Spectral Analysis and Detection of Extreme Vocal Effects

ORIOL NIETO

RESEARCH SEMINAR
MUSIC TECHNOLOGY GROUP - UNIVERSITAT POMPEU FABRA
NOVEMBER 13, 2019

pandora®

OUTLINE

Motivation and Background: Extreme Vocal Effects

Spectral Analysis of EVEs

Detection of EVEs with Neural Networks

OUTLINE

Motivation and Background: Extreme Vocal Effects

Spectral Analysis of EVEs

Detection of EVEs with Neural Networks

Extreme Vocal Effects

- Vocal techniques that enhance expressivity
- Common in music:
  - Metal
  - Tuvan Throat Singing
- Also known as:
  - distorted voices
  - screams
  - growls
  - …

pandora®

Extreme Vocal Effects

- Vocal techniques that enhance expressivity

- Common in music:
  - Metal
  - Tuvan Throat Singing

- Also known as:
  - distorted voices
  - screams
  - growls
  - …

pandora®

EVEs Classification

- Pitched:
  - Roughness / Distortion
- Un-pitched:
  - Death Growl / Grunt
  - Fry Scream
  - Inhale / Inward Scream
- Hybrid
  - Tuva Throat Singing

pandora®

(Nieto, 2008; Nieto, 2013)

EVEs Synthesis

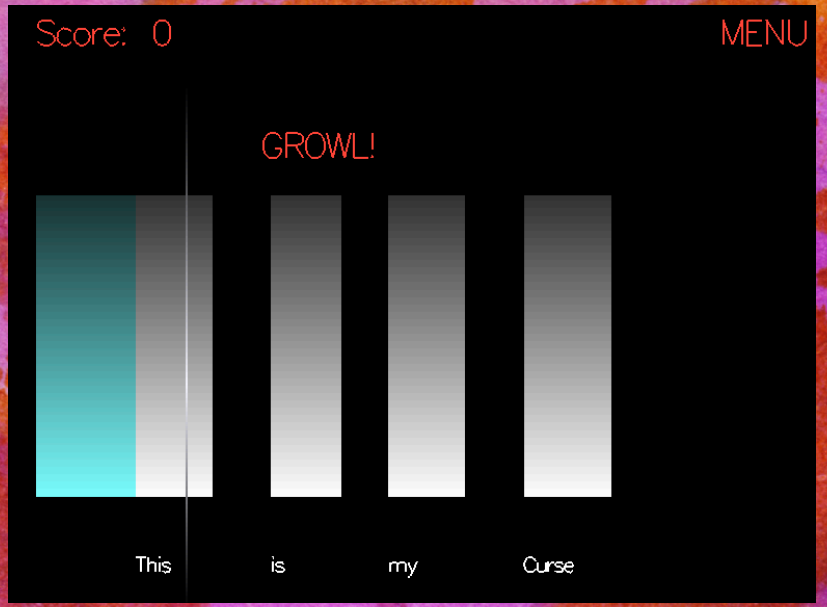
- Pedals / Plugins
  - TC Helicon - Voice Live
  - Auburn - Graillon
- Spectral Synthesis:
  - Aggressive Blues Vocals Synthesis (Loscos, 2004)
  - KaleiVoiceCope (Mayor, 2010)
  - Growl Spectral Synthesis (Bonada, 2013)
- Deep Learning:
  - SampleRNN trained on death growls (Carr, 2018)

pandora®

EVEs Applications

Growl Hero

- Like Rock Band but with 3 types of EVE detection:
  - Fry Scream
  - Roughness
  - Growl



https://ccrma.stanford.edu/~urinieto/256/growlhero/

pandora®

EVEs Applications

Screaminator

- Screaming into your phone to get a score, compete, and share with friends

http://screaminator.urinieto.com/

pandora®

EVEs Applications

Relentless Doppelganger (Carr, 2018)

- “Neural network generating technical death metal, via livestream 24/7 to infinity”

https://www.youtube.com/watch?v=CNNmBtNcccE

pandora®

OUTLINE

Motivation and Background: Extreme Vocal Effects

Spectral Analysis of EVEs

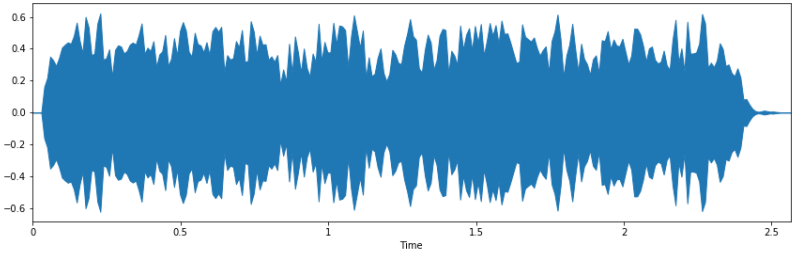
Detection of EVEs with Neural Networks

Spectral Analysis

DISSECTING EVES

- Explore frequency components of EVEs
  - Better understanding of sound qualities

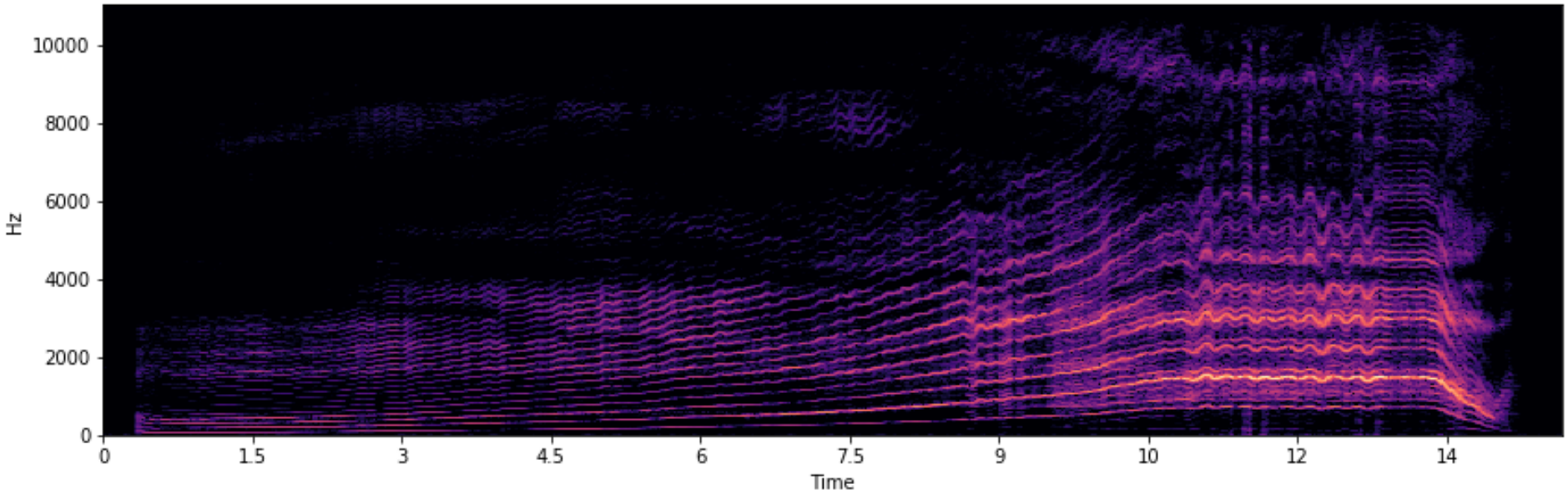
- Could foster better EVEs:
  - synthesis
  - classification
  - detection



pandora®

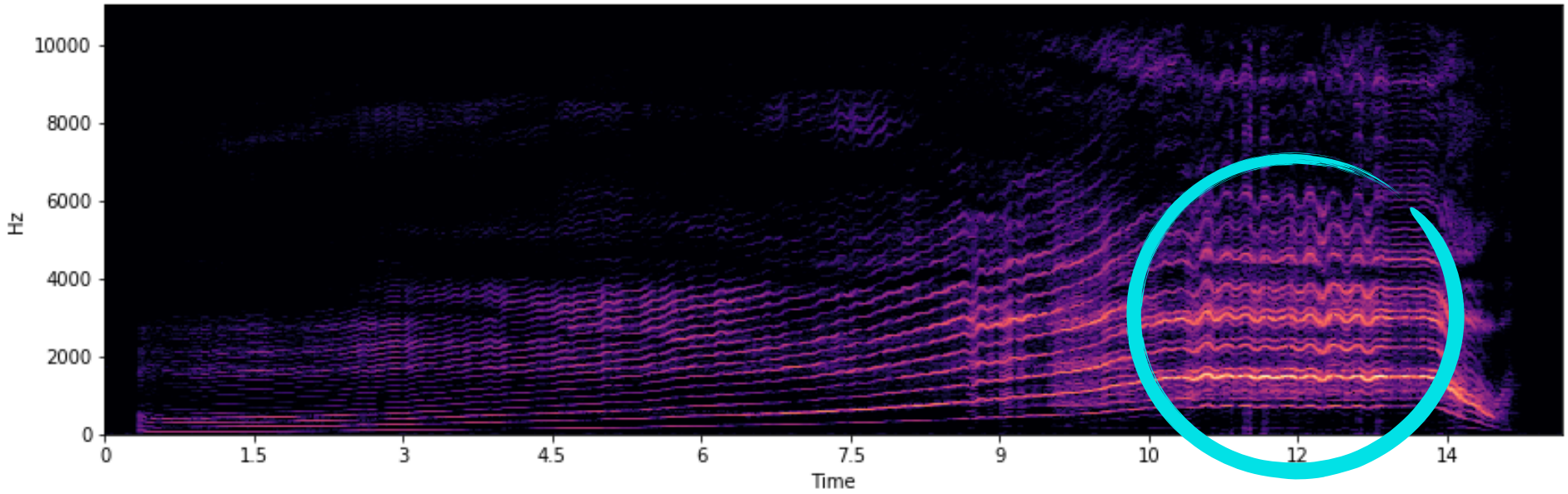
Spectral Analysis

- Visual representation of energy for a given frequency and time frame of an audio signal



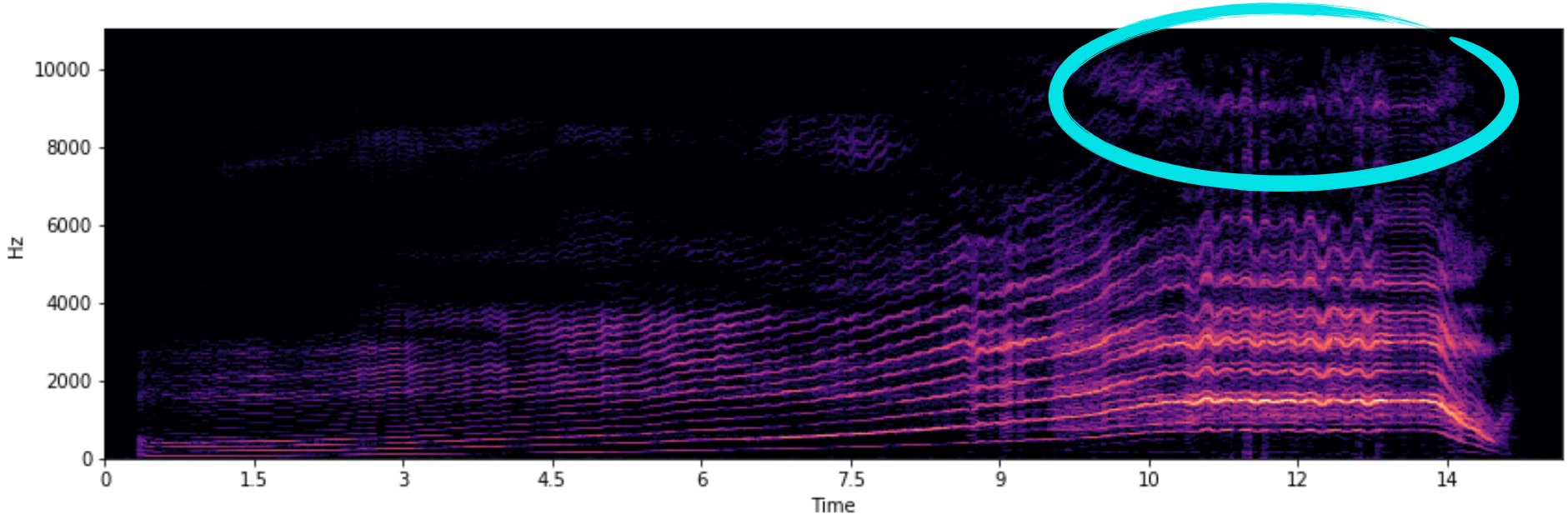
Spectral Analysis

- Visual representation of energy for a given frequency and time frame of an audio signal



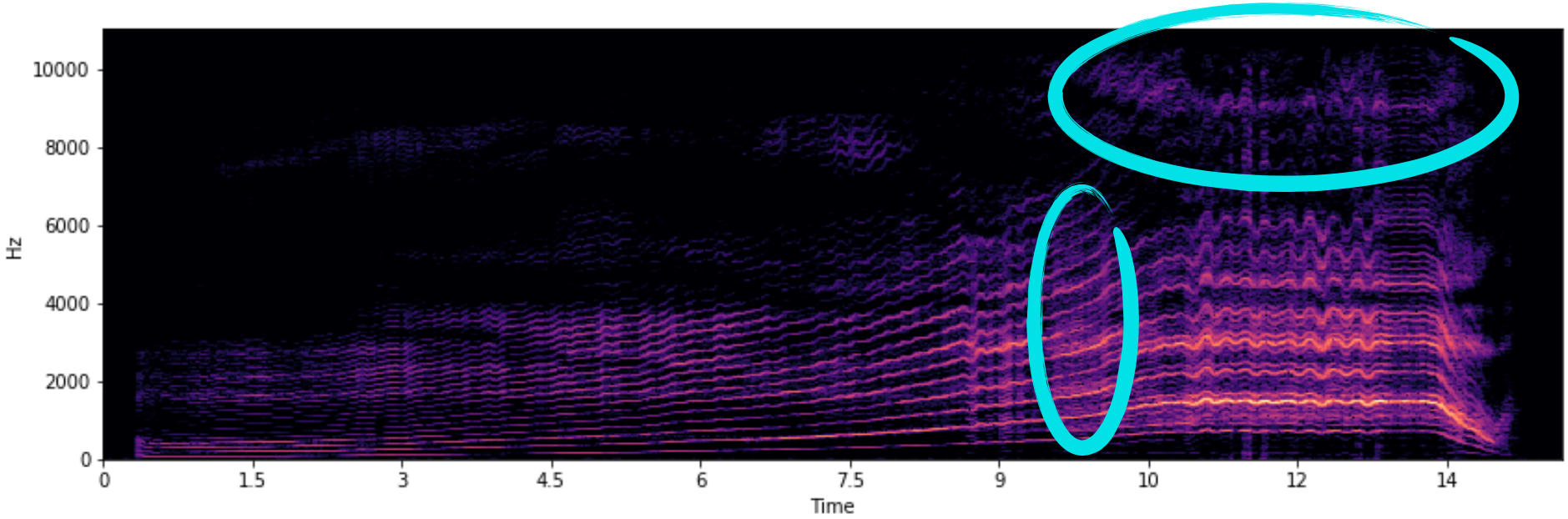
Spectral Analysis

- Visual representation of energy for a given frequency and time frame of an audio signal



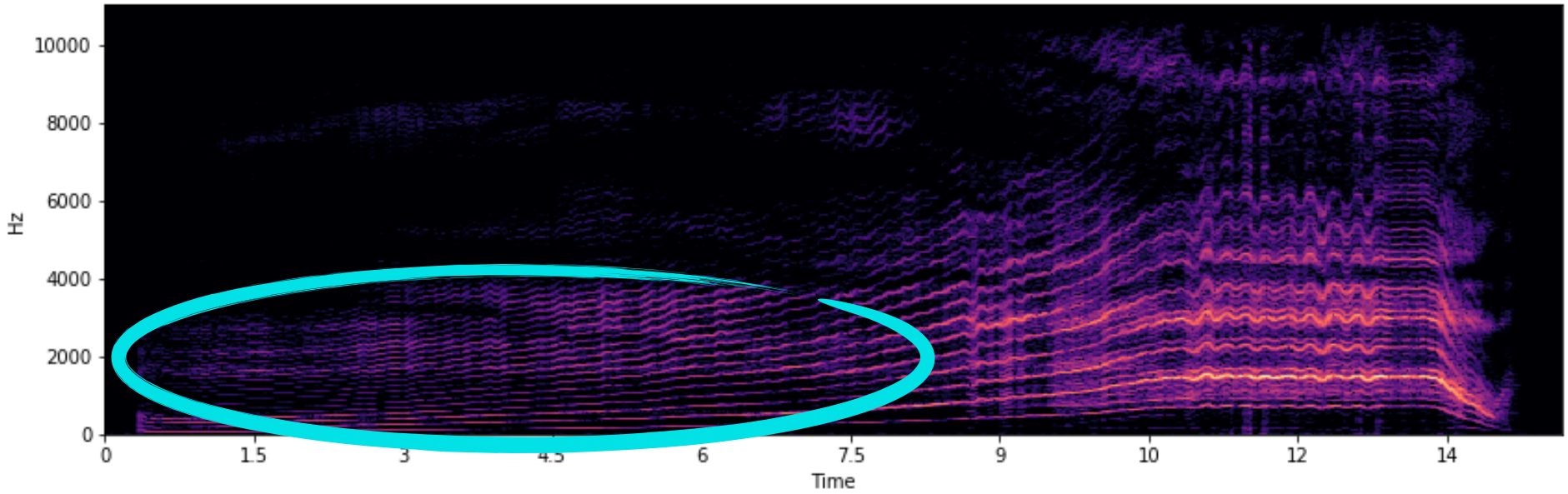
Spectral Analysis

- Visual representation of energy for a given frequency and time frame of an audio signal



Spectral Analysis

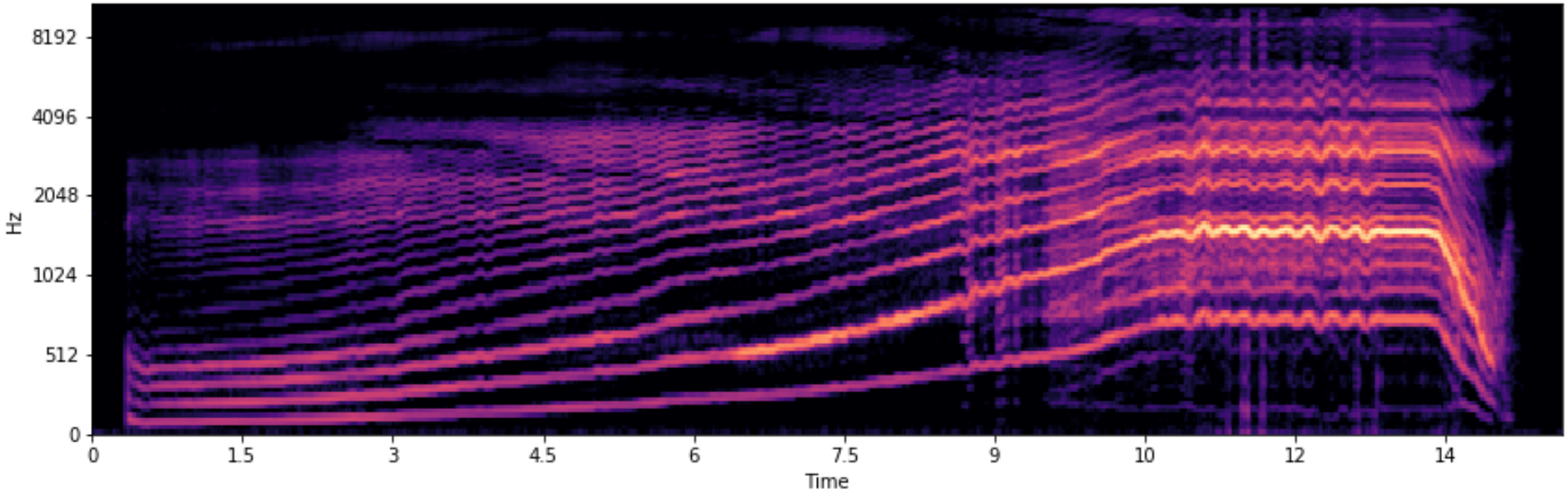
- Visual representation of energy for a given frequency and time frame of an audio signal



Spectral Analysis

MEL-SPECTROGRAM

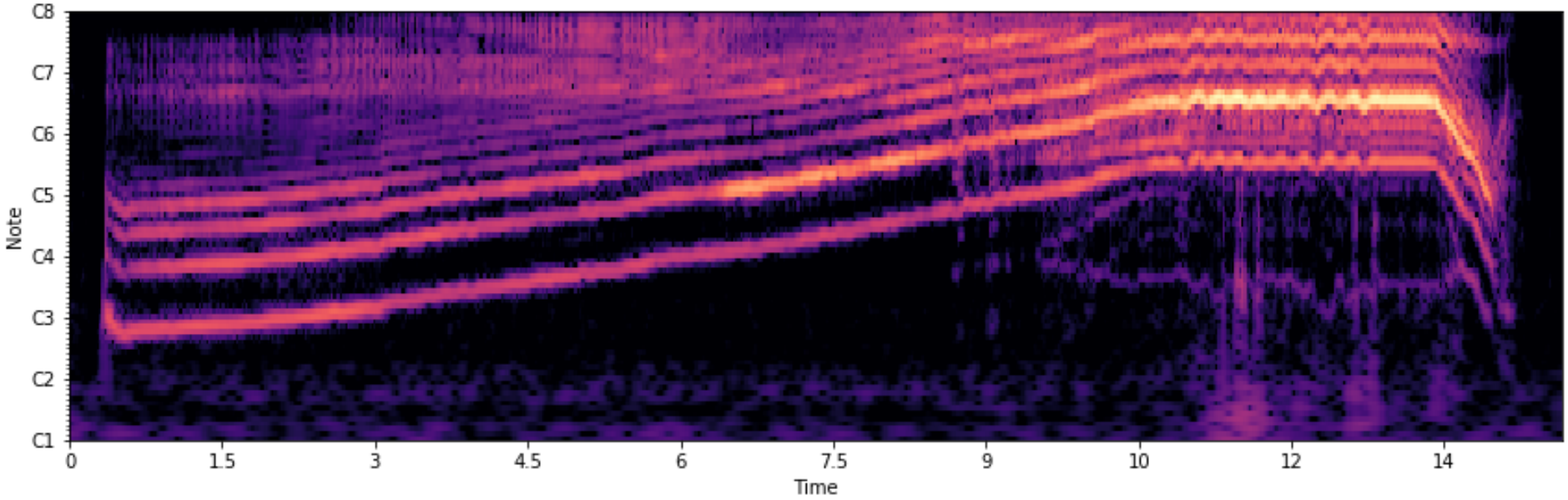
- Compact representation using a perceptually inspired scale (mel scale)
- 128 frequency bins (instead of 1024 of the original Spectrogram)



Spectral Analysis

CONSTANT-Q TRANSFORM

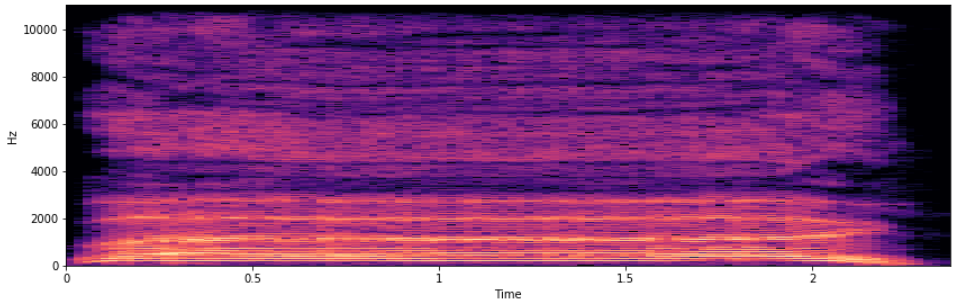
- Compact representation using a constant (linear) frequency scale
- 84 frequency bins



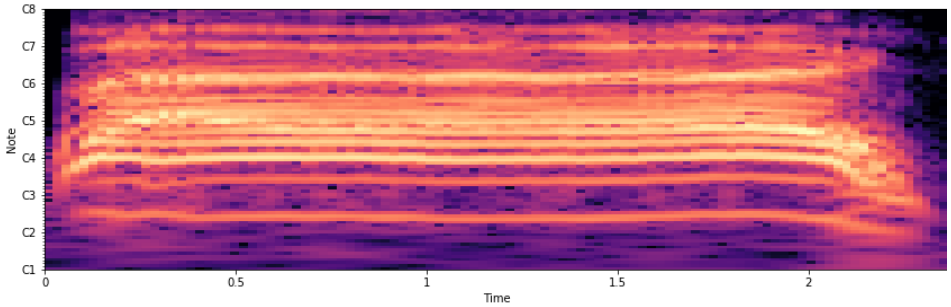
Spectral Analysis

ROUGHNESS / DISTORTION

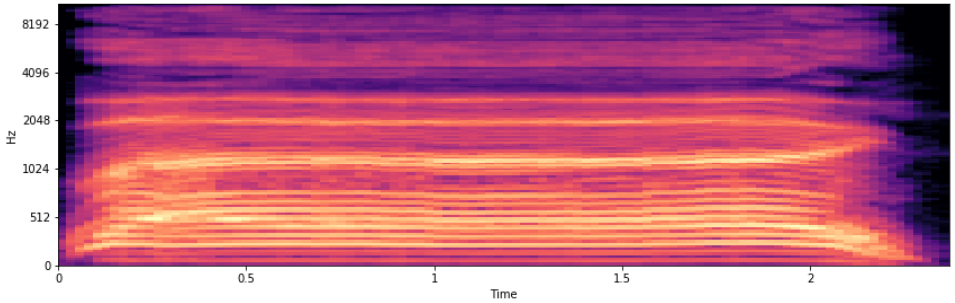
Spectrogram



Constant-Q Transform



Mel Spec



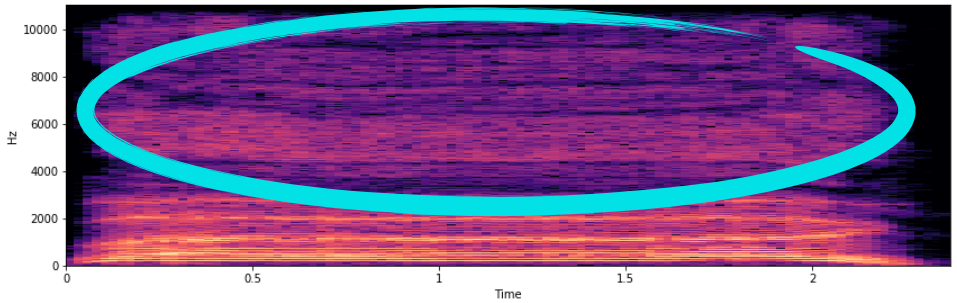
pandora®

Spectral Analysis

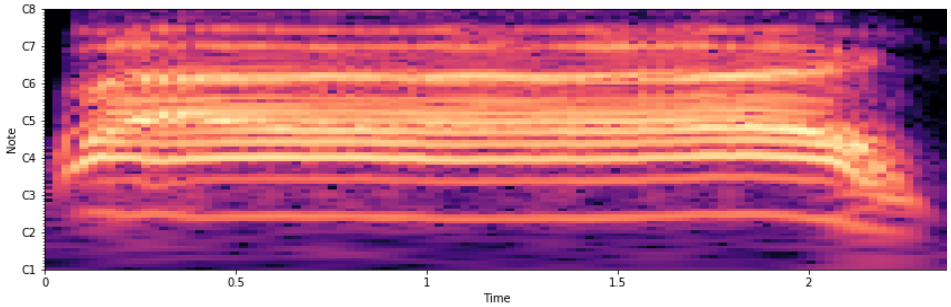
ROUGHNESS / DISTORTION

- **Chaotic higher frequencies**

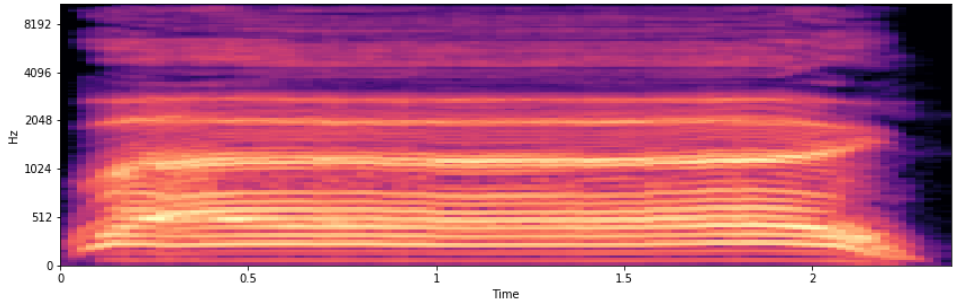
Spectrogram



Constant-Q Transform



Mel Spec



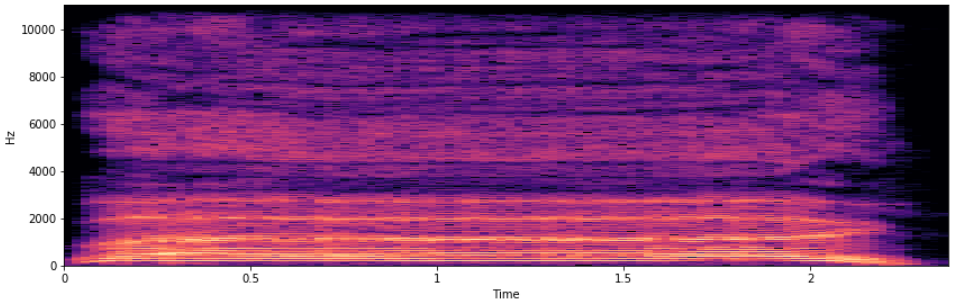
pandora®

Spectral Analysis

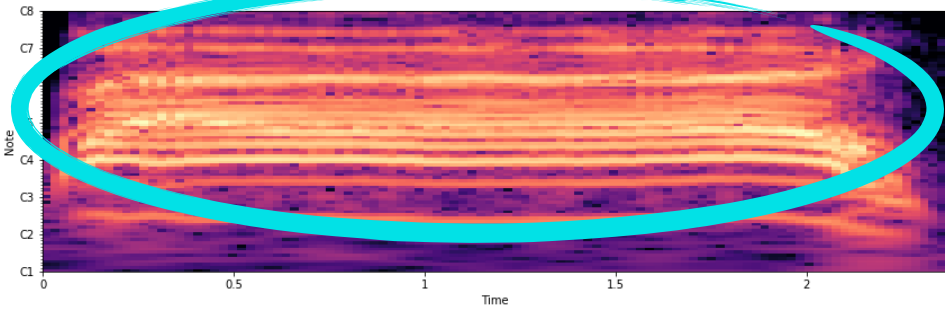
ROUGHNESS / DISTORTION

- Chaotic higher frequencies
- **Noisy but with perceivable pitch**

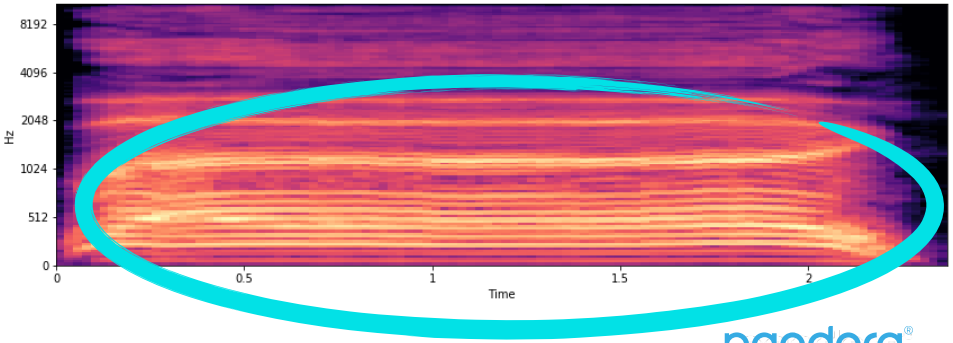
Spectrogram



Constant-Q Transform



Mel Spec



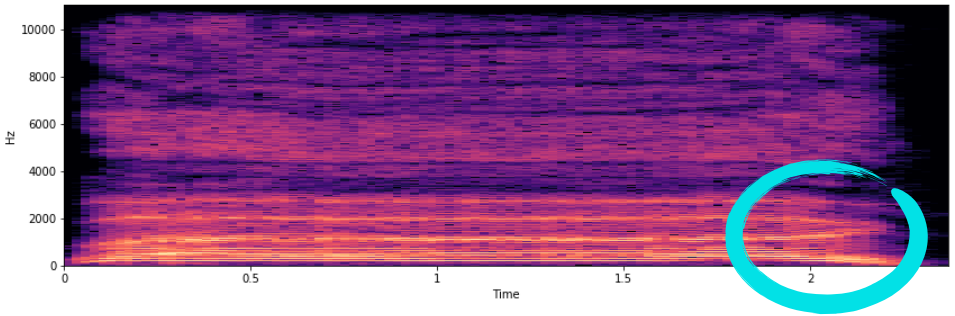
pandora®

Spectral Analysis

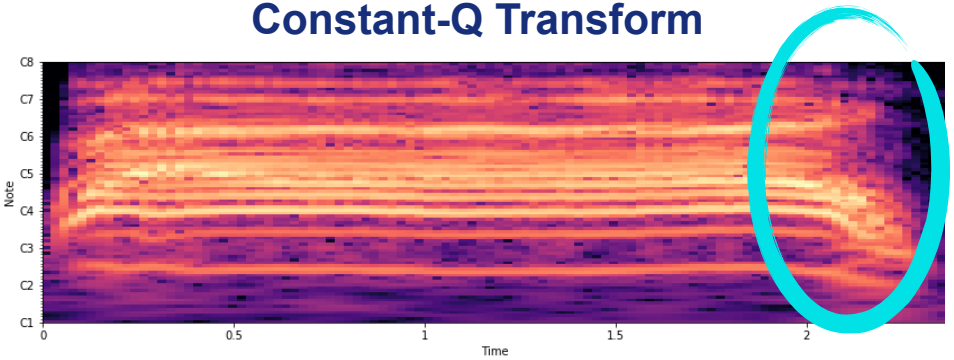
ROUGHNESS / DISTORTION

- Chaotic higher frequencies
- Noisy but with perceivable pitch
- **Overtones may harmonically diverge**

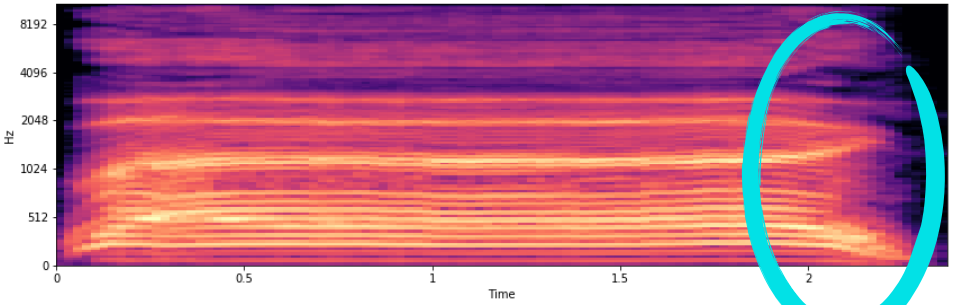
Spectrogram



Constant-Q Transform



Mel Spec

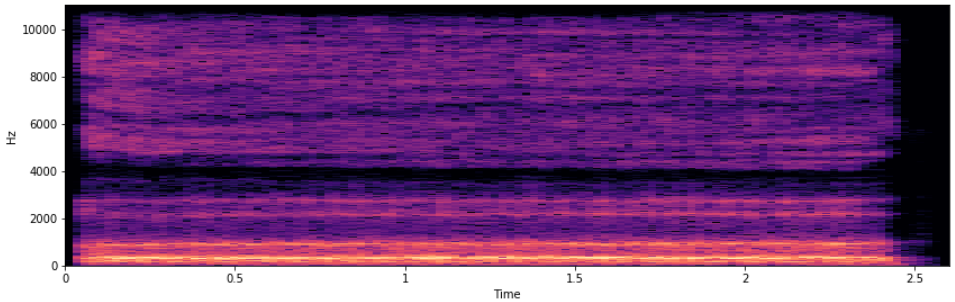


panaora®

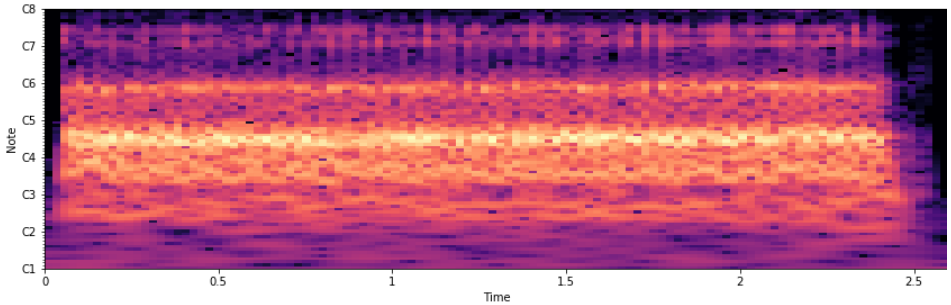
Spectral Analysis

DEATH GROWL / GRUNT

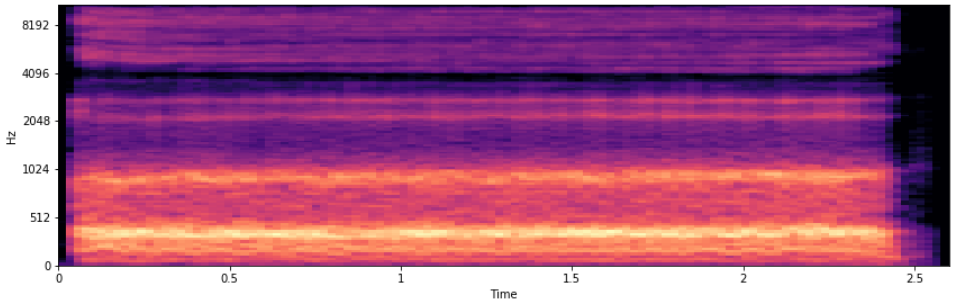
Spectrogram



Constant-Q Transform



Mel Spec



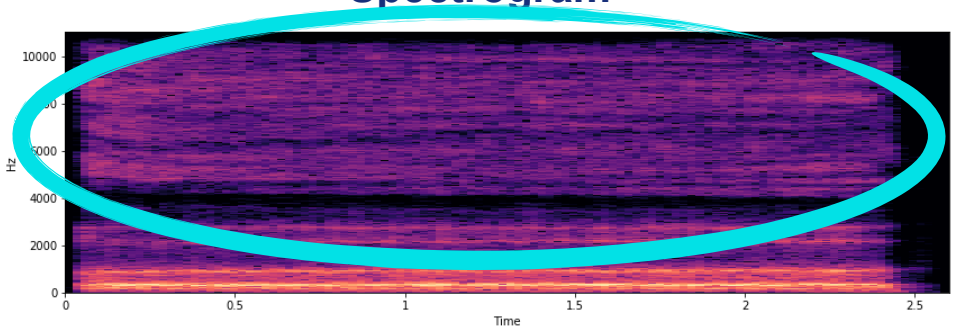
pandora®

Spectral Analysis

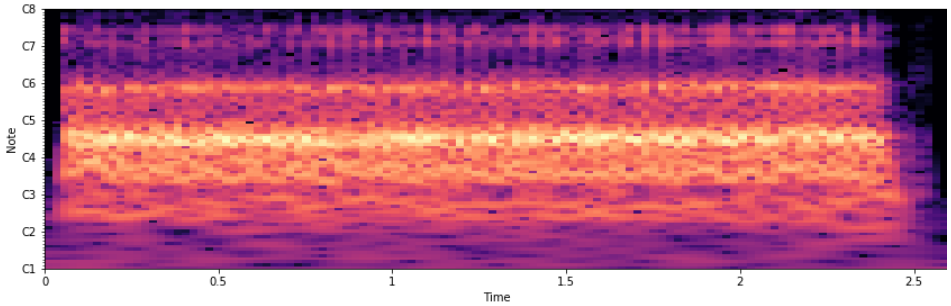
DEATH GROWL / GRUNT

- **Chaotic higher frequencies**

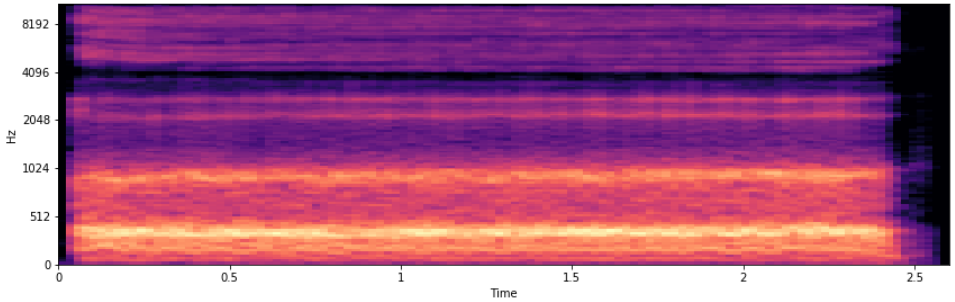
Spectrogram



Constant-Q Transform



Mel Spec



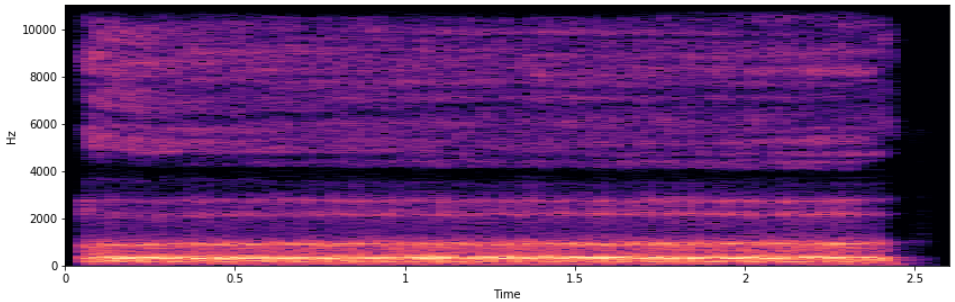
pandora®

Spectral Analysis

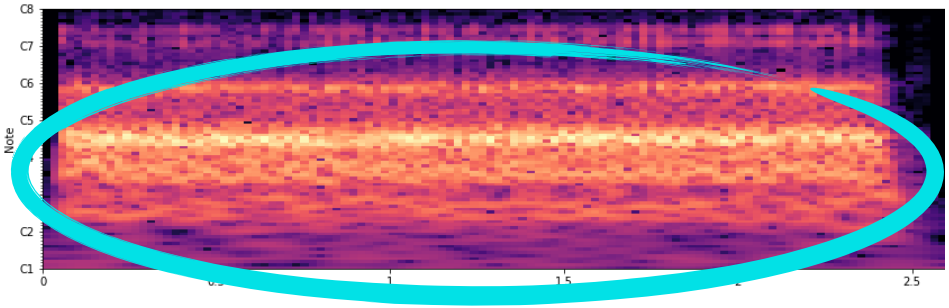
DEATH GROWL / GRUNT

- Chaotic higher frequencies
- **Not perceivable pitch**

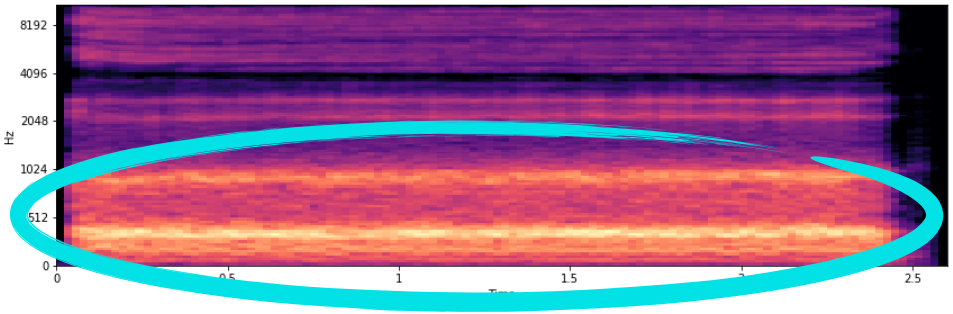
Spectrogram



Constant-Q Transform



Mel Spec



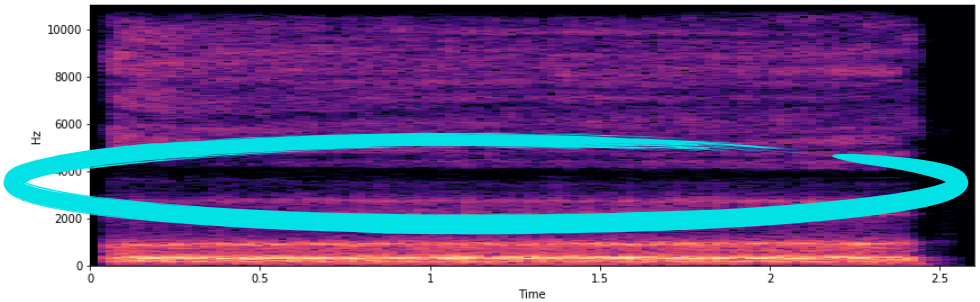
pandora®

Spectral Analysis

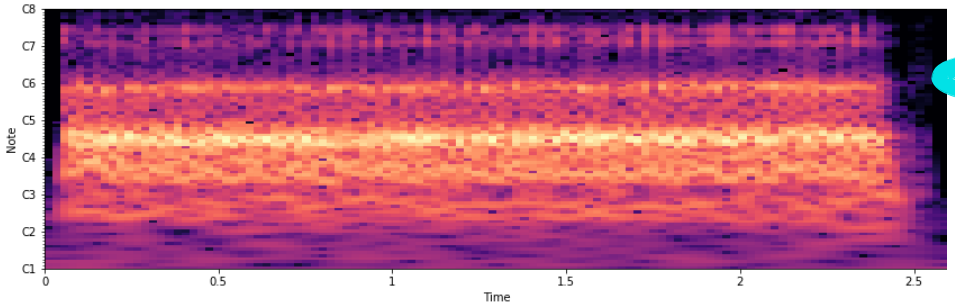
DEATH GROWL / GRUNT

- Chaotic higher frequencies
- Not perceivable pitch
- **Quiet frequency band around 4kHz**

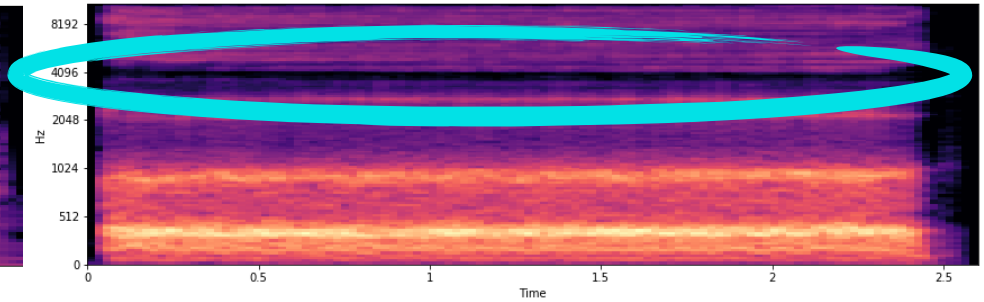
Spectrogram



Constant-Q Transform



Mel Spec

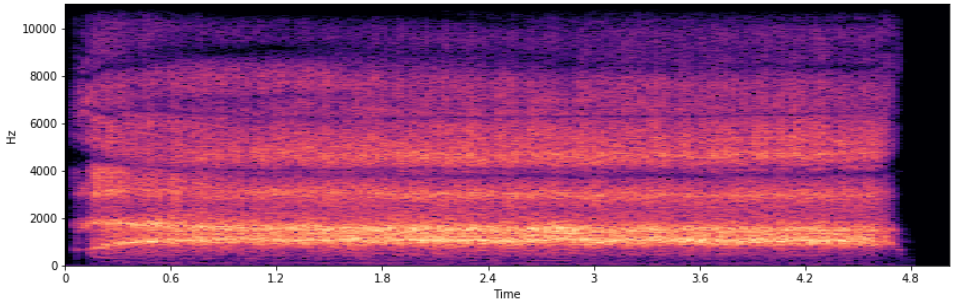


pandora®

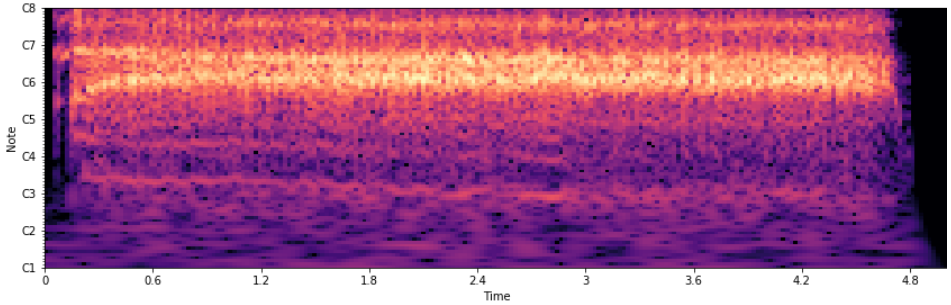
Spectral Analysis

FRY SCREAM

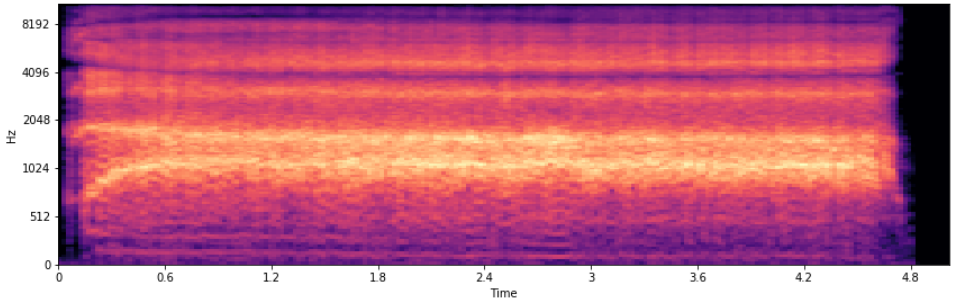
Spectrogram



Constant-Q Transform



Mel Spec

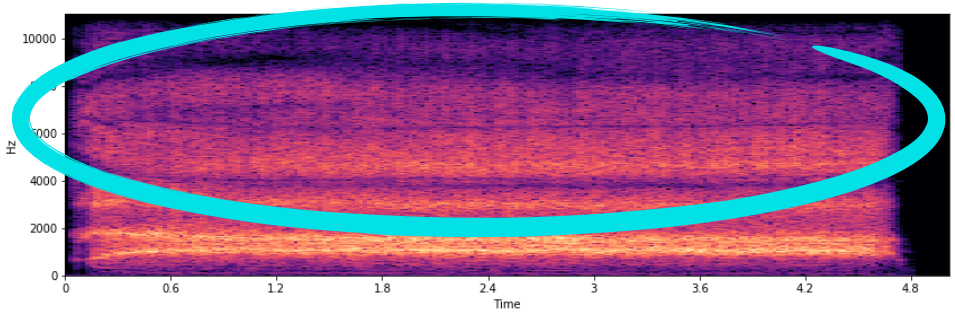


Spectral Analysis

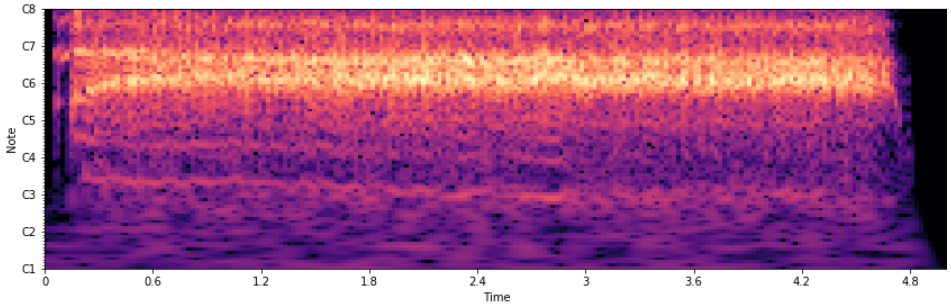
FRY SCREAM

- **Wide-band high energy (very noisy)**

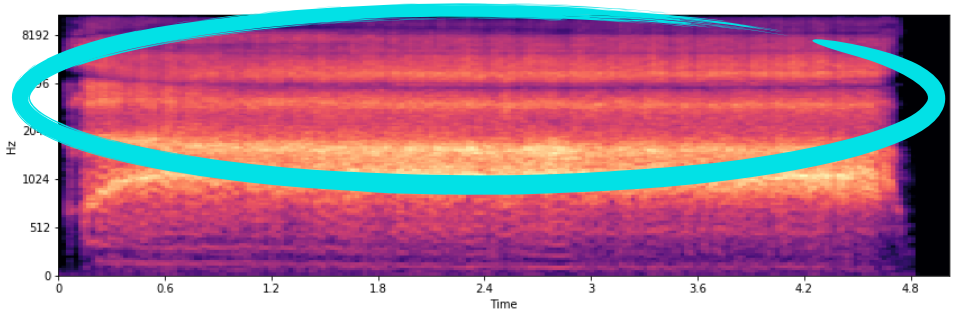
Spectrogram



Constant-Q Transform



Mel Spec

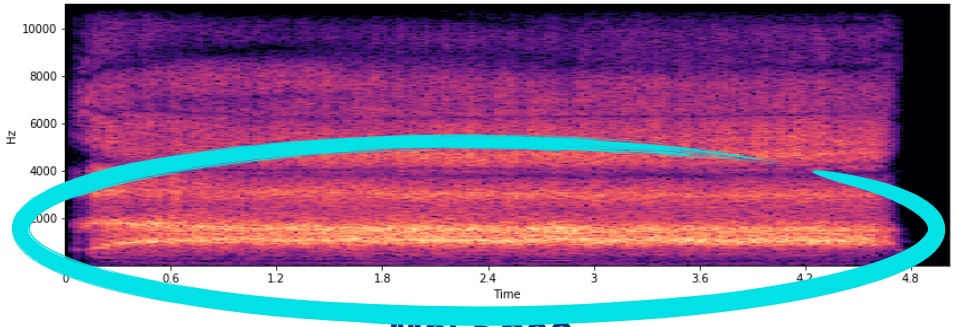


Spectral Analysis

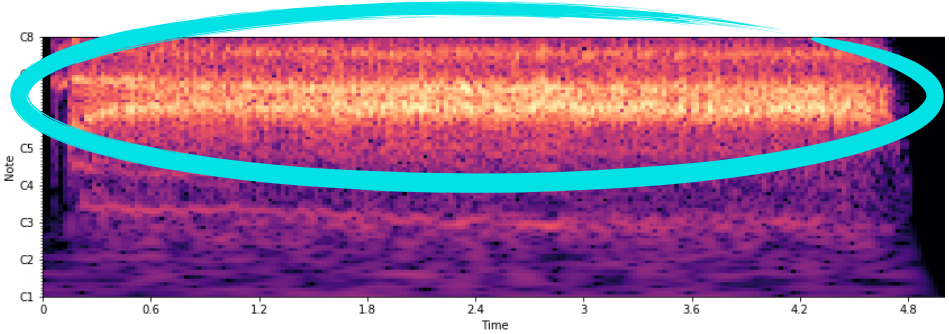
FRY SCREAM

- Wide-band high energy (very noisy)
- **Not perceivable pitch**

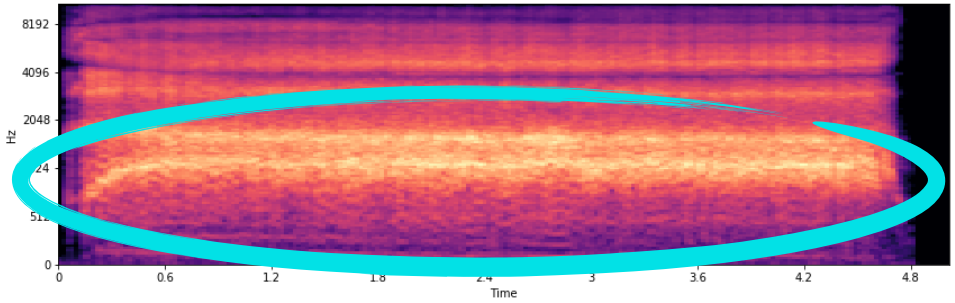
Spectrogram



Constant-Q Transform



Mel Spec

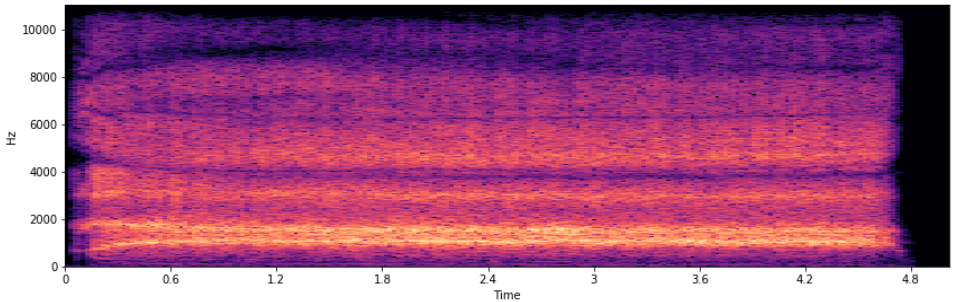


Spectral Analysis

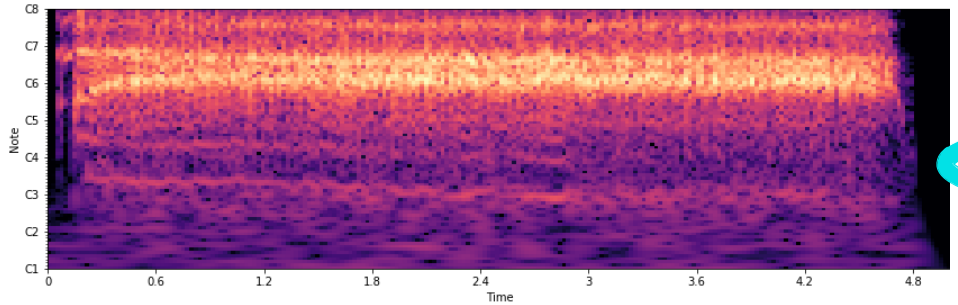
FRY SCREAM

- Wide-band high energy (very noisy)
- Not perceivable pitch
- **Fundamental frequency around 1kHz (very high)**

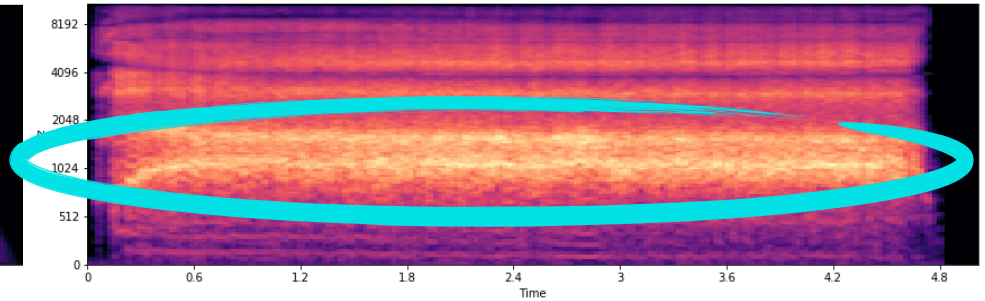
Spectrogram



Constant-Q Transform



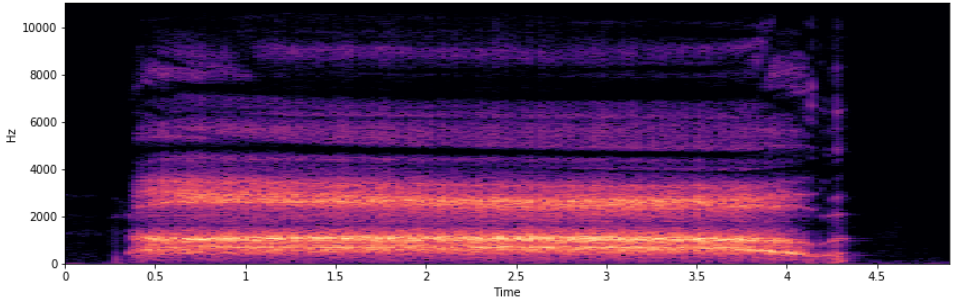
Mel Spec



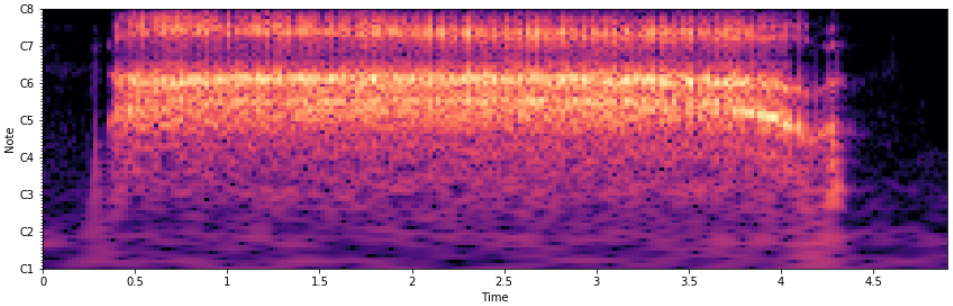
Spectral Analysis

INHALE SCREAM

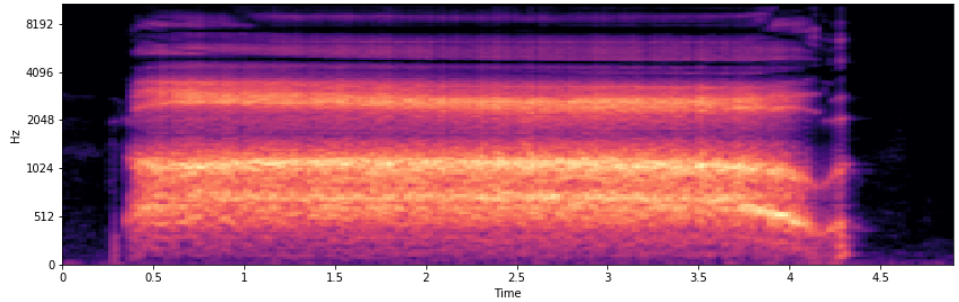
Spectrogram



Constant-Q Transform



Mel Spec

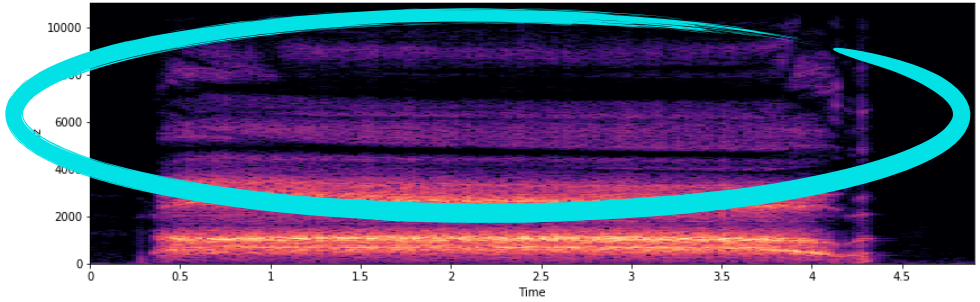


Spectral Analysis

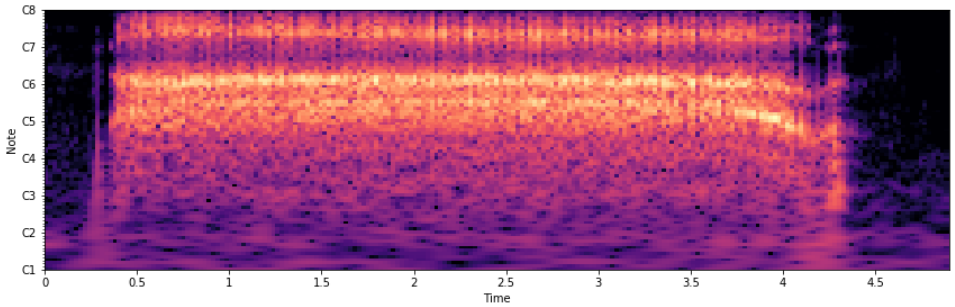
INHALE SCREAM

- Slightly less energy in high frequencies than fry scream (still noisy)

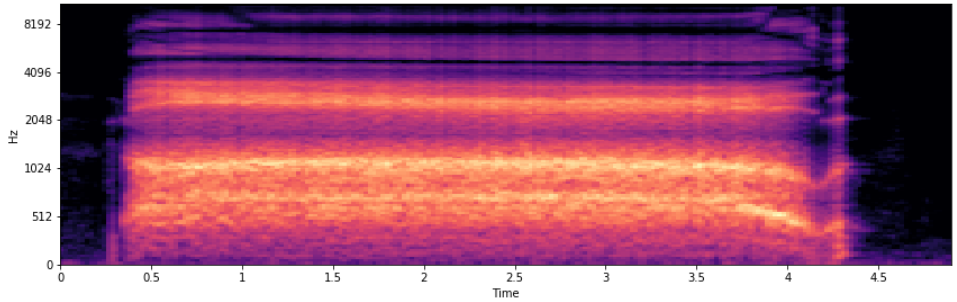
Spectrogram



Constant-Q Transform



Mel Spec

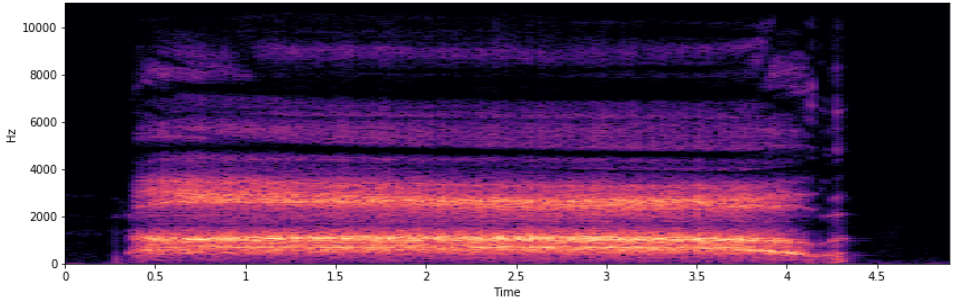


Spectral Analysis

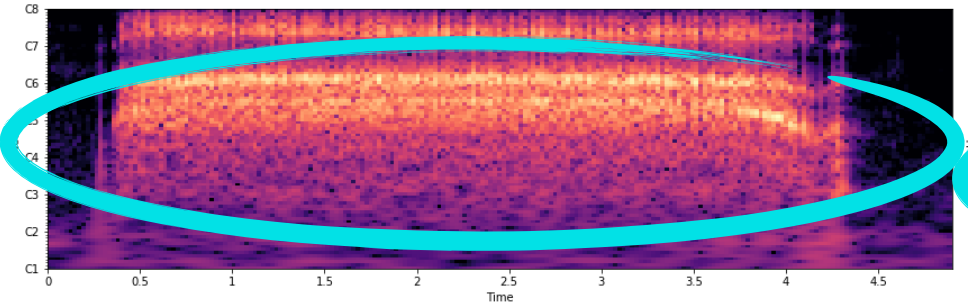
INHALE SCREAM

- Slightly less energy in high frequencies than fry scream (still noisy)
- **Not perceivable pitch**

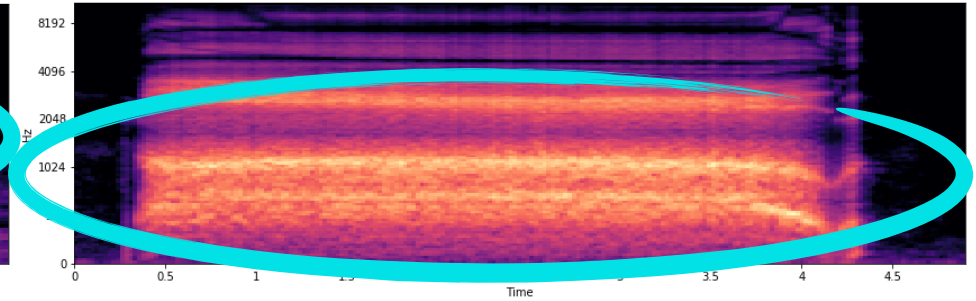
Spectrogram



Constant-Q Transform



Mel Spec

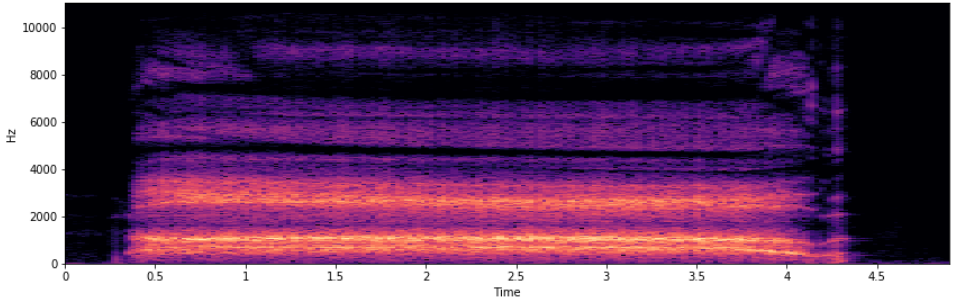


Spectral Analysis

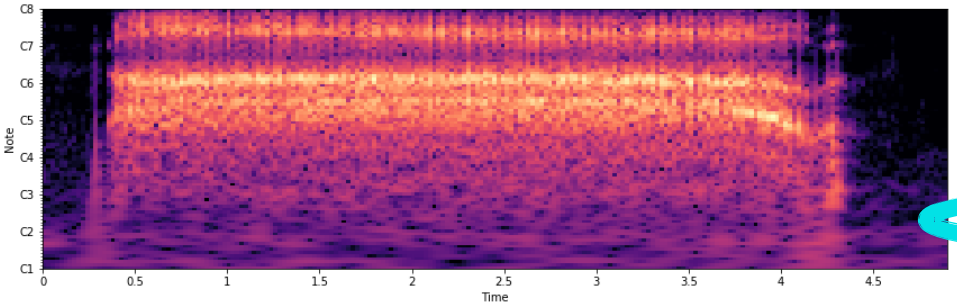
INHALE SCREAM

- Slightly less energy in high frequencies than fry scream (still noisy)
- Not perceivable pitch
- **Fundamental frequency around 600Hz**

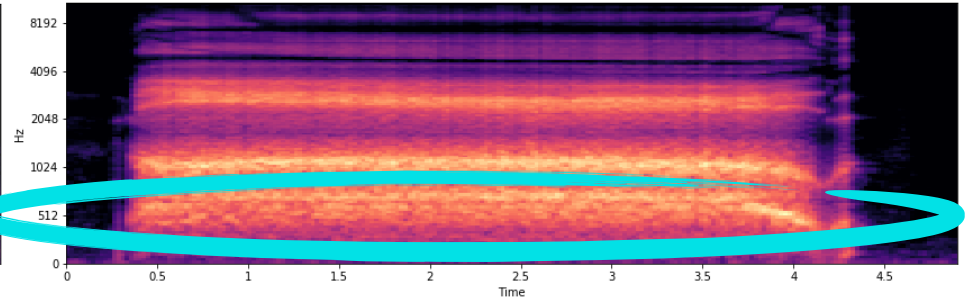
Spectrogram



Constant-Q Transform



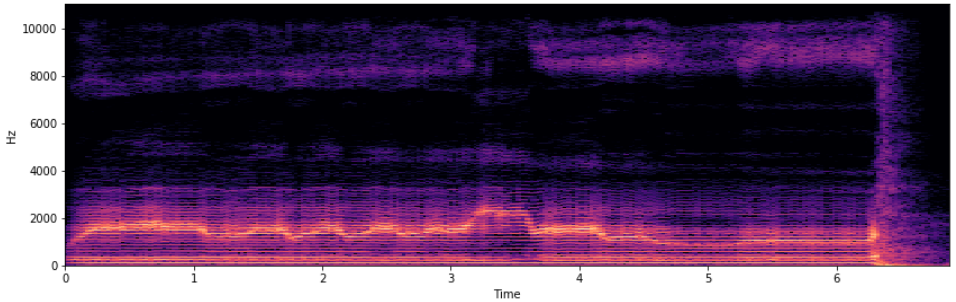
Mel Spec



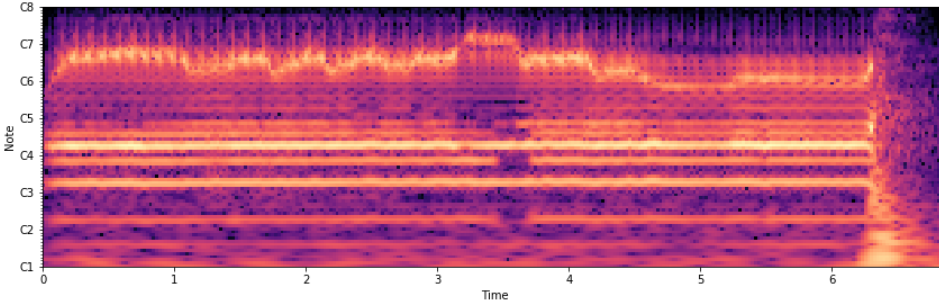
Spectral Analysis

TUVA THROAT SINGING

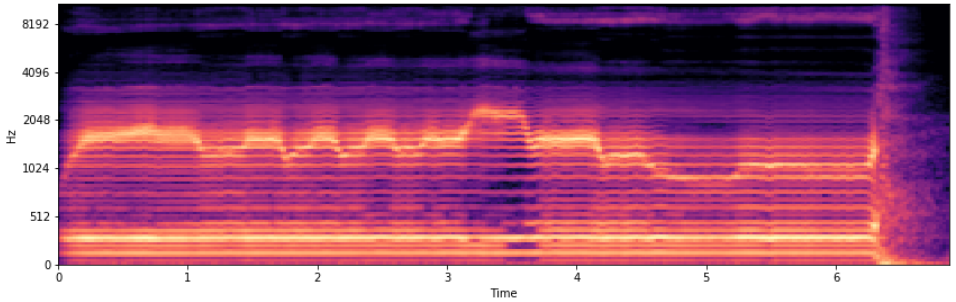
Spectrogram



Constant-Q Transform



Mel Spec

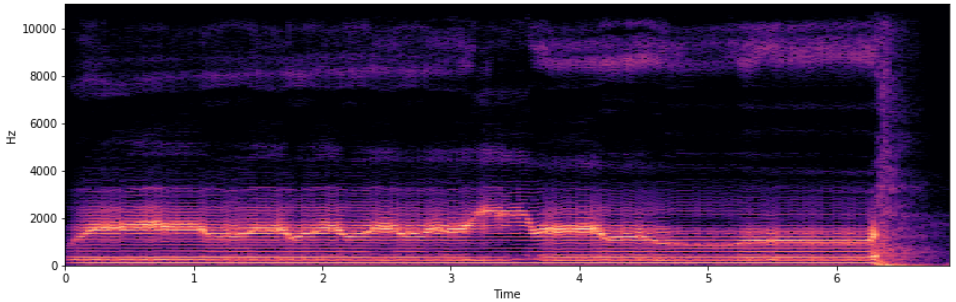


Spectral Analysis

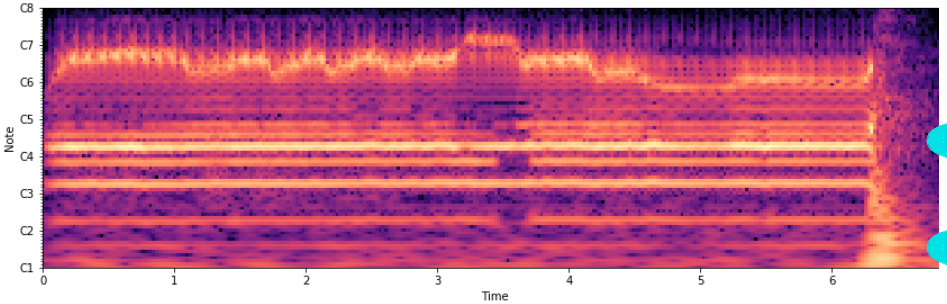
TUVA THROAT SINGING

- **Two pitches simultaneously**

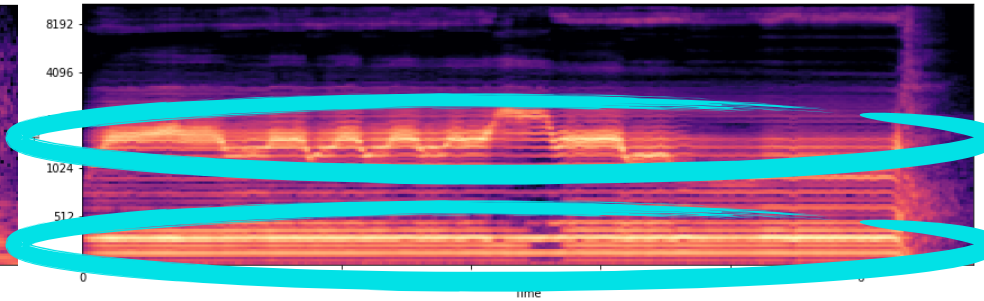
Spectrogram



Constant-Q Transform



Mel Spec

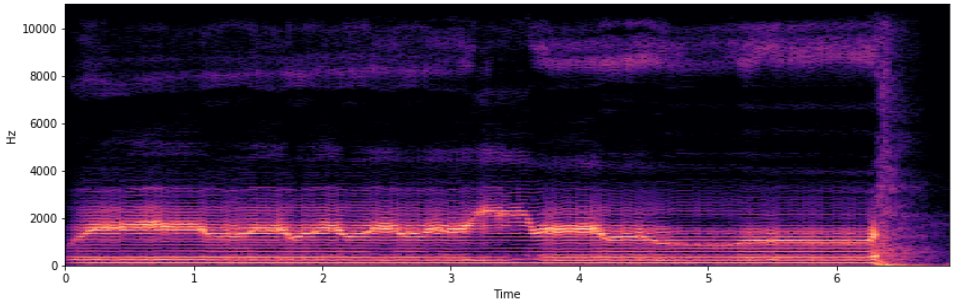


Spectral Analysis

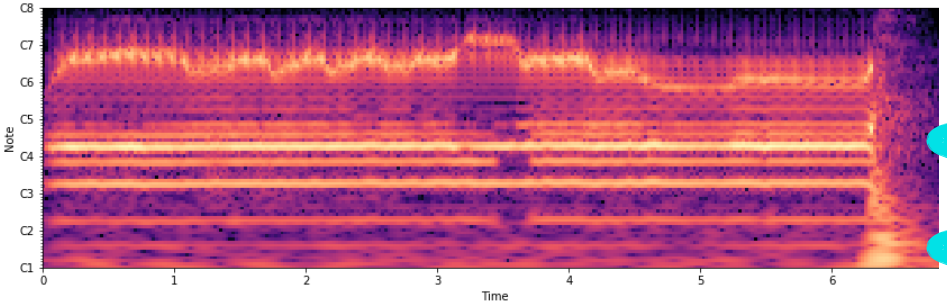
TUVA THROAT SINGING

- Two pitches simultaneously
- **One low and constant, noisy, other high and variable**

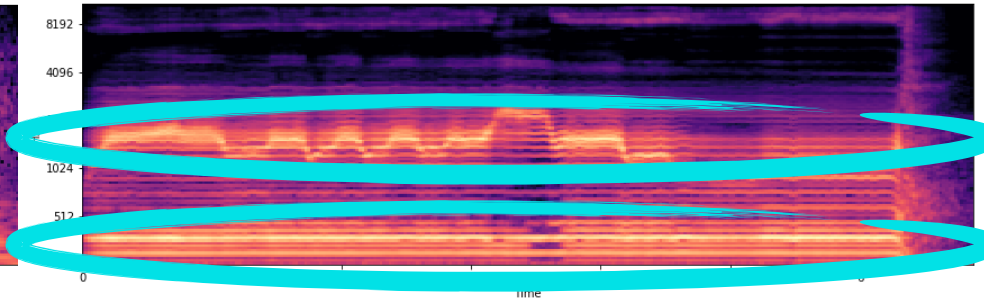
Spectrogram



Constant-Q Transform



Mel Spec

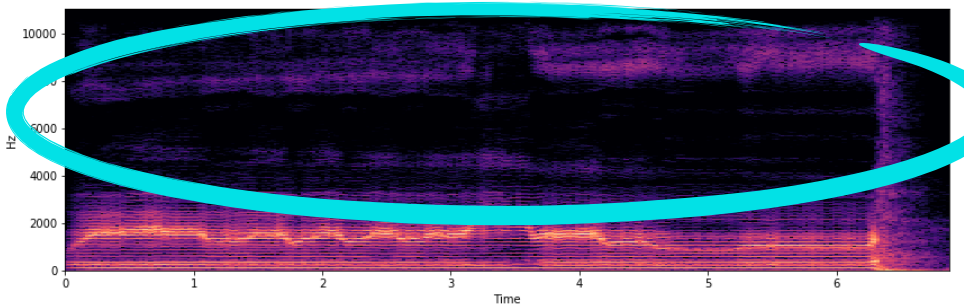


Spectral Analysis

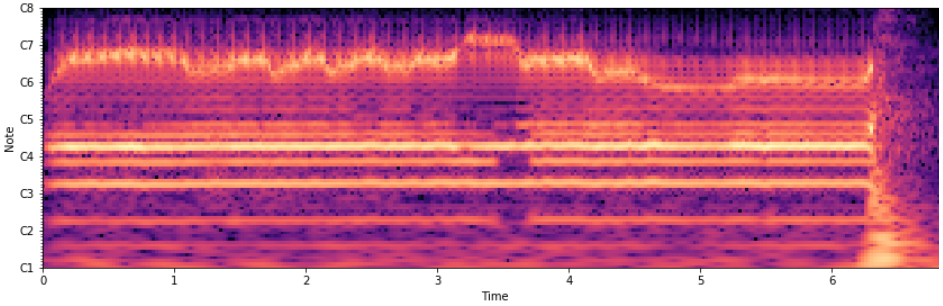
TUVA THROAT SINGING

- Two pitches simultaneously
- One low and constant, noisy, other high and variable
- **Relatively clean spectra (but still noisy overtones)**

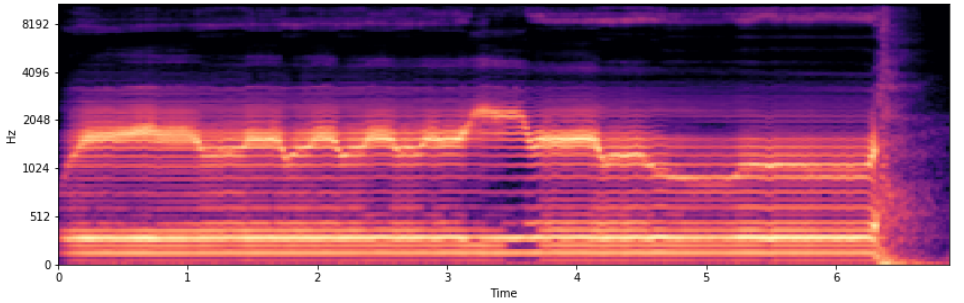
Spectrogram



Constant-Q Transform

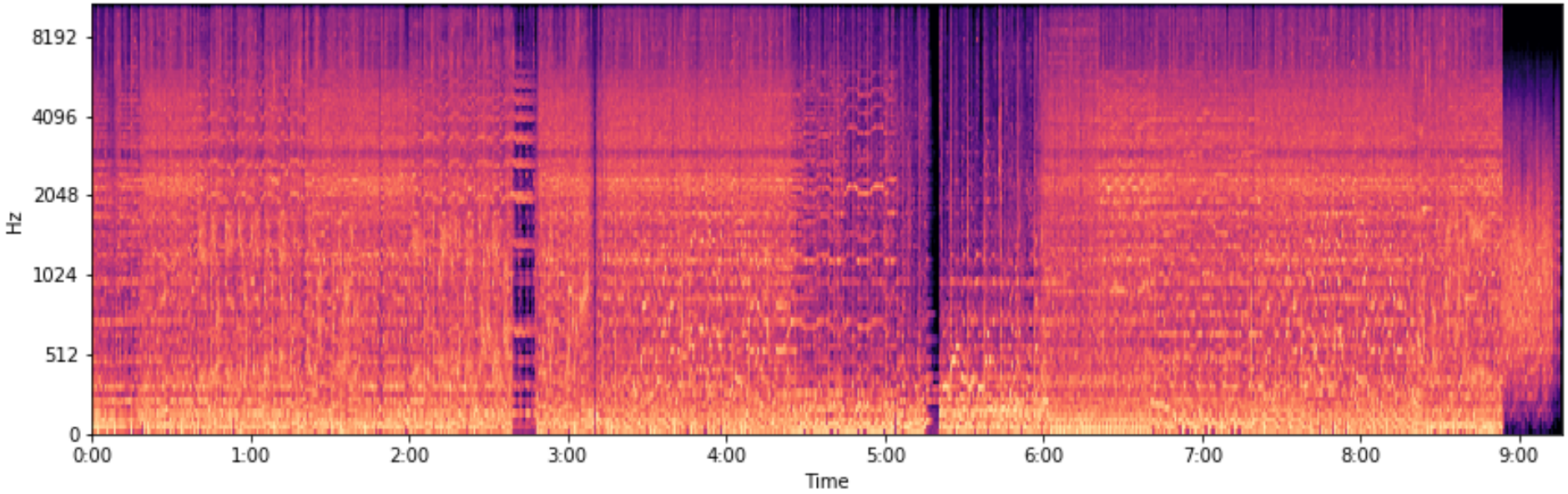


Mel Spec



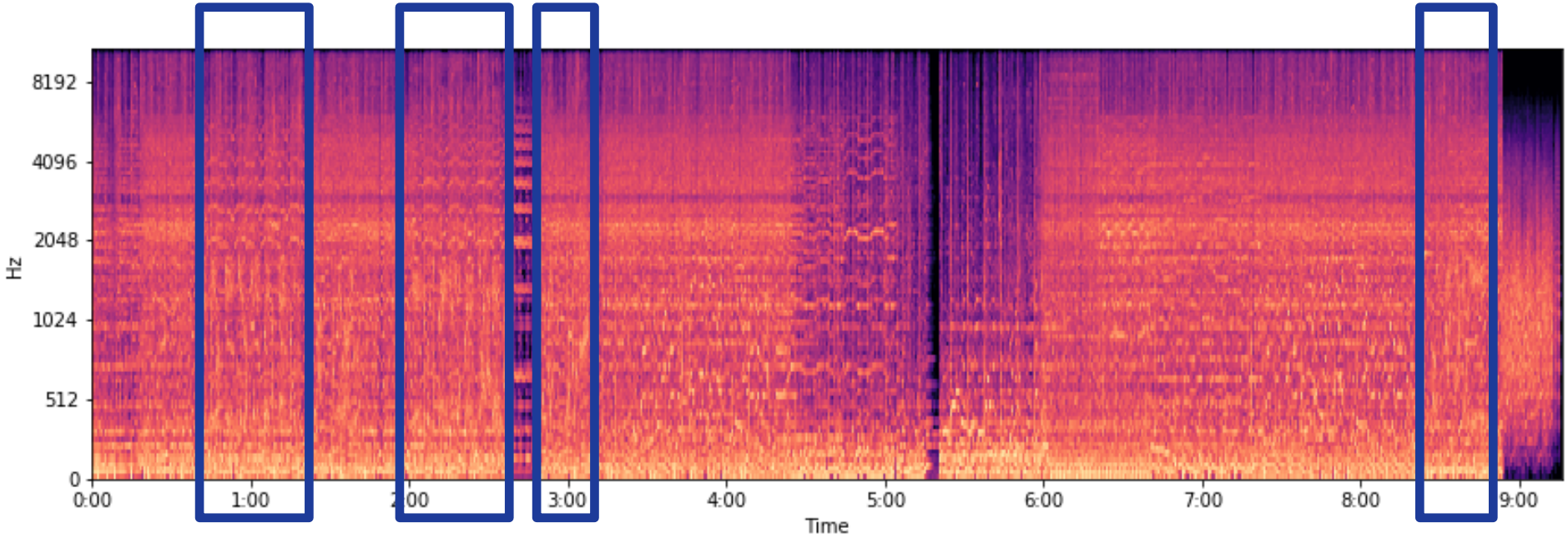
Spectral Analysis

OF BLEAK BY OPETH



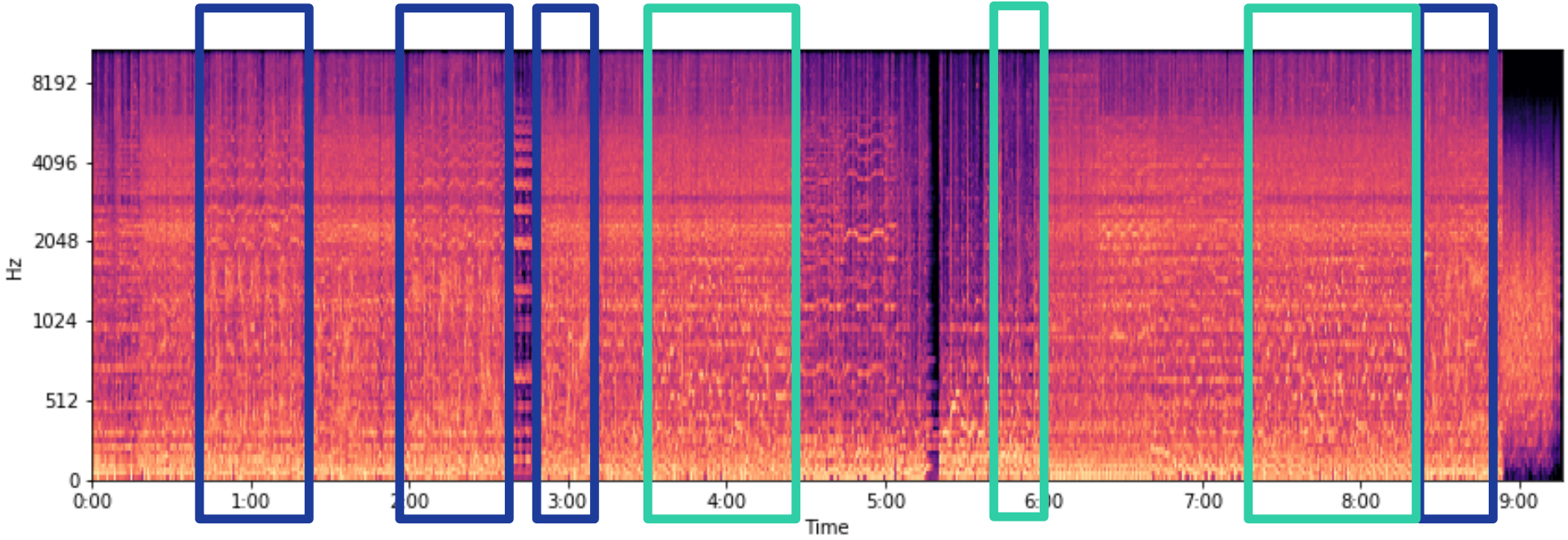
Spectral Analysis

OF BLEAK BY OPETH



Spectral Analysis

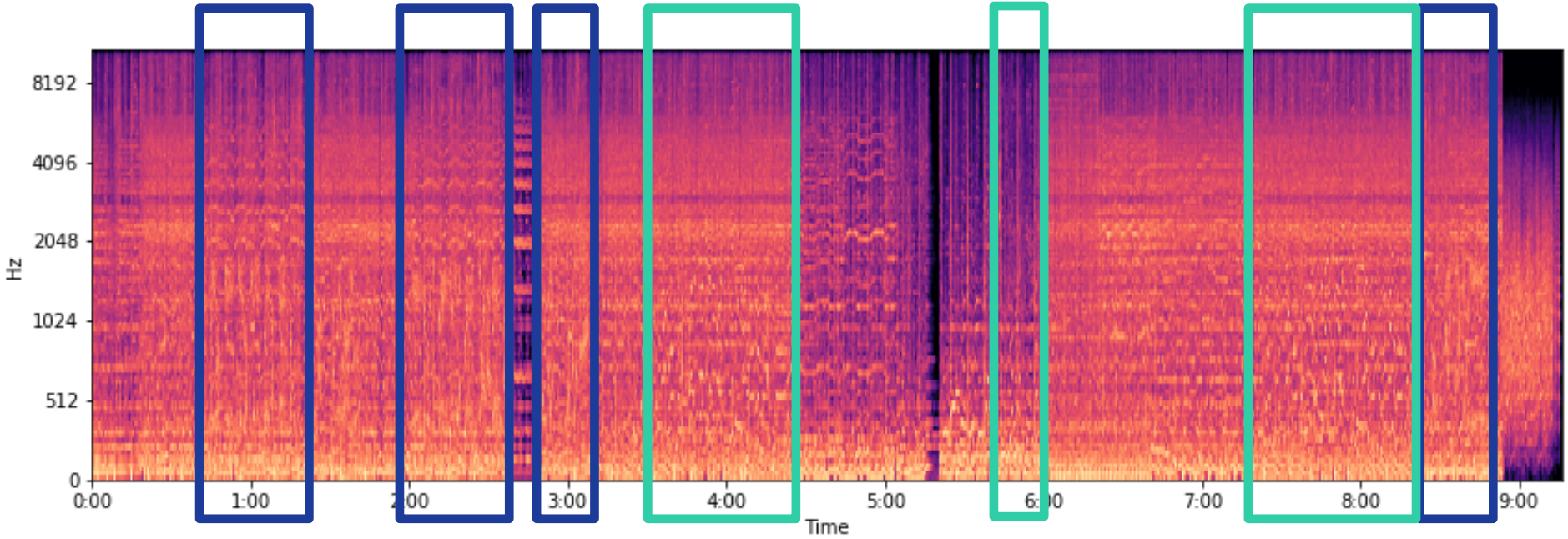
OF BLEAK BY OPETH



Spectral Analysis

OF BLEAK BY OPETH

- Hard to distinguish just visually
- Can we train a mathematical model to learn the difference?



OUTLINE

Motivation and Background: Extreme Vocal Effects

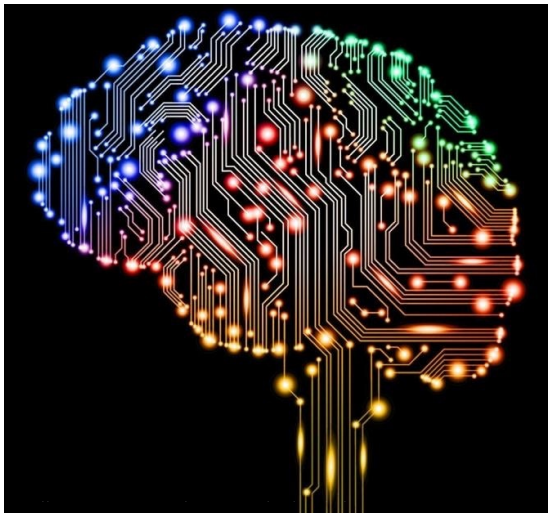
Spectral Analysis of EVEs

Detection of EVEs with Neural Networks

Neural Networks

A.K.A. “UNIVERSAL APPROXIMATORS”

$$g(\mathbf{x}) = \mathbf{y}$$

$\mathbf{x}$ →  → $\hat{\mathbf{y}}$

$$f(\mathbf{x};\theta) \approx \mathbf{y} = \hat{\mathbf{y}}$$

pandora®

Neural Networks

TRAINING PROCESS

Dataset of N pairs of observations and labels:

Observations: $\mathbf{x} \in X$

Labels: $\mathbf{y} \in Y$

$$g(\mathbf{x}) = \mathbf{y}$$

Goal: Minimize the difference between the labels and the network predictions by updating the parameters:

$$\operatorname{argmin}_\theta \frac{1}{N} \sum_{\mathbf{x} \in X, \mathbf{y} \in Y} (f(\mathbf{x}; \theta) - \mathbf{y})^2$$

(Mean Squared Error)

Neural Networks

TRAINING PROCESS

Dataset of N pairs of observations and labels:

Observations: $\mathbf{x} \in X$

Labels: $\mathbf{y} \in Y$

$$g(\mathbf{x}) = \mathbf{y}$$

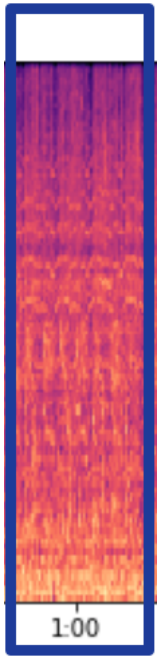
Goal: Minimize the difference between the labels and the network predictions by updating the parameters:

$$\text{argmin}_\theta - \sum_{\mathbf{x} \in X, \mathbf{y} \in Y} \mathbf{y} \log(\mathbf{x}) + (1 - \mathbf{y}) \log(1 - \mathbf{x})$$

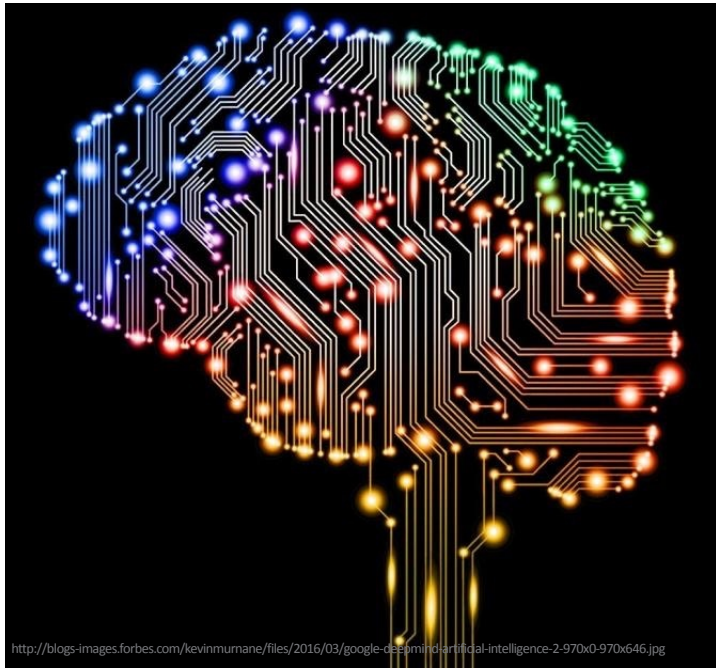
(Binary Crossentropy)

Neural Networks

DETECTING SCREAMS



$\mathbf{x}$



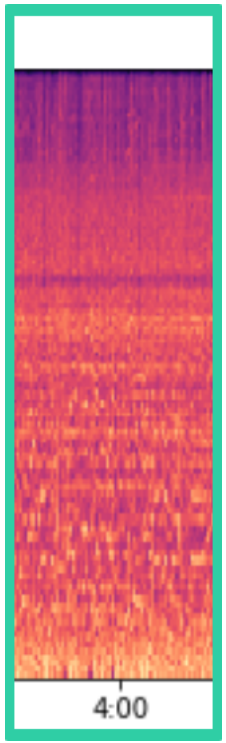
$\hat{\mathbf{y}}$

$$f(\mathbf{x};\theta) \approx \mathbf{y} = \hat{\mathbf{y}}$$

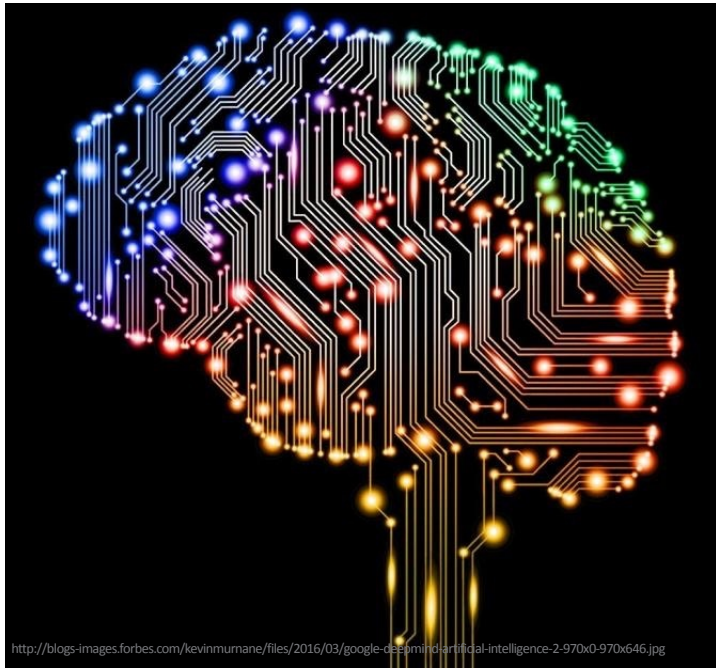
pandora®

Neural Networks

VERY (VERY) BRIEFLY



$\mathbf{x}$



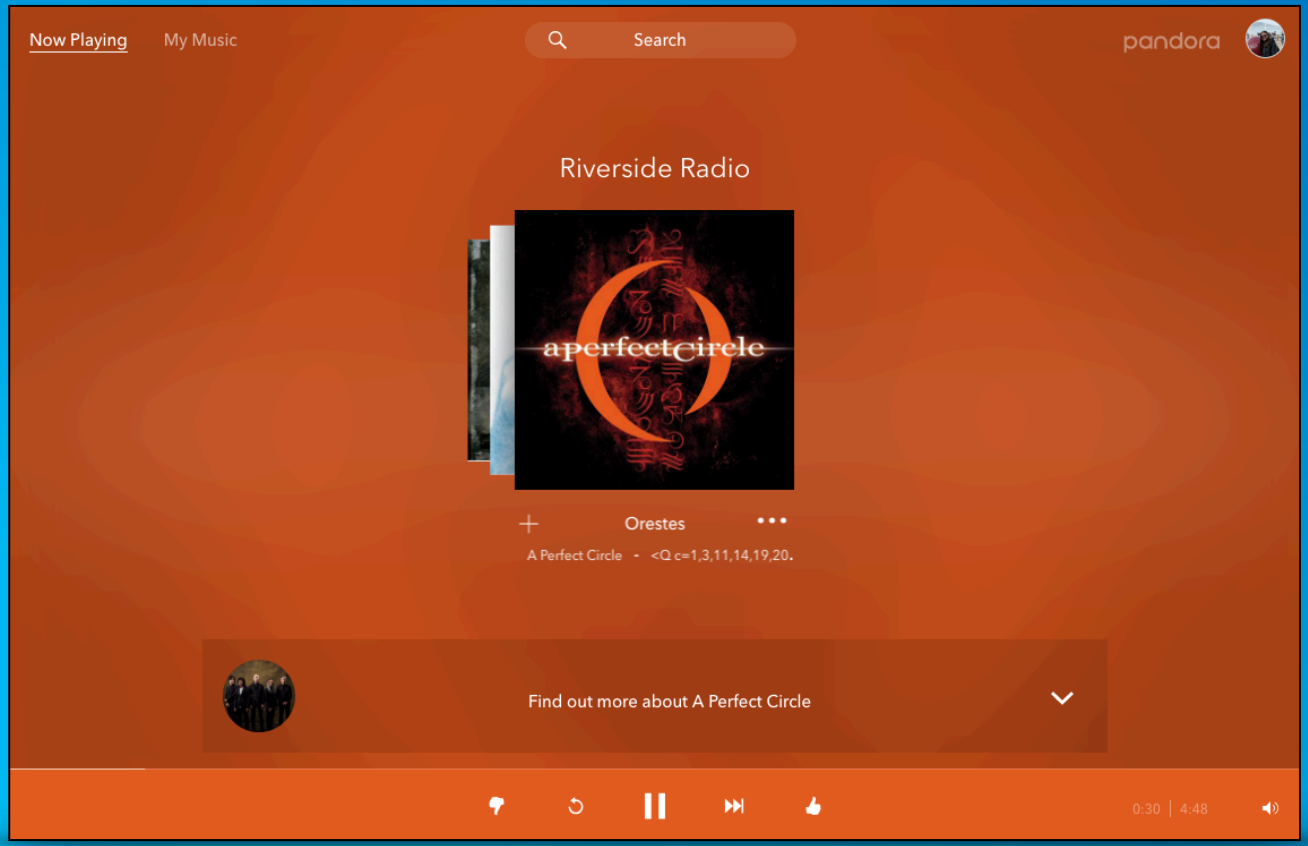
$\hat{\mathbf{y}}$

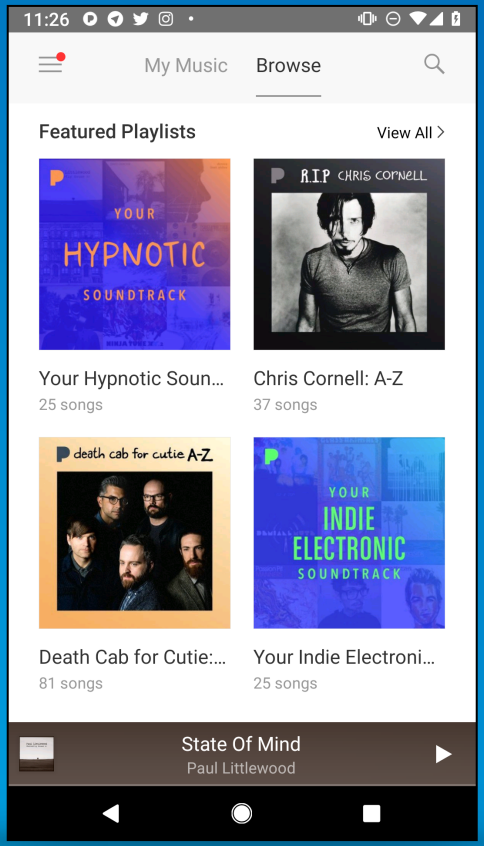
$$f(\mathbf{x};\theta) \approx \mathbf{y} = \hat{\mathbf{y}}$$

pandora®


PANDORA

pandora®


Now Playing
My Music
Search
pandora
Riverside Radio
aperfectcircle
Orestes
A Perfect Circle - <Q c=1,3,11,14,19,20.
Find out more about A Perfect Circle
0:30 | 4:48


11:26
My Music
Browse
Featured Playlists
View All
YOUR HYPNOTIC SOUNDTRACK
Your Hypnotic Soun...
25 songs
R.I.P CHRIS CORNELL
Chris Cornell: A-Z
37 songs
death cab for cutie A-Z
Death Cab for Cutie:...
81 songs
YOUR INDIE ELECTRONIC SOUNDTRACK
Your Indie Electroni...
25 songs
State Of Mind
Paul Littlewood

pandora


Now Playing
My Mus
pandora
0:30 | 4:48


11:26
My Music
Browse
Featured Playlists
View All
State Of Mind
Paul Littlewood

pandora

The Music Genome Project

LARGE-SCALE HUMAN ANNOTATED DATASET

Attribute Examples

Breathy Voice

Nasal Voice

Odd Meter

Has Banjo

Joyful Lyrics

...

Up to ~400 attributes per track

pandora®

The Music Genome Project

LARGE-SCALE HUMAN ANNOTATED DATASET

Attribute Examples

Breathy Voice

Nasal Voice

Odd Meter

Has Banjo

Joyful Lyrics

...

Gritty or Gravely voice

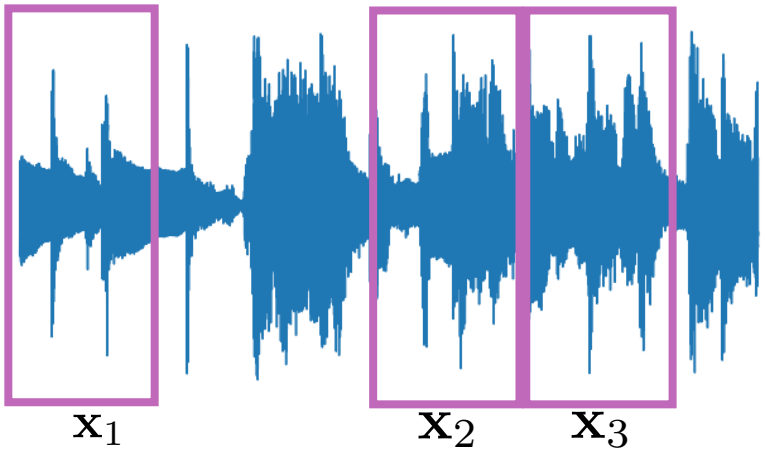
Up to ~400 attributes per track

pandora®

Gathering Data

USING THE MUSIC GENOME PROJECT

- Data set:
  - N= 8k tracks (4k with EVEs, 4k without)
  - All Hard Rock / Punk / Metal tracks
  - Sample ~2 second patches from each track:
    - 10 patches per track
    - Total of ~80k patches
    - 10% for testing



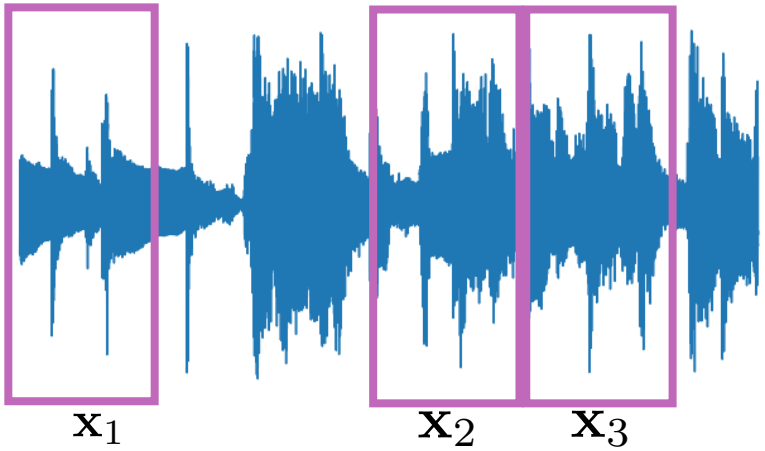
$$\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y}_3$$

pandora®

Gathering Data

USING THE MUSIC GENOME PROJECT

- Data set:
  - N = 8k tracks (4k with EVEs, 4k without)
  - All Hard Rock / Punk / Metal tracks
  - Sample ~2 second patches from each track:
    - 10 patches per track
    - Total of ~80k patches
    - 10% for testing



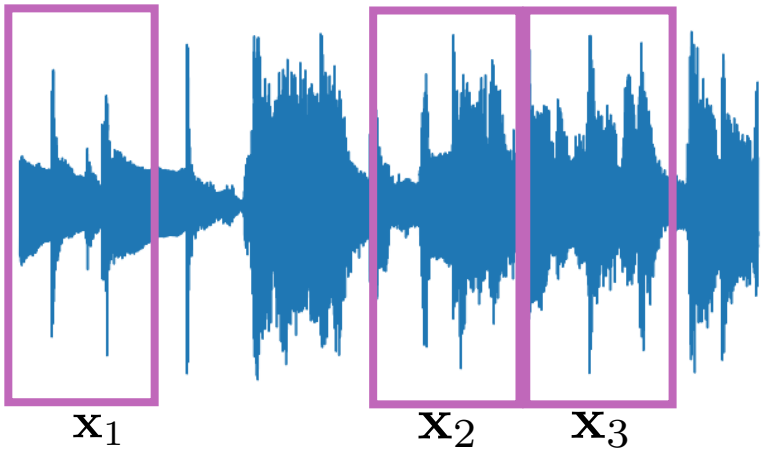
$$\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y}_3$$

pandora®

Gathering Data

USING THE MUSIC GENOME PROJECT

- Data set:
  - N= 8k tracks (4k with EVEs, 4k without)
  - All Hard Rock / Punk / Metal tracks
  - Sample ~2 second patches from each track:
    - 10 patches per track
    - Total of ~80k patches
    - 10% for testing

