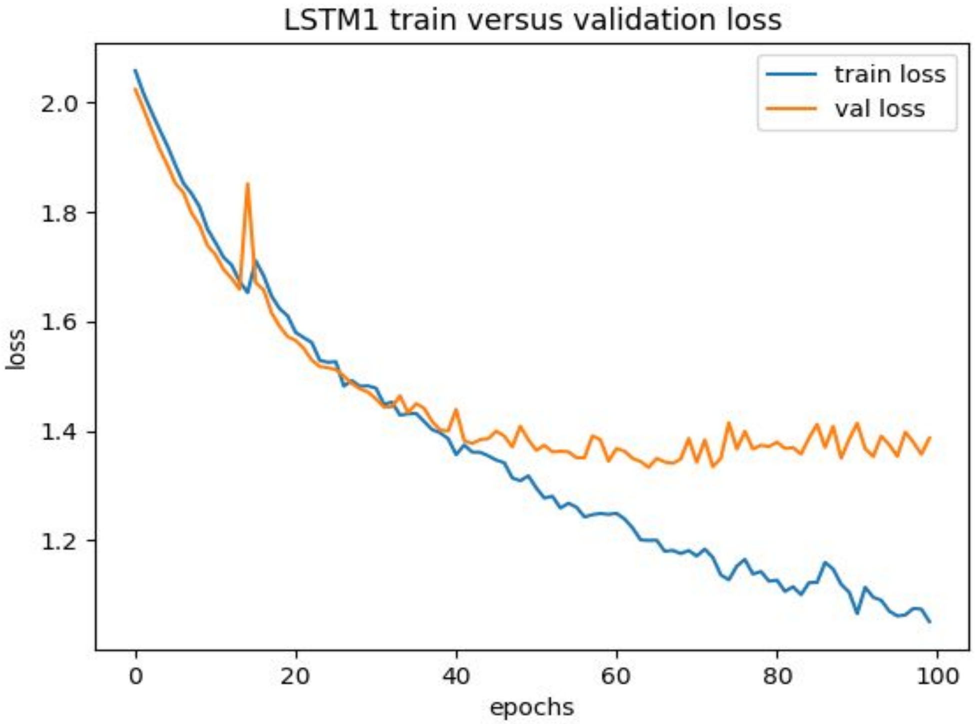
Example 2 Female Fourier Spectogram Suprise.png




Count of Samples per Emotion
count
2000
1750
1500
1250
1000
750
500
250
0
fearful
calm
suprised
sad
disgust
angry
happy
neutral
labels

