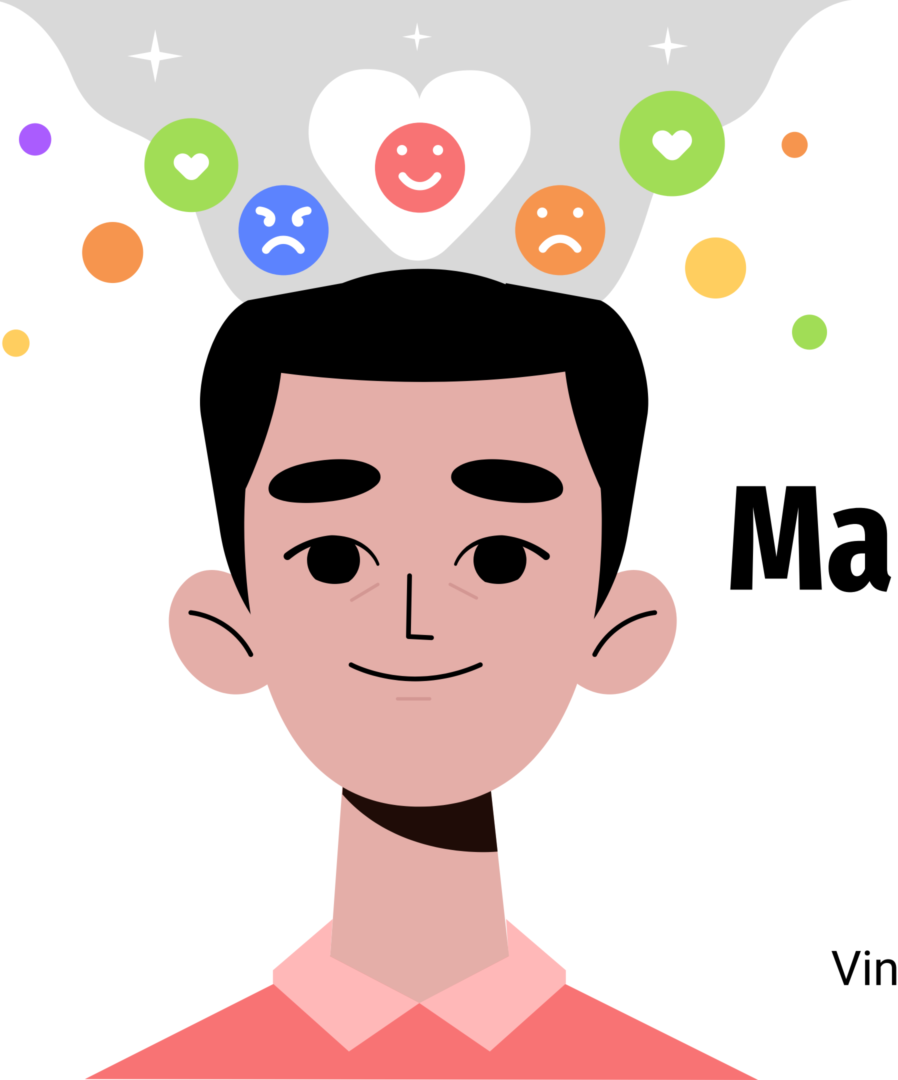


Classifying Emotions with Machine Learning

Team 26

Cooper Bosch, Conrad Hock,
Vincent Hock, Tomas Arevalo, Jake Pappo

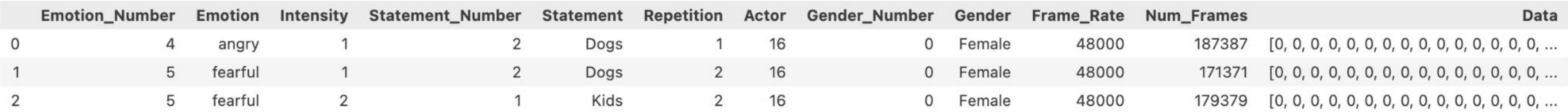
Problem Statement

to develop a sentiment analysis classifier capable of accurately identifying the emotional state of a speaker based on an audio clip of their speech

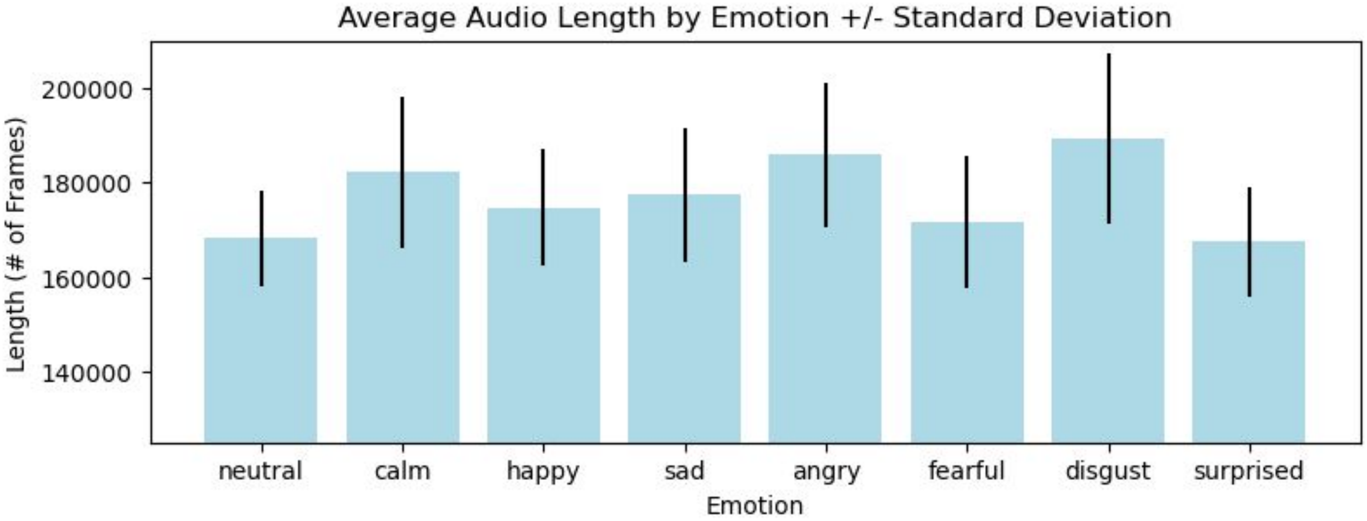
Dataset Description

- Sourced from Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
- 24 voice actors tasked with recording phrases with a specific emotional mood
- About 1500 recordings total containing neutral voicings, and 7 different emotions:
  - Calm
  - Happy
  - Sad
  - Angry
  - Fearful
  - Disgusted
  - Surprised

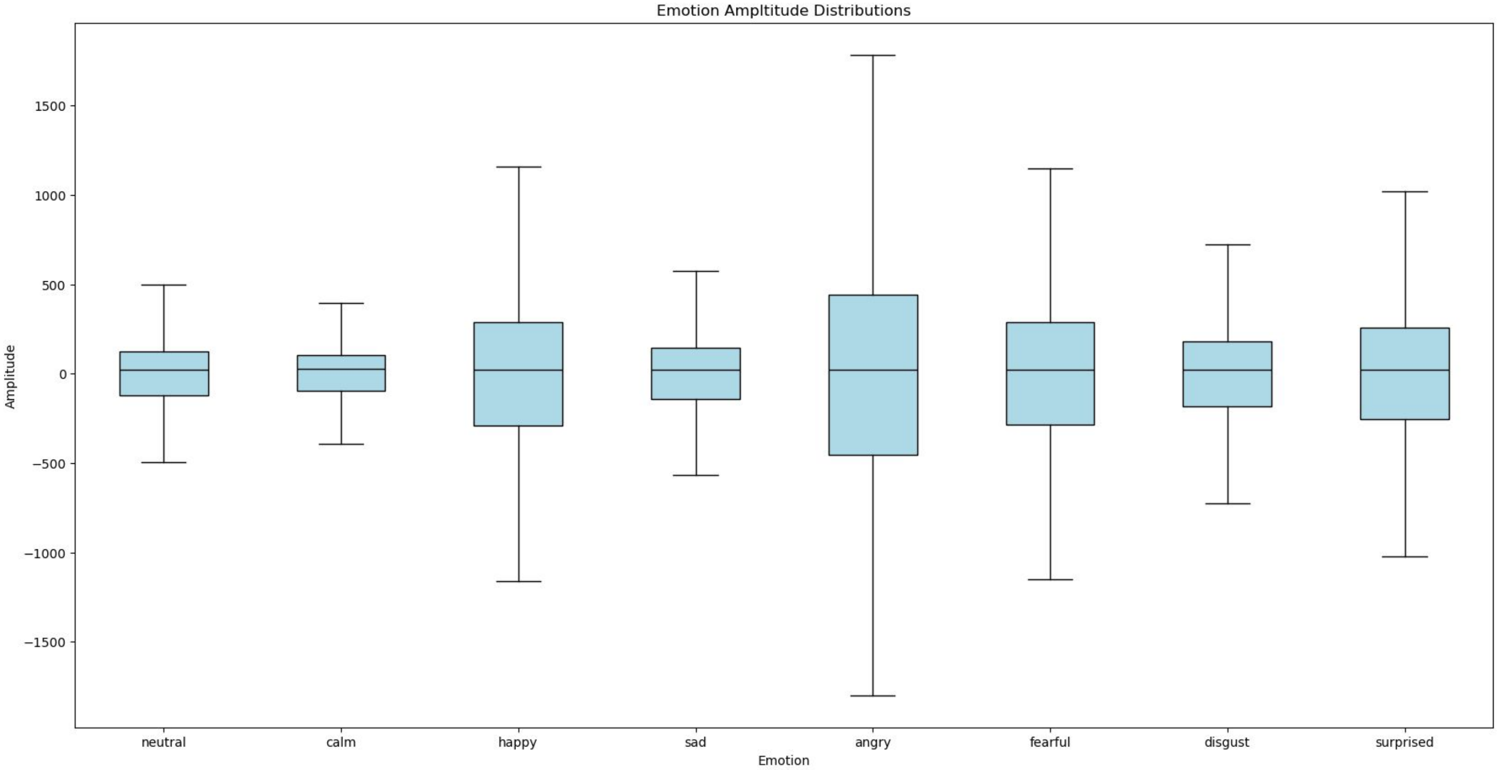
EDA



	Emotion_Number	Emotion	Intensity	Statement_Number	Statement	Repetition	Actor	Gender_Number	Gender	Frame_Rate	Num_Frames	Data
0	4	angry	1	2	Dogs	1	16	0	Female	48000	187387	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
1	5	fearful	1	2	Dogs	2	16	0	Female	48000	171371	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
2	5	fearful	2	1	Kids	2	16	0	Female	48000	179379	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...



EDA

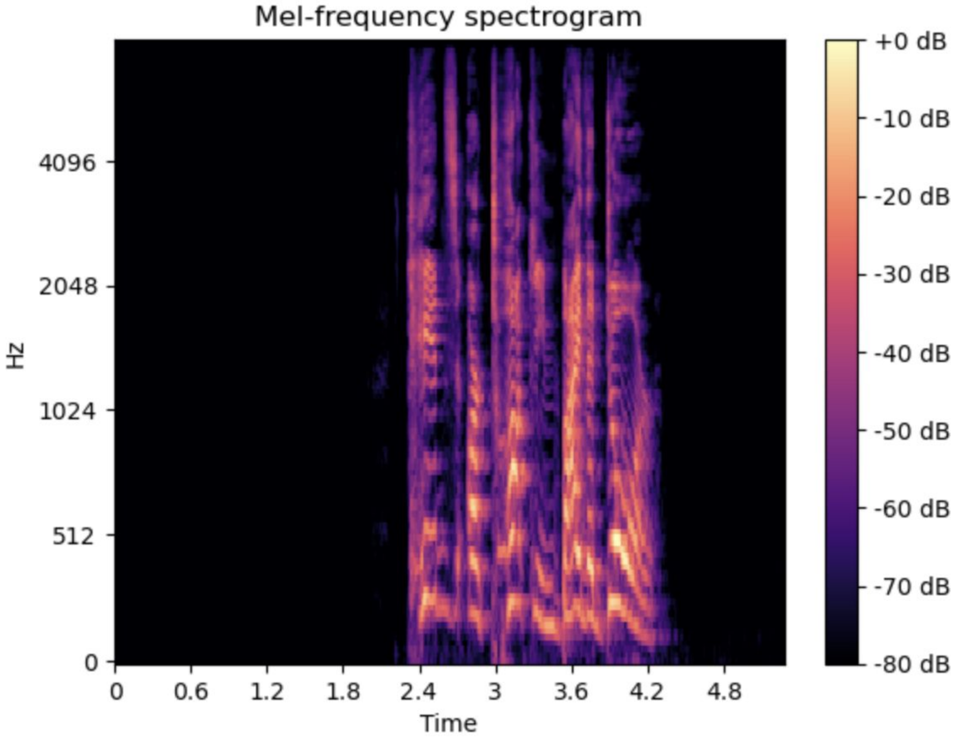

Emotion Ampltitude Distributions
Amplitude
1500
1000
500
0
−500
−1000
−1500
neutral
calm
happy
sad
angry
fearful
disgust
surprised
Emotion

Mel Spectrogram

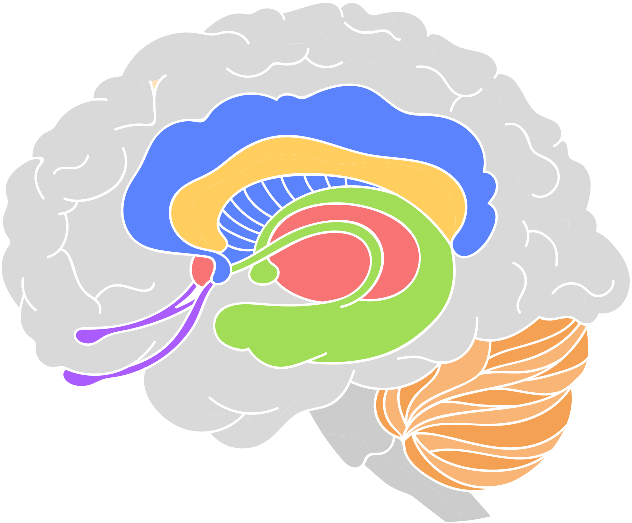
Important preprocessing step with 2 main components:

1. Run FFT on small windows of signal to convert to pitch data

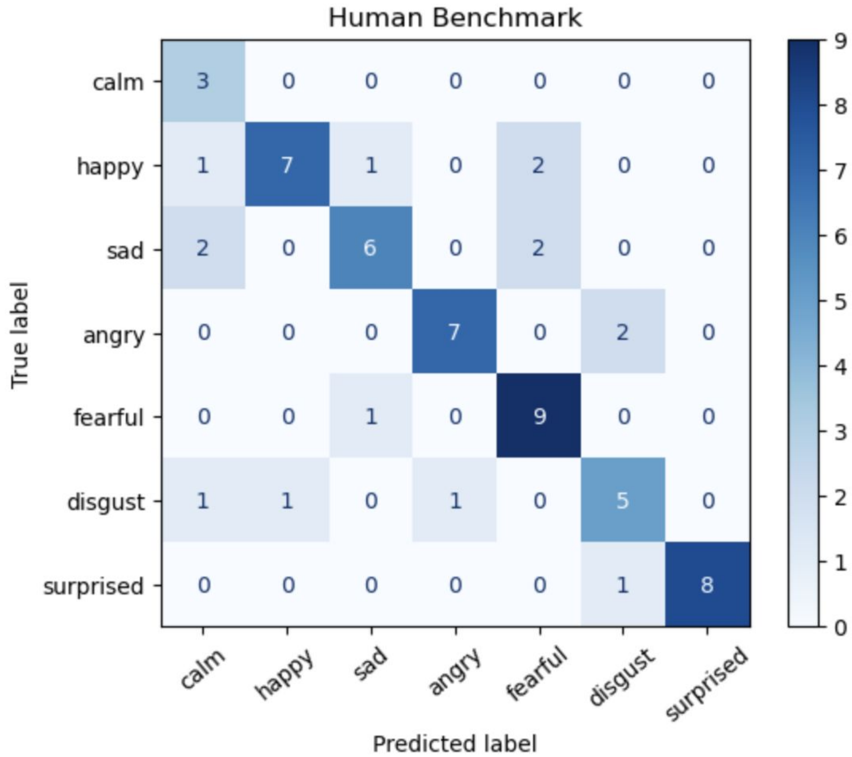
2. Transform pitches using mel scale, which is specifically designed to sound linear to humans



Model Implementation and Analysis



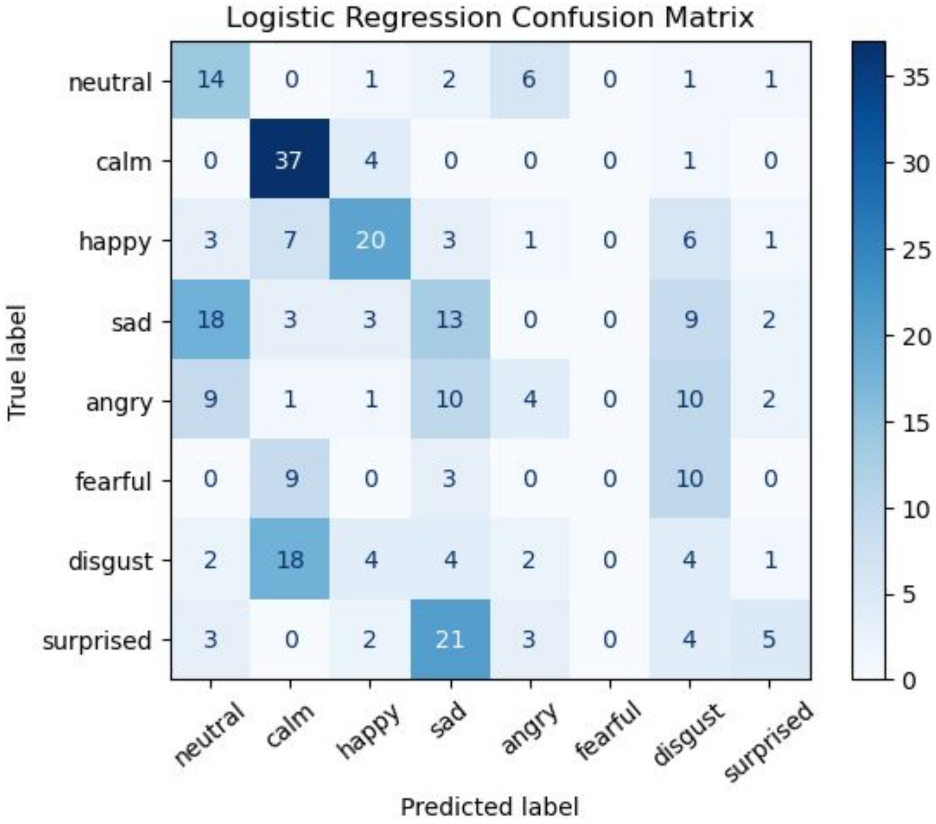
Human Benchmark



	calm	happy	sad	angry	fearful	disgust	surprised
calm	3	0	0	0	0	0	0
happy	1	7	1	0	2	0	0
sad	2	0	6	0	2	0	0
angry	0	0	0	7	0	2	0
fearful	0	0	1	0	9	0	0
disgust	1	1	0	1	0	5	0
surprised	0	0	0	0	0	1	8

- 75% accuracy
- Worst at classifying sadness

Logistic Regression - Baseline Model



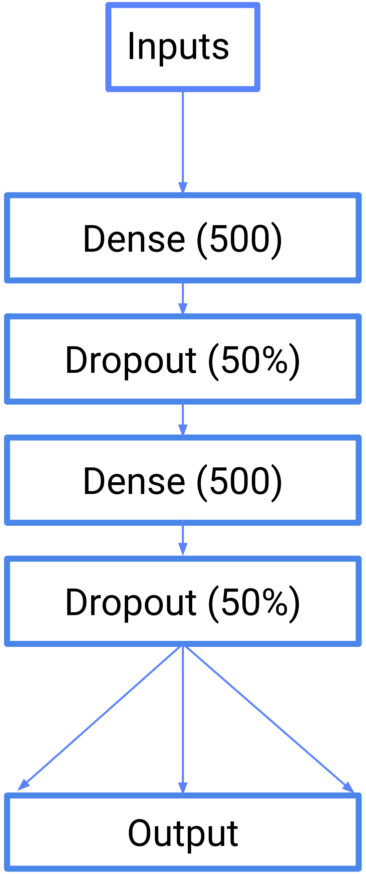
Logistic Regression Confusion Matrix

True label \ Predicted label	neutral	calm	happy	sad	angry	fearful	disgust	surprised
neutral	14	0	1	2	6	0	1	1
calm	0	37	4	0	0	0	1	0
happy	3	7	20	3	1	0	6	1
sad	18	3	3	13	0	0	9	2
angry	9	1	1	10	4	0	10	2
fearful	0	9	0	3	0	0	10	0
disgust	2	18	4	4	2	0	4	1
surprised	3	0	2	21	3	0	4	5

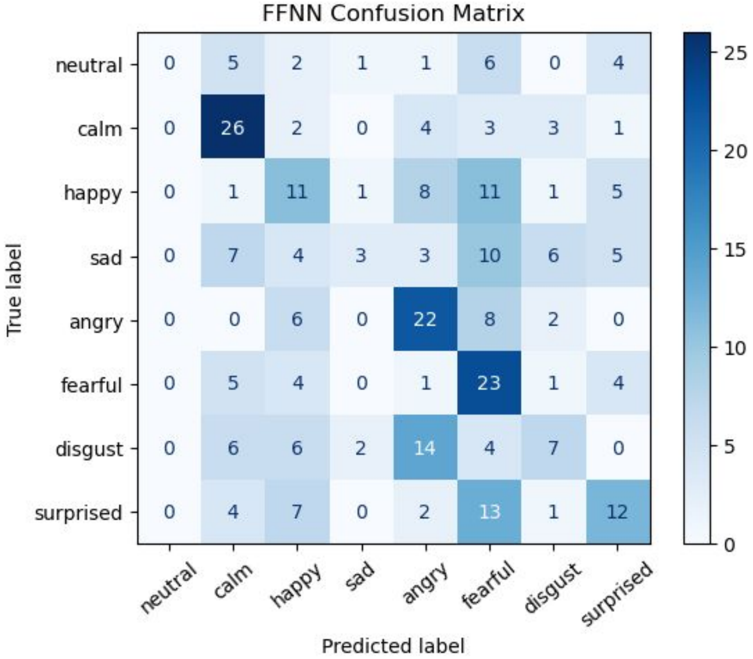
- 34% Accuracy
- Takes as input:
    - Standard deviation of amplitudes
    - Length of clip
- Best at classifying calm emotion, which has lowest STD

FFNN

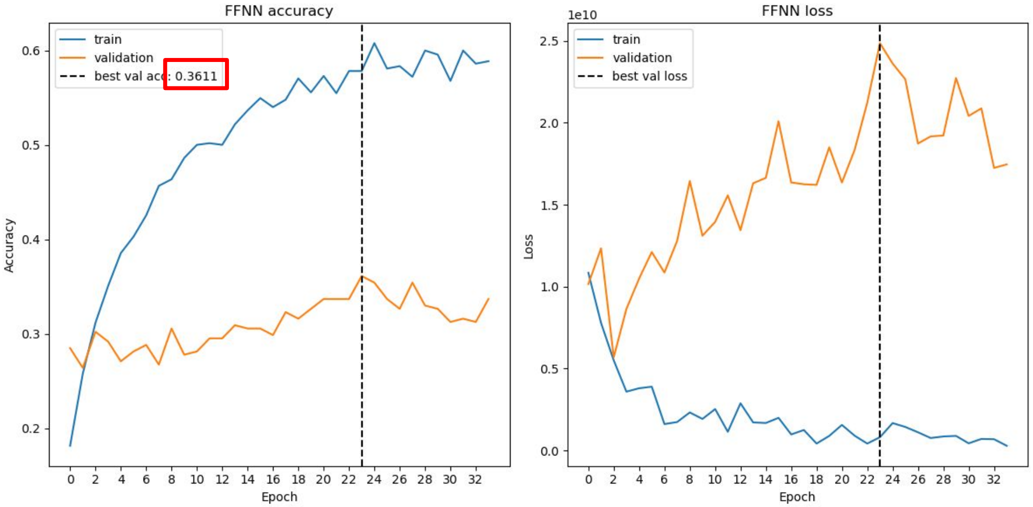
Parameters: 31,935,008
Train time: ~4 min
Epochs: 34
Val Acc: 36.1%



FFNN

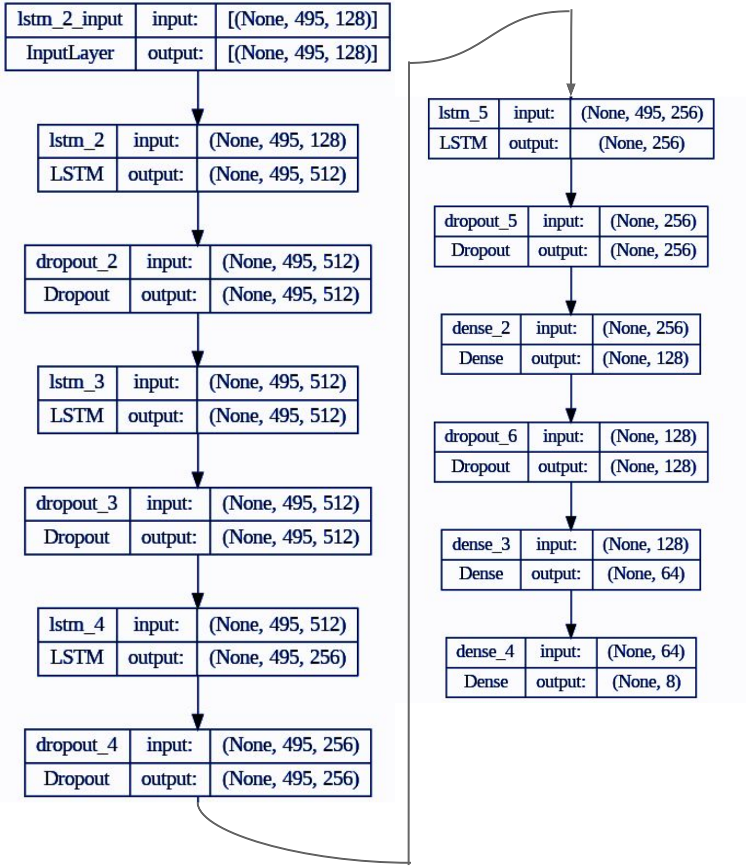


FFNN Confusion Matrix
True label
Predicted label
neutral 0 5 2 1 1 6 0 4
calm 0 26 2 0 4 3 3 1
happy 0 1 11 1 8 11 1 5
sad 0 7 4 3 3 10 6 5
angry 0 0 6 0 22 8 2 0
fearful 0 5 4 0 1 23 1 4
disgust 0 6 6 2 14 4 7 0
surprised 0 4 7 0 2 13 1 12
neutral calm happy sad angry fearful disgust surprised
25
20
15
10
5
0



FFNN accuracy
train
validation
best val acc: 0.3611
Accuracy
0.6
0.5
0.4
0.3
0.2
Epoch
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
FFNN loss
1e10
train
validation
best val loss
Loss
2.5
2.0
1.5
1.0
0.5
0.0
Epoch
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32

LSTM



lstm_2_input	input:	[(None, 495, 128)]
InputLayer	output:	[(None, 495, 128)]

lstm_2	input:	(None, 495, 128)
LSTM	output:	(None, 495, 512)

dropout_2	input:	(None, 495, 512)
Dropout	output:	(None, 495, 512)

lstm_3	input:	(None, 495, 512)
LSTM	output:	(None, 495, 512)

dropout_3	input:	(None, 495, 512)
Dropout	output:	(None, 495, 512)

lstm_4	input:	(None, 495, 512)
LSTM	output:	(None, 495, 256)

dropout_4	input:	(None, 495, 256)
Dropout	output:	(None, 495, 256)

lstm_5	input:	(None, 495, 256)
LSTM	output:	(None, 256)

dropout_5	input:	(None, 256)
Dropout	output:	(None, 256)

dense_2	input:	(None, 256)
Dense	output:	(None, 128)

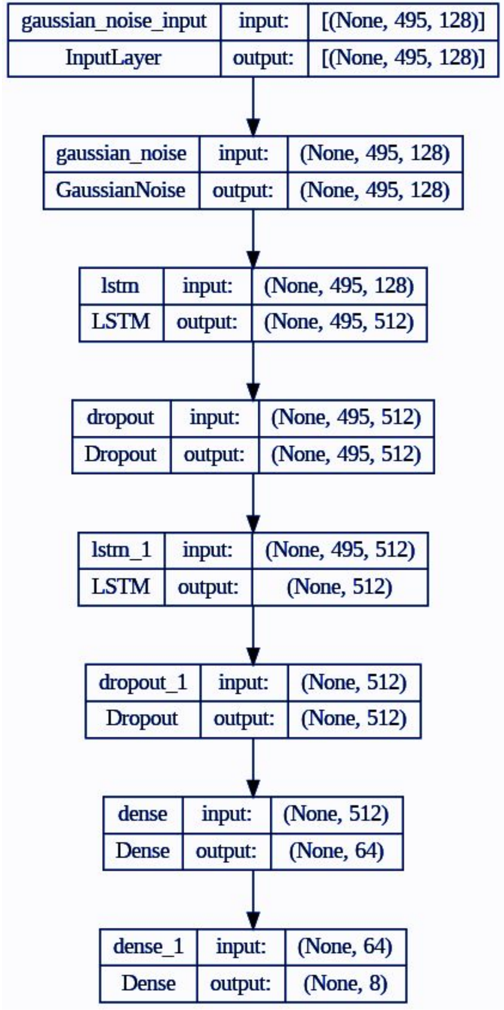
dropout_6	input:	(None, 128)
Dropout	output:	(None, 128)

dense_3	input:	(None, 128)
Dense	output:	(None, 64)

dense_4	input:	(None, 64)
Dense	output:	(None, 8)

SIMPLE
is
better

Parameters: 3,445,320
Train time: ~13.5 min
Epochs: 100
Val Acc: 55.9%



gaussian_noise_input	input:	[(None, 495, 128)]
InputLayer	output:	[(None, 495, 128)]

gaussian_noise	input:	(None, 495, 128)
GaussianNoise	output:	(None, 495, 128)

lstm	input:	(None, 495, 128)
LSTM	output:	(None, 495, 512)

dropout	input:	(None, 495, 512)
Dropout	output:	(None, 495, 512)

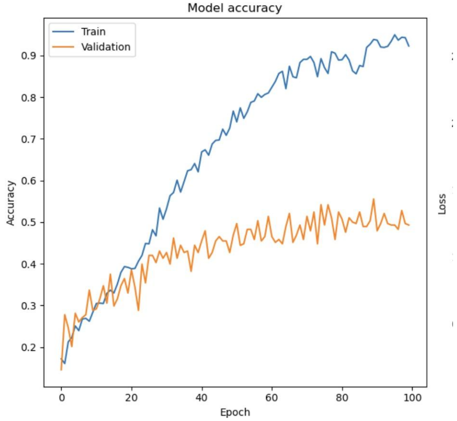
lstm_1	input:	(None, 495, 512)
LSTM	output:	(None, 512)

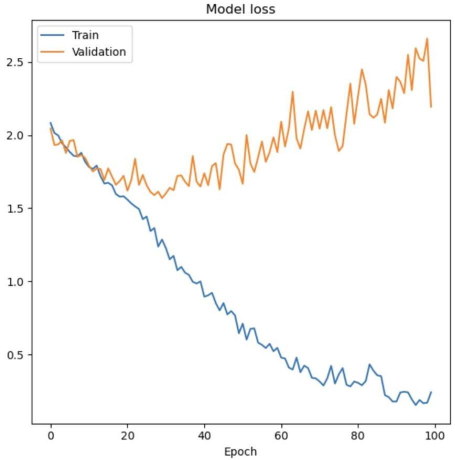
dropout_1	input:	(None, 512)
Dropout	output:	(None, 512)

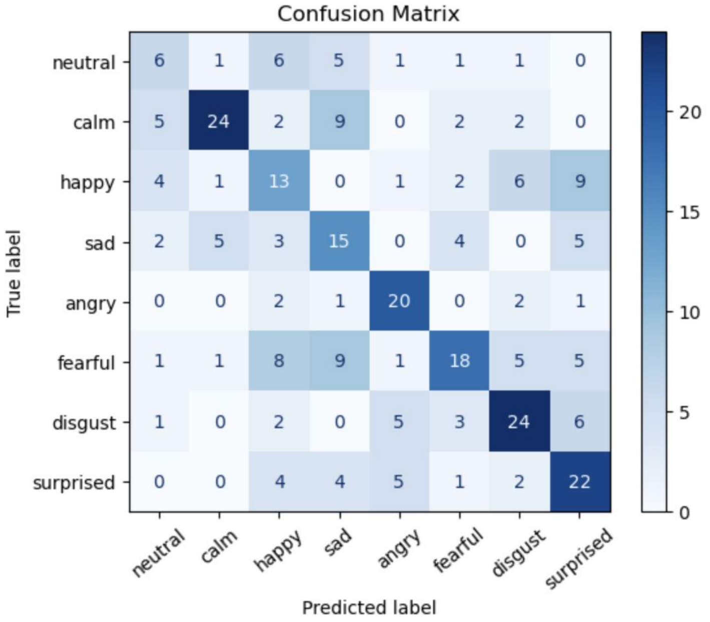
dense	input:	(None, 512)
Dense	output:	(None, 64)

dense_1	input:	(None, 64)
Dense	output:	(None, 8)

LSTM







	neutral	calm	happy	sad	angry	fearful	disgust	surprised
neutral	6	1	6	5	1	1	1	0
calm	5	24	2	9	0	2	2	0
happy	4	1	13	0	1	2	6	9
sad	2	5	3	15	0	4	0	5
angry	0	0	2	1	20	0	2	1
fearful	1	1	8	9	1	18	5	5
disgust	1	0	2	0	5	3	24	6
surprised	0	0	4	4	5	1	2	22

Ranking accuracy

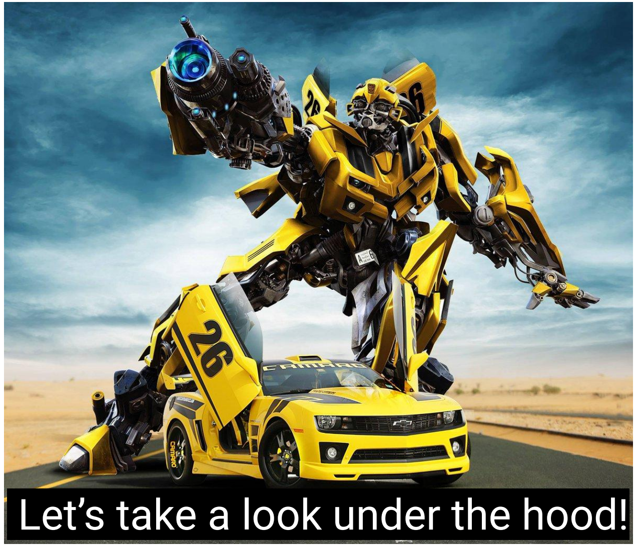
1. Angry: 0.769
2. Disgust: 0.585
3. Surprised: 0.579
4. Calm: 0.545
5. Sad: 0.441
6. Happy: 0.361
7. Fearful: 0.375
8. Neutral: 0.286

Ranking precision:

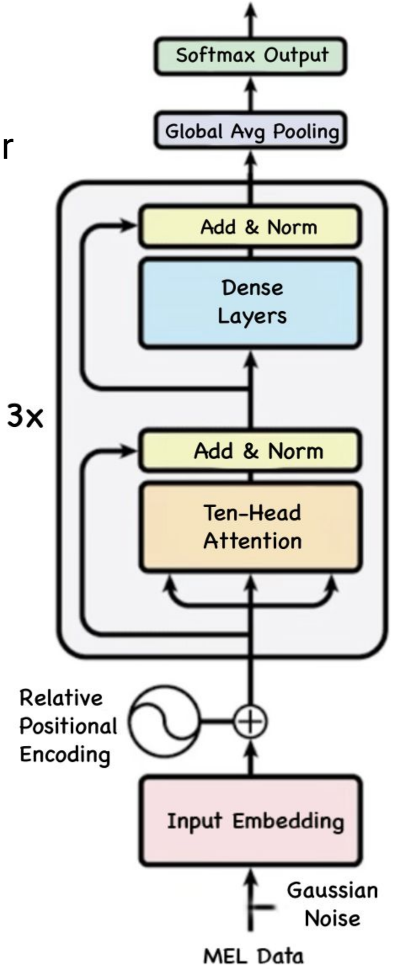
1. Calm: 0.889
2. Angry: 0.606
3. Fearful: 0.581
4. Disgust: 0.632
5. Sad: 0.349
6. Neutral: 0.286
7. Surprised: 0.458
8. Happy: 0.325

Transformer

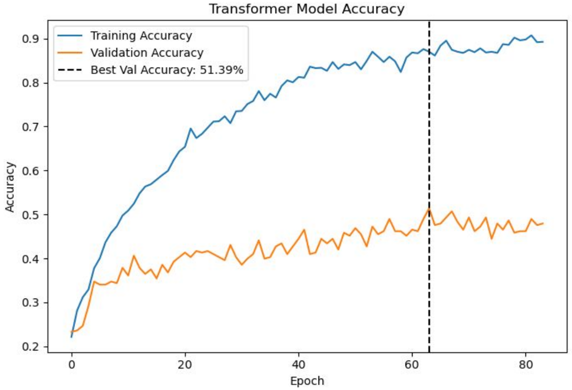
Specific Architecture: Bidirectional encoder

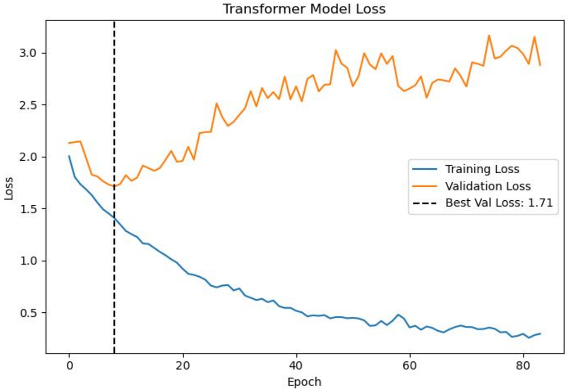

Let’s take a look under the hood!

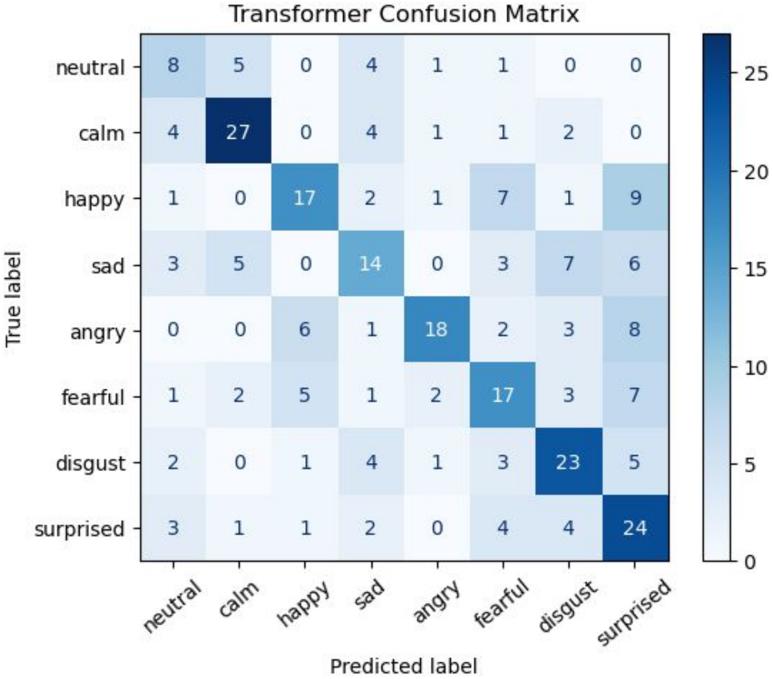
Parameters: 376,834
Train time: ~7 minutes
Epochs: 83
Embed Dimen: 50



Transformer





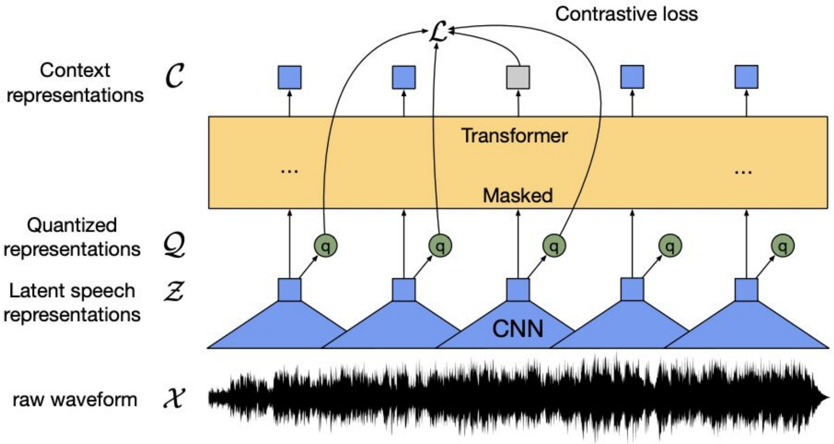


	neutral	calm	happy	sad	angry	fearful	disgust	surprised
neutral	8	5	0	4	1	1	0	0
calm	4	27	0	4	1	1	2	0
happy	1	0	17	2	1	7	1	9
sad	3	5	0	14	0	3	7	6
angry	0	0	6	1	18	2	3	8
fearful	1	2	5	1	2	17	3	7
disgust	2	0	1	4	1	3	23	5
surprised	3	1	1	2	0	4	4	24

- **Validation Accuracy:** 51.4%
- “Calm” accuracy: 69.2%
- “Sad” accuracy: 36.8%
- “Surprised” precision: 40.7% (35 FPs)
- Relatively distributed error

