UNIVERSITY OF CALGARY

Multi-Label Speech Emotion Recognition Using 2D CNNs

Using machine learning to identify multiple emotions in speech

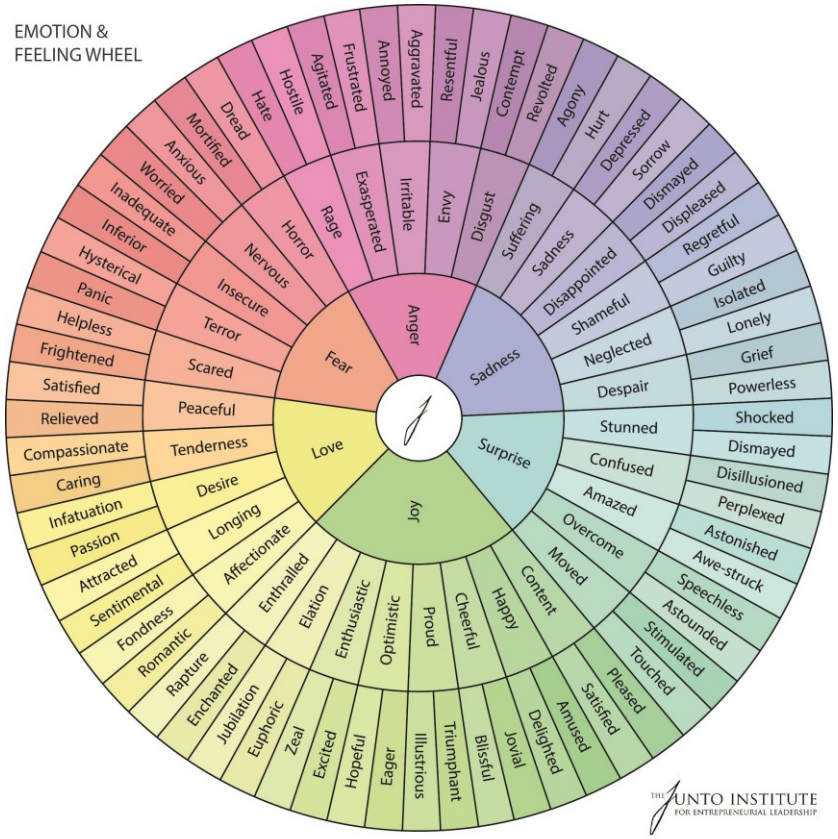
Brian Pho
BSc Software Engineering
Department of Electrical and Computer Engineering

Friday, August 23, 2019

Table of Contents

Table of Contents

Literature Review – Big Picture



- Affective computing
  - "The study and development of systems and devices that can recognize, interpret, process, and simulate human affects."
- Affect
  - "The experiencing of feeling or emotion."
- Affective computing has applications in voice assistants, kiosks, healthcare, education, and video games.

Literature Review – Speech Emotion Recognition

Speech Recognition Using Deep Neural Networks: A Systematic Review

ALI BOU NASSIF[1], ISMAIL SHAHIN[1], IMTINAN ATTILI[1], MOHAMMAD AZZEH[2], AND KHALED SHAALAN[3]
[1]Department of Electrical and Computer Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates
[2]Department of Software Engineering, Applied Science Private University, Amman 163, Jordan
[3]Faculty of Engineering and IT, The British University in Dubai, Dubai 345015, United Arab Emirates
Corresponding author: Ali Bou Nassif (anassif@sharjah.ac.ae)

This work was supported by the University of Sharjah through the Competitive Research Project "Emotion Recognition in each of Stressful and Emotional Talking Environments Using Artificial Models" under Grant 1602040348-P. The work of M. Azzeh was supported by the Applied Science Private University, Amman, Jordan.

ABSTRACT Over the past decades, a tremendous amount of research has been done on the use of machine learning for speech processing applications, especially speech recognition. However, in the past few years, research has focused on utilizing deep learning for speech-related applications. This new area of machine learning has yielded far better results when compared to others in a variety of applications including speech, and thus became a very attractive area of research. This paper provides a thorough examination of the different studies that have been conducted since 2006, when deep learning first arose as a new area of machine learning, for speech applications. A thorough statistical analysis is provided in this review which was conducted by extracting specific information from 174 papers published between the years 2006 and 2018. The results provided in this paper shed light on the trends of research in this area as well as bring focus to new research topics.

INDEX TERMS Speech recognition, deep neural network, systematic review.

I. INTRODUCTION

Since the last decade, deep learning has arisen as a new attractive area of machine learning, and ever since has been examined and utilized in a range of different research topics [1]. Deep learning consists of a multiple of machine learning algorithms fed with inputs in the form of multiple layered models. These models are usually neural networks consisting of different levels of non-linear operations. The machine learning algorithms attempt to learn from these deep neural networks by extracting specific features and information [2]. Prior to 2006, searching deep architecture inputs was not a predictable straight forward task; however, the development of deep learning algorithms helped resolve this issue and simplified the process of searching the parameter space of deep architectures [2]. Deep learning models can also operate as a greedy layerwise unsupervised pre-training. This means that it will learn hierarchy from extracted features from each layer at a time. Feature learning is achieved by training each layer with an unsupervised learning algorithm, which takes the features extracted from the previous layer and uses it as an input for the next layer. Thus, feature learning will attempt to learn the transformation of the previously learned features at each new layer. Each iteration feature learning adds one layer of weights to a deep neural network. The resulted layers with learned weights can eventually be loaded to initialize a deep supervised predictor [2], [3]. Using deep architectures has proven to be more efficient in representing non-linear functions in comparison to shallower architectures. Studies have shown that fewer parameters are required to represent a certain non-linear function in a deep architecture in comparison with the large number of parameters needed to represent the same function in a shallower architecture. This shows that deeper architectures are more efficient from a statistical point of view [2], [3].

Deep learning algorithms have been mostly used to further enhance the capabilities of computers so that it understands what humans can do, which includes speech recognition. Speech in particular, being the main method of communication among human beings, received much interest for the

The associate editor coordinating the review of this manuscript and approving it for publication was Malik Jahan Khan.

- The goal is recognizing which emotions are present in speech.
- Two main approaches
  - Feature engineering (pitch, log energy, MFCC)
  - Machine learning (CNN, LSTM, HMM)
- Machine learning has been the dominant approach for the last 5 years due to how accurate it has been.

Table of Contents

Summary of Results

UNIVERSITY OF CALGARY

- We extend upon current work by
  - Increasing the number of labels from single to multiple
    - E.g. From [“happy”] to [“happy”, “surprised”].
  - Increasing the number of databases the neural network is trained on
    - E.g. From one or two databases to four databases.
- We achieved an accuracy of 52.57% using an eight-layer convolutional neural network trained on four databases.