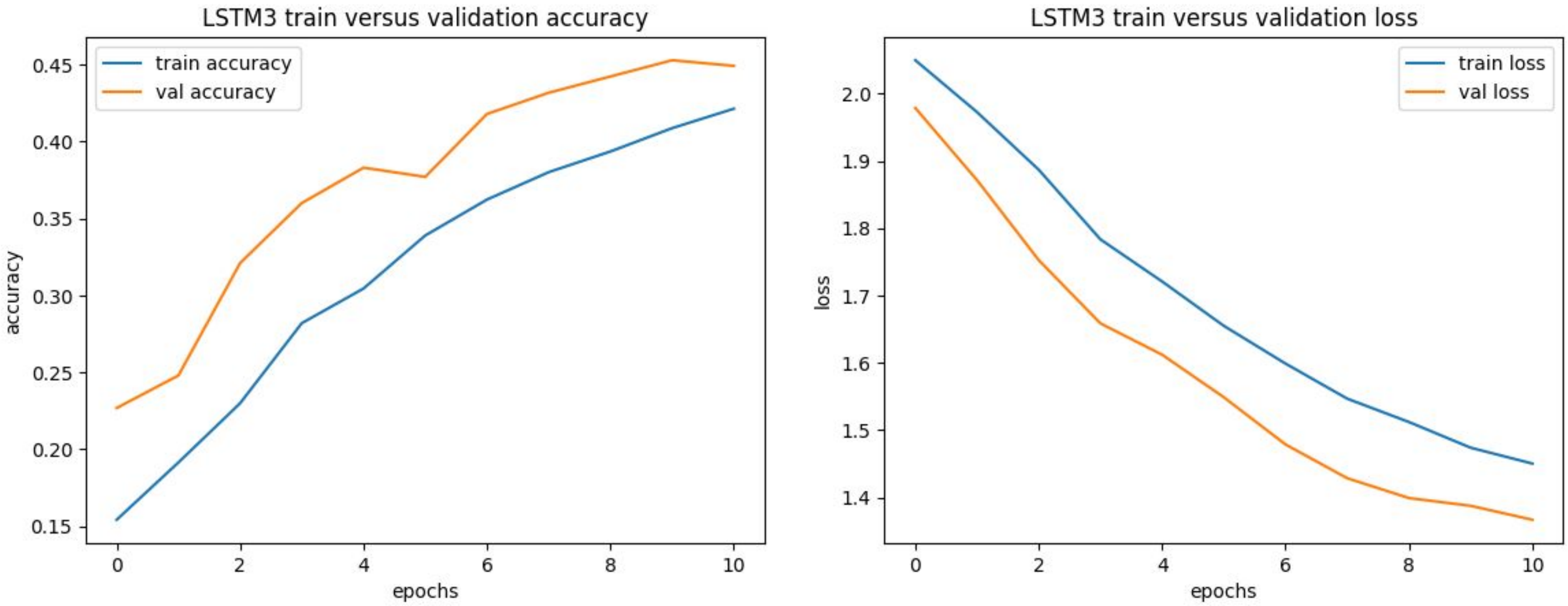
Performance



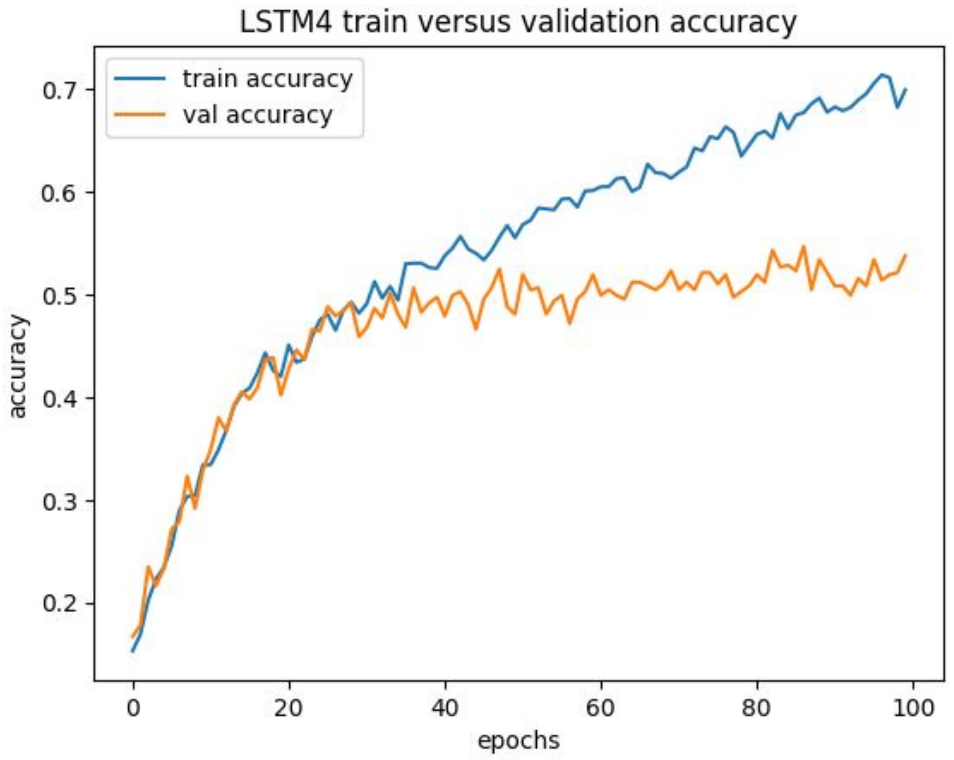


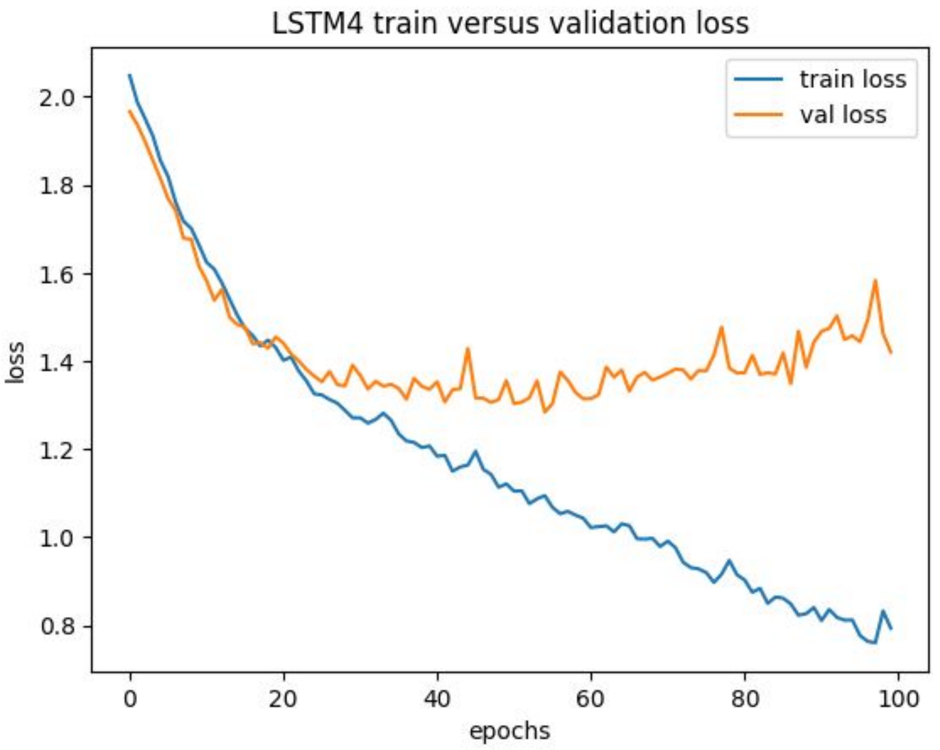
Performance

Performance



Performance





Outcomes

- Best model:

  model_LSTM, 'first_lstm_model.h5'

- Best score:

  accuracy: 0.6710 - loss: 0.8293 - val_accuracy: 0.6022 - val_loss: 1.1673

Future work:

- address class imbalances
- continue to tune and iterate or at least identify performance decline.
- Access other possible standardisations techniques, feature extraction and tuning methods.
- Utilize pretrained models such as Whisper, WavLM, and Wav2Vec 2.0, which can be fine-tuned for SER tasks.

Sources: