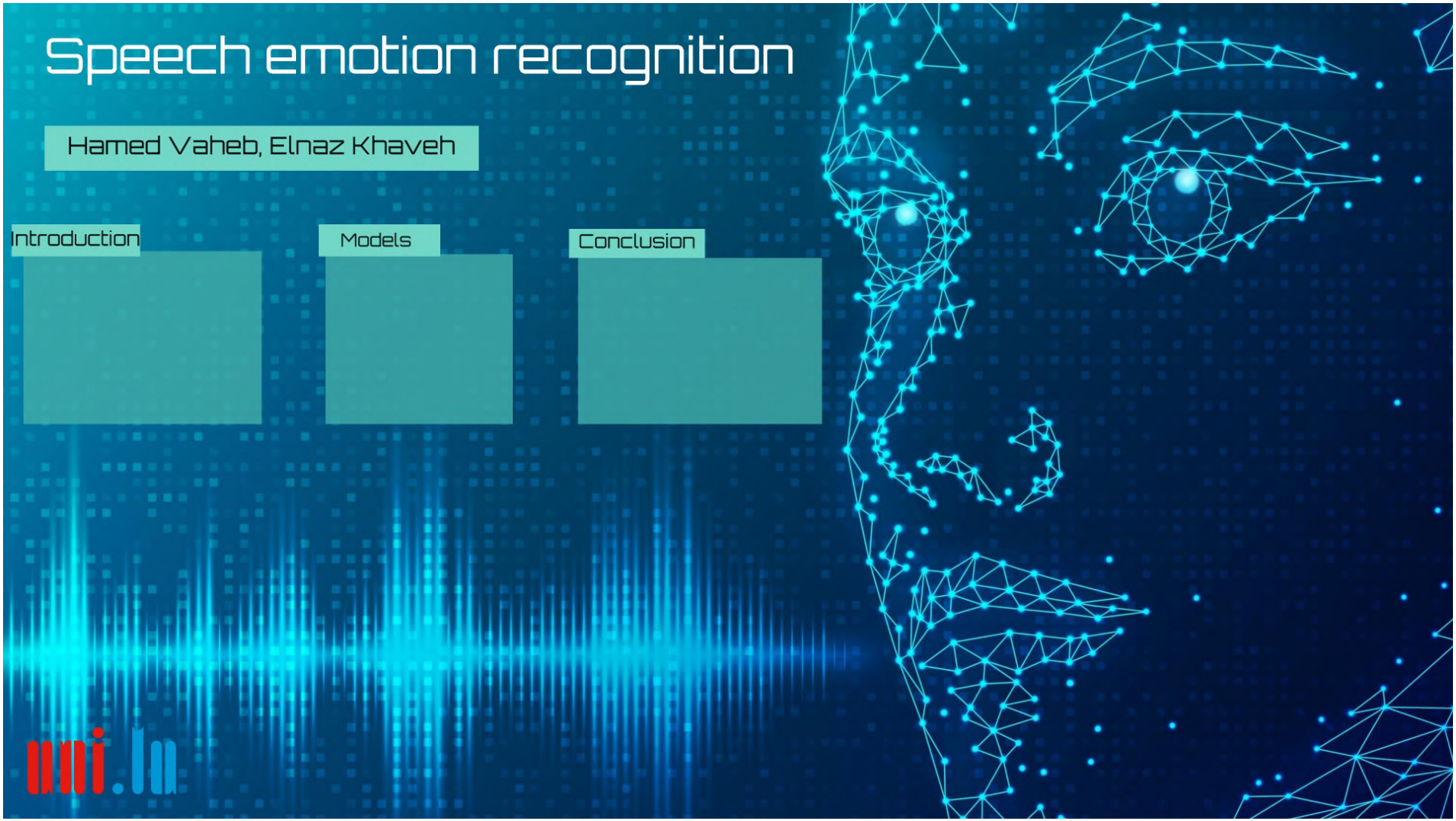

Speech emotion recognition
Hamed Vaheb, Elnaz Khaveh
Introduction
Models
Conclusion


Introduction
Applications
Data set


Introduction
Importance:
Applications
Data set

Introduction

Importance:

- The most natural way to express ourselves

Introduction

Importance:


The Evolution of Human Vocal Emotion
Gregory A. Bryant
First Published June 24, 2020 | Research Article
https://doi.org/10.1177/1754073920930791
Article information

- The most natural way to express ourselves

Applications

Data set

Introduction

Importance:


The Evolution of Human Vocal Emotion
Gregory A. Bryant
First Published June 24, 2020 | Research Article
https://doi.org/10.1177/1754073920930791
Article information

- The most natural way to express ourselves
- Influence over judgement
  - "go home!"

Applications

Data set

Introduction

Importance:


The Evolution of Human Vocal Emotion
Gregory A. Bryant
First Published June 24, 2020 | Research Article
https://doi.org/10.1177/1754073920930791
Article information

- The most natural way to express ourselves
- Influence over judgement
  - "go home!"
- Human-computer interaction (HCI)
  - Personal Assistants
  - Next Generation AI

Applications

Data set


Applications

Applications

- Psychological consultation

Applications

- Psychological consultation

Applications

- Psychological consultation

Predicting couple therapy outcomes based on speech acoustic features

Md Nasir, Brian Robert Baucom, Panayiotis Georgiou, Shrikanth Narayanan

Published: September 21, 2017 • https://doi.org/10.1371/journal.pone.0185123

Applications

- Psychological consultation

Predicting couple therapy outcomes based on speech acoustic features

Md Nasir, Brian Robert Baucom, Panayiotis Georgiou, Shrikanth Narayanan

Published: September 21, 2017 • https://doi.org/10.1371/journal.pone.0185123

Interspeech. 2018 Sep; 2018: 3413–3417. PMID:

doi: 10.21437/interspeech.2018-2551

Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla,[1] Zhuohao Chen,[1] Nikolaos Flemotomos,[1] James Gibson,[1] Dogan Can,[1] David C. Atkins,[2] and Shrikanth Narayanan[1]

Applications

Predicting couple therapy outcomes based on speech acoustic features

Md Nasir, Brian Robert Baucom, Panayiotis Georgiou, Shrikanth Narayanan

Published: September 21, 2017 • https://doi.org/10.1371/journal.pone.0185123

Interspeech. 2018 Sep; 2018: 3413–3417. PMID:
doi: 10.21437/interspeech.2018-2551

Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla,[1] Zhuohao Chen,[1] Nikolaos Flemotomos,[1] James Gibson,[1] Dogan Can,[1] David C. Atkins,[2] and Shrikanth Narayanan[1]

Applications

Interspeech. 2018 Sep; 2018: 3413–3417. PMID:
doi: 10.21437/interspeech.2018-2551

Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla,[1] Zhuohao Chen,[1] Nikolaos Flemotomos,[1] James Gibson,[1] Dogan Can,[1] David C. Atkins,[2] and Shrikanth Narayanan[1]

Applications

Interspeech. 2018 Sep; 2018: 3413–3417. PMID:
doi: 10.21437/interspeech.2018-2551

Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla,[1] Zhuohao Chen,[1] Nikolaos Flemotomos,[1] James Gibson,[1] Dogan Can,[1] David C. Atkins,[2] and Shrikanth Narayanan[1]

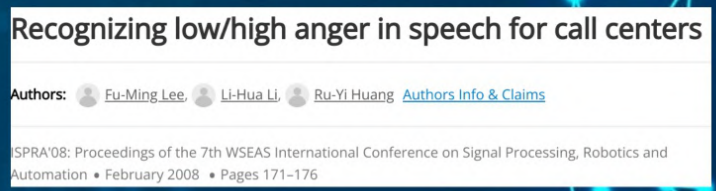
Applications

Applications

- Call centers

Applications

- Call centers



Introduction

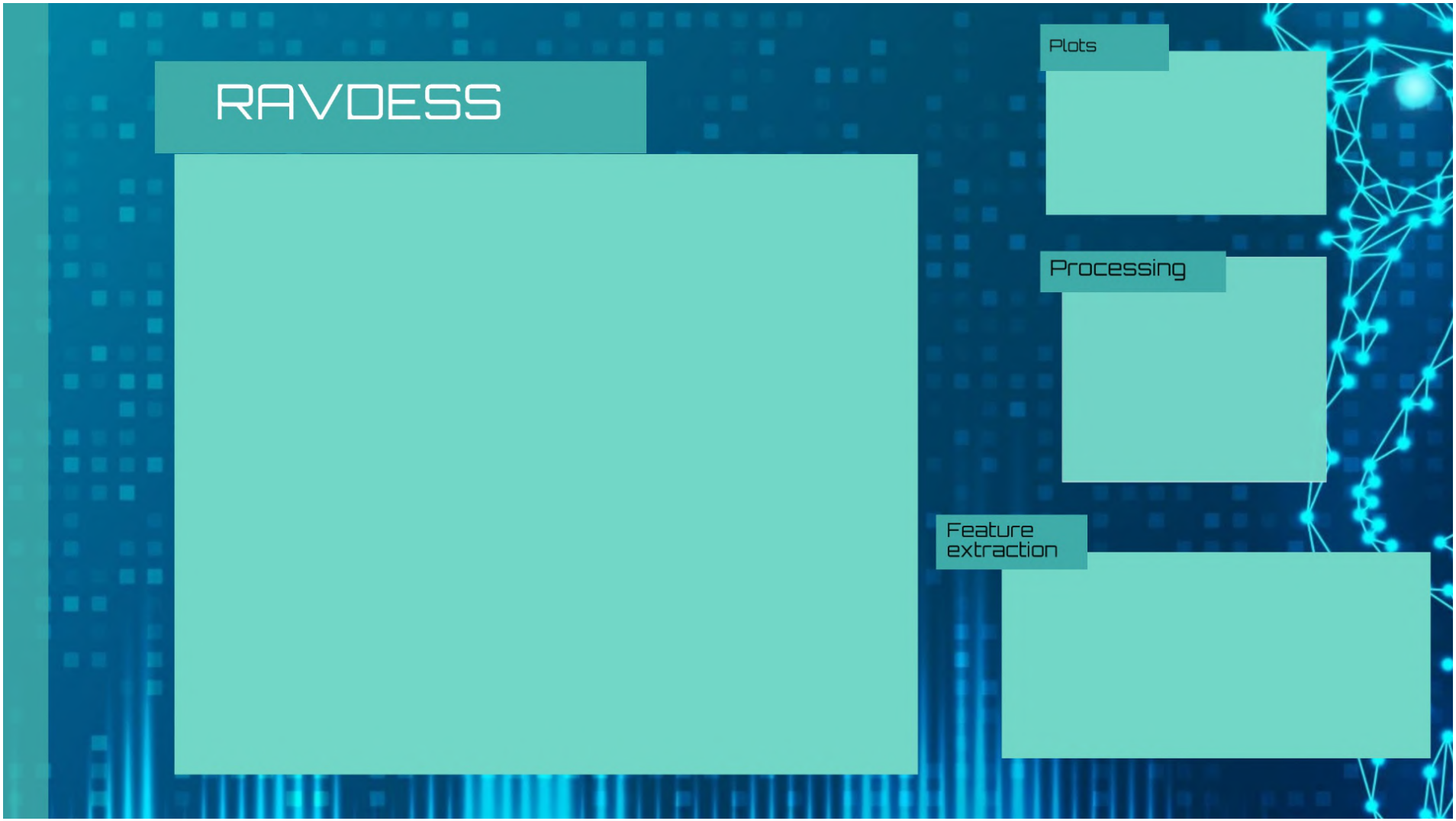
Importance:


The Evolution of Human Vocal Emotion
Gregory A. Bryant
First Published June 24, 2020 | Research Article
https://doi.org/10.1177/1754073920930791
Article information

- The most natural way to express ourselves
- Influence over judgement
  - "go home!"
- Human-computer interaction (HCI)
  - Personal Assistants
  - Next Generation AI

Applications

Data set


RAVDESS
Plots
Processing
Feature extraction

RAVDESS

- A validated multimodal database

Plots

Processing

Feature extraction

RAVDESS

- A validated multimodal database
- 24 professional actors

Plots

Processing

Feature extraction

RAVDESS

- A validated multimodal database
- 24 professional actors
- Gender-balanced

Plots

Processing

Feature extraction

RAVDESS

- 24 professional actors
- Gender-balanced

Plots

Processing

Feature extraction

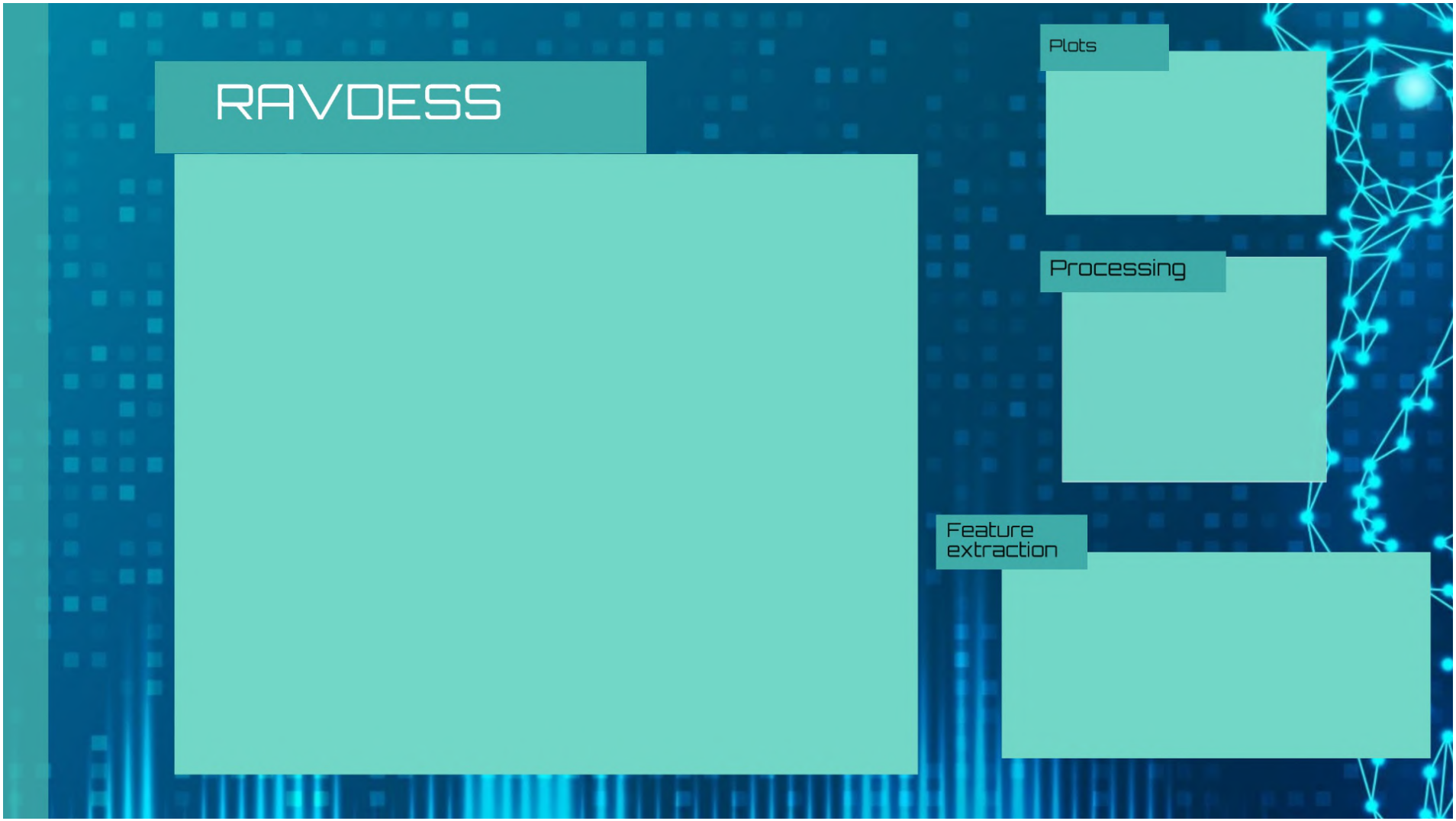
RAVDESS

- Gender-balanced

Plots

Processing

Feature extraction


RAVDESS
Plots
Processing
Feature extraction

RAVDESS

- 8 classes :

Plots

Processing

Feature extraction

RAVDESS

- 8 classes :

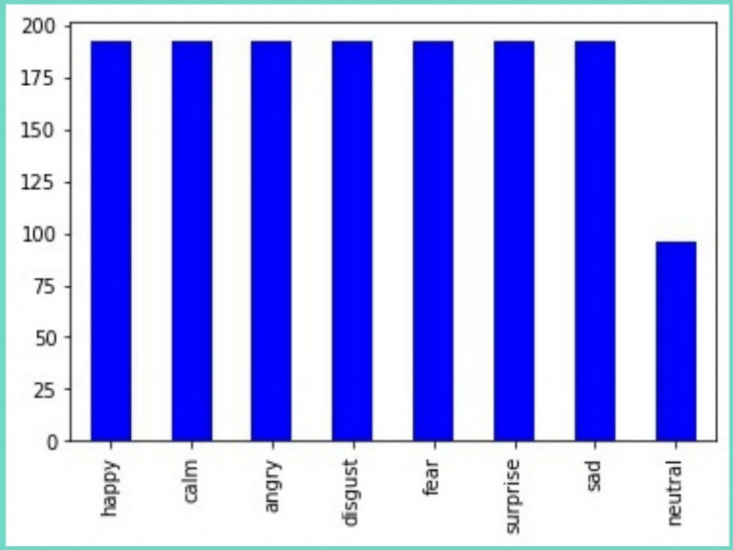
  Neutral, calm, happy, sad, angry, fearful, disgust, surprise

Plots

Processing

Feature extraction

RAVDESS



- 8 classes :

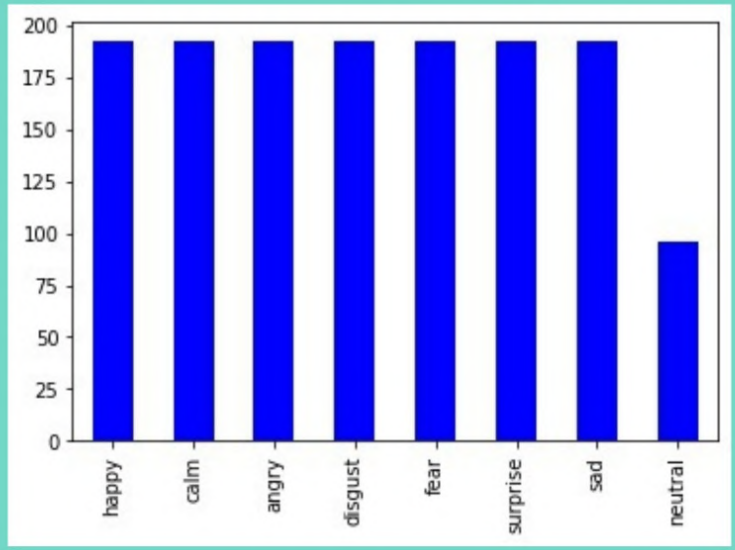
  Neutral, calm, happy, sad, angry, fearful, disgust, surprise

Plots

Processing

Feature extraction

RAVDESS



Neutral, calm, happy, sad, angry, fearful, disgust, surprise

Plots

Processing

Feature extraction

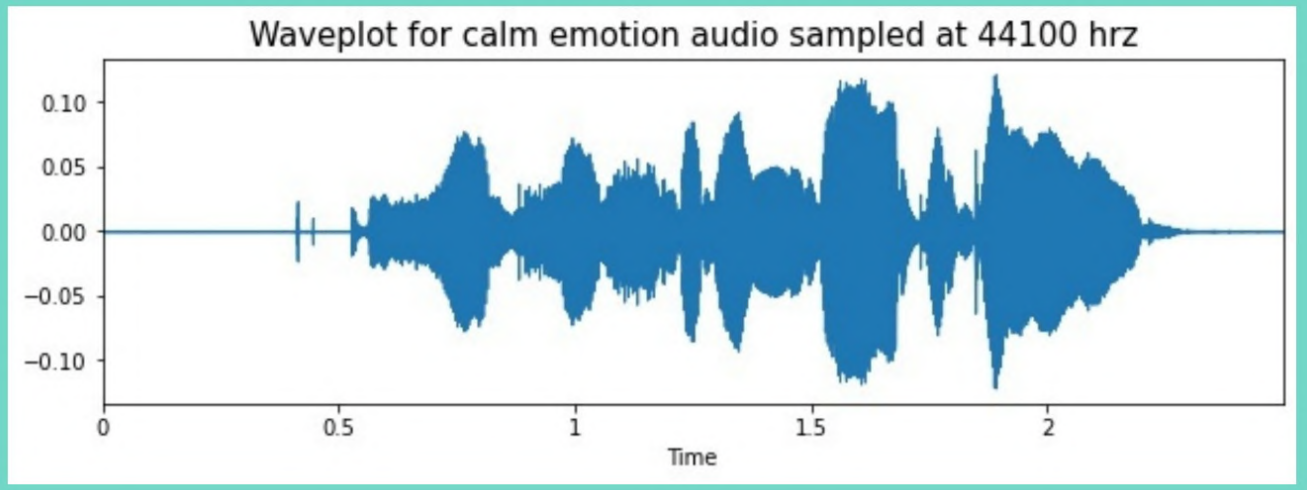
RAVDESS

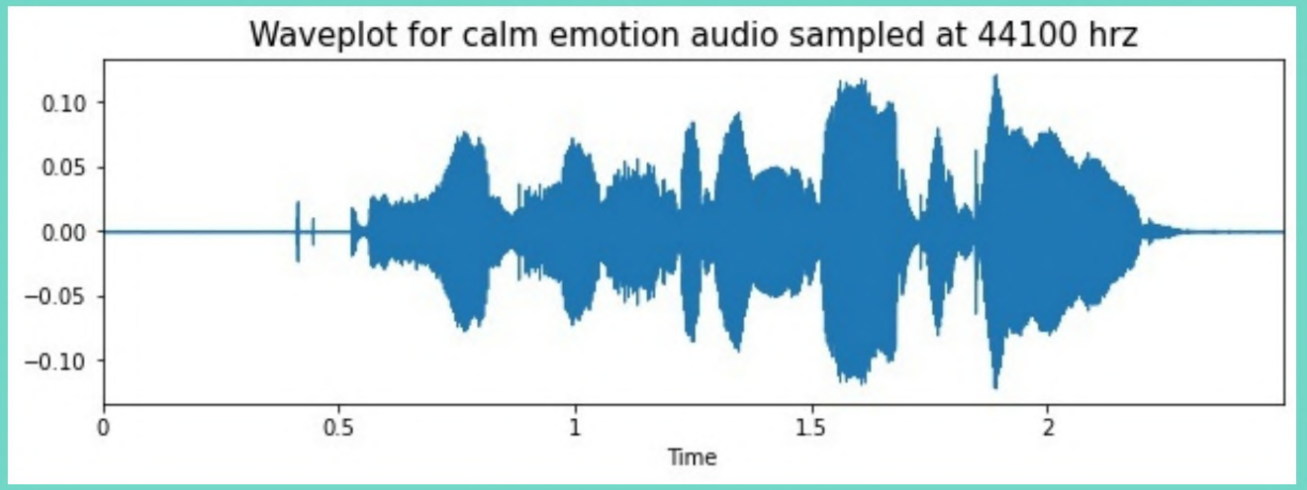
Neutral, calm, happy, sad, angry, fearful, disgust, surprise

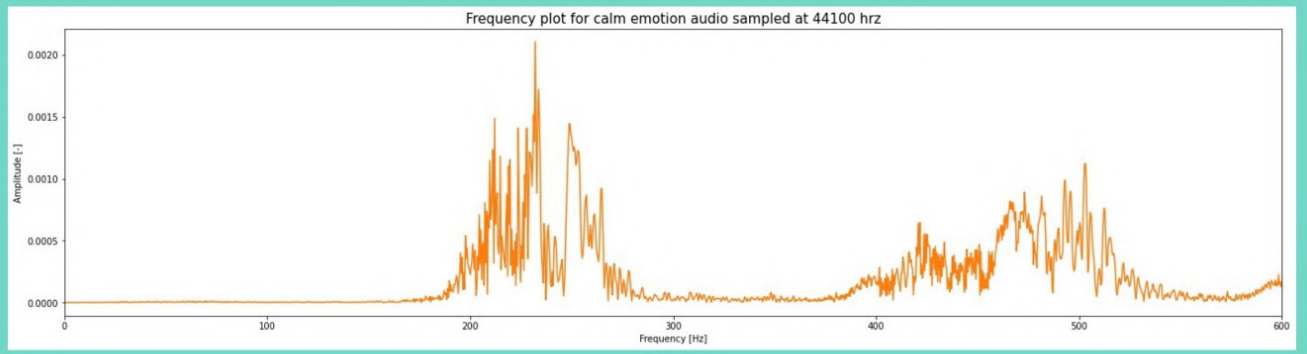
Plots

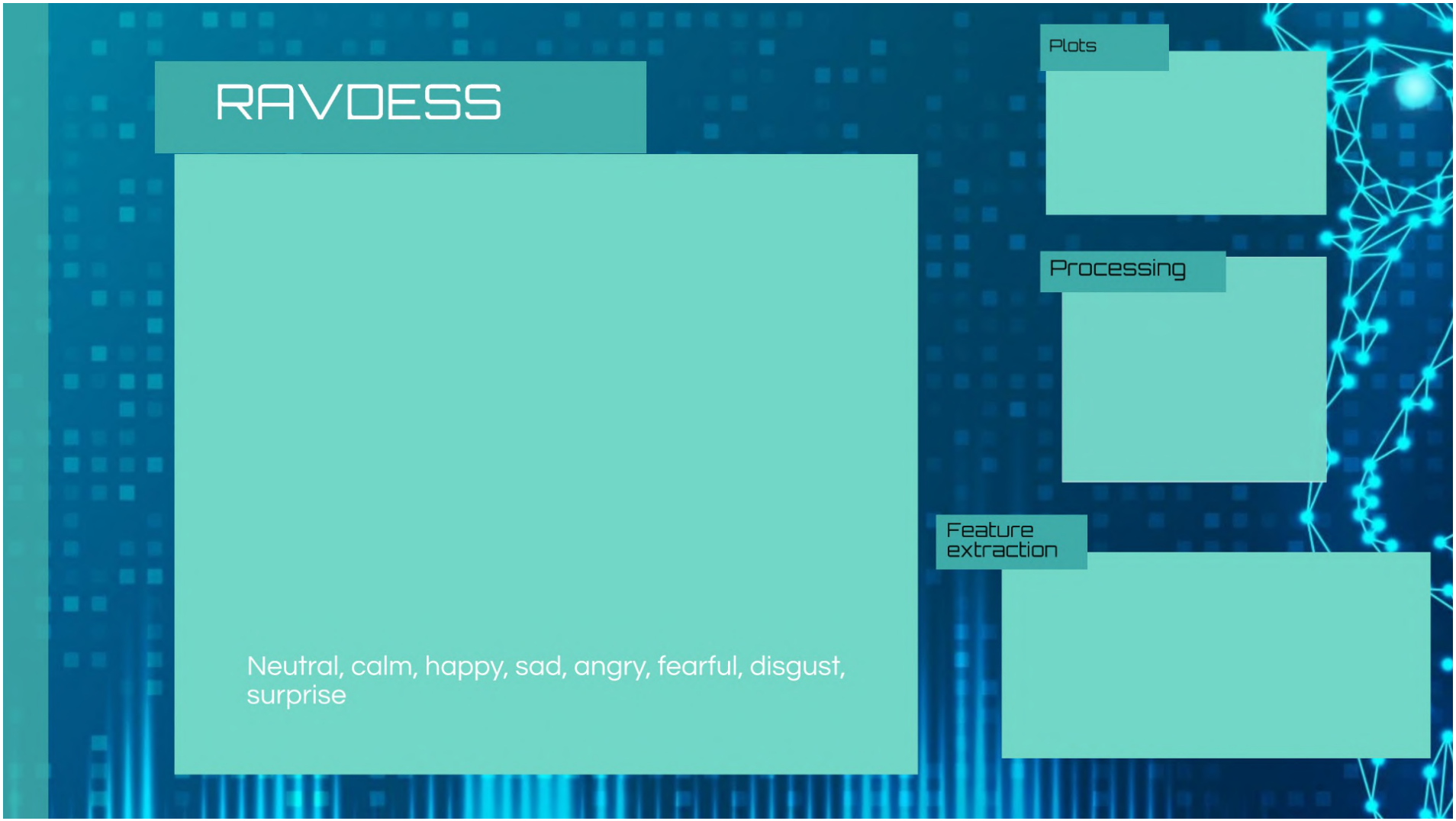
Processing

Feature extraction

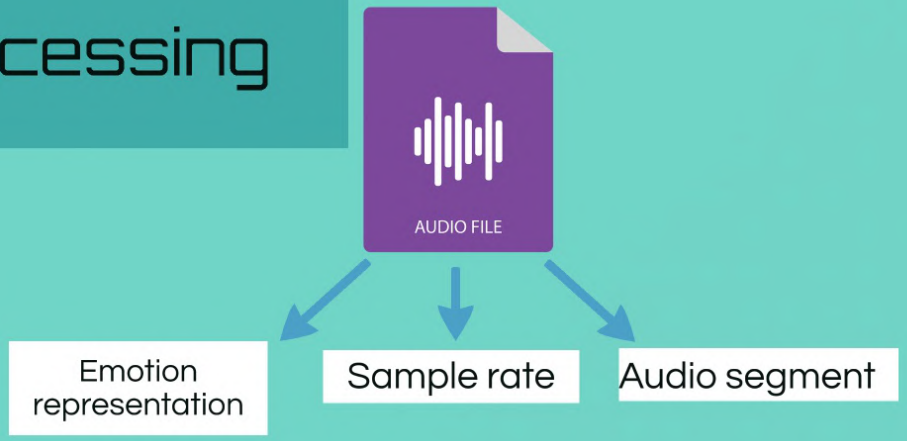

Waveplot for calm emotion audio sampled at 44100 hrz
0.10
0.05
0.00
−0.05
−0.10
0
0.5
1
1.5
2
Time

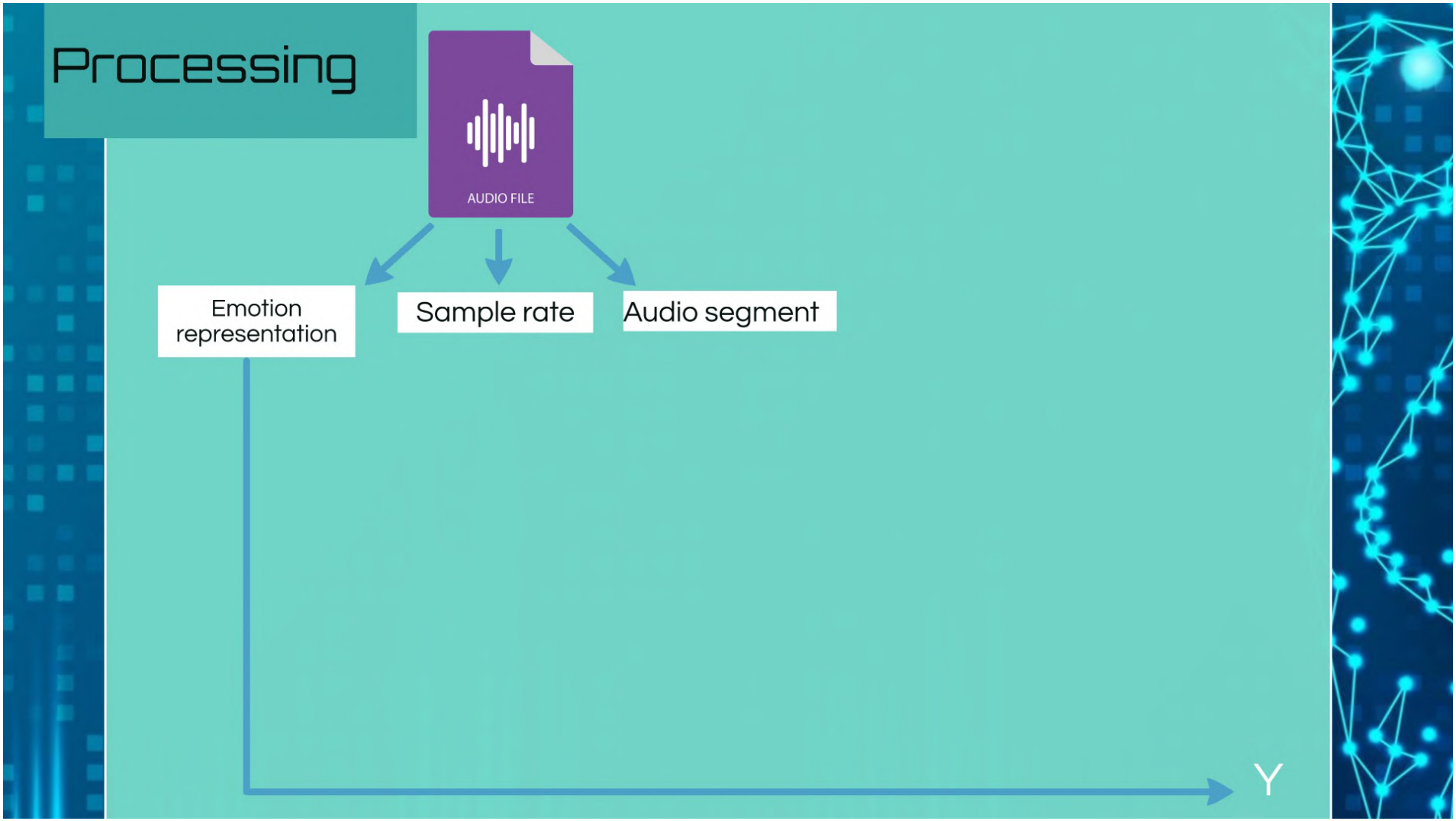

Waveplot for calm emotion audio sampled at 44100 hrz
0.10
0.05
0.00
−0.05
−0.10
0
0.5
1
1.5
2
Time

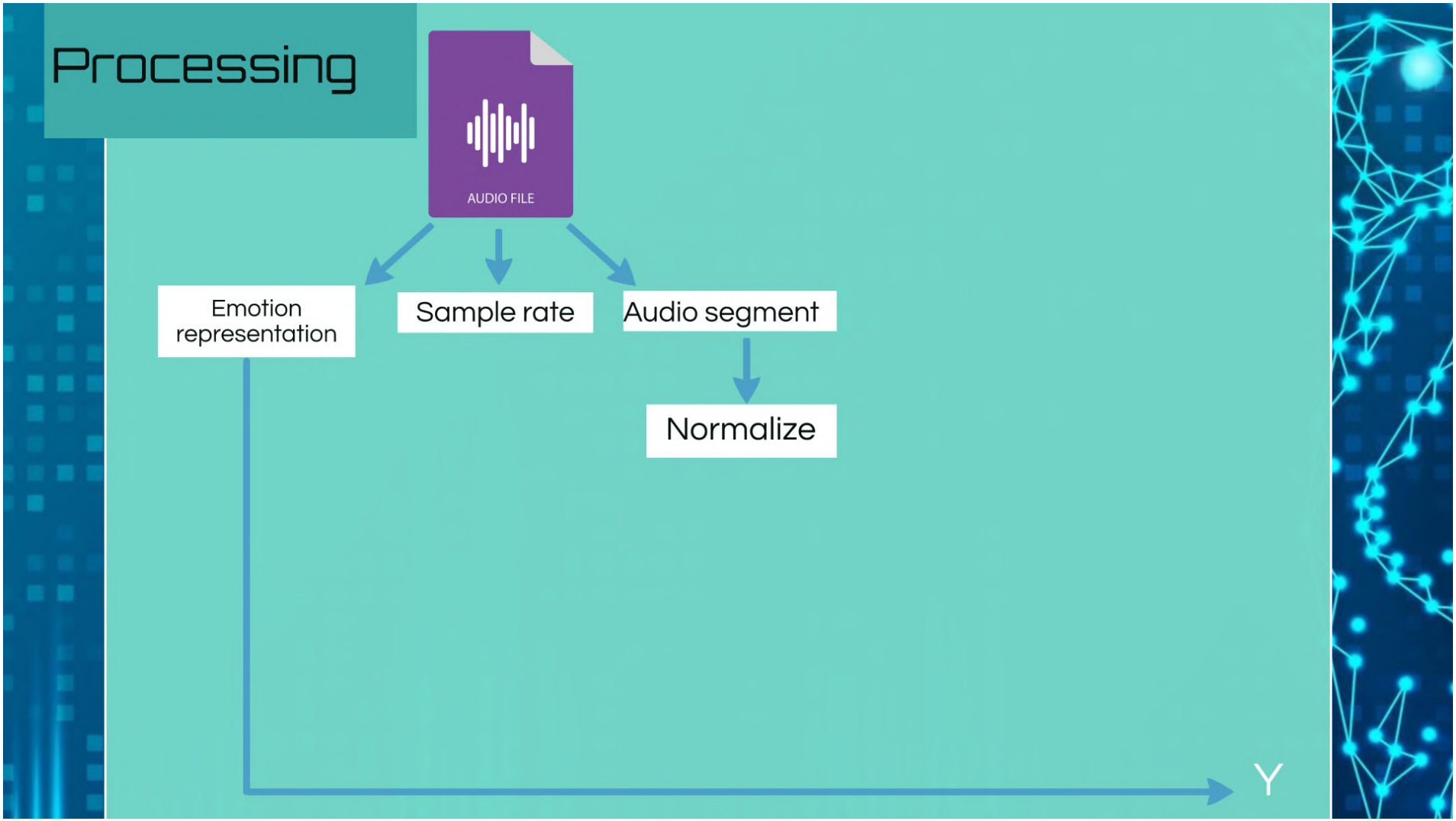

Frequency plot for calm emotion audio sampled at 44100 hrz
Amplitude [-]
0.0020
0.0015
0.0010
0.0005
0.0000
0
100
200
300
400
500
600
Frequency [Hz]

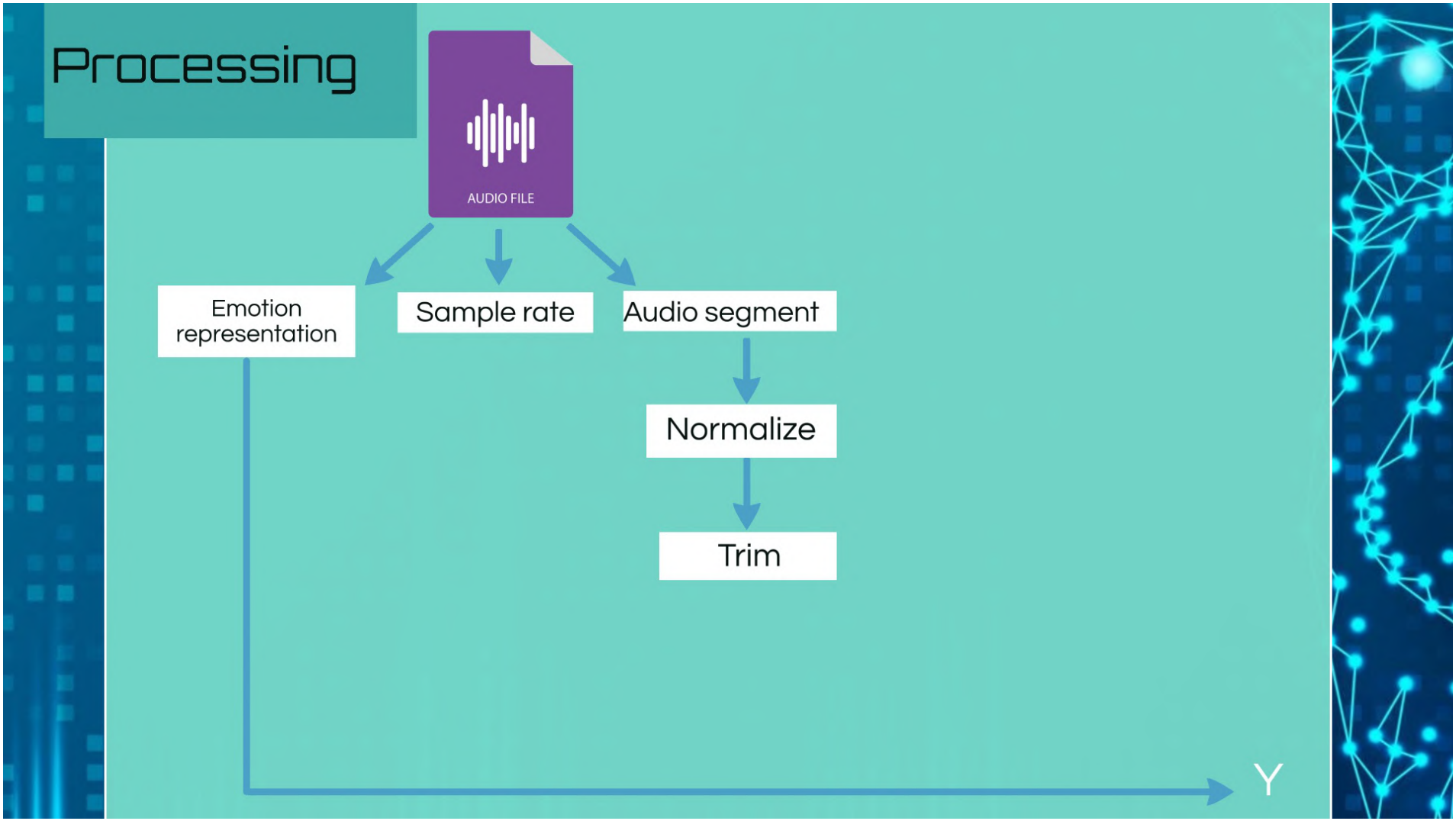

RAVDESS
Neutral, calm, happy, sad, angry, fearful, disgust, surprise
Plots
Processing
Feature extraction

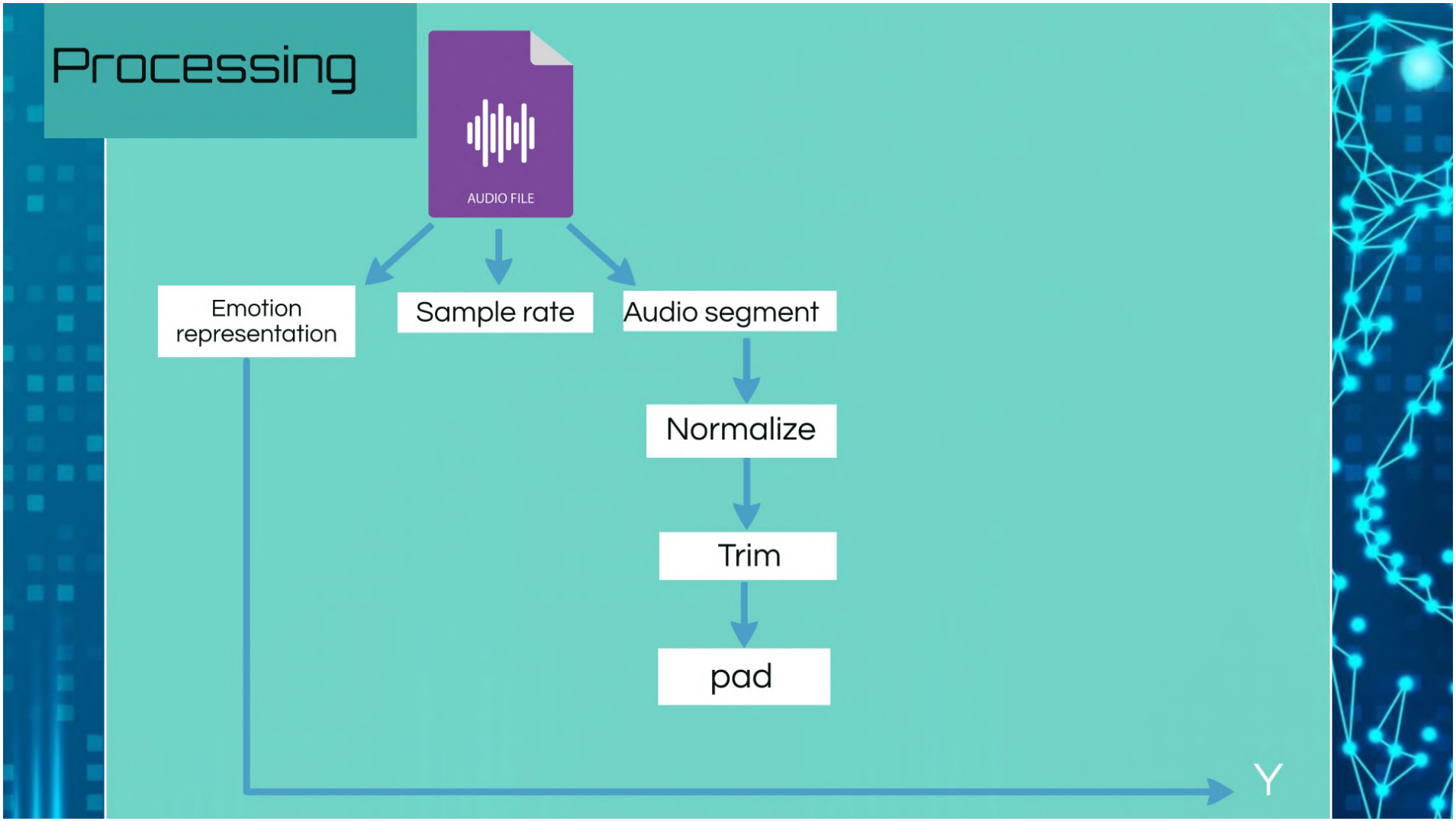
Processing

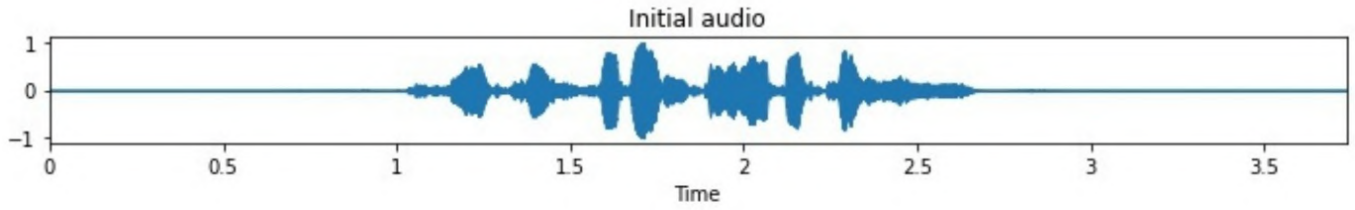

AUDIO FILE
Emotion representation
Sample rate
Audio segment

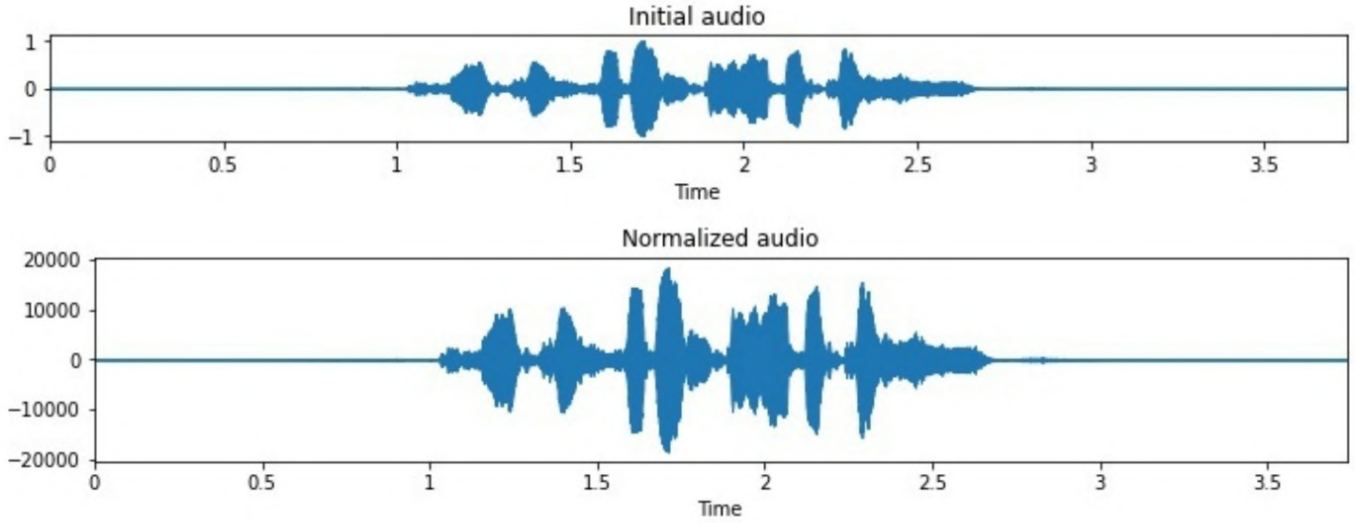

Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Y

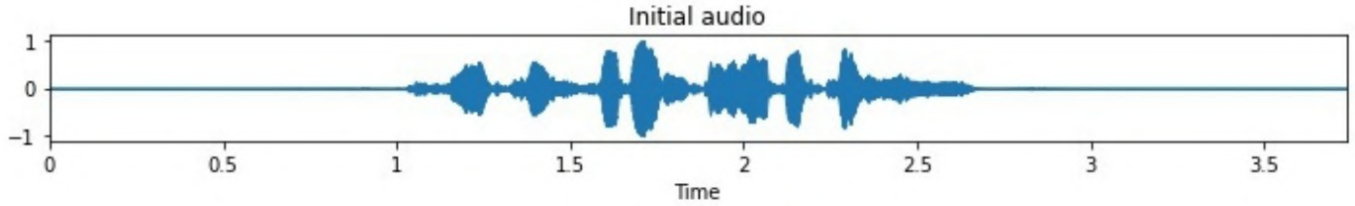

Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Y

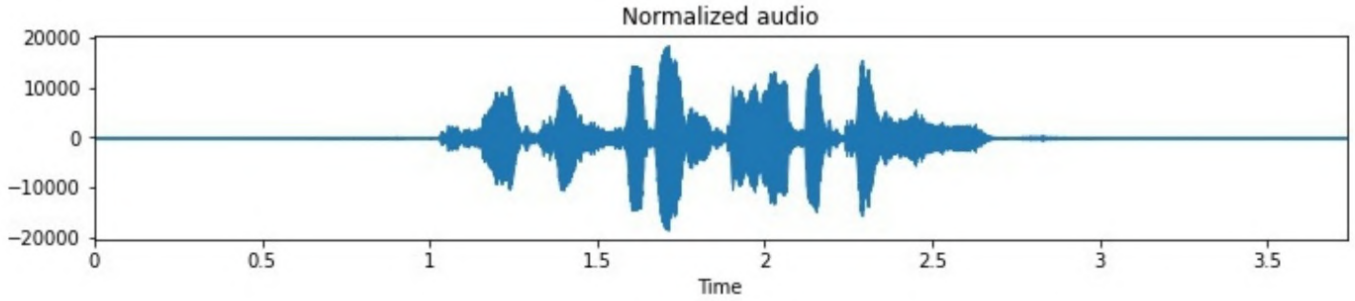

Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Trim
Y

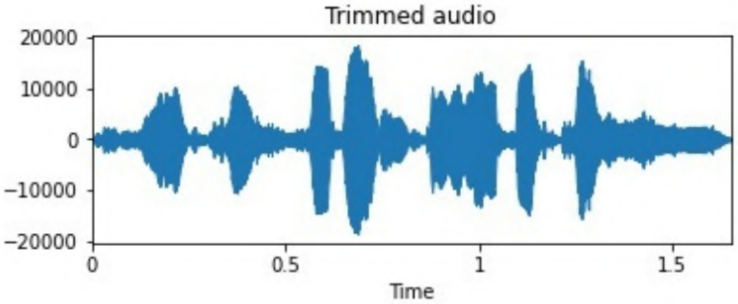

Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Trim
pad
Y

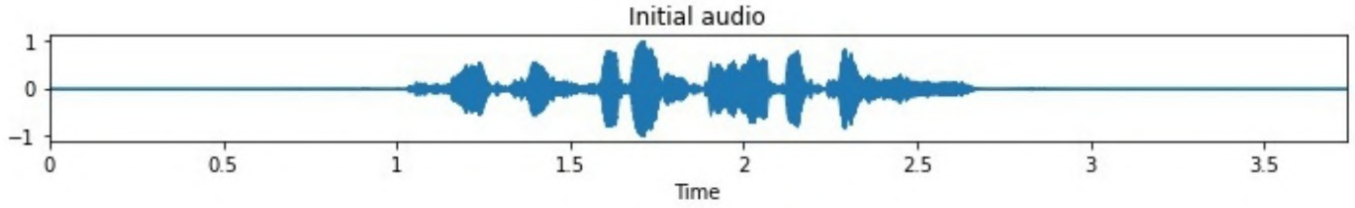

Initial audio
1
0
−1
0
0.5
1
1.5
2
2.5
3
3.5
Time

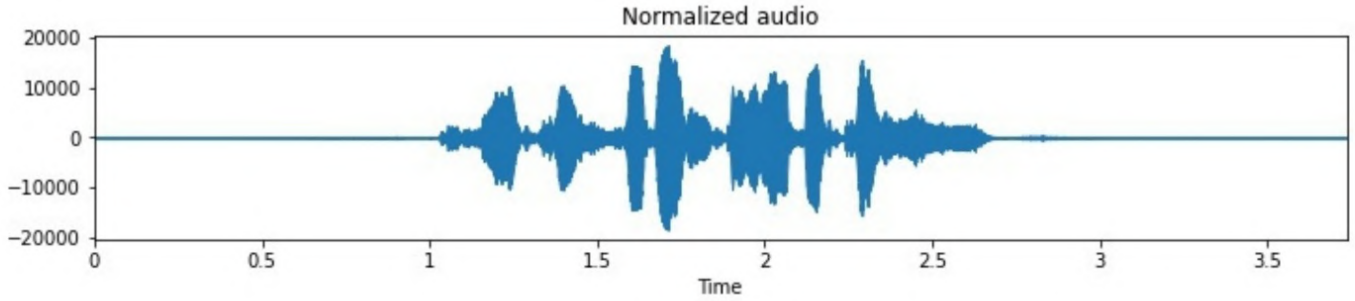

Initial audio
1
0
−1
0
0.5
1
1.5
2
2.5
3
3.5
Time
Normalized audio
20000
10000
0
−10000
−20000
0
0.5
1
1.5
2
2.5
3
3.5
Time

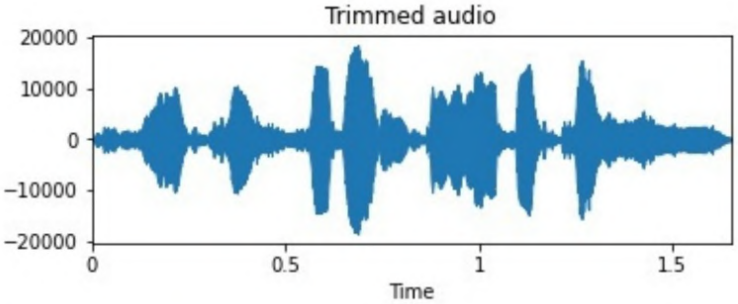

Initial audio
1
0
−1
0
0.5
1
1.5
2
2.5
3
3.5
Time

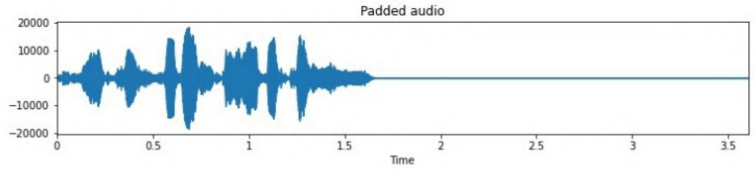

Normalized audio
20000
10000
0
−10000
−20000
0
0.5
1
1.5
2
2.5
3
3.5
Time

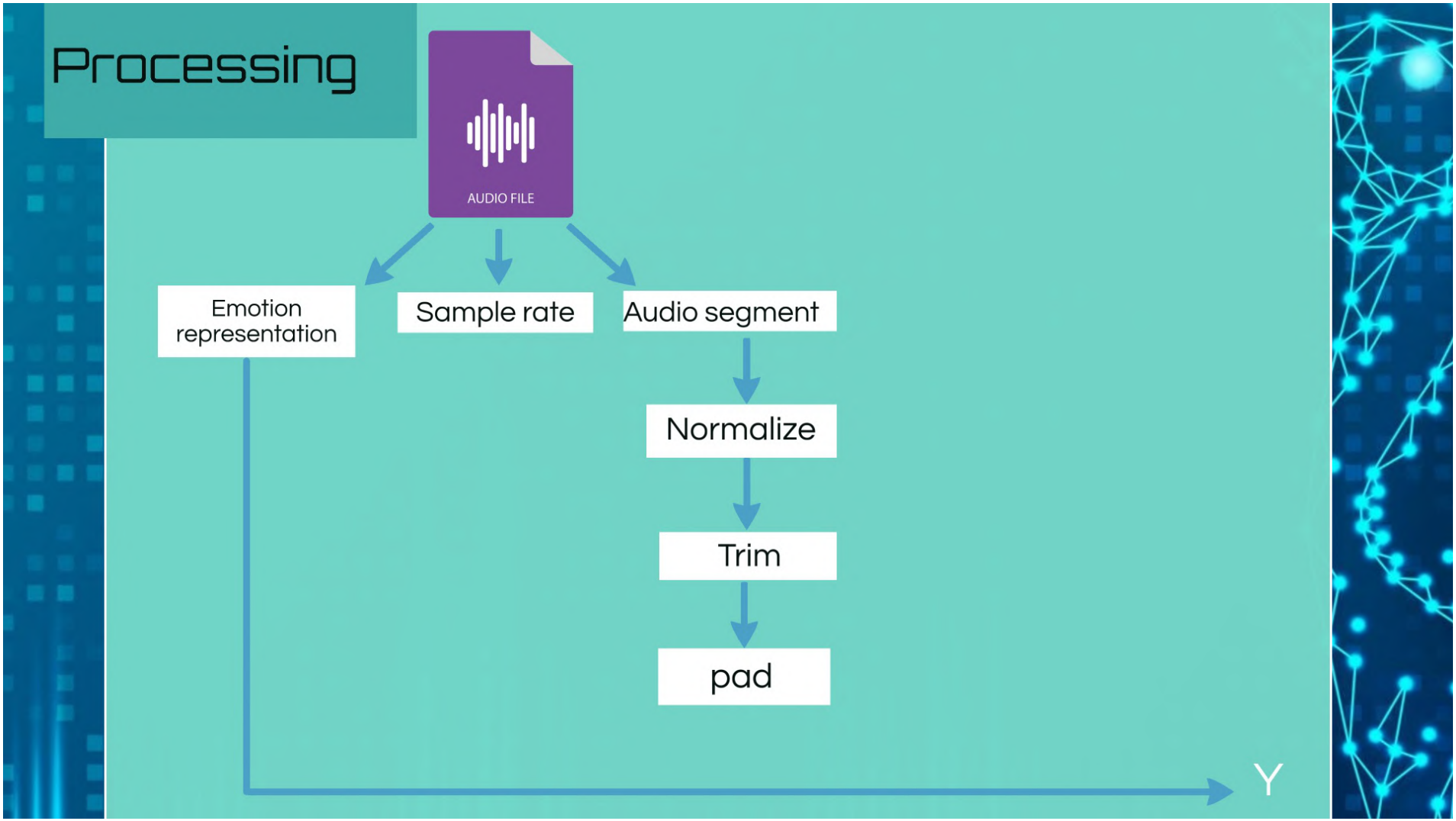

Trimmed audio
20000
10000
0
−10000
−20000
0
0.5
1
1.5
Time

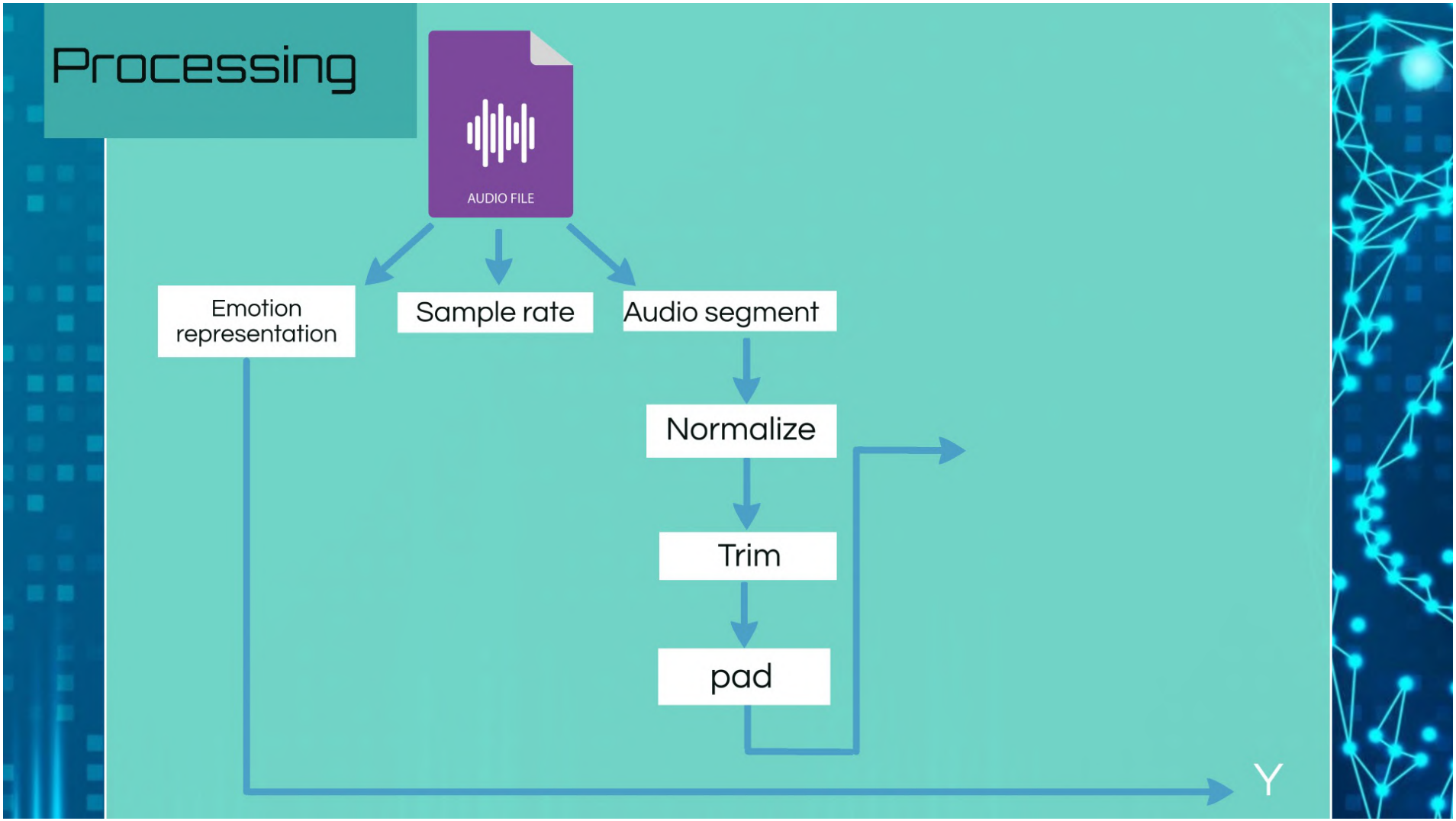

Initial audio
1
0
−1
0
0.5
1
1.5
2
2.5
3
3.5
Time

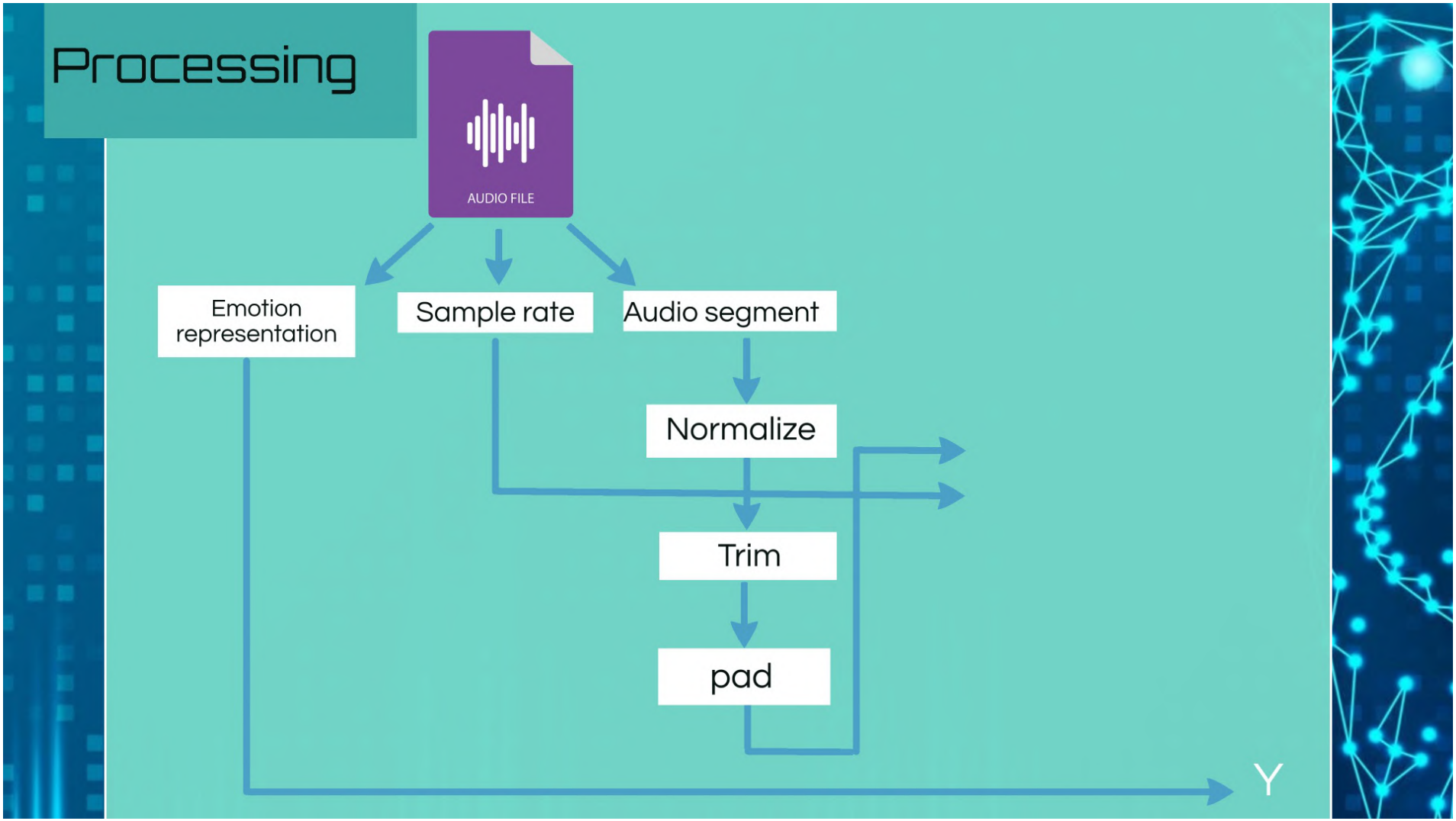

Normalized audio
20000
10000
0
−10000
−20000
0
0.5
1
1.5
2
2.5
3
3.5
Time

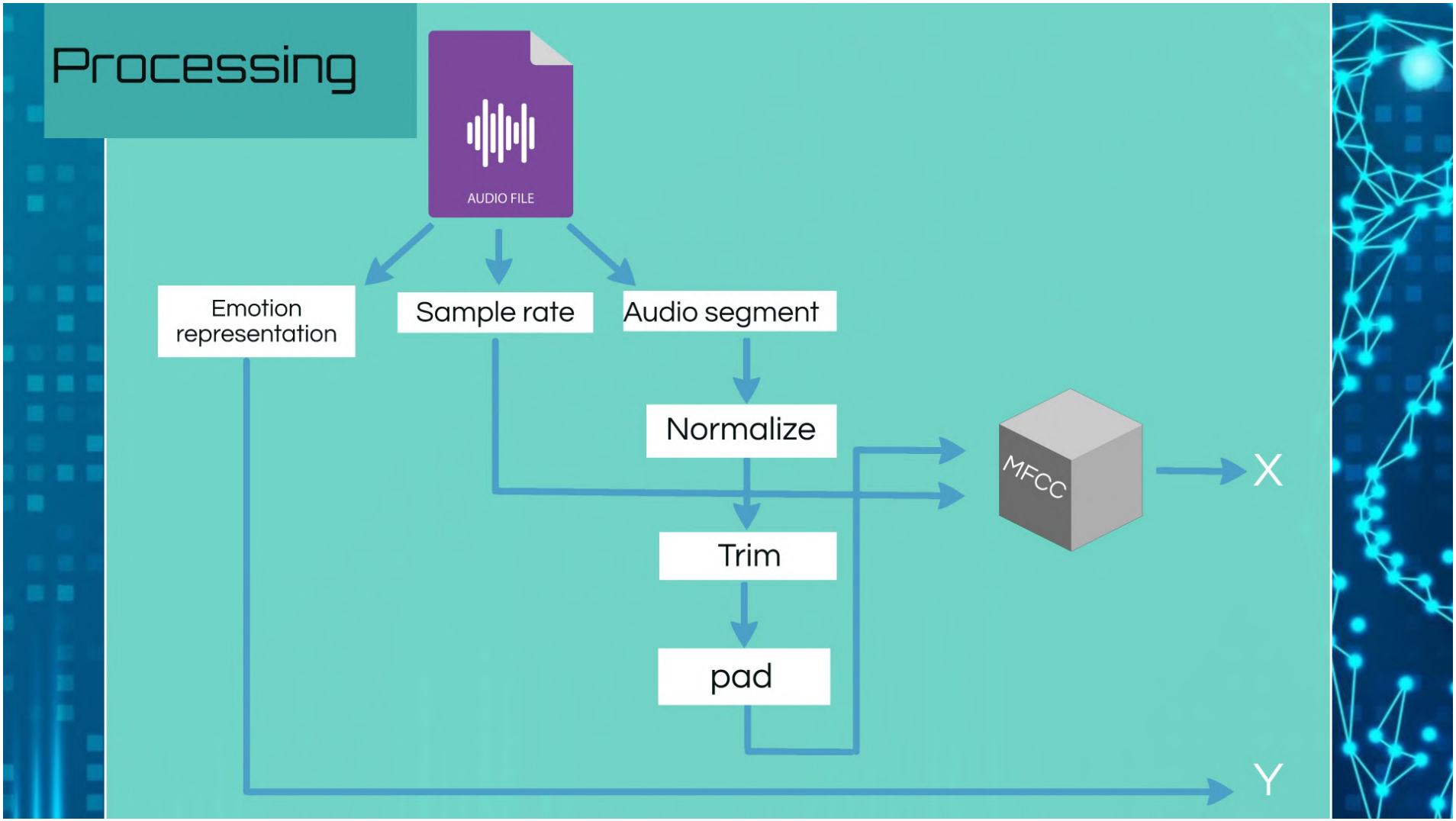

Trimmed audio
20000
10000
0
−10000
−20000
0
0.5
1
1.5
Time

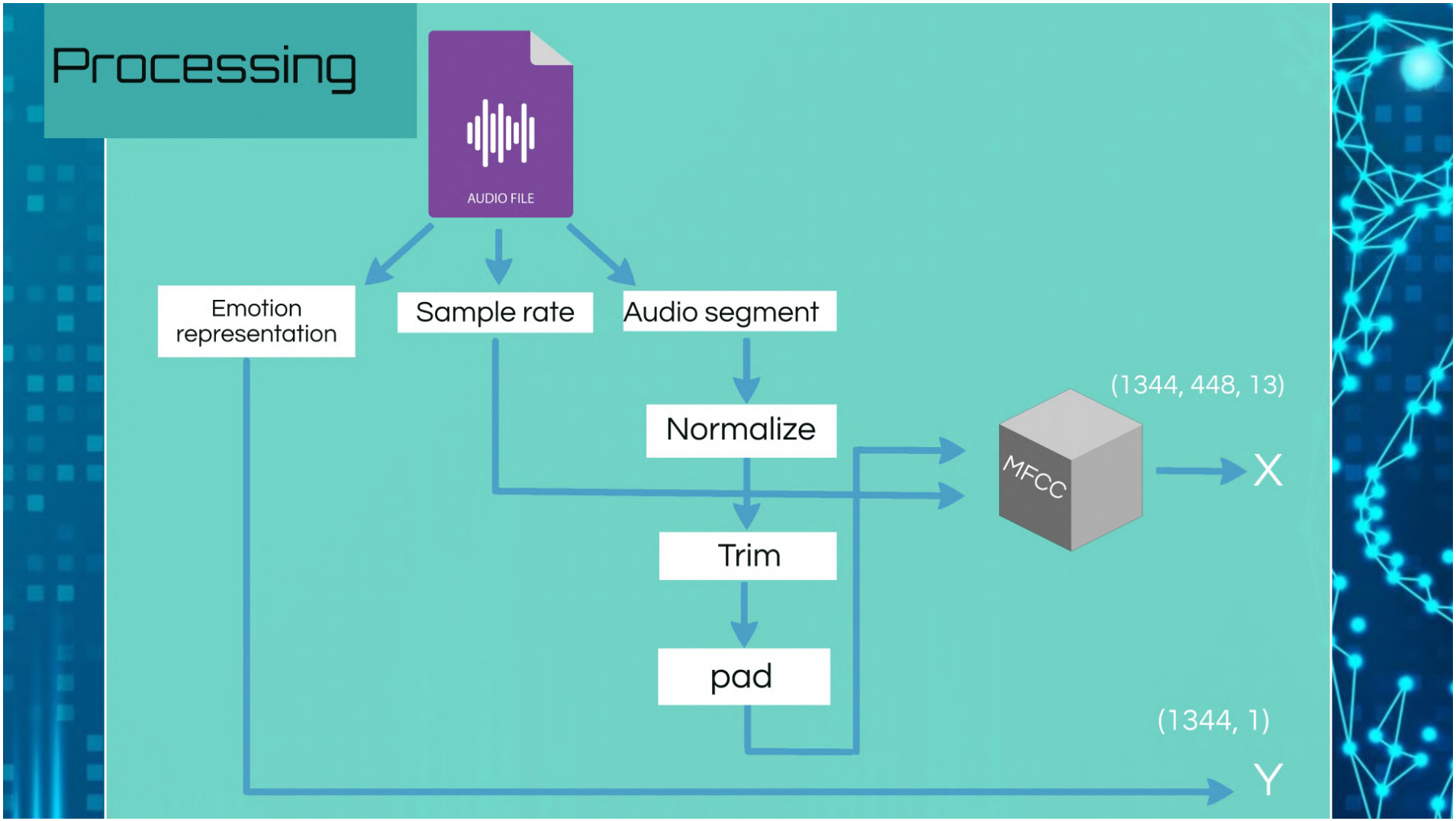

Padded audio
20000
10000
0
−10000
−20000
0
0.5
1
1.5
2
2.5
3
3.5
Time

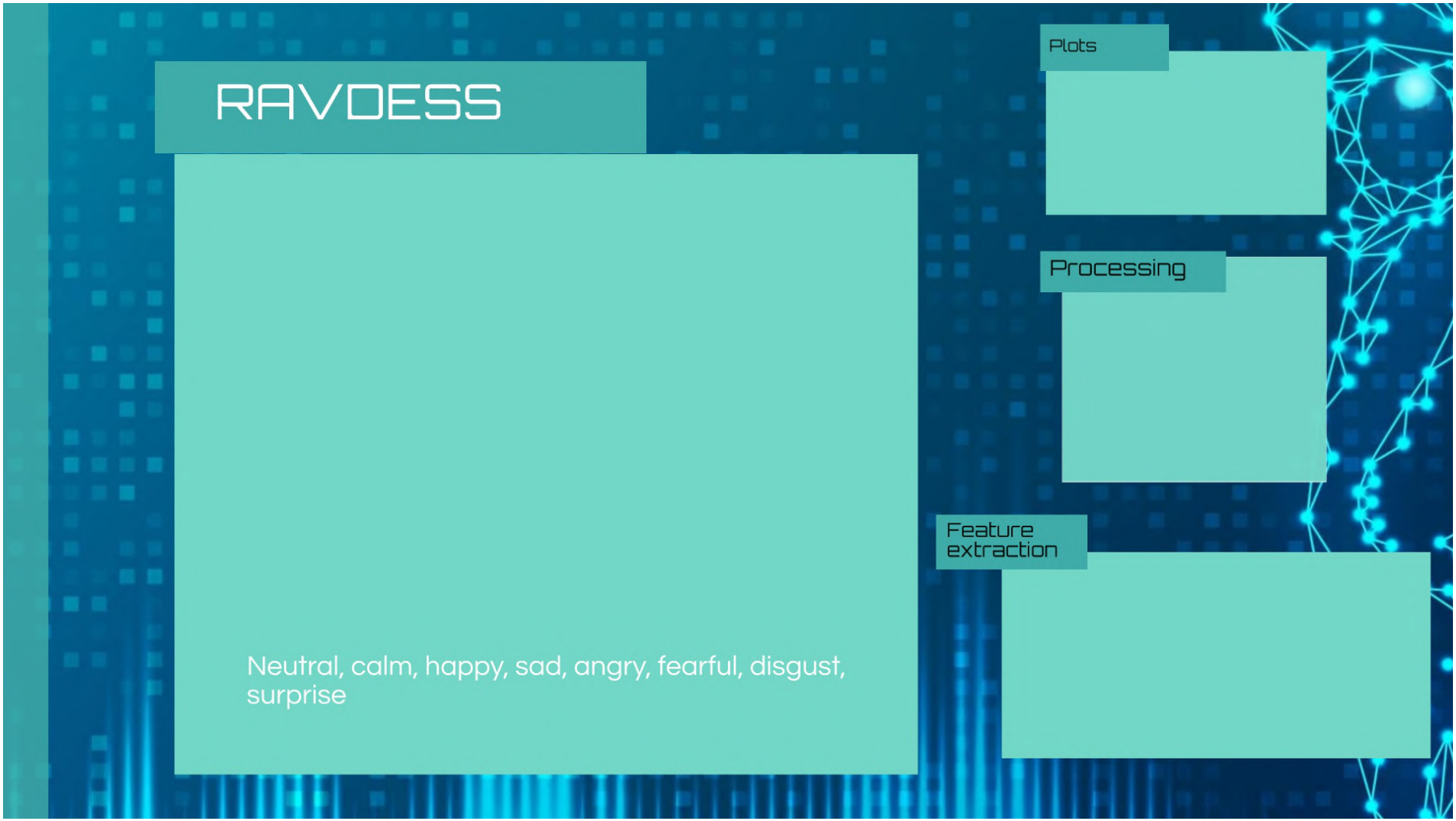

Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Trim
pad
Y


Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Trim
pad
Y


Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Trim
pad
Y


Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
MFCC
X
Trim
pad
Y

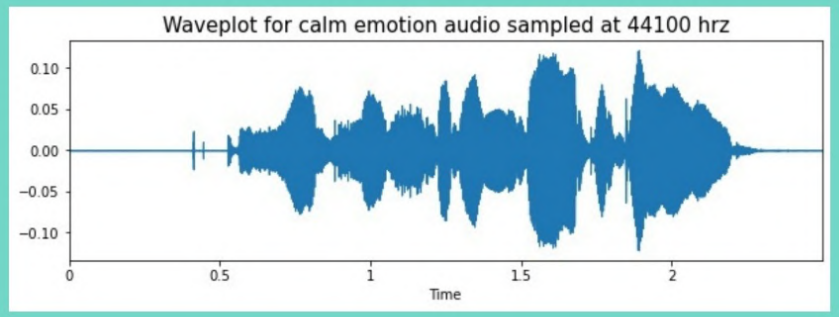

Processing
AUDIO FILE
Emotion representation
Sample rate
Audio segment
Normalize
Trim
pad
MFCC
(1344, 448, 13)
X
(1344, 1)
Y


RAVDESS
Neutral, calm, happy, sad, angry, fearful, disgust, surprise
Plots
Processing
Feature extraction

Feature extraction

Feature extraction

- Time-Domain
  - RMSE, ZCR

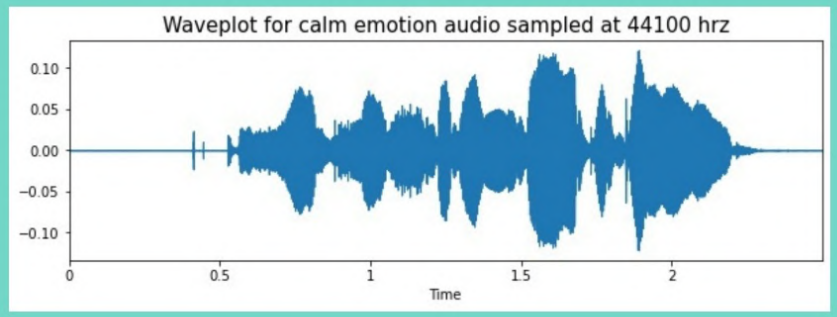


Mel-Spectrogram

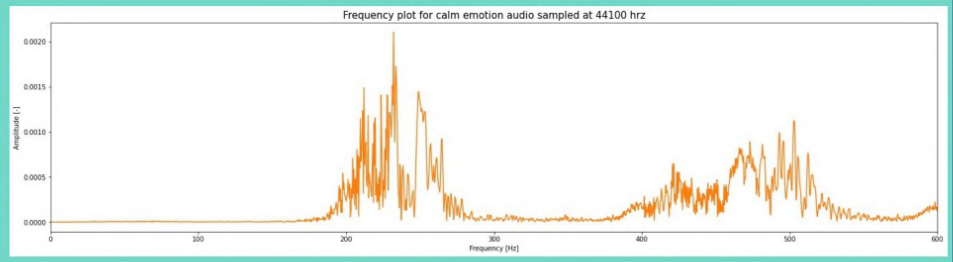
MFCC

Feature extraction

- Time-Domain
  - RMSE, ZCR



- Frequency Domain
  - Spectral Centroid



Mel-Spectrogram

MFCC

Mel Spectrogram

Mel Spectrogram

1. Extract STFT

Mel Spectrogram

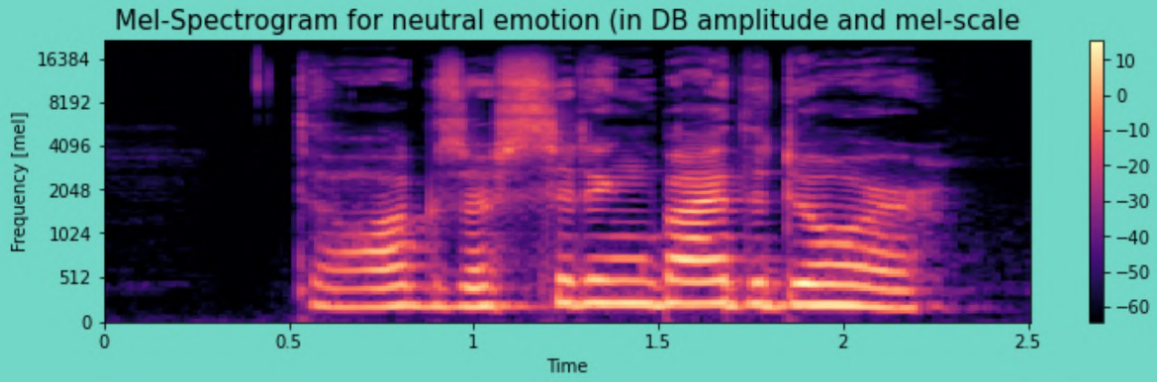
1. Extract STFT
2. Convert amplitude to DBs

Mel Spectrogram

1. Extract STFT
2. Convert amplitude to DBs
3. Convert frequencies to Mel Scale

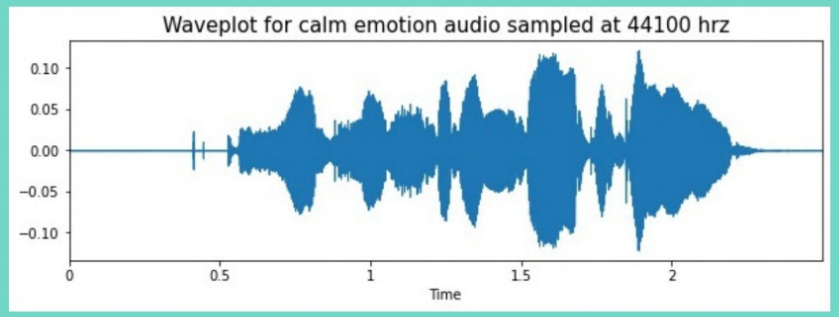
Mel Spectrogram

1. Extract STFT
2. Convert amplitude to DBs
3. Convert frequencies to Mel Scale

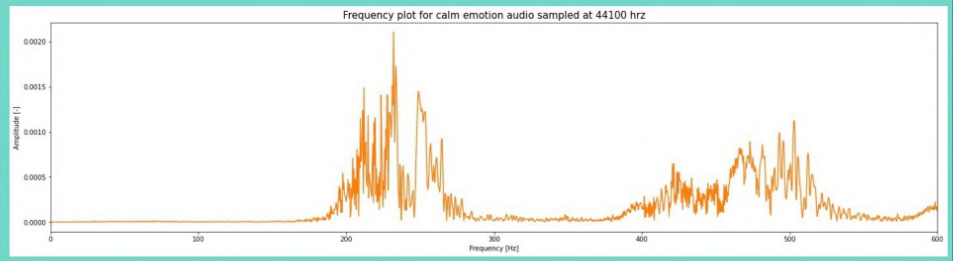


Feature extraction

- Time-Domain
  - RMSE, ZCR



- Frequency Domain
  - Spectral Centroid



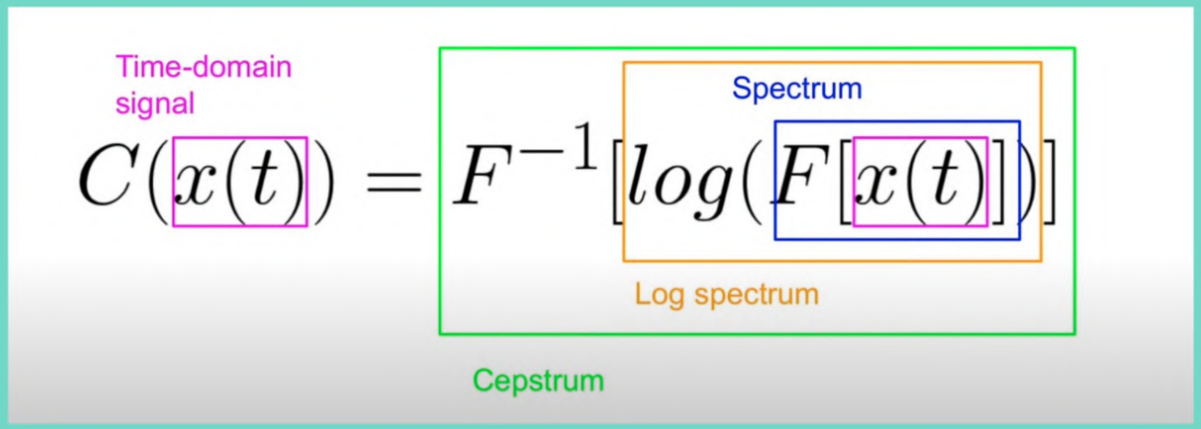
Mel-Spectrogram

MFCC

Mel-Frequecy Cepstrum Coefficients (MFCCs)

Mel-Frequecy Cepstrum Coefficients (MFCCs)

Cepstrum



Visualize

Speech Generation

Cepstrum and Speech

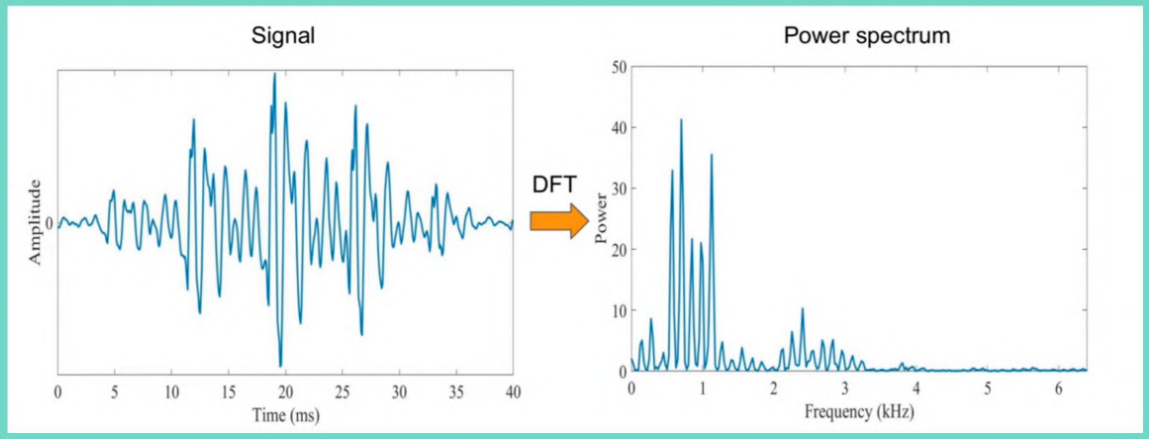
Compute

$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

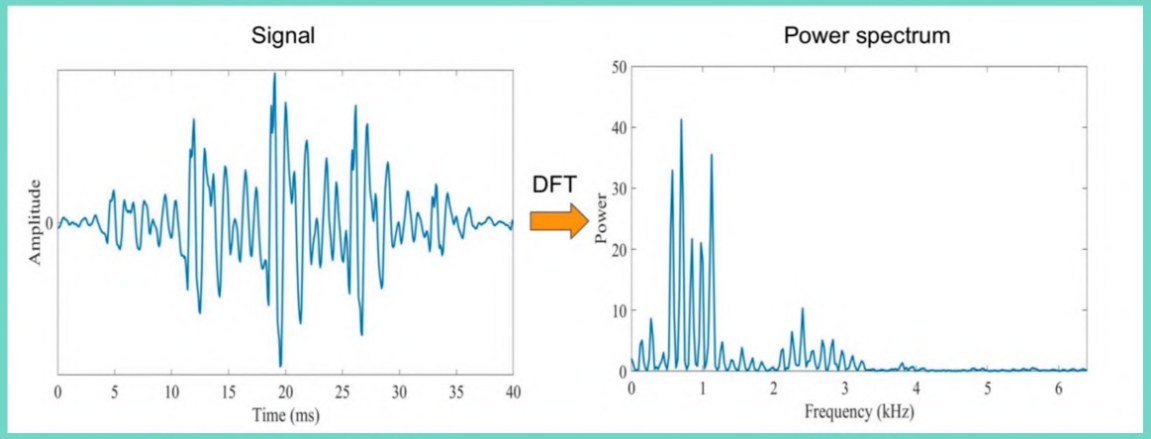
$x(t)$

$F[x(t)]$

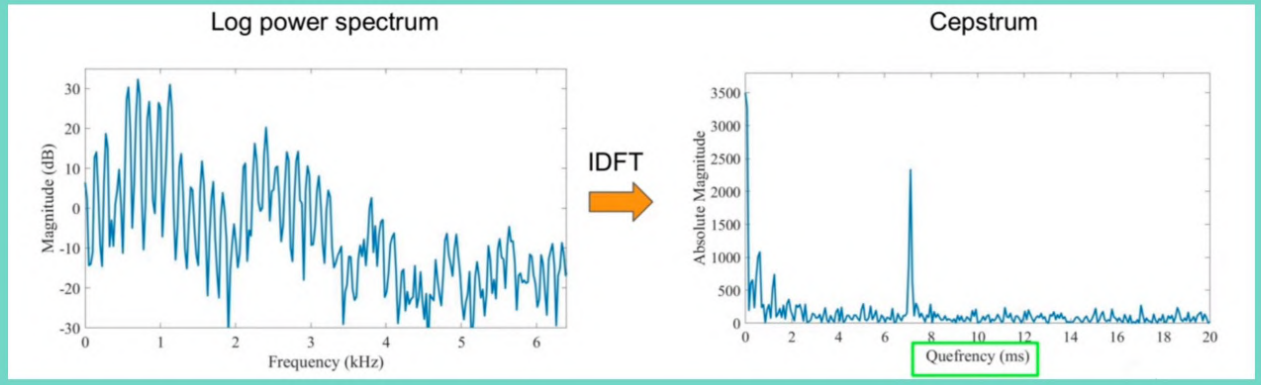


$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

$x(t)$ $F[x(t)]$



$log(F[x(t)])$ F^{-1}



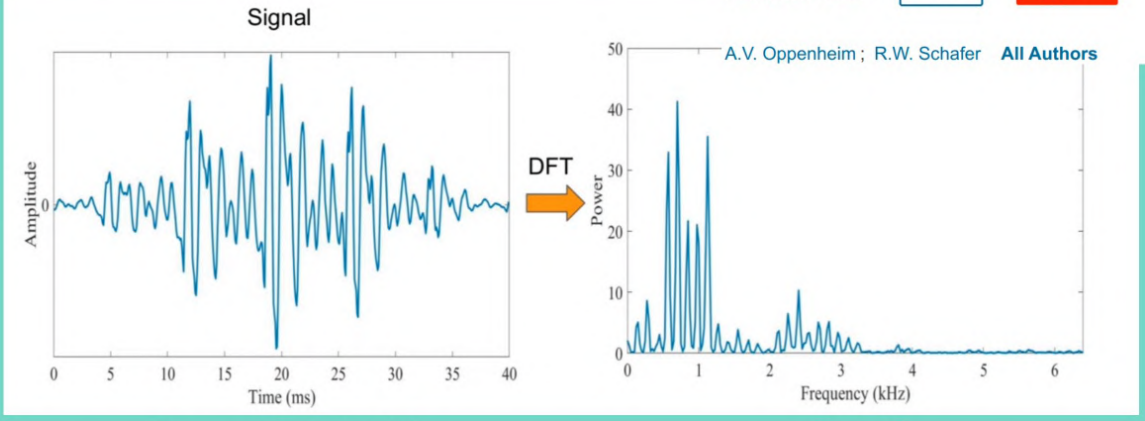
$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

$x(t)$

From frequency to quefrency: a history of the cepstrum

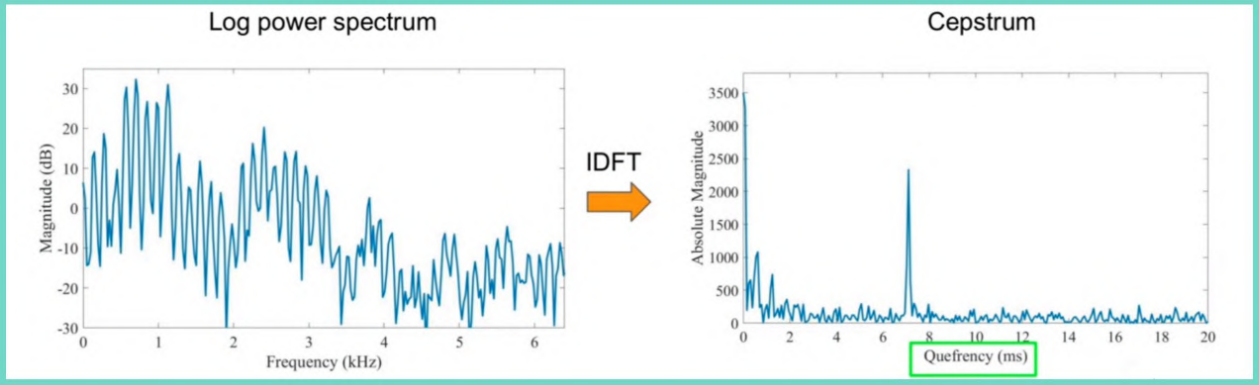
Publisher: IEEE | Cite This | PDF

A.V. Oppenheim ; R.W. Schafer All Authors



$log(F[x(t)])$

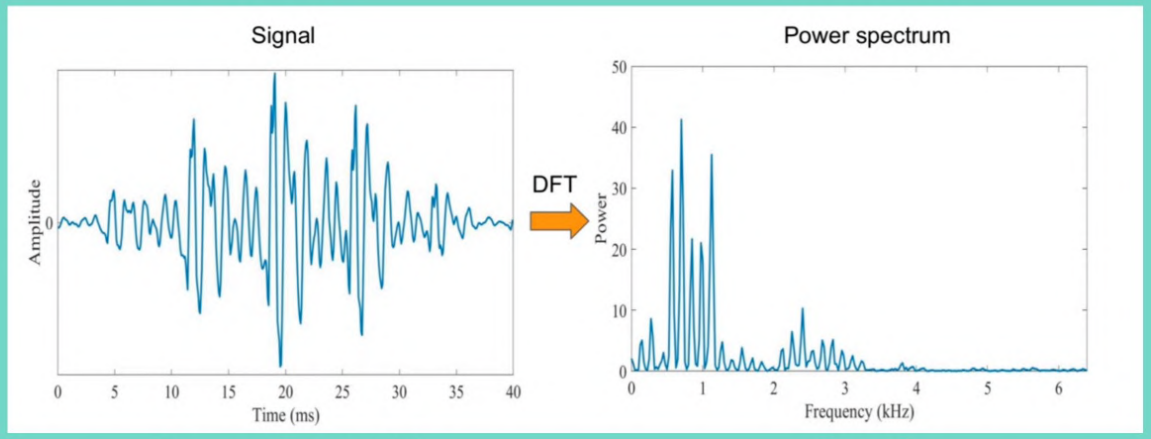
F^{-1}



$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

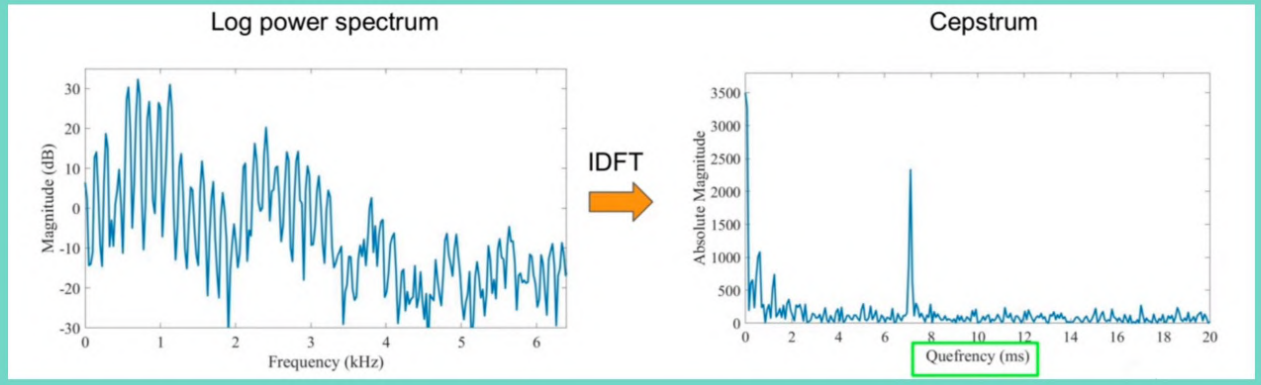
$x(t)$

$F[x(t)]$



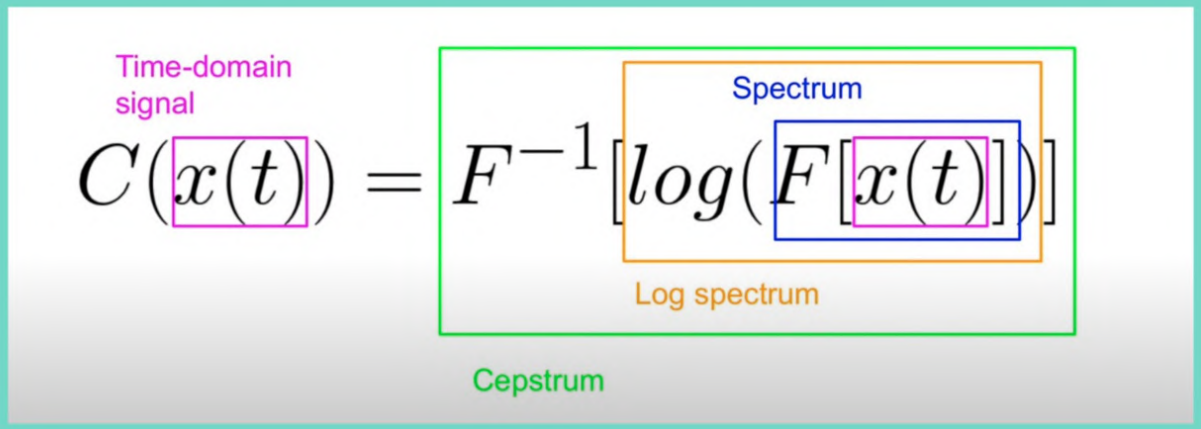
$log(F[x(t)])$

F^{-1}



Mel-Frequecy Cepstrum Coefficients (MFCCs)

Cepstrum

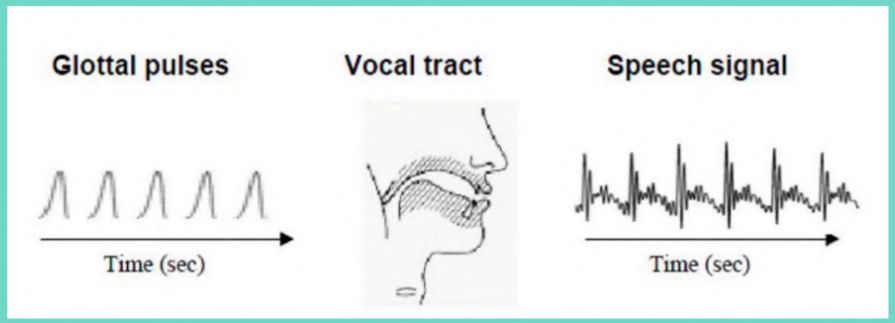


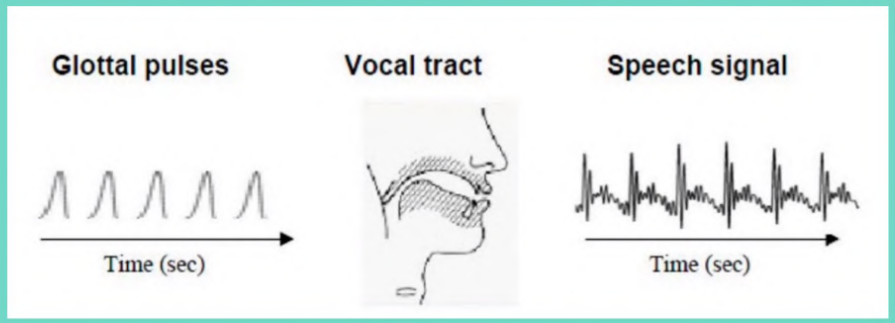
Visualize

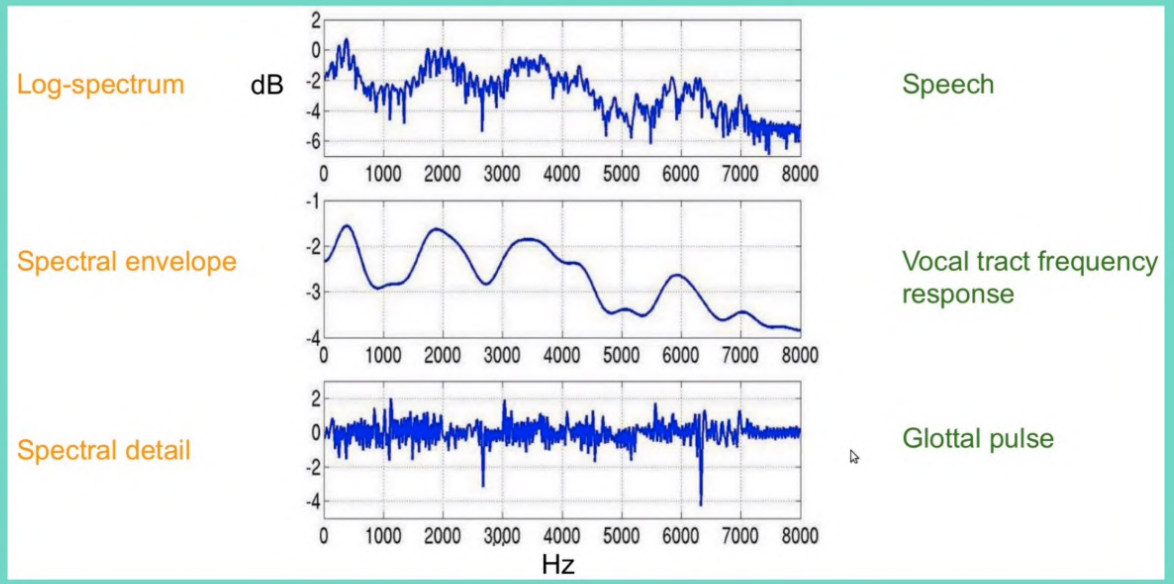
Speech Generation

Cepstrum and Speech

Compute

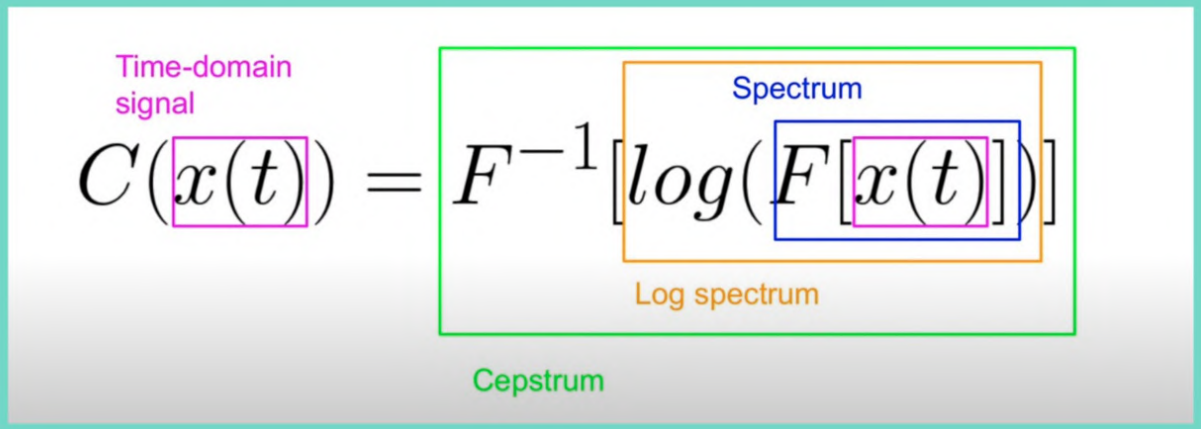

Glottal pulses
Vocal tract
Speech signal
Time (sec)
Time (sec)


Glottal pulses
Vocal tract
Speech signal
Time (sec)
Time (sec)


Log-spectrum
dB
Spectral envelope
Spectral detail
Hz
Speech
Vocal tract frequency response
Glottal pulse

Mel-Frequecy Cepstrum Coefficients (MFCCs)

Cepstrum



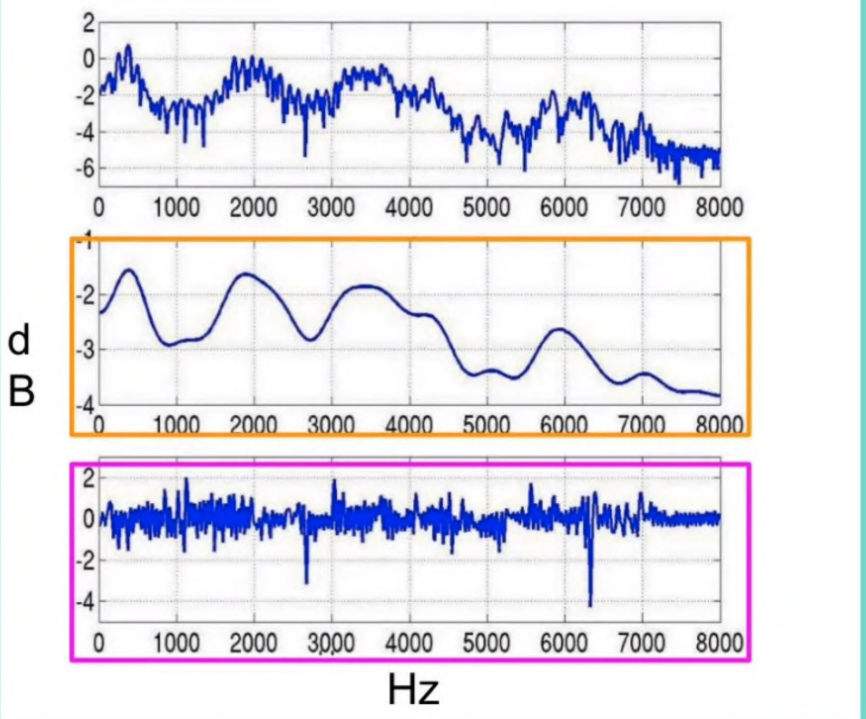
Visualize

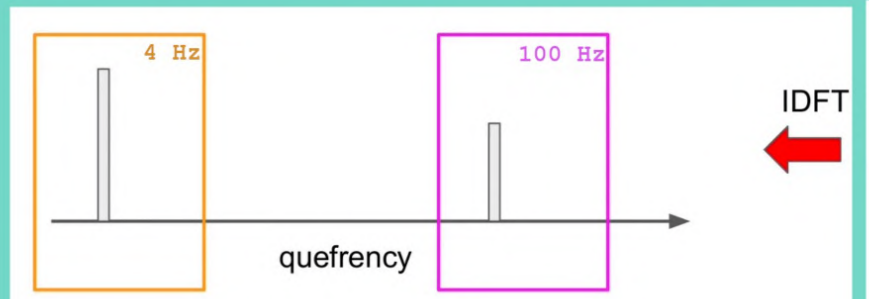
Speech Generation

Cepstrum and Speech

Compute

$$log(X(t)) = log(E(t)) + log(H(t))$$

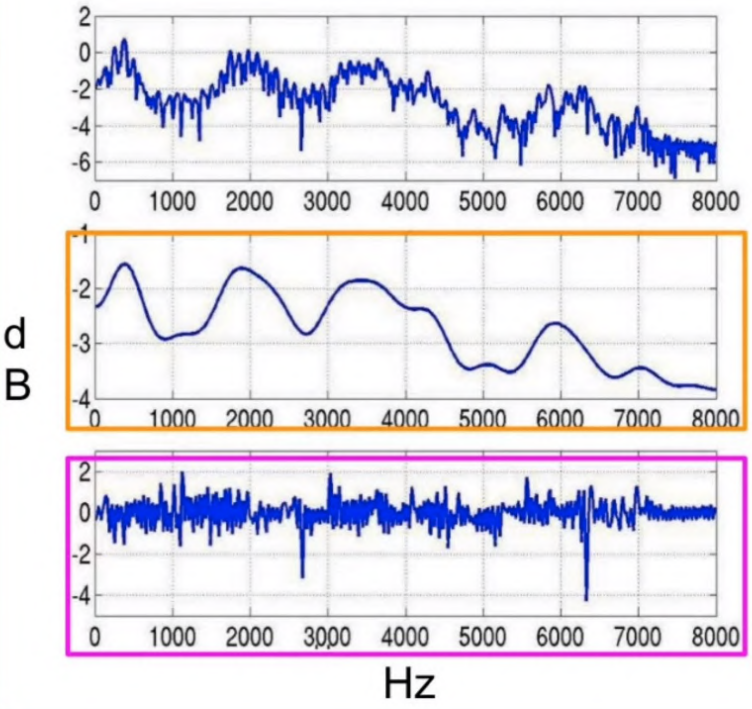




IDFT

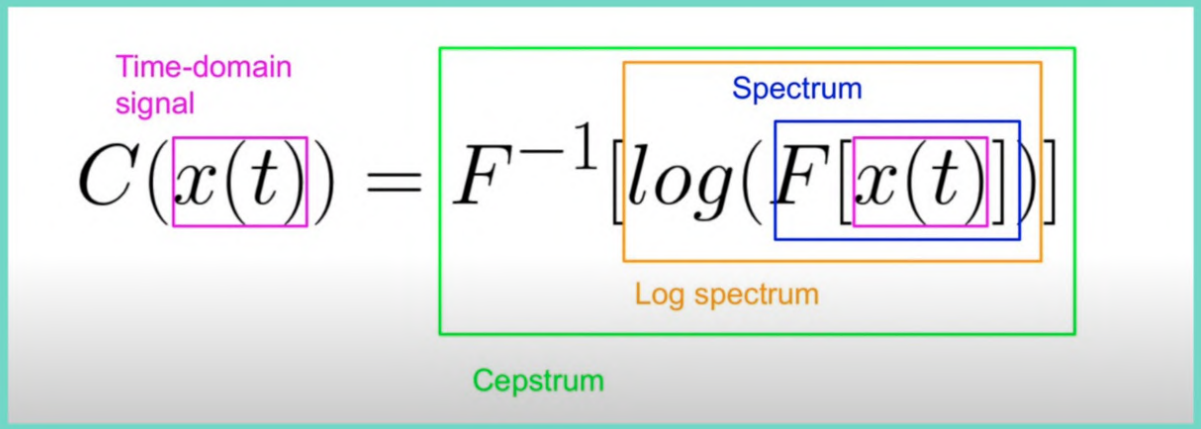
$$X(t) = E(t) \cdot H(t)$$

$$log(X(t)) = log(E(t)) + log(H(t))$$



Mel-Frequecy Cepstrum Coefficients (MFCCs)

Cepstrum

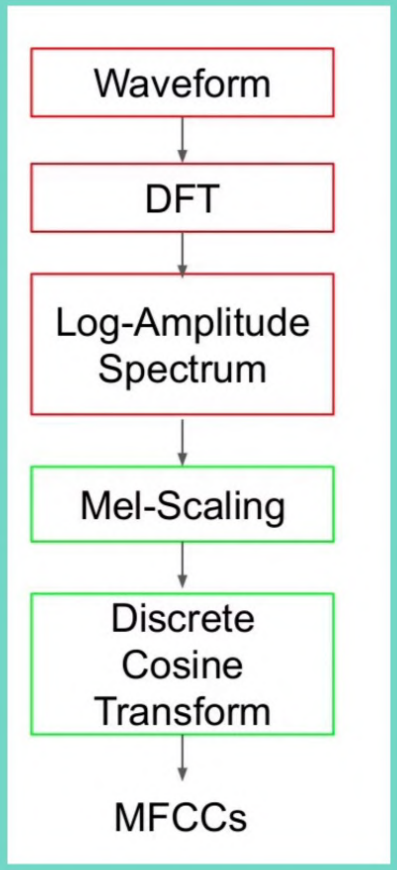


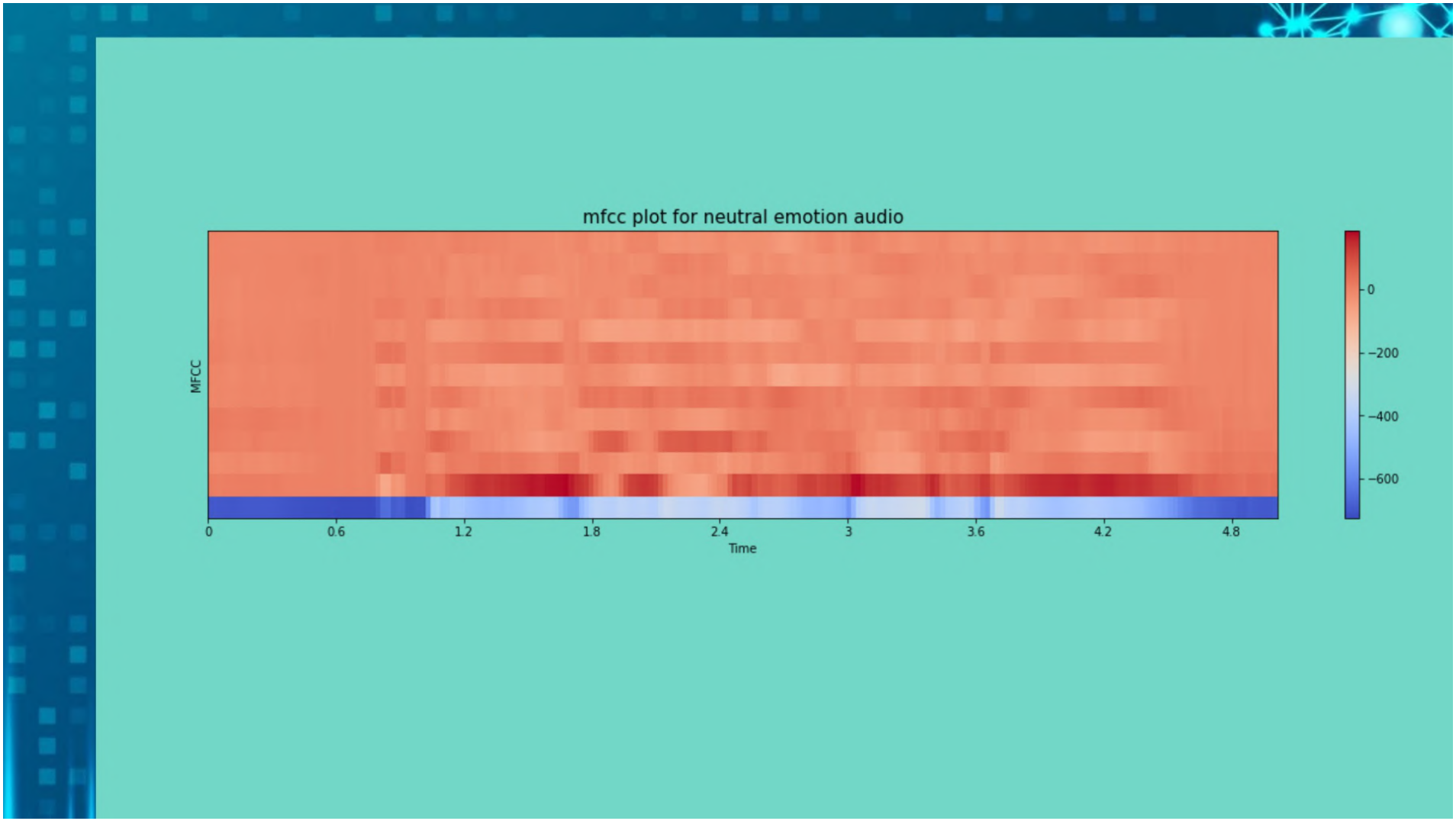
Visualize

Speech Generation

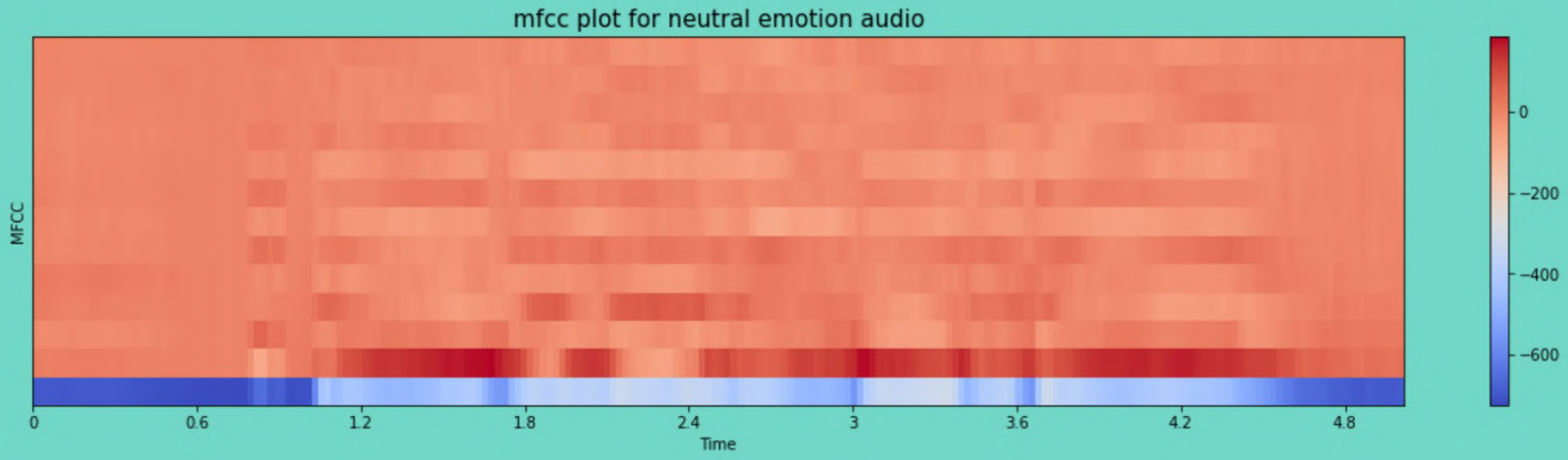
Cepstrum and Speech

Compute


Waveform
DFT
Log-Amplitude Spectrum
Mel-Scaling
Discrete Cosine Transform
MFCCs

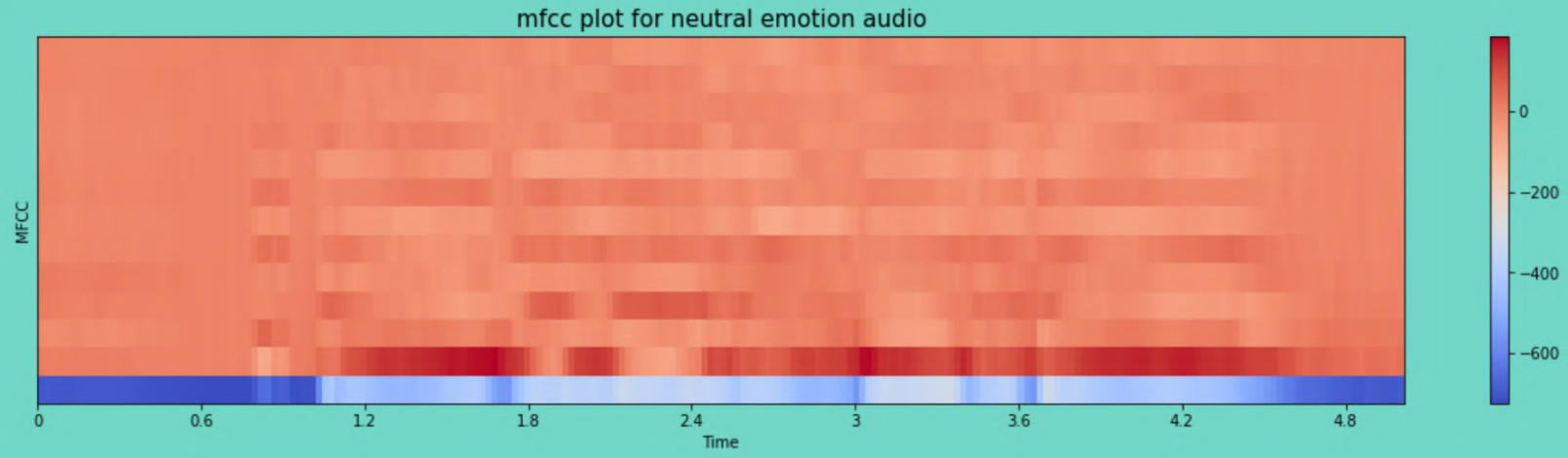

mfcc plot for neutral emotion audio
MFCC
Time
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
0
−200
−400
−600

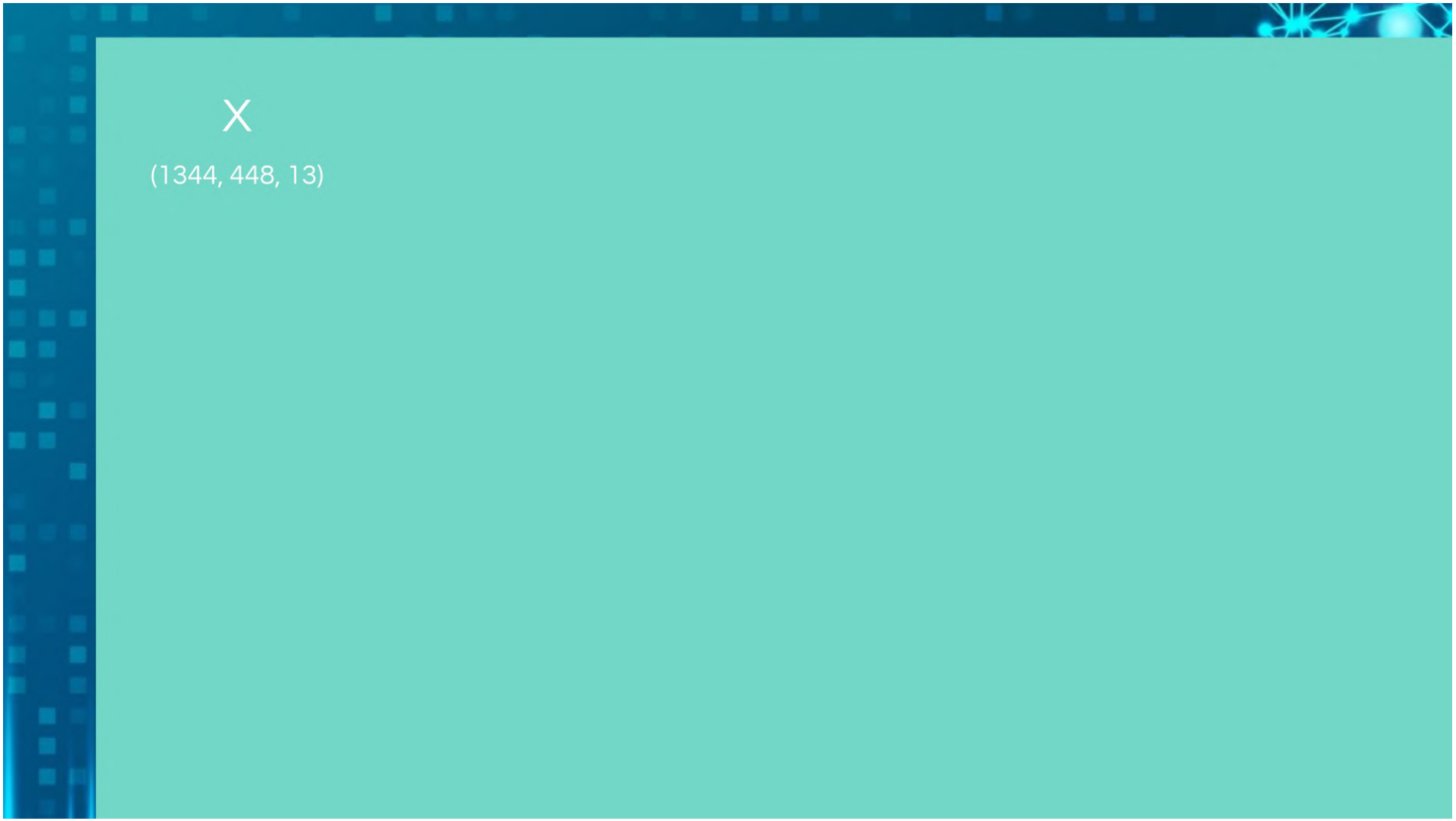
X


mfcc plot for neutral emotion audio
MFCC
Time
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
0
−200
−400
−600

X

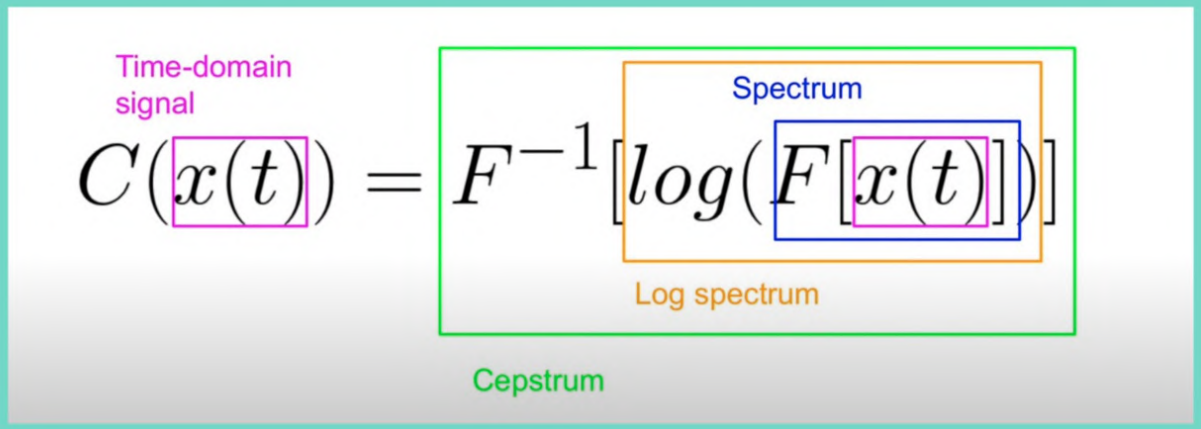
(1344, 448, 13)


mfcc plot for neutral emotion audio
MFCC
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time
0
−200
−400
−600


X
(1344, 448, 13)

Mel-Frequecy Cepstrum Coefficients (MFCCs)

Cepstrum



Visualize

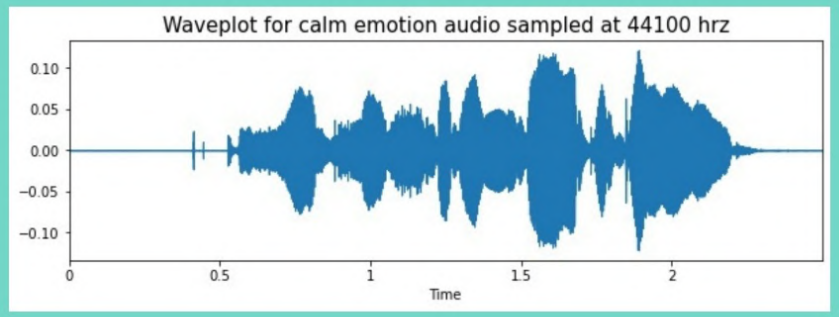
Speech Generation

Cepstrum and Speech

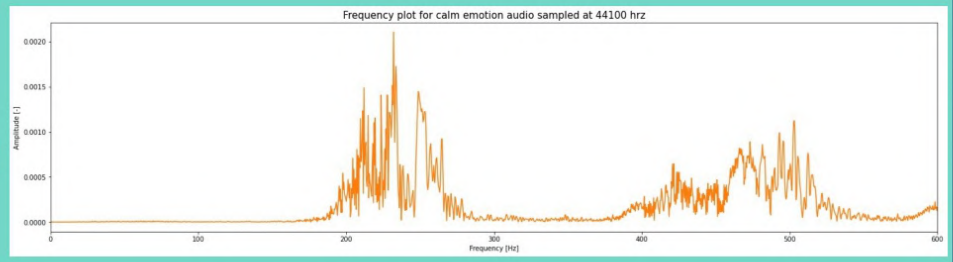
Compute

Feature extraction

- Time-Domain
  - RMSE, ZCR

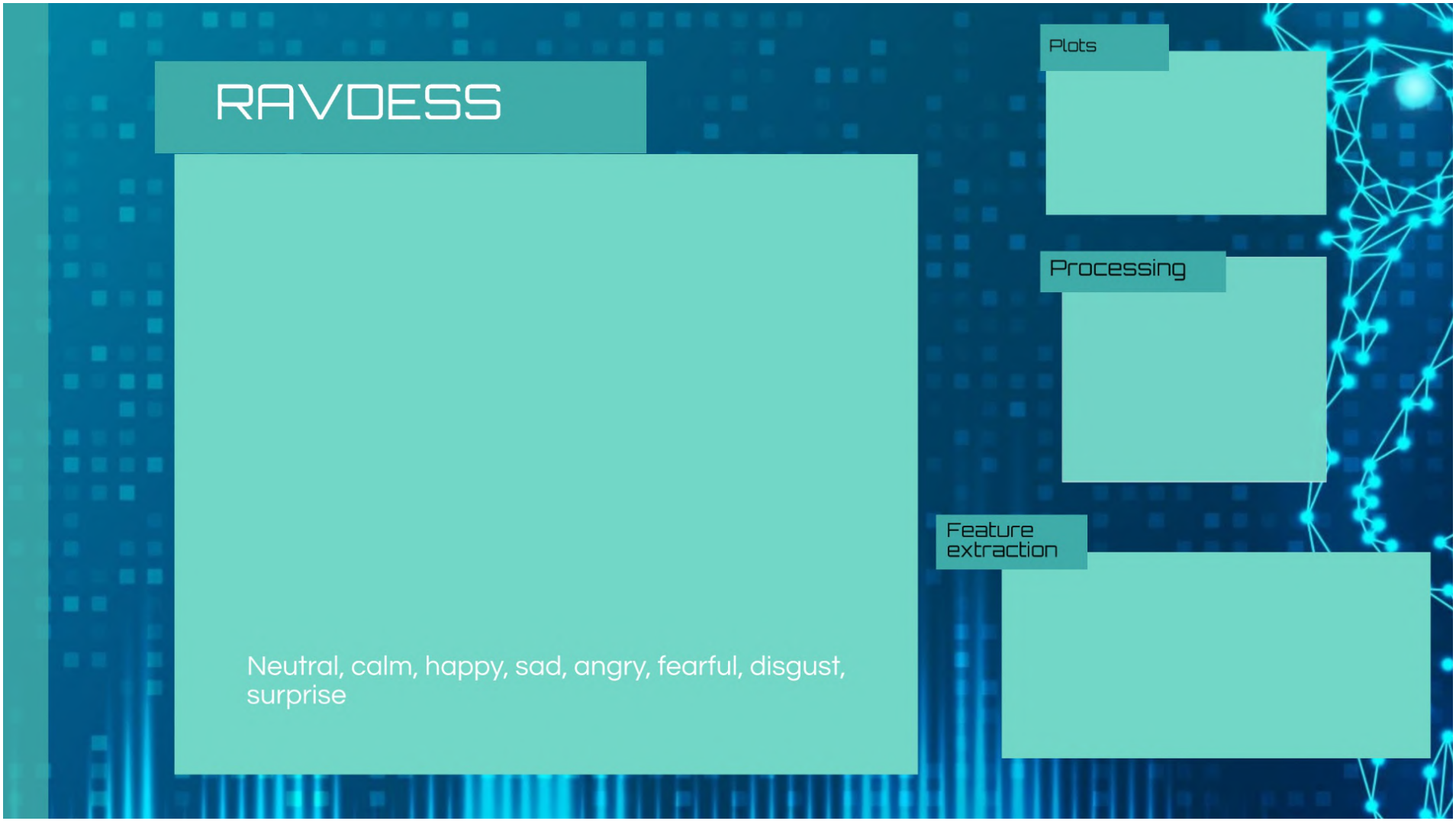


- Frequency Domain
  - Spectral Centroid



Mel-Spectrogram

MFCC


RAVDESS
Neutral, calm, happy, sad, angry, fearful, disgust, surprise
Plots
Processing
Feature extraction

Introduction

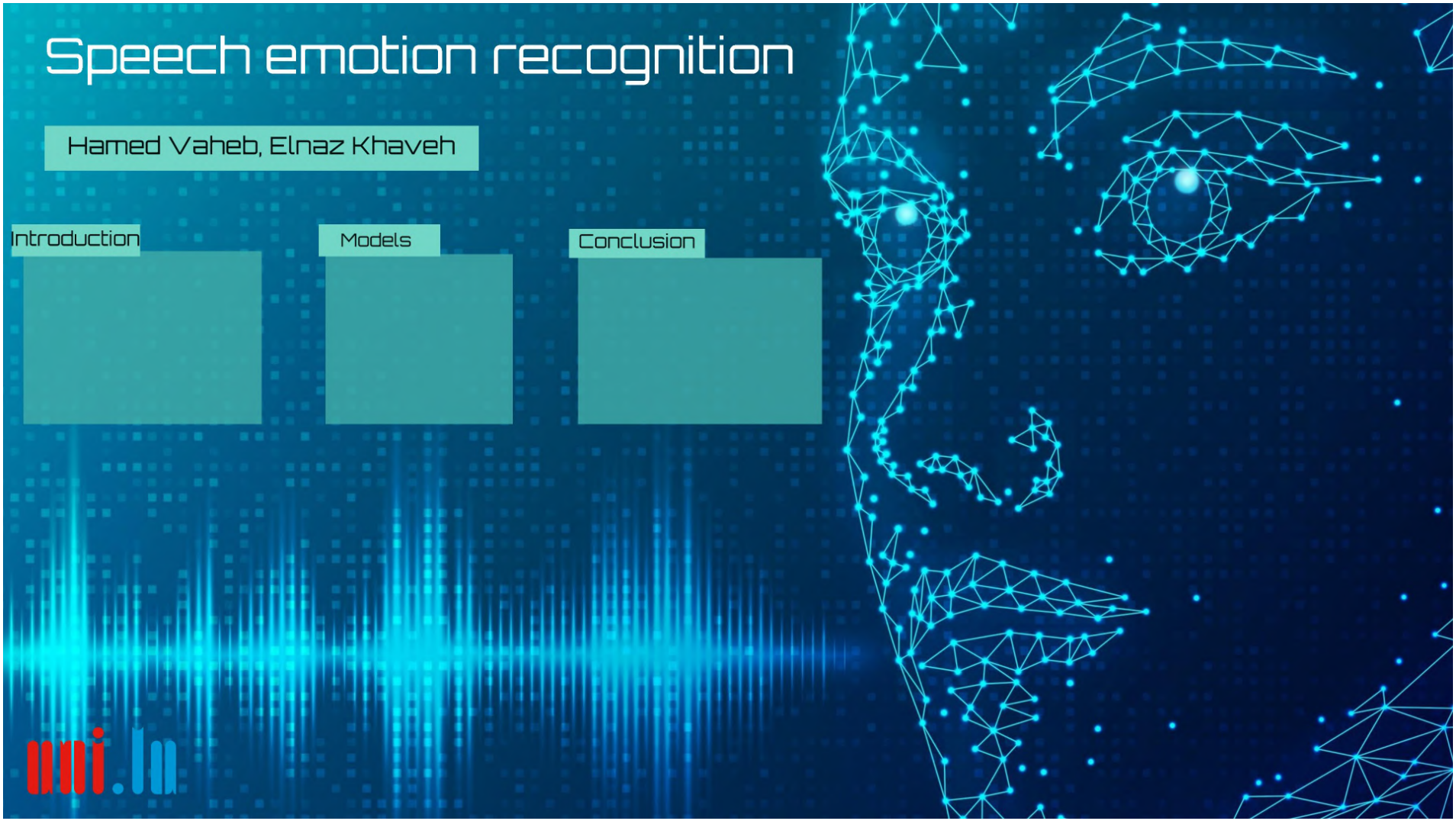
Importance:


The Evolution of Human Vocal Emotion
Gregory A. Bryant
First Published June 24, 2020 | Research Article
https://doi.org/10.1177/1754073920930791
Article information

- The most natural way to express ourselves
- Influence over judgement
  - "go home!"
- Human-computer interaction (HCI)
  - Personal Assistants
  - Next Generation AI

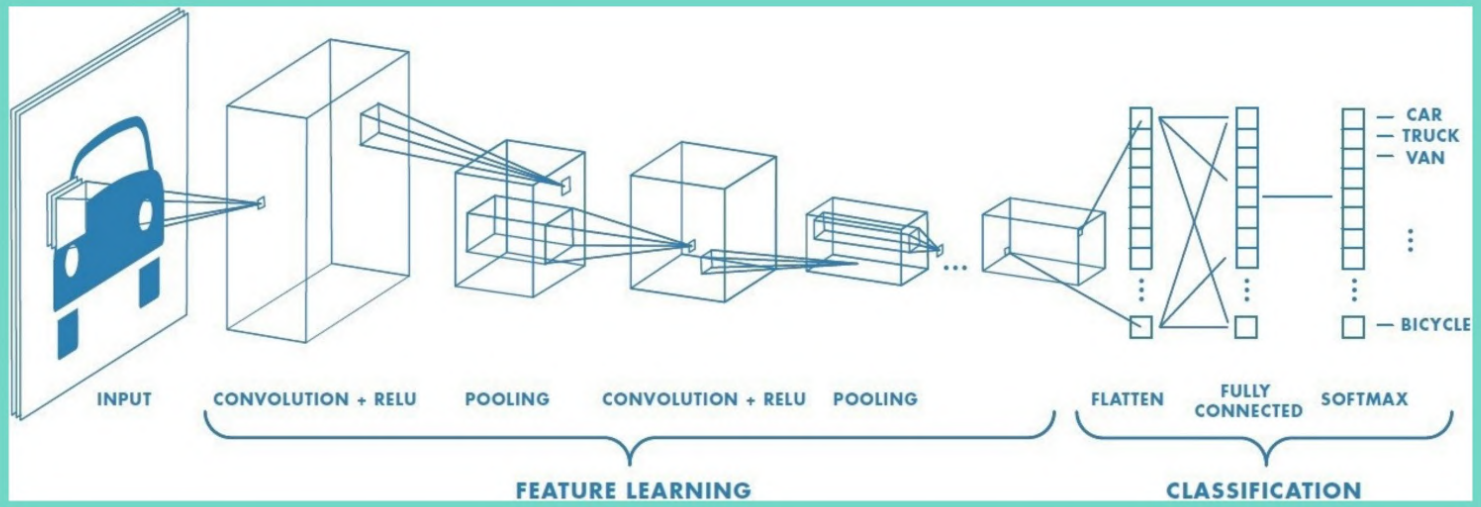
Applications

Data set

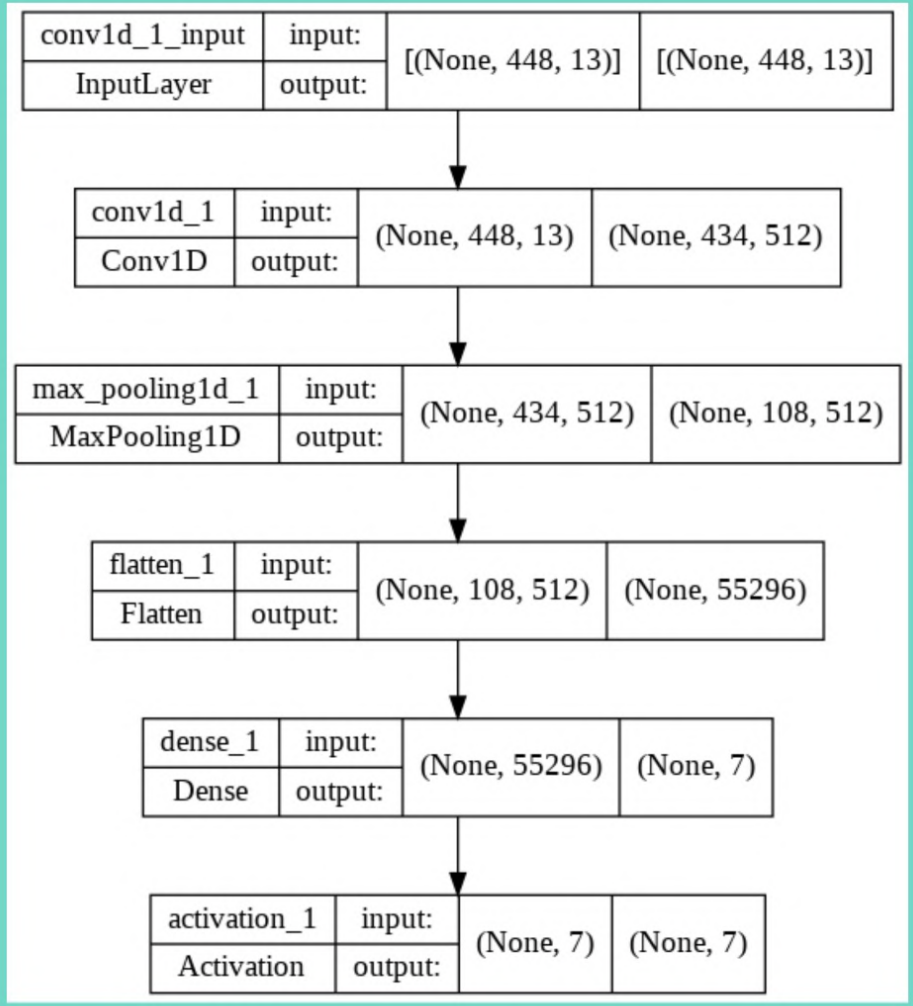

Speech emotion recognition
Hamed Vaheb, Elnaz Khaveh
Introduction
Models
Conclusion


Models
CNN
RNN
Results

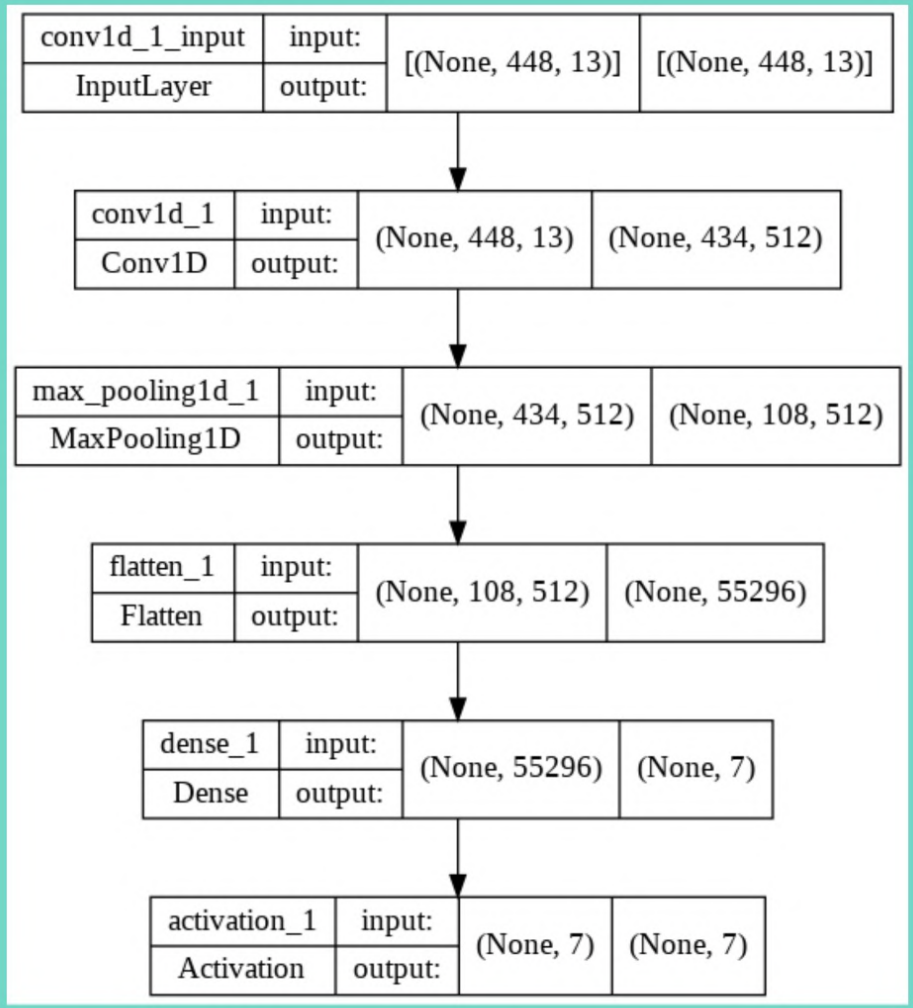
CNN



architecture

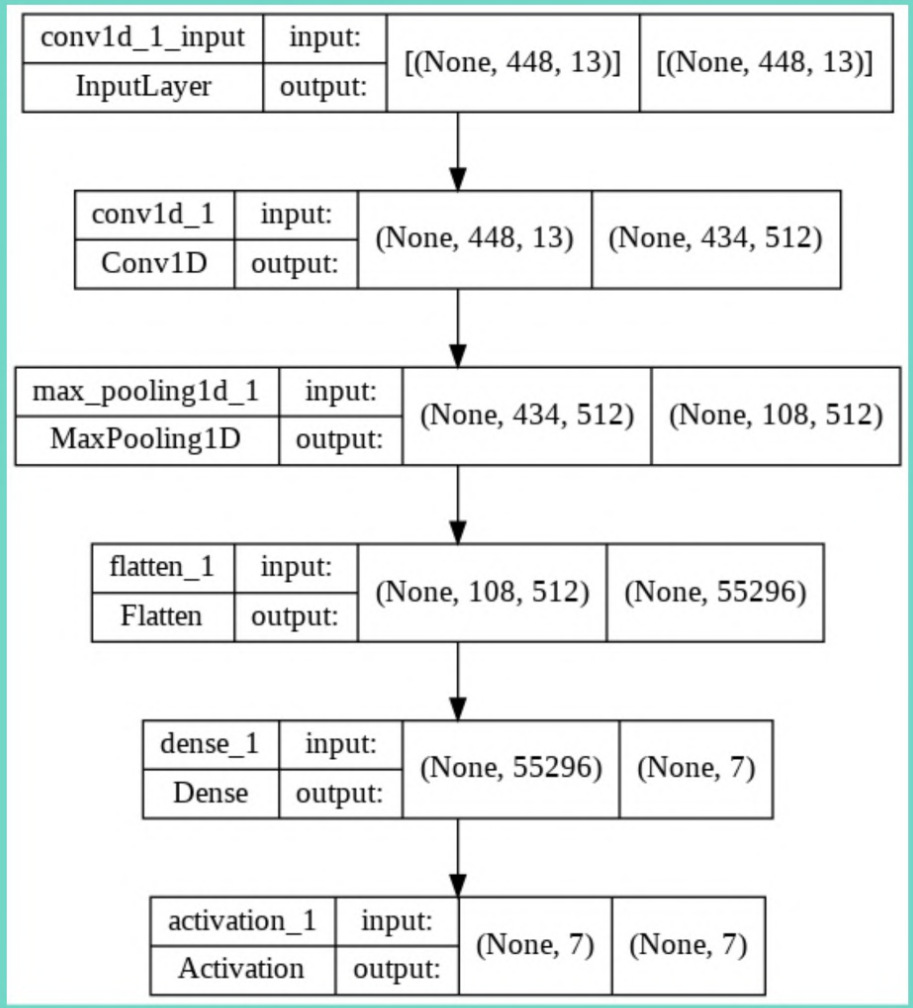

conv1d_1_input
InputLayer
input:
output:
[(None, 448, 13)]
[(None, 448, 13)]
conv1d_1
Conv1D
input:
output:
(None, 448, 13)
(None, 434, 512)
max_pooling1d_1
MaxPooling1D
input:
output:
(None, 434, 512)
(None, 108, 512)
flatten_1
Flatten
input:
output:
(None, 108, 512)
(None, 55296)
dense_1
Dense
input:
output:
(None, 55296)
(None, 7)
activation_1
Activation
input:
output:
(None, 7)
(None, 7)

X

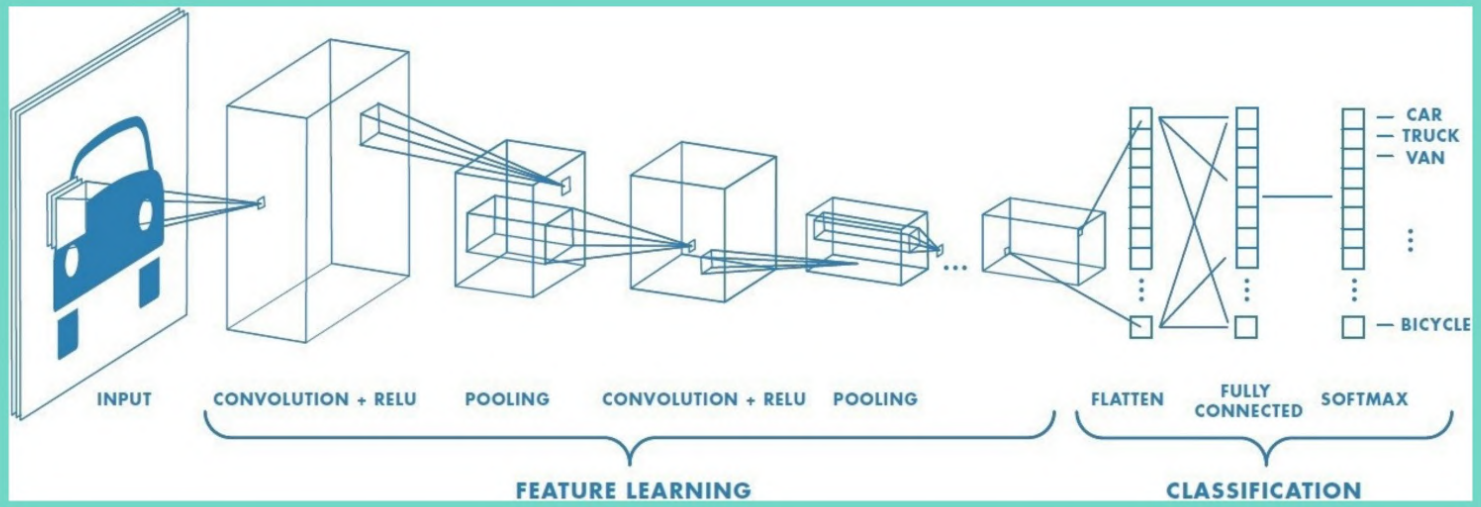

conv1d_1_input
input:
InputLayer
output:
[(None, 448, 13)]
[(None, 448, 13)]
conv1d_1
input:
Conv1D
output:
(None, 448, 13)
(None, 434, 512)
max_pooling1d_1
input:
MaxPooling1D
output:
(None, 434, 512)
(None, 108, 512)
flatten_1
input:
Flatten
output:
(None, 108, 512)
(None, 55296)
dense_1
input:
Dense
output:
(None, 55296)
(None, 7)
activation_1
input:
Activation
output:
(None, 7)
(None, 7)

X

(1344, 448, 13)


conv1d_1_input
InputLayer
input:
output:
[(None, 448, 13)]
[(None, 448, 13)]
conv1d_1
Conv1D
input:
output:
(None, 448, 13)
(None, 434, 512)
max_pooling1d_1
MaxPooling1D
input:
output:
(None, 434, 512)
(None, 108, 512)
flatten_1
Flatten
input:
output:
(None, 108, 512)
(None, 55296)
dense_1
Dense
input:
output:
(None, 55296)
(None, 7)
activation_1
Activation
input:
output:
(None, 7)
(None, 7)

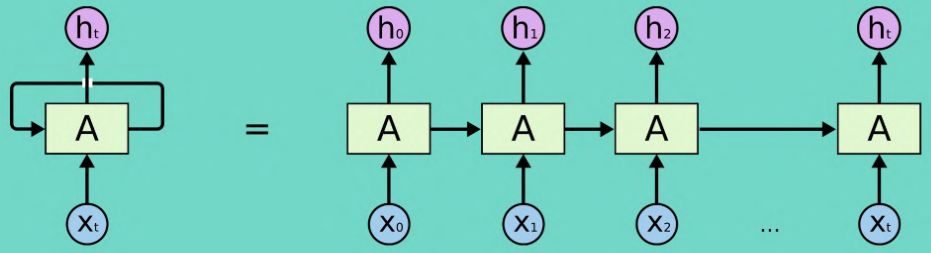
CNN



architecture


Models
CNN
RNN
Results

RNN



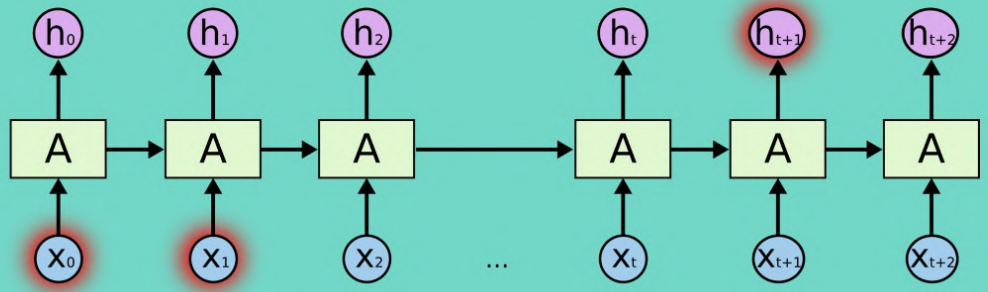
GRU

architecture

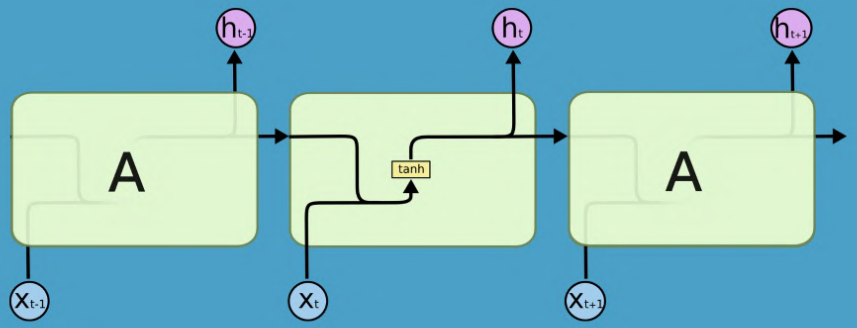

RNN
GRU
architecture

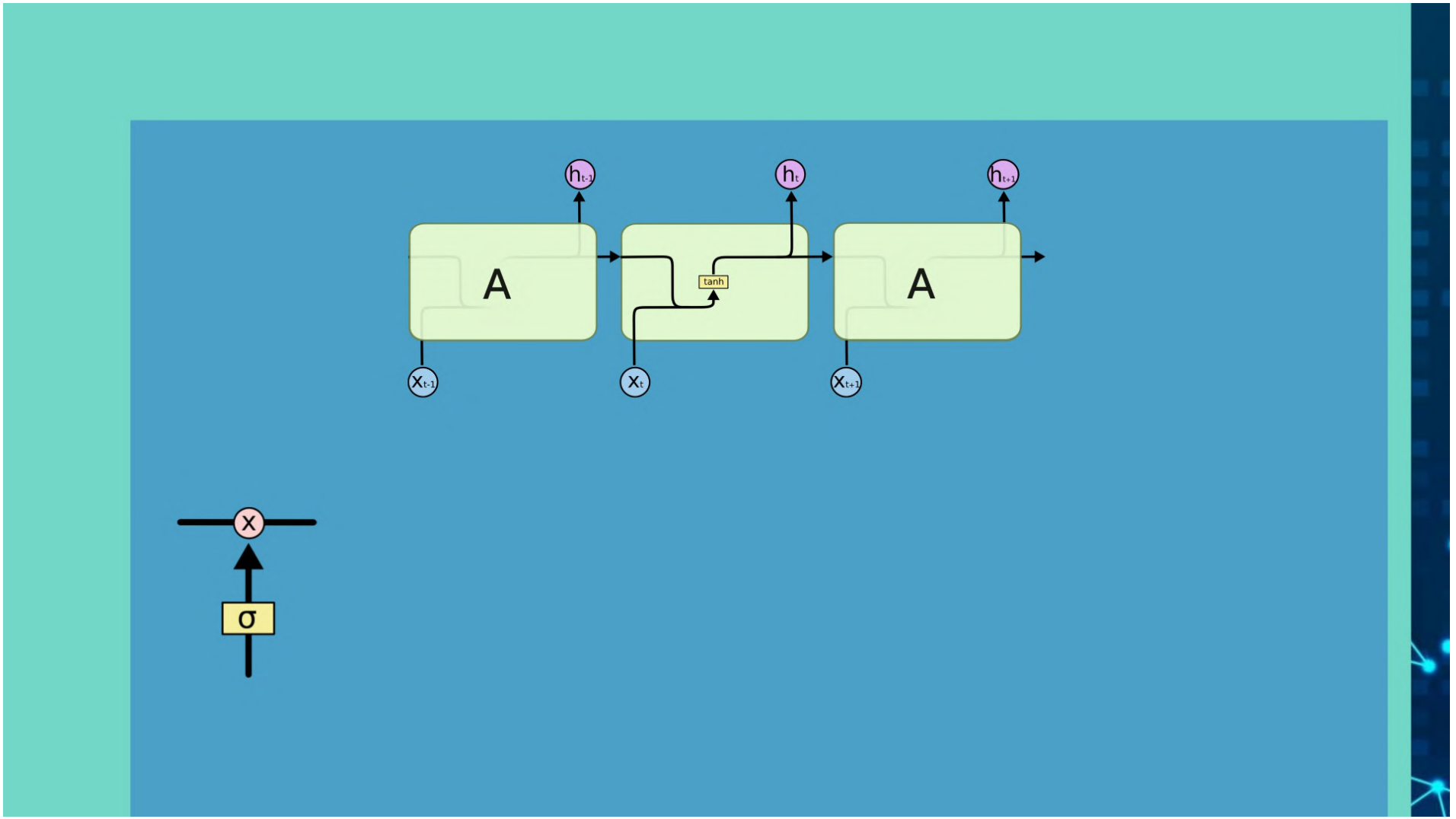
RNN

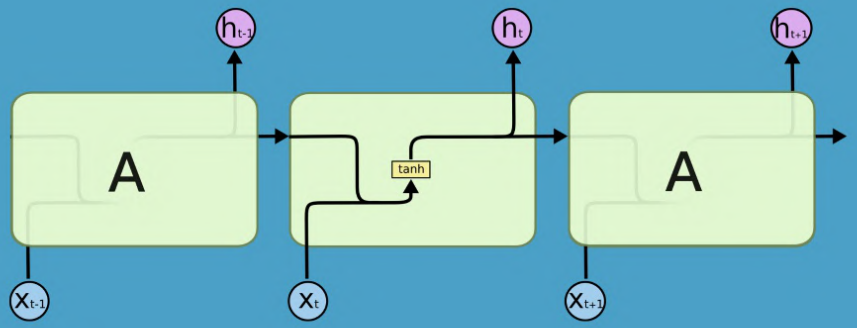
GRU

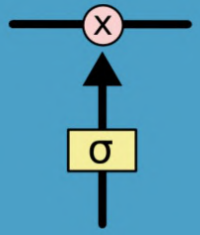


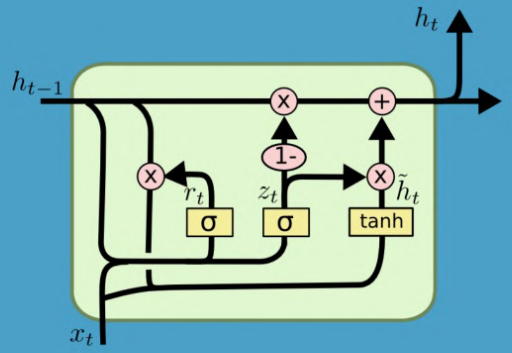
architecture


h_{t-1}
h_t
h_{t+1}
A
tanh
A
x_{t-1}
x_t
x_{t+1}


h_{t-1}
h_t
h_{t+1}
A
tanh
A
x_{t-1}
x_t
x_{t+1}
x
σ







$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

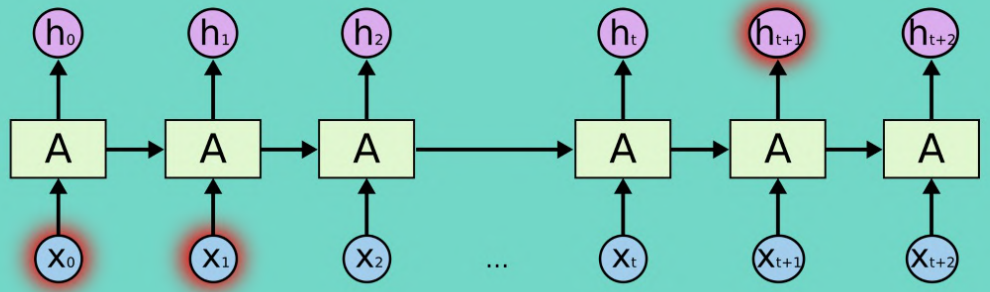
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

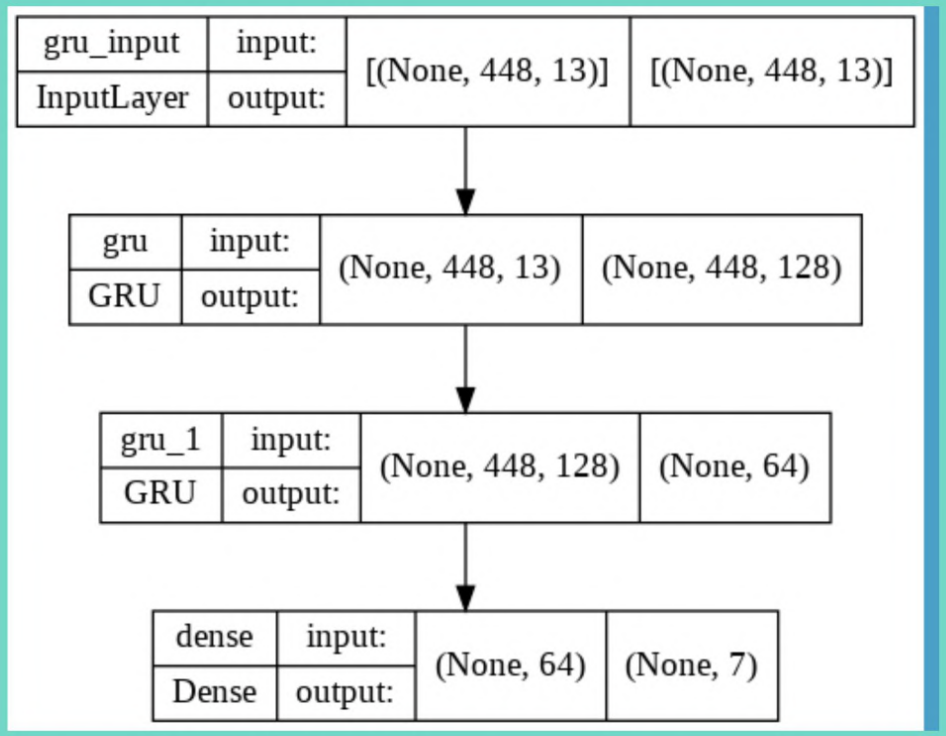
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

RNN

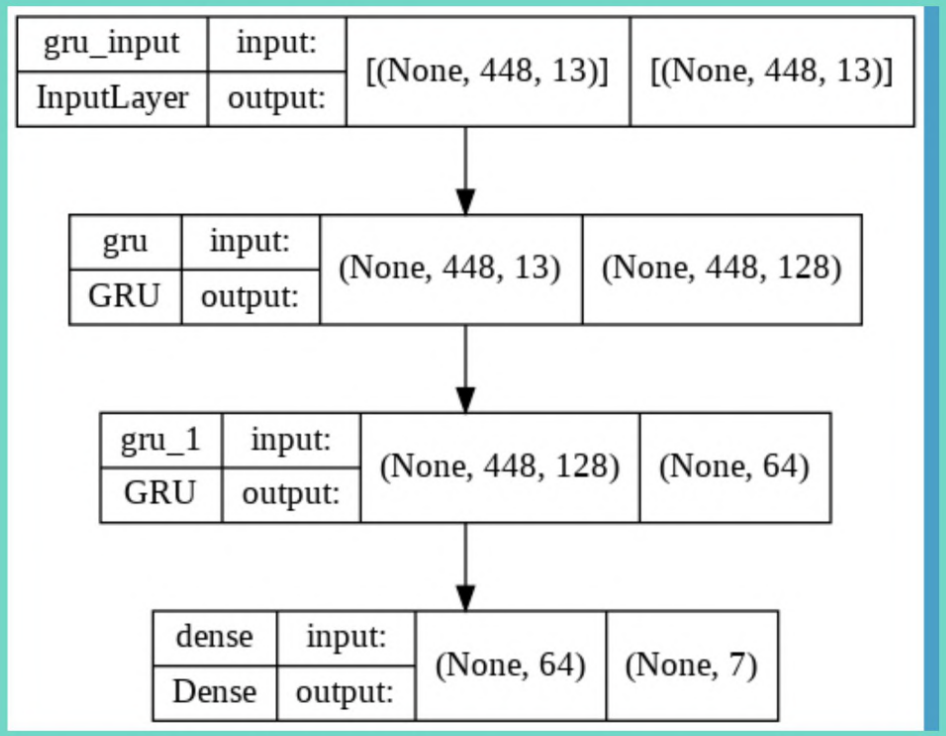
GRU



architecture

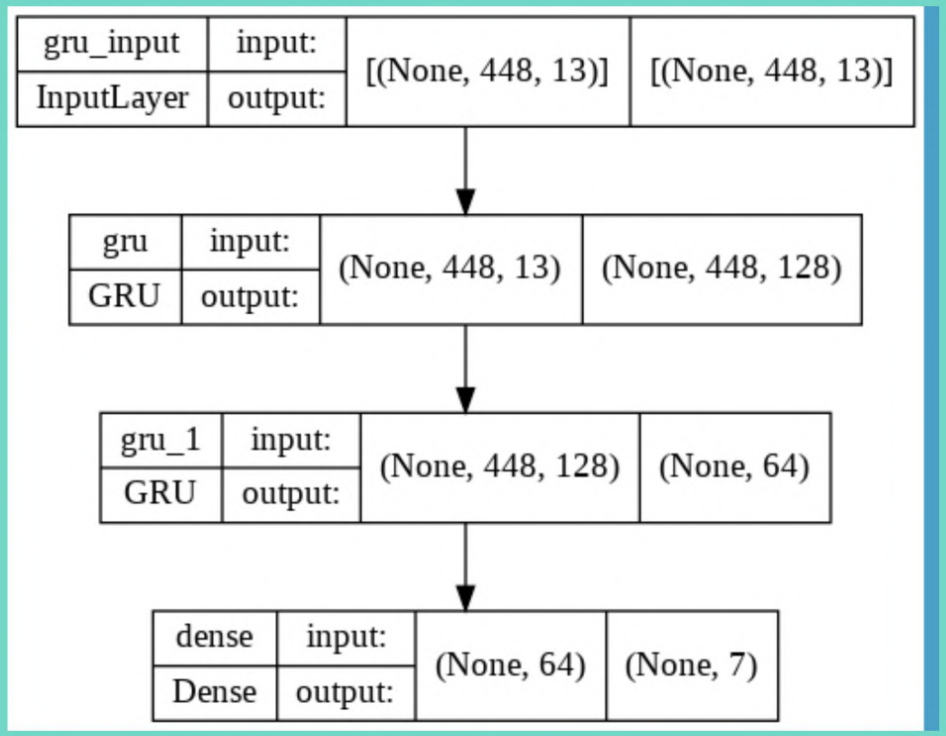

gru_input
InputLayer
input:
output:
[(None, 448, 13)]
[(None, 448, 13)]
gru
GRU
input:
output:
(None, 448, 13)
(None, 448, 128)
gru_1
GRU
input:
output:
(None, 448, 128)
(None, 64)
dense
Dense
input:
output:
(None, 64)
(None, 7)

X


gru_input
InputLayer
input:
output:
[(None, 448, 13)]
[(None, 448, 13)]
gru
GRU
input:
output:
(None, 448, 13)
(None, 448, 128)
gru_1
GRU
input:
output:
(None, 448, 128)
(None, 64)
dense
Dense
input:
output:
(None, 64)
(None, 7)

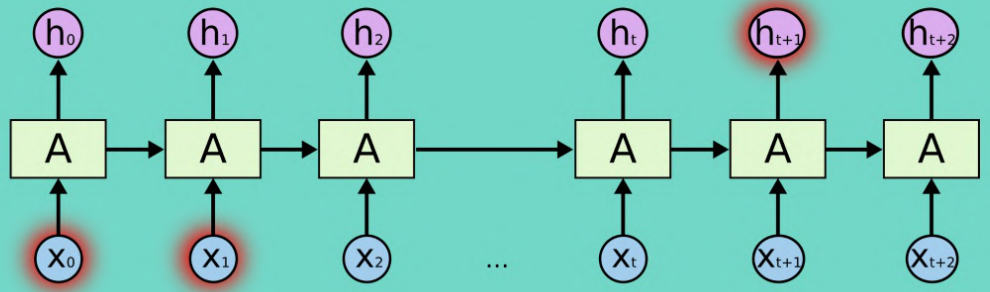
X

(1344, 448, 13)


gru_input
InputLayer
input:
output:
[(None, 448, 13)]
[(None, 448, 13)]
gru
GRU
input:
output:
(None, 448, 13)
(None, 448, 128)
gru_1
GRU
input:
output:
(None, 448, 128)
(None, 64)
dense
Dense
input:
output:
(None, 64)
(None, 7)

RNN

GRU

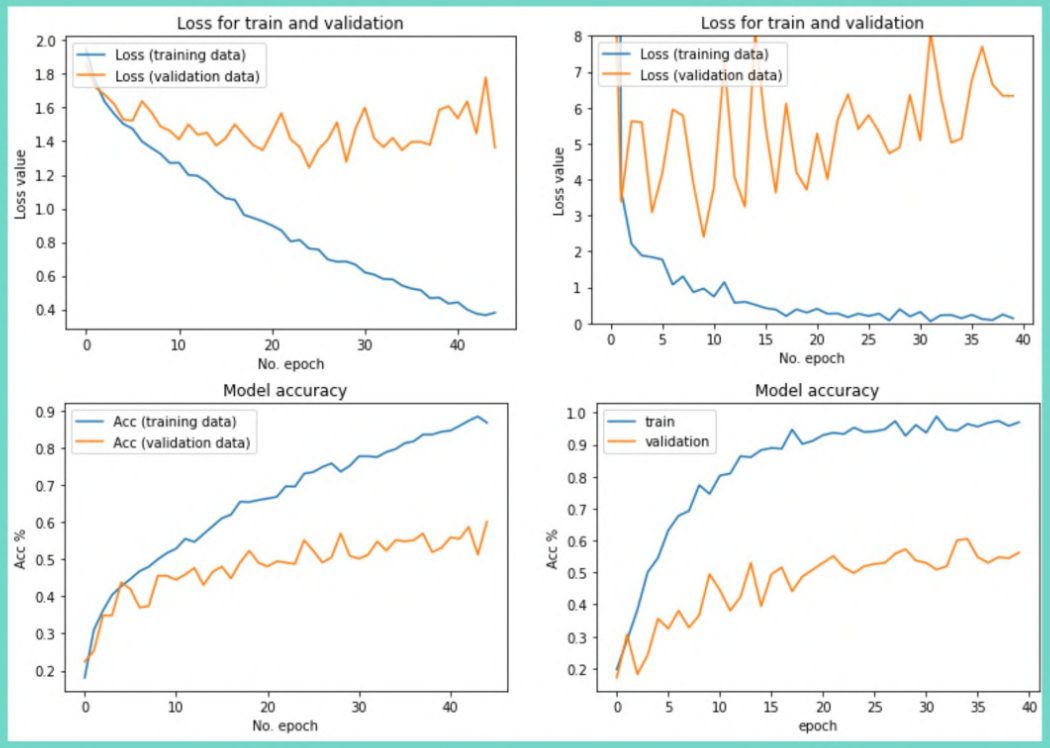


architecture


Models
CNN
RNN
Results


Compare the results
Loss-accuracy
Table
Confusion matrix

RNN CNN


Loss for train and validation
Loss (training data)
Loss (validation data)
Loss value
No. epoch
Loss for train and validation
Loss (training data)
Loss (validation data)
Loss value
No. epoch
Model accuracy
Acc (training data)
Acc (validation data)
Acc %
No. epoch
Model accuracy
train
validation
Acc %
epoch

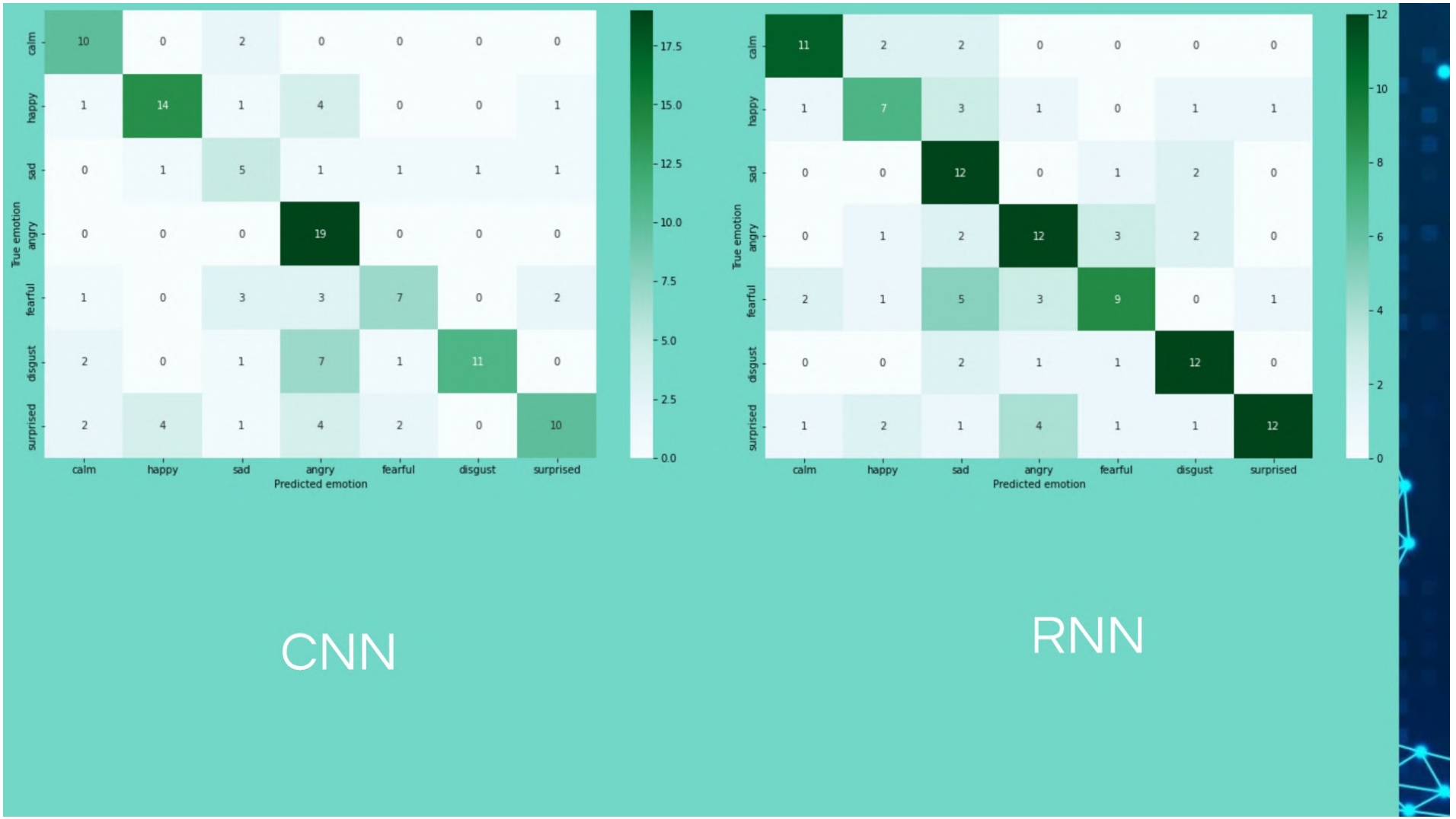

Compare the results
Loss-accuracy
Table
Confusion matrix

	RNN	CNN
Accuracy	60%	61%
Precision	63%	64%
Recall	62%	62%
F1 score	61%	60%
loss	1.39	6.58

The training time for RNN: 18 minutes

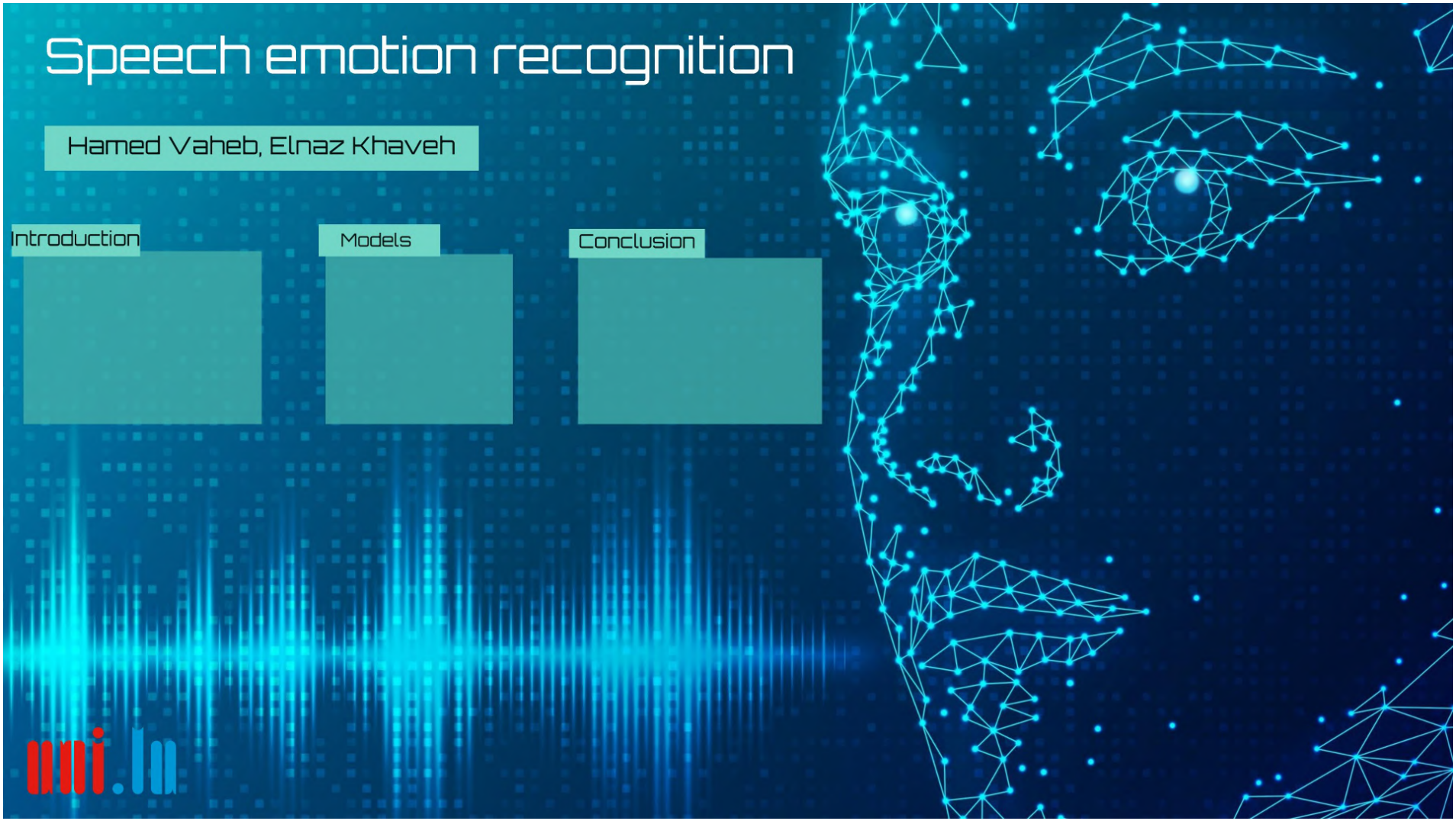
The training time for CNN: 3 minutes


Compare the results
Loss-accuracy
Table
Confusion matrix


True emotion
calm
happy
sad
angry
fearful
disgust
surprised
Predicted emotion
CNN
RNN


Compare the results
Loss-accuracy
Table
Confusion matrix


Models
CNN
RNN
Results


Speech emotion recognition
Hamed Vaheb, Elnaz Khaveh
Introduction
Models
Conclusion

Conclusion

- CNN and RNN perform almost the same
- Time complexity of CNN is slightly better
- MFCC was the most determinant feature

Future work

Future work

- Real-time
- continuous representation of emotions
- Small data set
- hyper-parameter
- ensemble

Future work

- Real-time
- continuous representation of emotions
- Small data set
- hyper-parameter
- ensemble

https://github.com/berserkhmdvhb/Project-SER

Future work

- Real-time
- continuous representation of emotions
- Small data set
- hyper-parameter
- ensemble

Thanks for your attention !

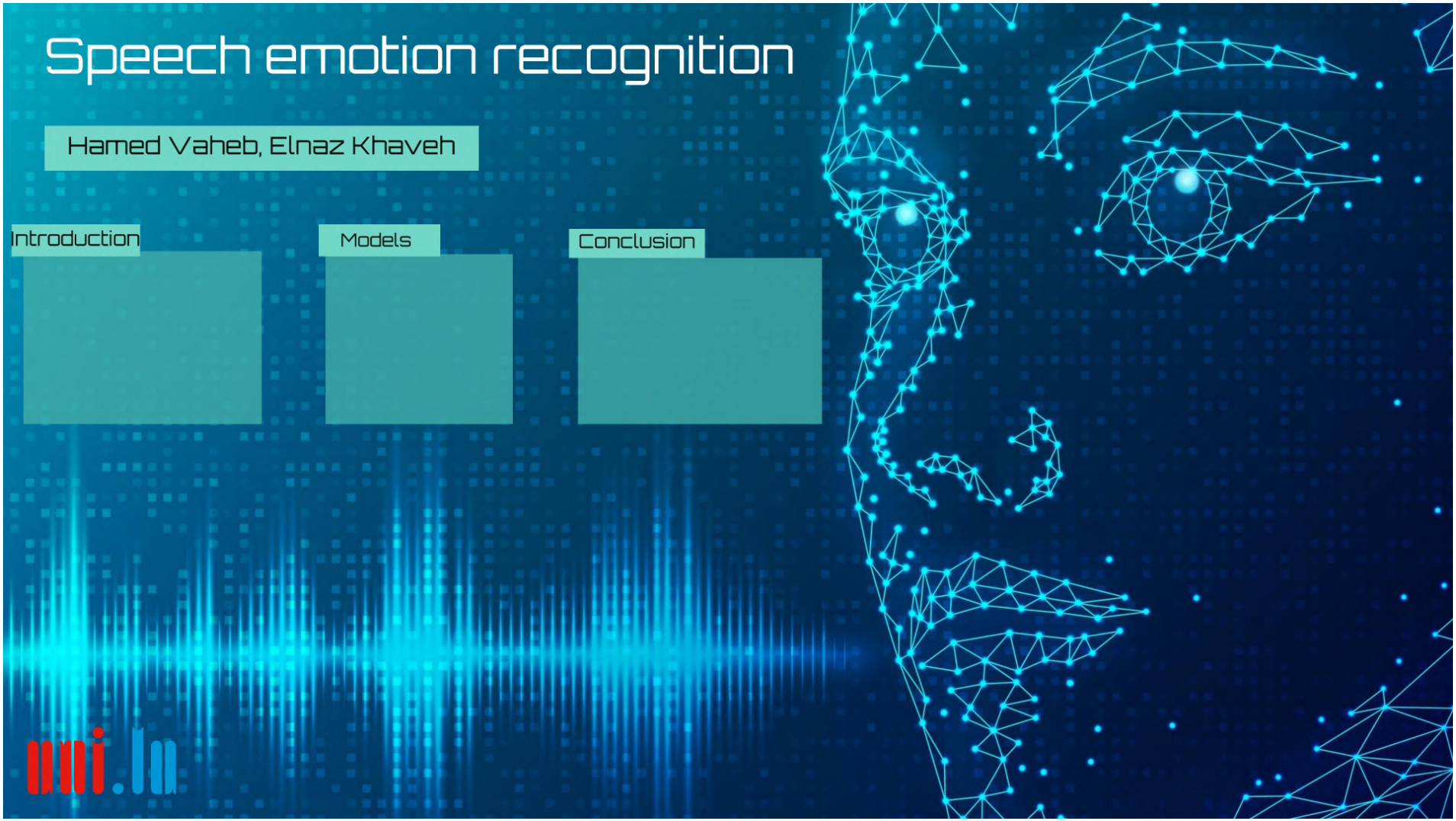
Future work

- Real-time
- continuous representation of emotions
- Small data set
- hyper-parameter
- ensemble

Conclusion

- CNN and RNN perform almost the same
- Time complexity of CNN is slightly better
- MFCC was the most determinant feature

Future work


Speech emotion recognition
Hamed Vaheb, Elnaz Khaveh
Introduction
Models
Conclusion