Table of Contents

Preprocessing – Emotions

- We considered seven emotions
  1. Neutral
  2. Anger
  3. Disgust
  4. Fear
  5. Happy
  6. Sad
  7. Surprised

Preprocessing – Databases

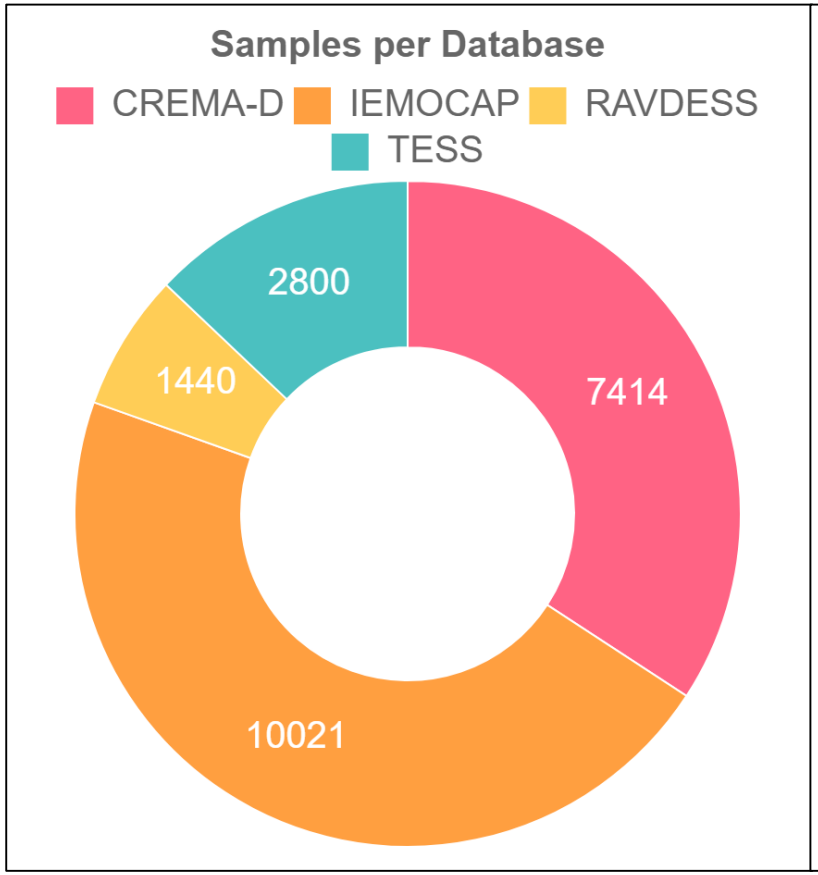
UNIVERSITY OF CALGARY

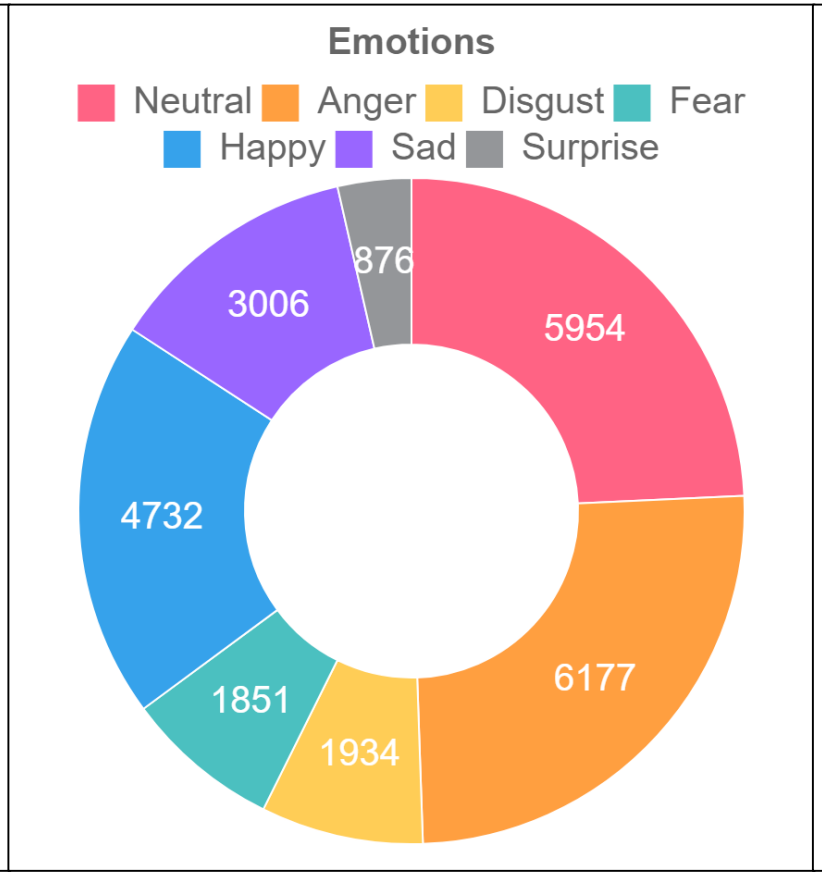
	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Calm	Excitement	Frustration	Amused	Sleepy	Bored
IEMOCAP	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
CREMA-D	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
TESS	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
RAVDESS	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
MSP-IMPROV	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
SAVEE	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Emo-DB	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓
EmoV-DB	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗

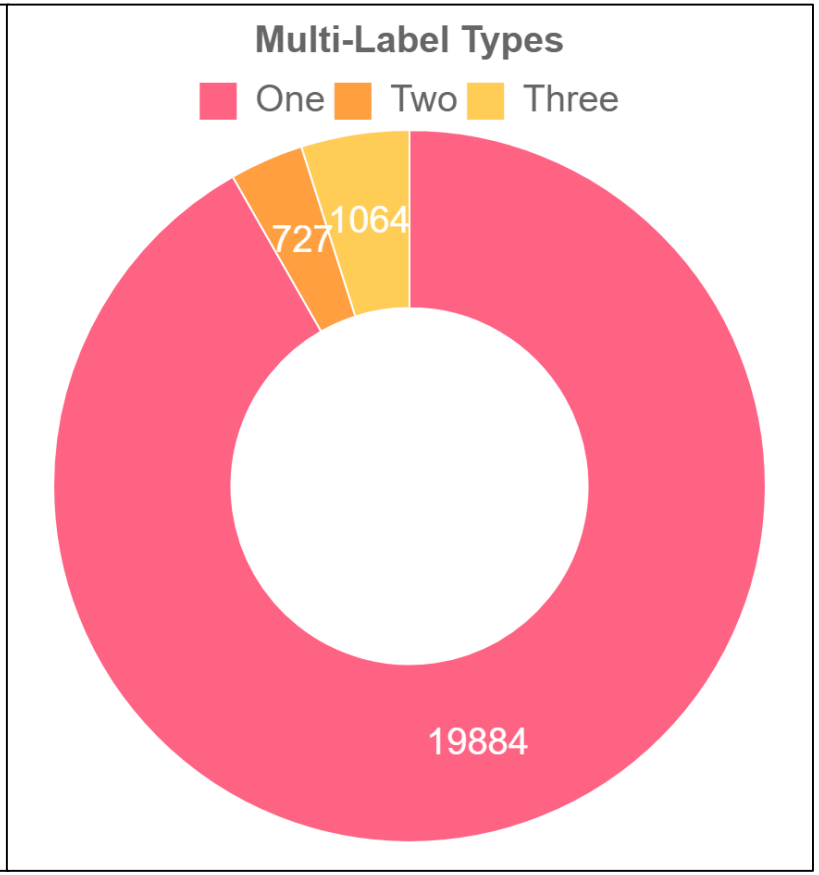
Preprocessing – Databases

- We used four databases
  - IEMOCAP
  - TESS
  - RAVDESS
  - CREMA-D
- We combined the databases and ended up with 21,675 samples (excluding four and five labels).

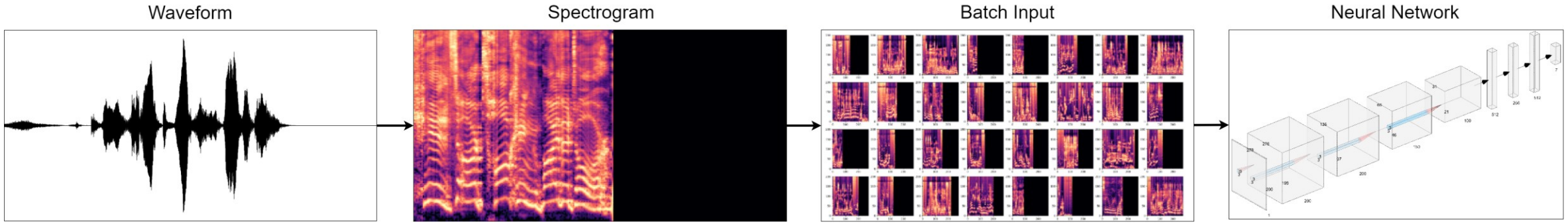
Preprocessing – Combined Database







Preprocessing – High Level System Overview



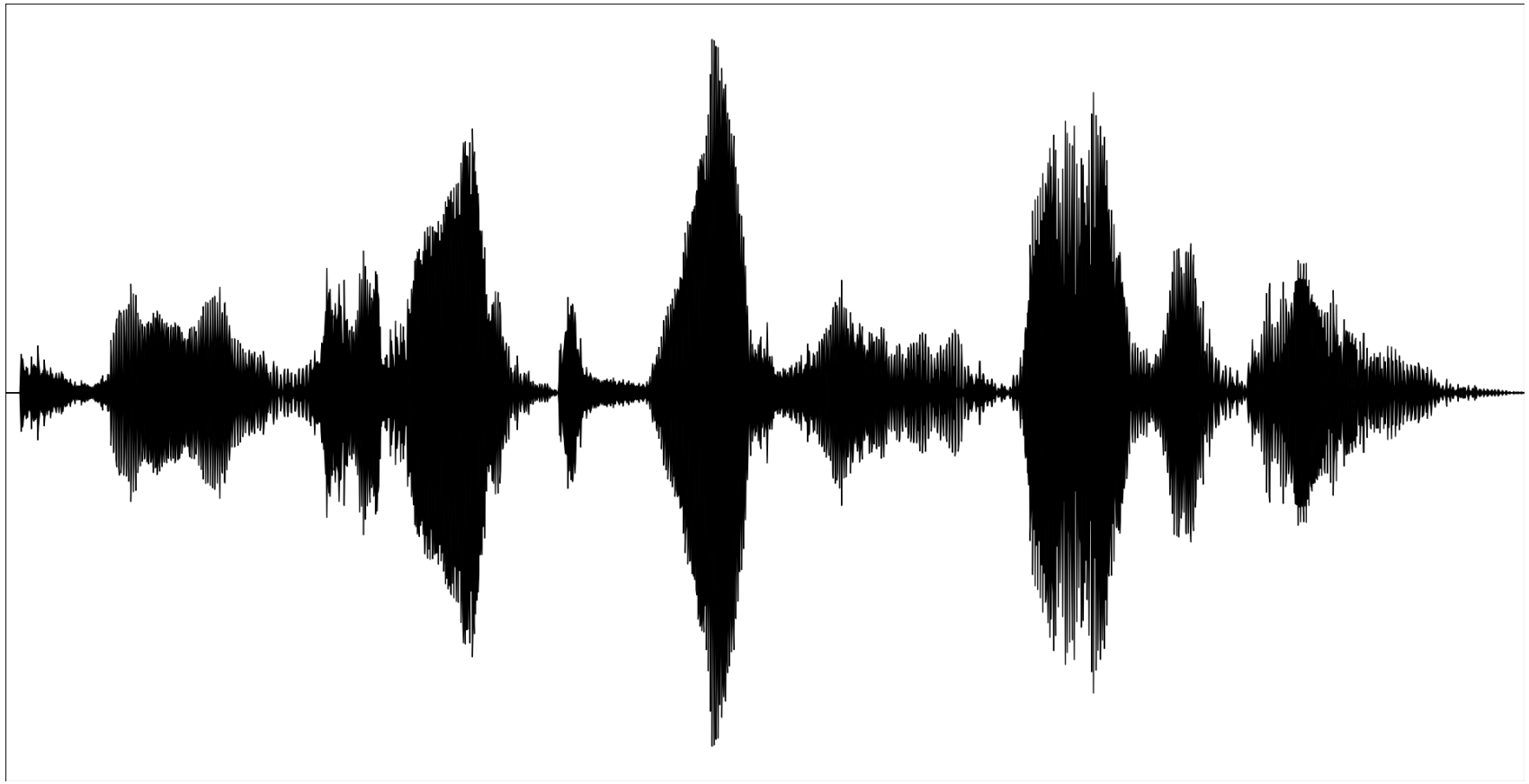
Preprocessing – Processing Samples

UNIVERSITY OF CALGARY

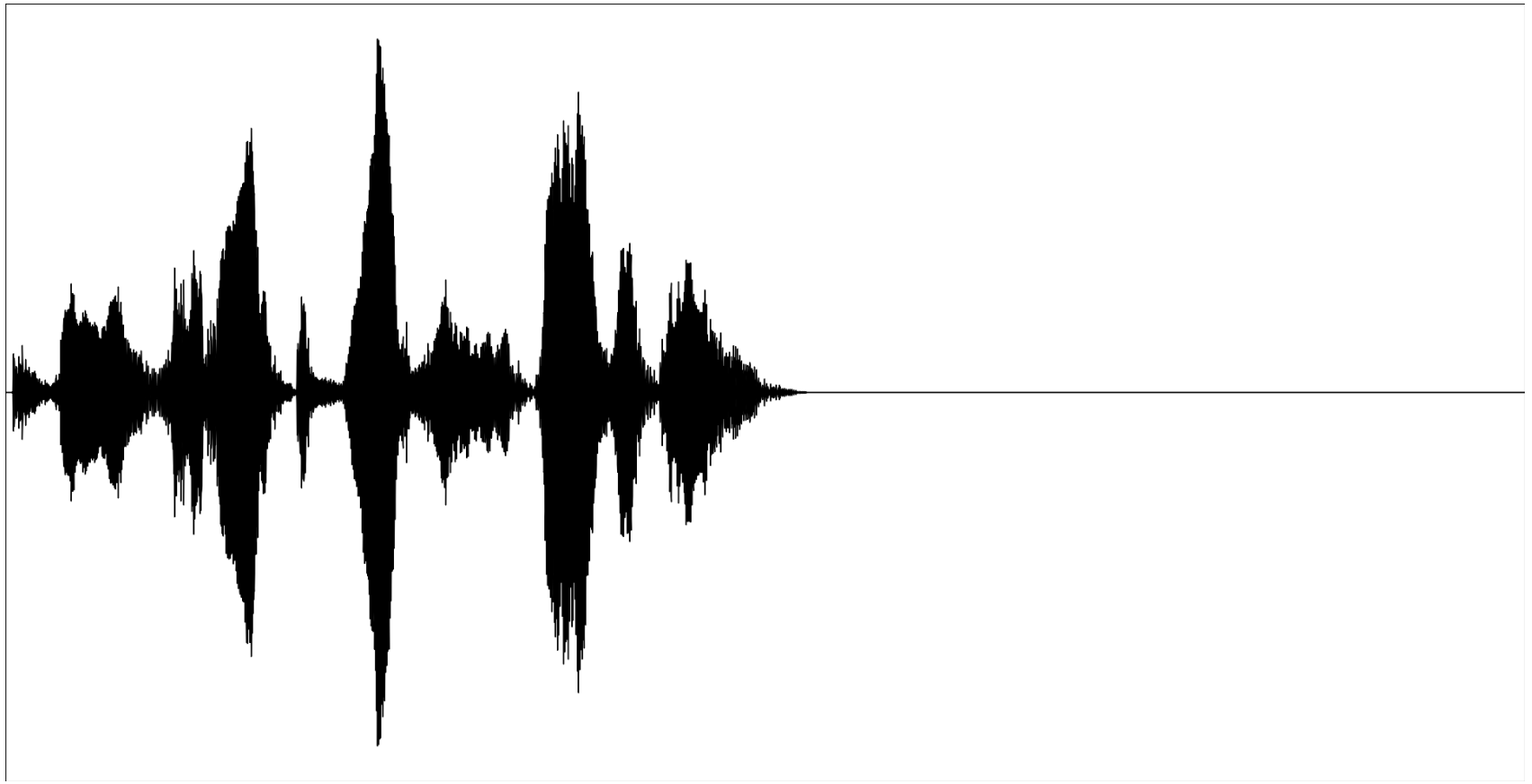
- Preprocessing steps for audio samples
  1. Raw waveform
  2. Cropped waveform
  3. Padded waveform
  4. Raw spectrogram
  5. Mel spectrogram
  6. Log-Mel spectrogram
  7. Scaled spectrogram

Preprocessing – Step 1. Raw waveform

Preprocessing – Step 2. Cropped waveform

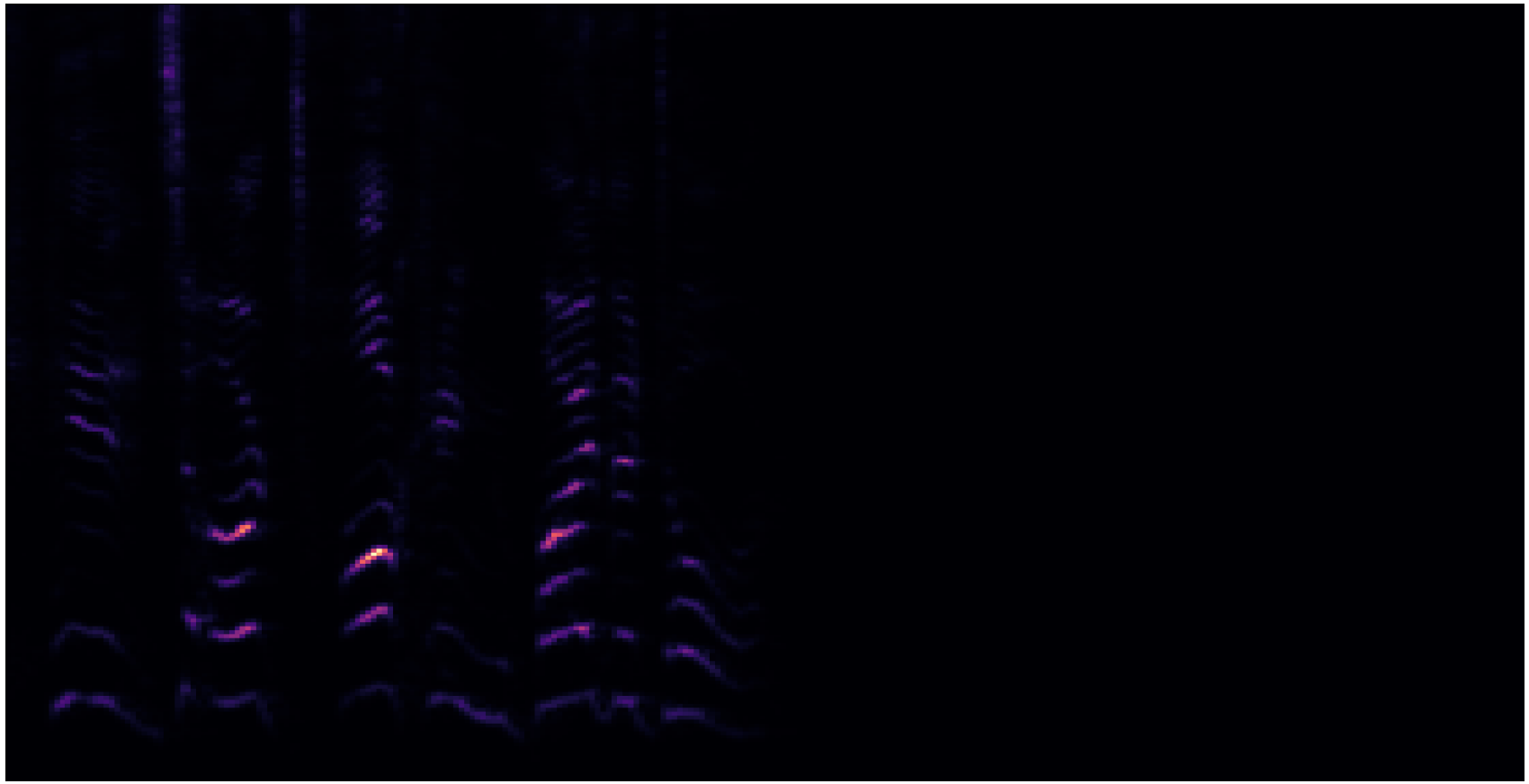


Preprocessing – Step 3. Padded waveform

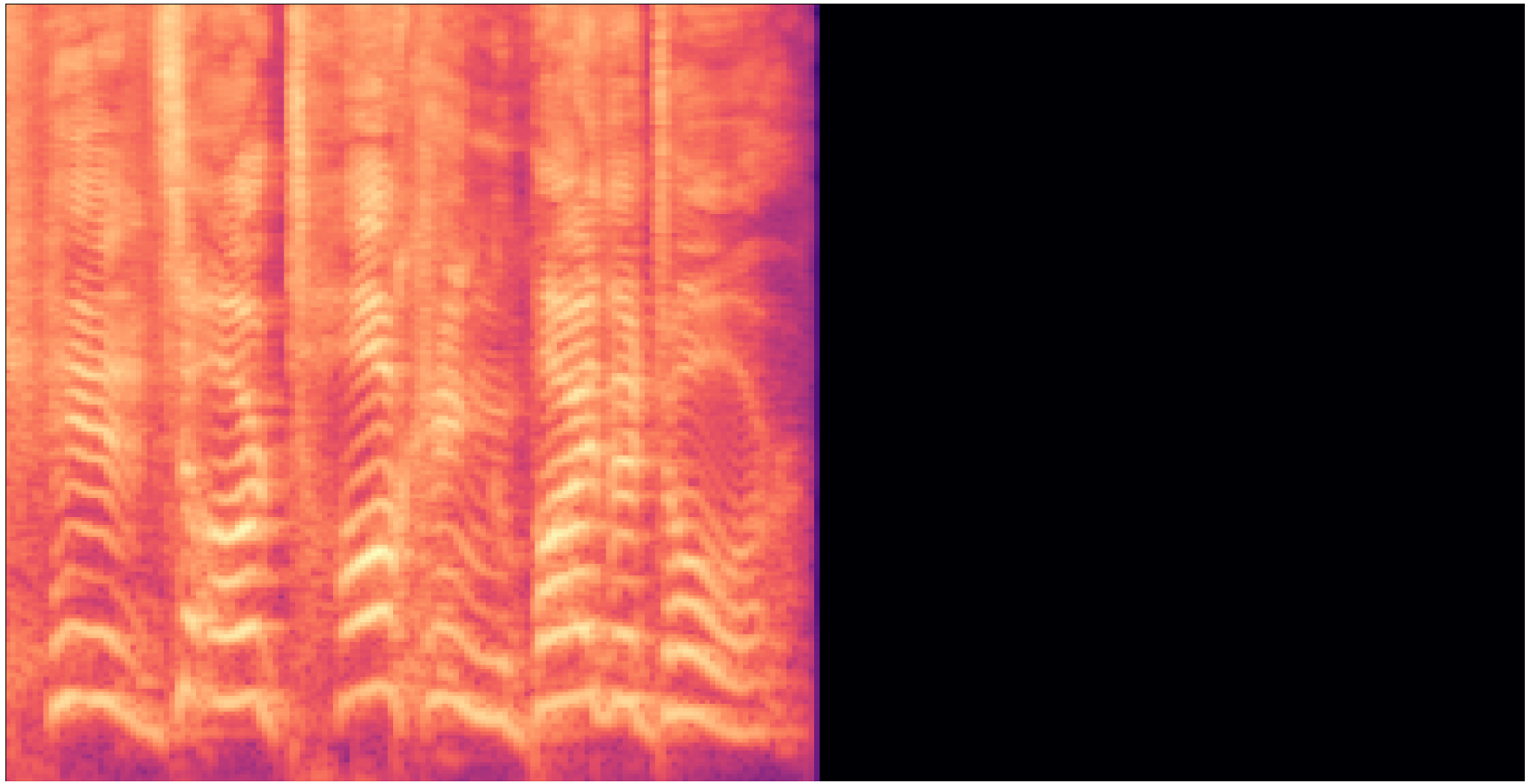


Preprocessing – Step 4. Raw spectrogram

Preprocessing – Step 5. Mel spectrogram



Preprocessing – Step 6. Log-Mel spectrogram



Preprocessing – Step 7. Scaled spectrogram

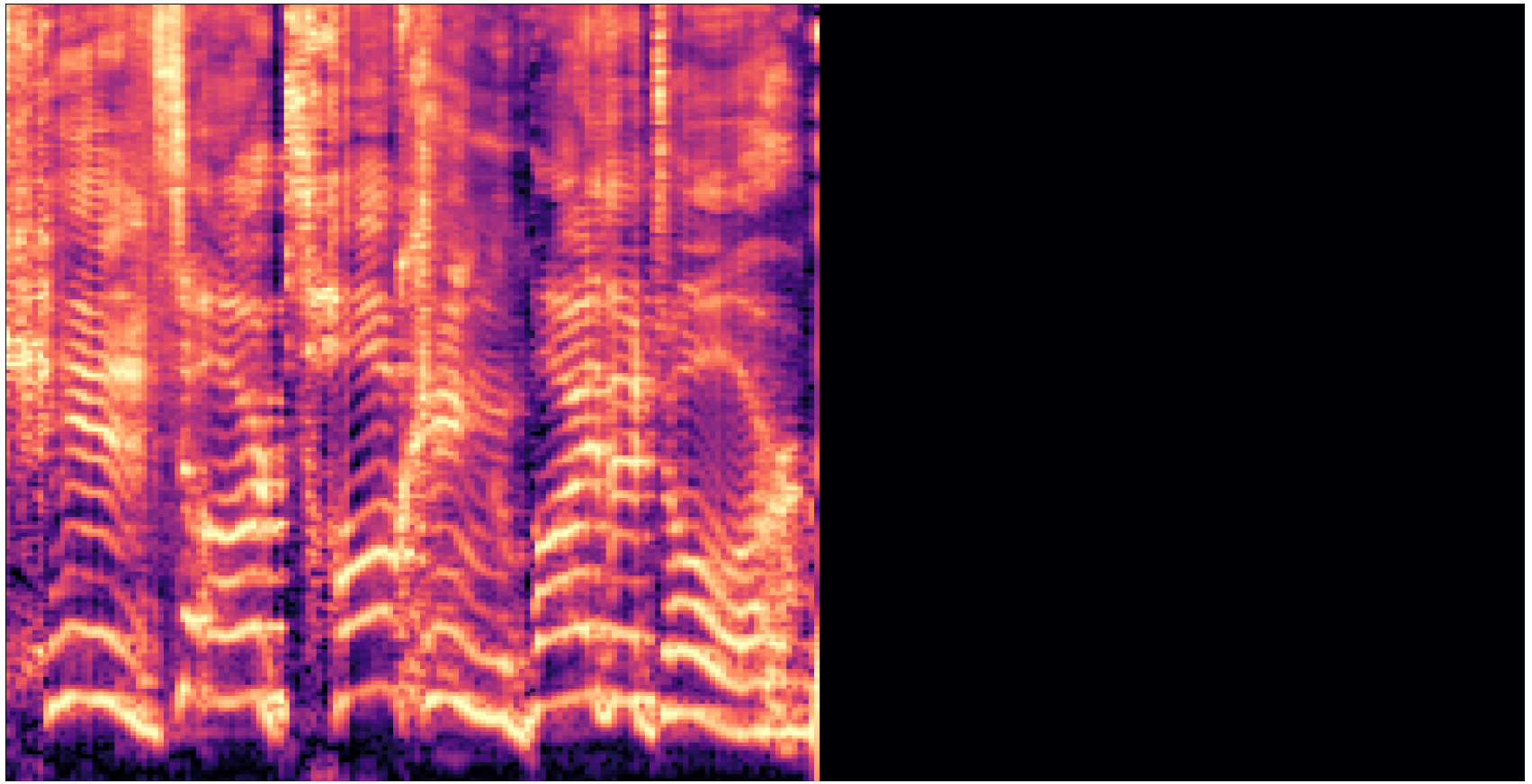
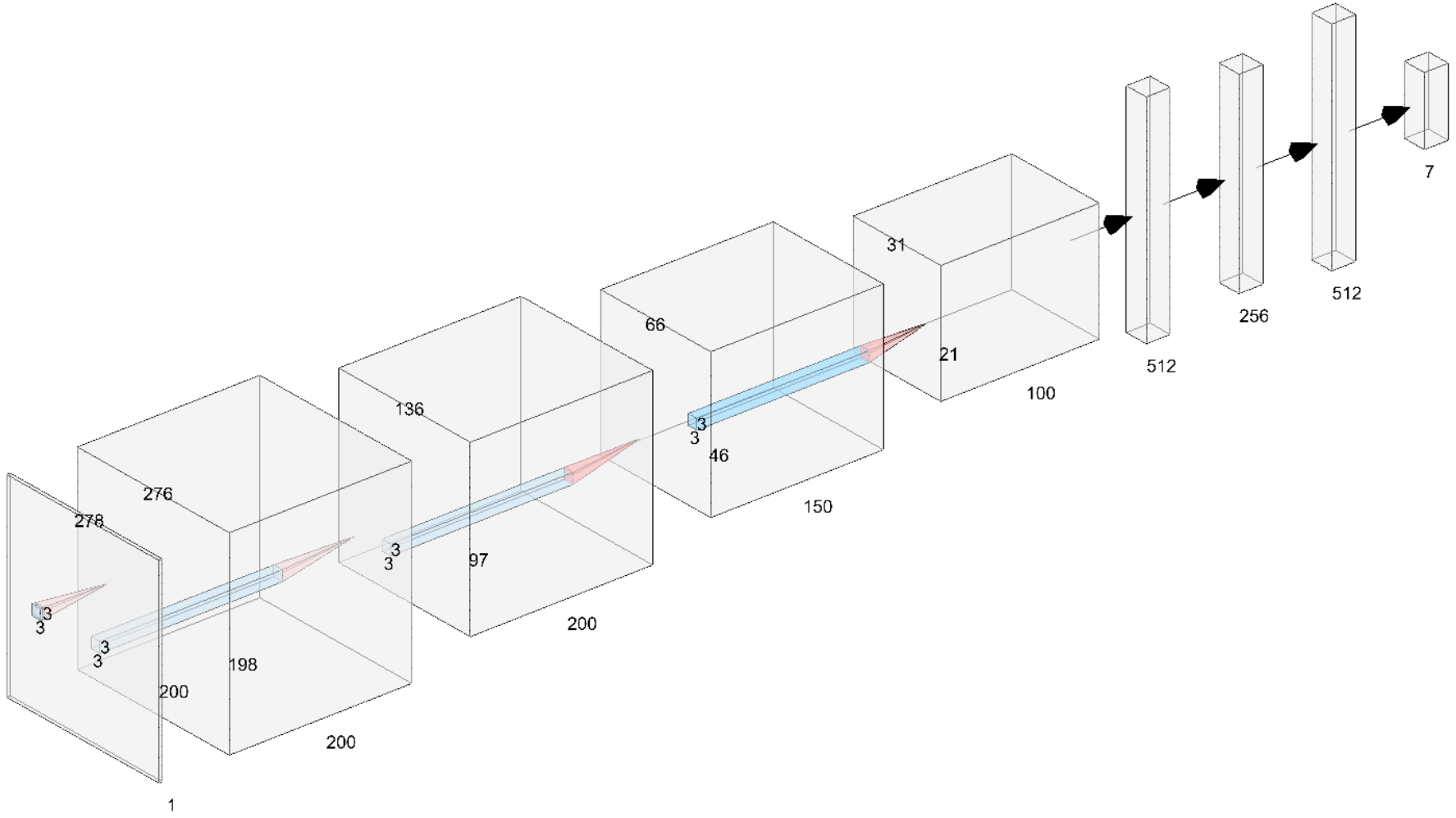


Table of Contents

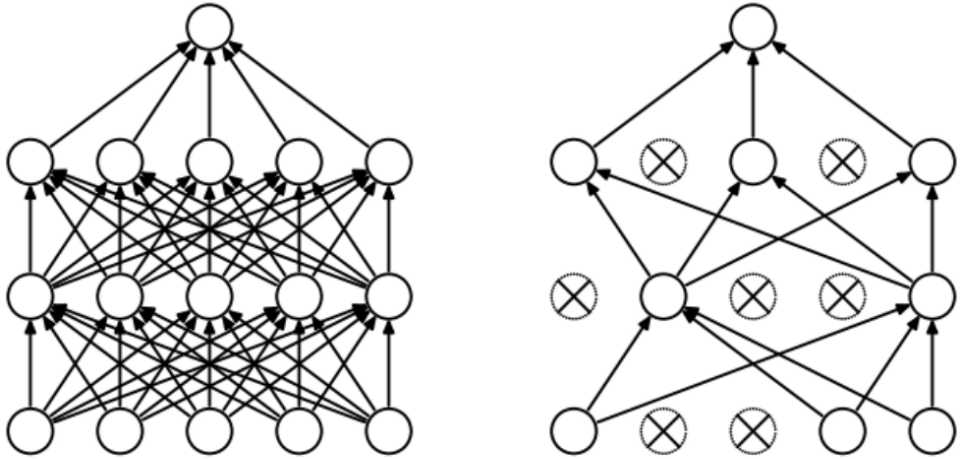
Neural Network – Architecture



Neural Network – Hyperparameters

Hyperparameter	Value
Input Dimensions	278w x 200h x 1d
Optimization Algorithm	Adam
Loss Measure	Binary Cross-entropy
Accuracy Metric	Categorical Cross-entropy
Activation Function	Sigmoid
Epochs	20
Batch Size	32
Training Set	17,341
Validation Set	2,167
Testing Set	2,167

Neural Network – Anti-overfit Methods



(a) Standard Neural Net

(b) After applying dropout.

- Dropout on dense layers

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

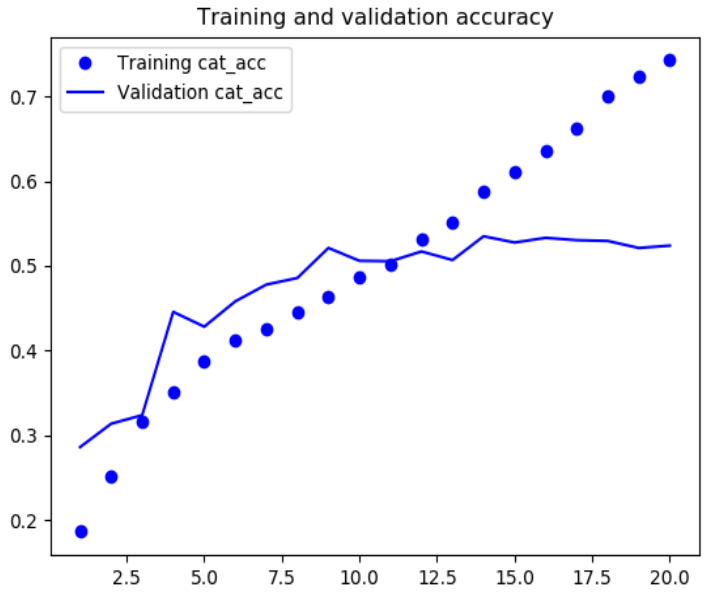
$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

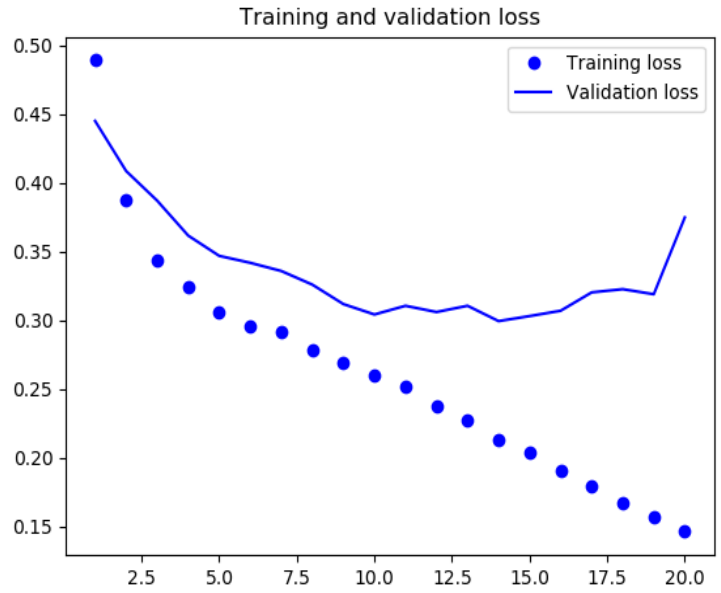
- Batch normalization on convolution layers

Table of Contents

Summary – Neural Network Training



- The accuracy training curve shows overfitting after 12 epochs.
- Validation accuracy ends at about 53.91%.

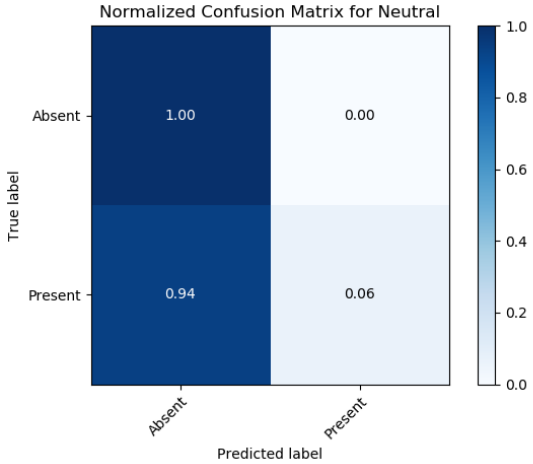


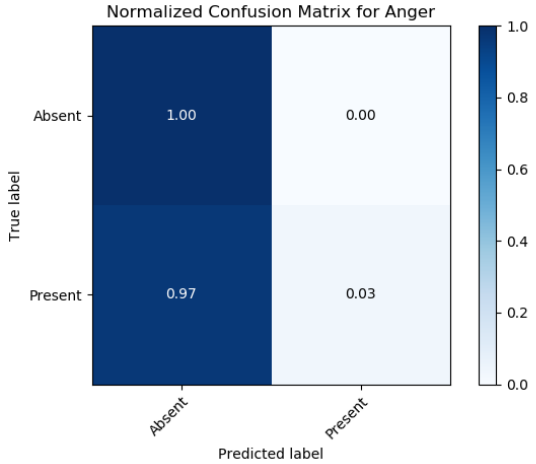
- The loss training curve also shows overfitting right from the start.
- The lowest (best) loss is achieved at around epoch 18 before it rises.

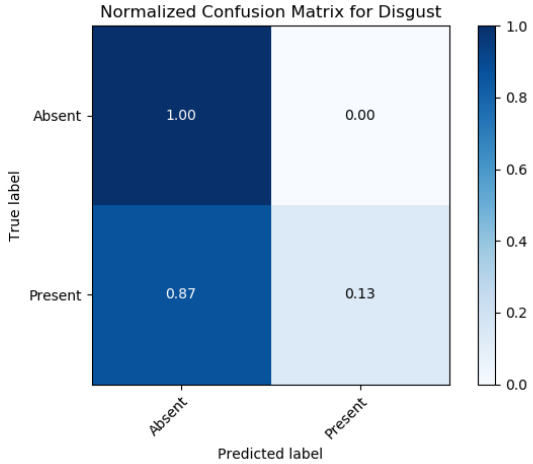
UNIVERSITY OF CALGARY

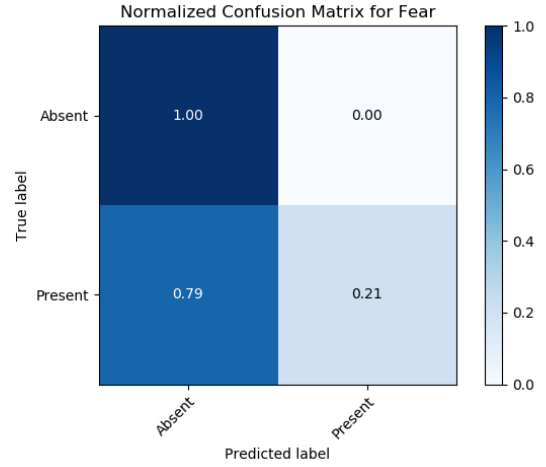
The neural network achieves a final test accuracy of 52.57%.

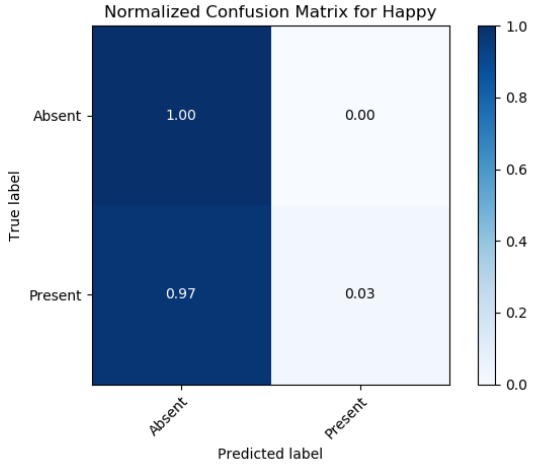
Summary – Confusion Matrices

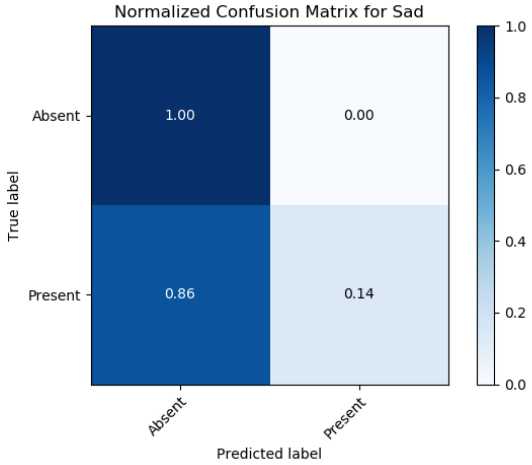












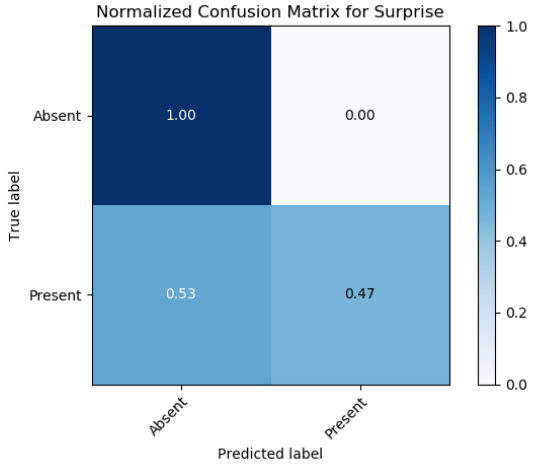
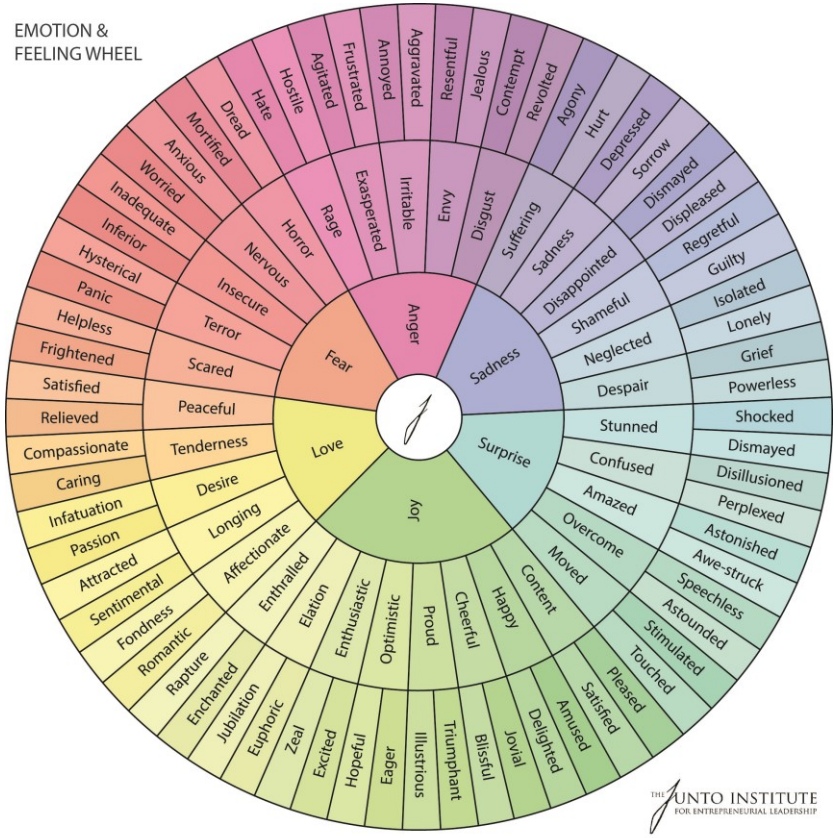


Table of Contents

Limitations

- Only considered seven emotions.
- Accuracy could be better as the human baseline is around 70%.
- Only considered samples in English.



Future Work

- Address the limitations.
- Use phase data from spectrograms.
- Wavelets.
- Binary relevance.
- Real-time emotion recognition.

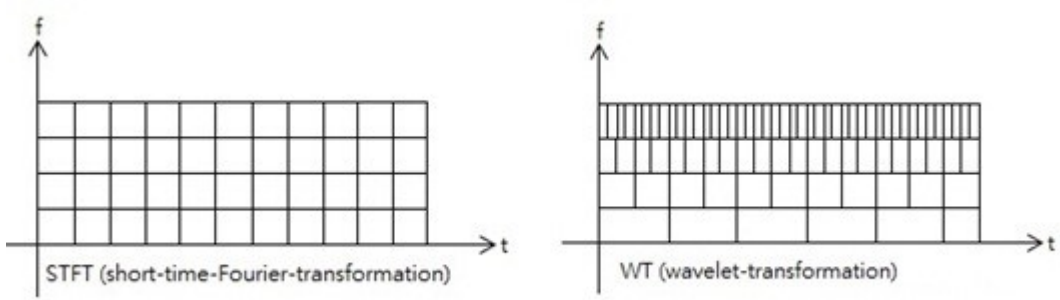


Table of Contents

Conclusion

UNIVERSITY OF CALGARY

- We applied machine learning to the problem of multi-class, multi-label speech emotion recognition.
- We achieved an accuracy of 52.57%.
- While this accuracy is a good start, we can do better by
  - Expanding the list of labeled emotions
  - Including the use of phase data from the STFT
  - Testing binary relevance against this problem

Acknowledgements

- I would like to thank the following people for their support over the course of this research project
  - Thomas Truong
    - For providing technical advice and the project idea
  - Svetlana Yanushkevich
    - For providing lab space and the (expensive) GPUs for neural network training
  - Ethan Sengsavang
    - For coding up functions that I used and manually labeling the databases
- I would also like to thank the Program for Undergraduate Research Experience (PURE) award for funding this research.

UNIVERSITY OF CALGARY

Thank You.
Questions?

For more information, a paper will be released soon and the code is available at https://github.com/Brian-Pho/RVST598_Speech-Emotion-Recognition

Brian Pho
brian.pho@ucalgary.ca

Appendix – Equations

- Mel scale: $m = 2595log_{10}(1 + \frac{f}{700})$
- Short-time Fourier transform: $STFT\{x(t)\}(\tau, w) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}dt$
- Softmax: $\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}\ for\ j = 1, \ldots, K.$
- Sigmoid: $sig(t) = \frac{1}{1+ e^{-t}}$