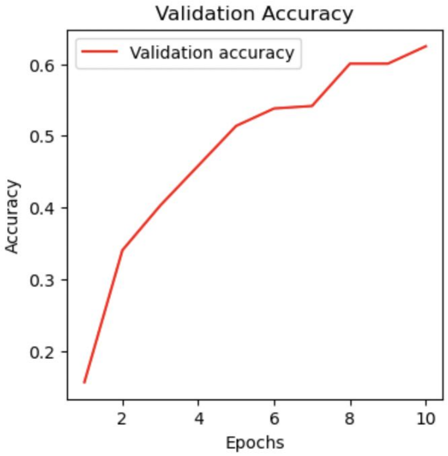
SOTA Model: Wav2Vec2

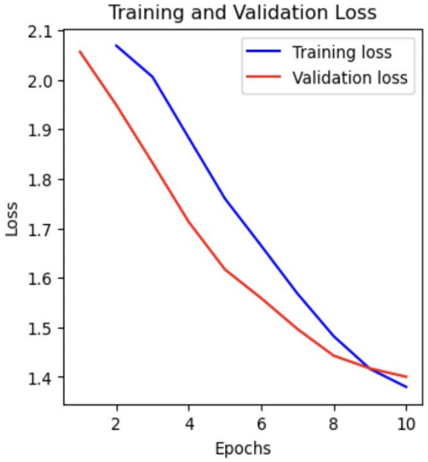
Parameters: 94,570,632
Train time: ~32 minutes
Epochs: 10
Val Acc: 62.5%

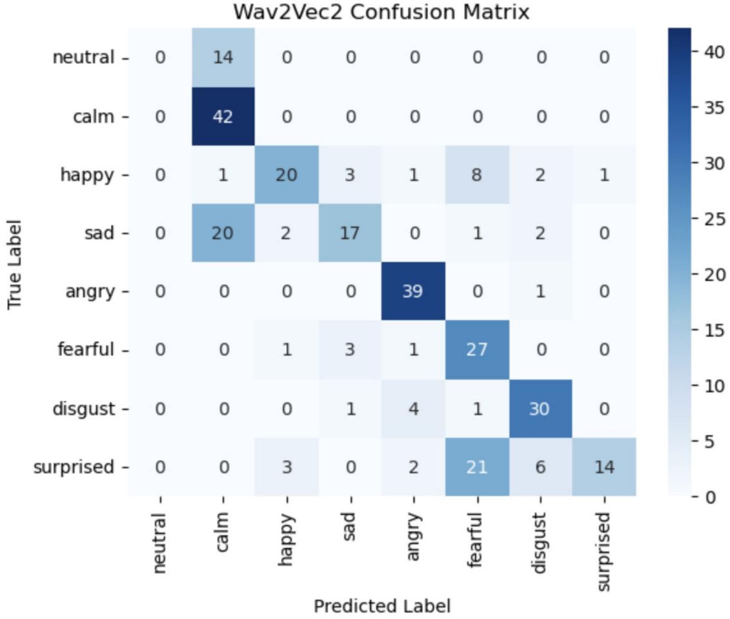


Baevski et al., 2020

SOTA Model: Wav2Vec2



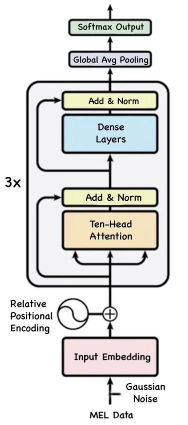




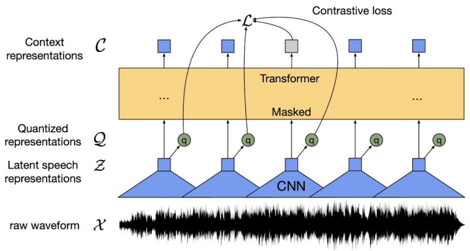
	neutral	calm	happy	sad	angry	fearful	disgust	surprised
neutral	0	14	0	0	0	0	0	0
calm	0	42	0	0	0	0	0	0
happy	0	1	20	3	1	8	2	1
sad	0	20	2	17	0	1	2	0
angry	0	0	0	0	39	0	1	0
fearful	0	0	1	3	1	27	0	0
disgust	0	0	0	1	4	1	30	0
surprised	0	0	3	0	2	21	6	14

Conclusion

Fundamentally difficult problem

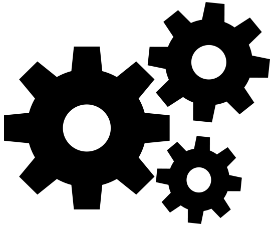


Don’t need a ton of parameters for decent accuracy with transformer

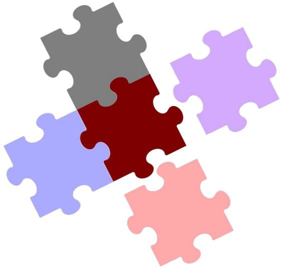

Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
...
Masked
Quantized representations $\mathcal{Q}$
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

Transfer learning is usually the way to go

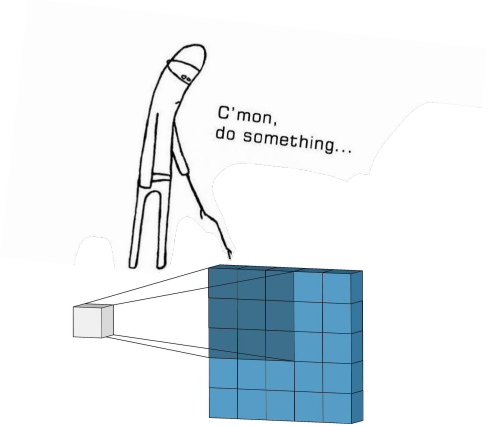
Future Steps



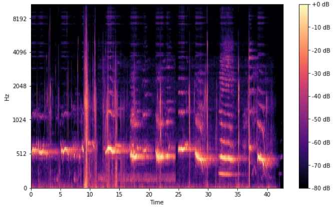
Tweak architectures & tune hyperparameters



Combining emotions or removing neutrals



Further exploration of CNNs



Mel spectrogram hyperparameters

Throw more compute at Wave2Vec2

MORE DATA!

Thanks!