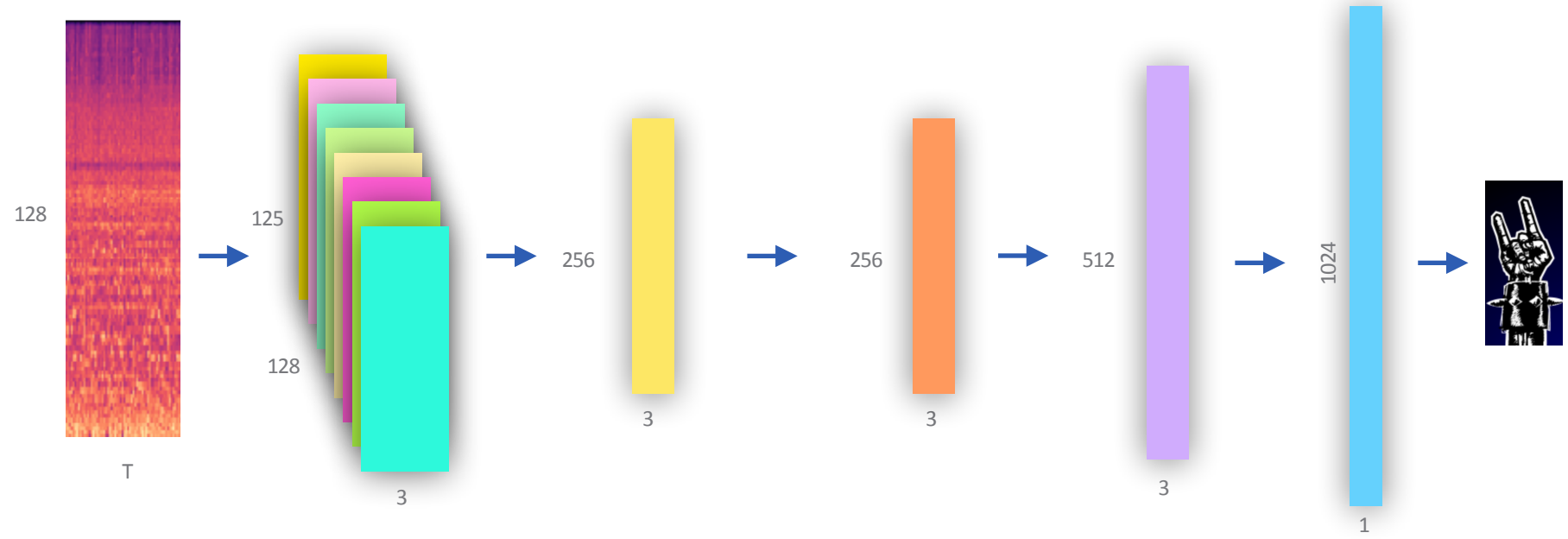


$$\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y}_3$$

pandora®

Deep Architecture

CONVOLUTIONAL NEURAL NETWORK

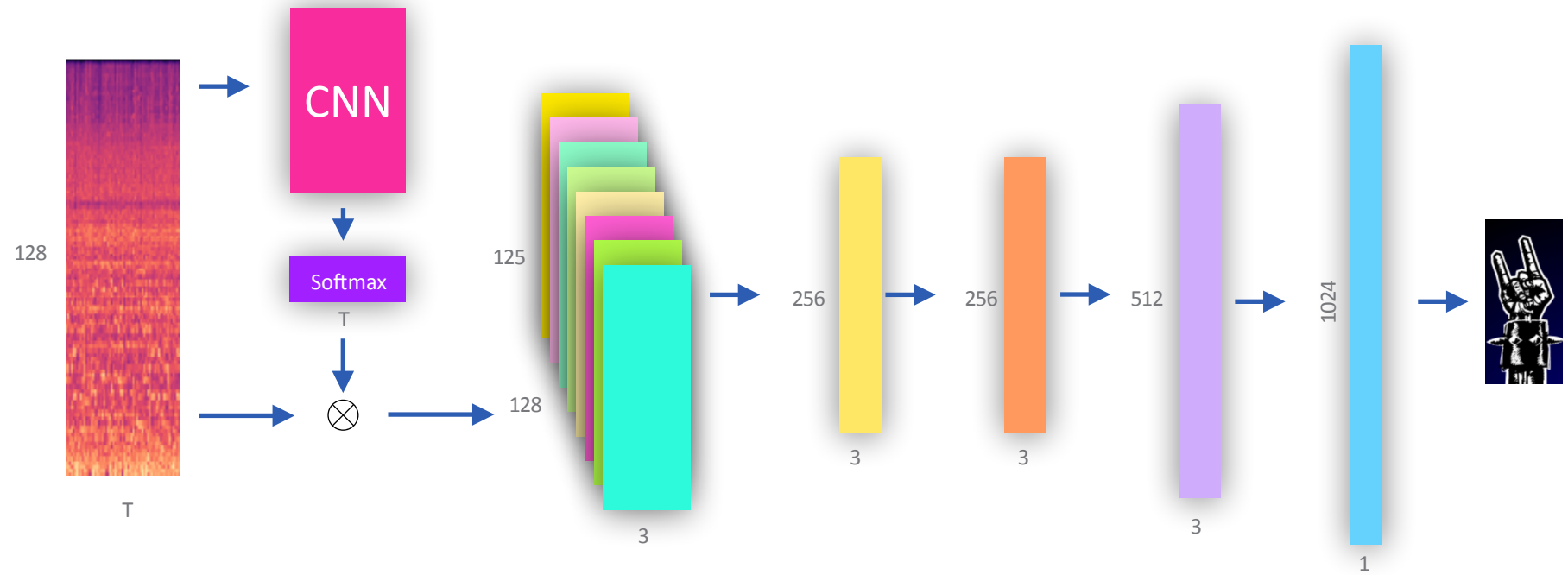


(Pons 2017)

pandora®

Deep Architecture

CONVOLUTIONAL NEURAL NETWORK W/ SOFT ATTENTION



Soft Attention | Convolutional Layers + Max Pooling + Batch Norm + ELU | Dense Layer

(Balke, 2019)

pandora®

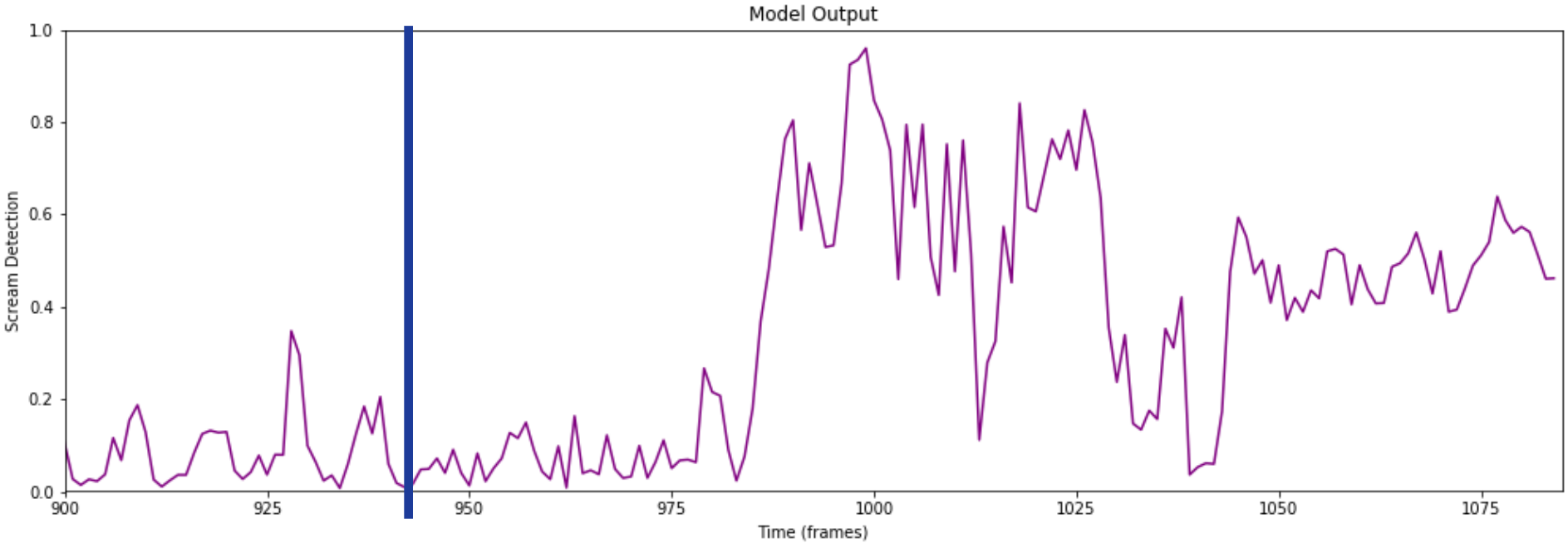
Training and Testing

- Loss: Binary Crossentropy
- Dropout: 20% in Dense Layer
- Adam Optimizer
  - 1e-4 learning rate
- Mini-batch: 16 observations
- After training (~30min):
  - 85.4% accuracy in test subset (w/o soft attention)
  - 80.2% accuracy in test subset (w/ soft attention)

$$\text{argmin}_\theta - \sum_{\mathbf{x} \in X, \mathbf{y} \in Y} \mathbf{y} \log(\mathbf{x}) + (1 - \mathbf{y}) \log(1 - \mathbf{x})$$

Examples

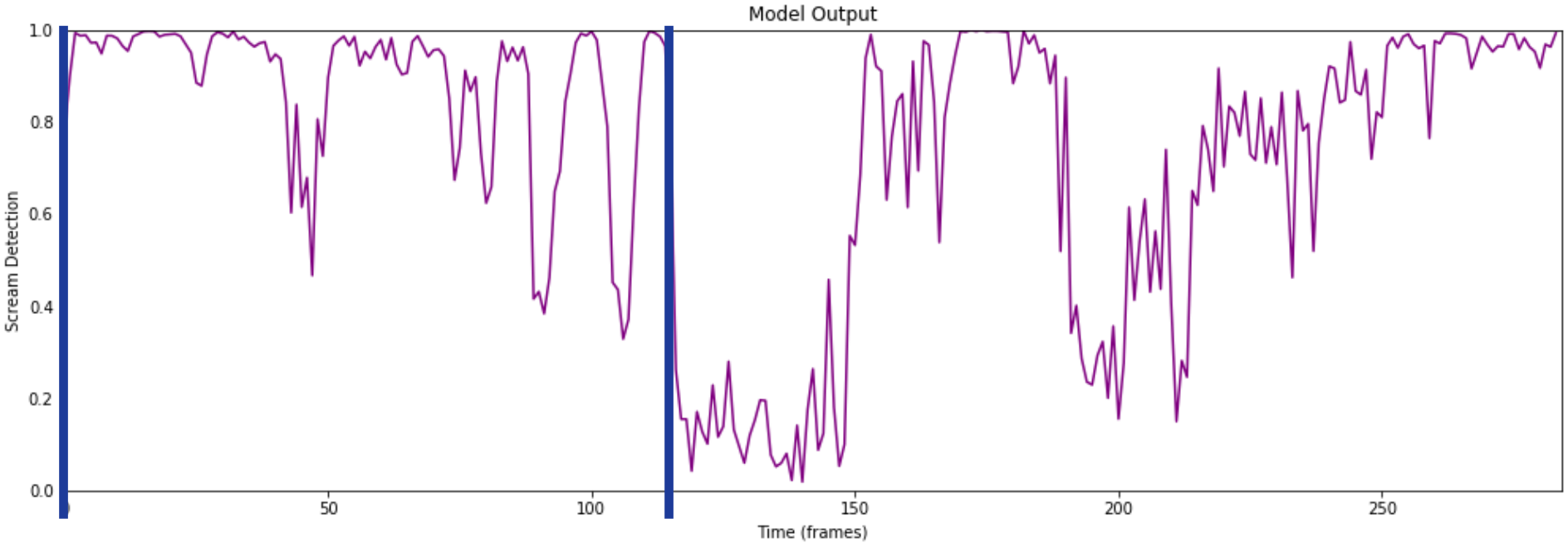
BLEAK - OPETH



pandora®

Examples

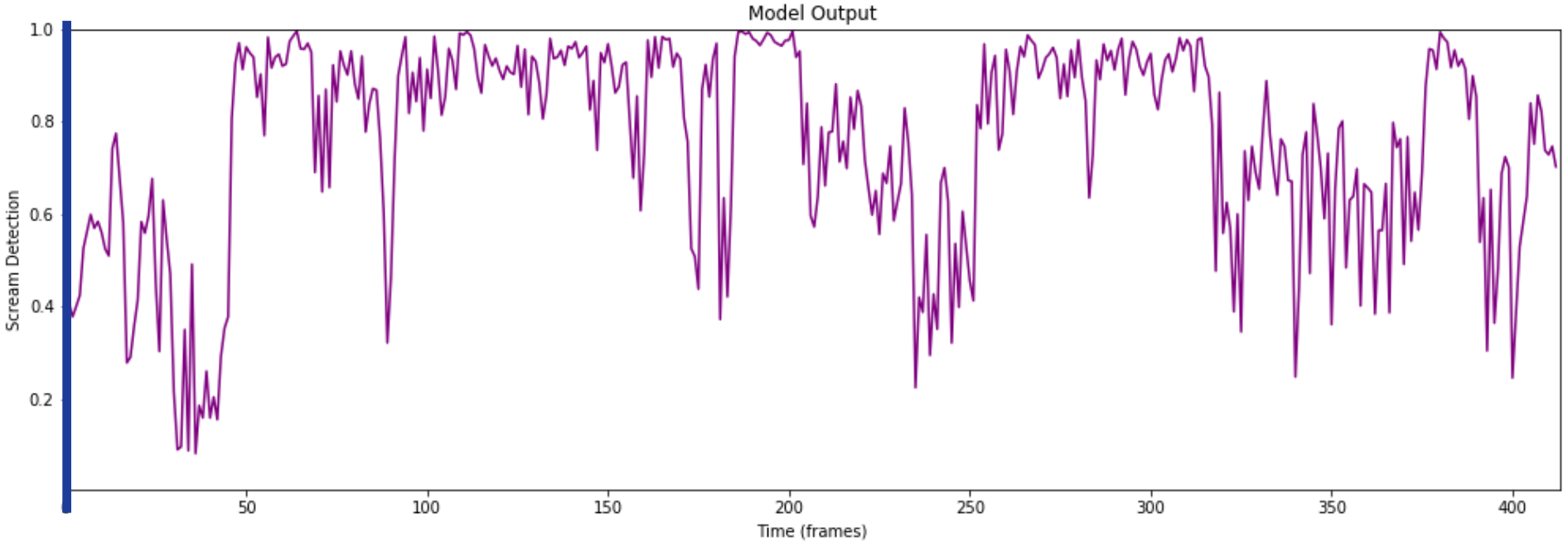
PANASONIC YOUTH - THE DILLINGER ESCAPE PLAN



pandora®

Examples

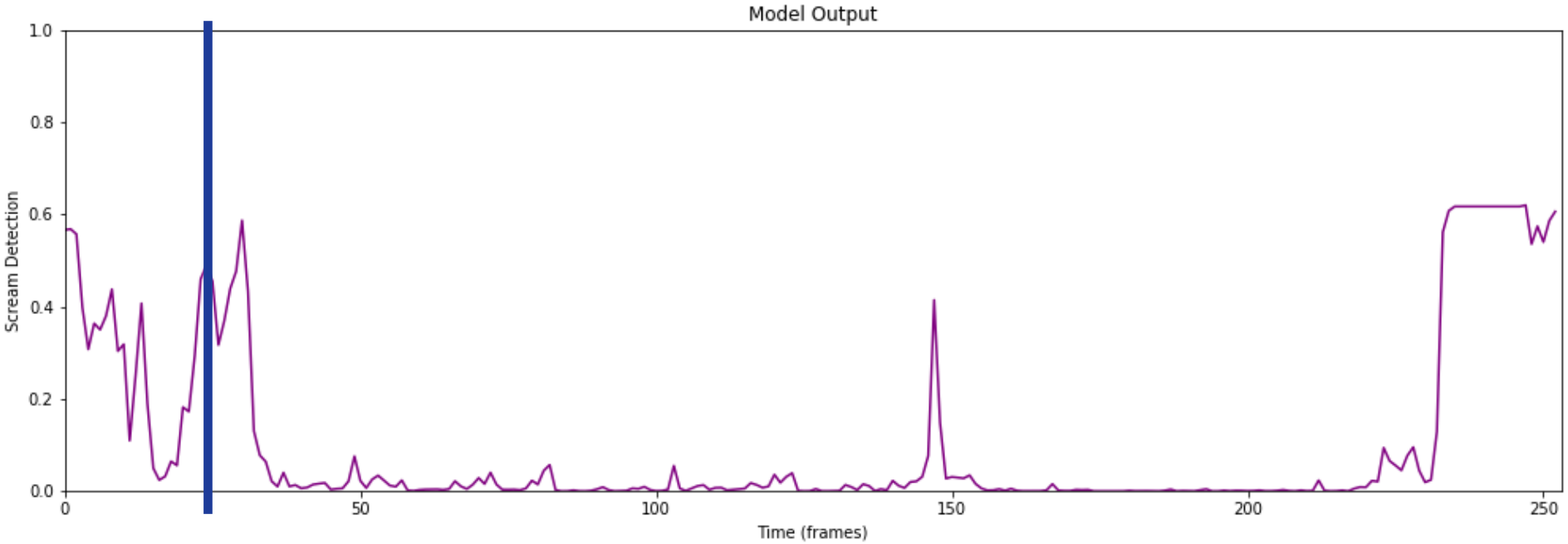
ROOTS, BLOODY ROOTS - SEPULTURA



pandora®

Examples

DARK MATTER - PORCUPINE TREE



pandora®

CONCLUSIONS

Discussed several spectra of different types of EVEs

Spectral analysis on full tracks might be too noisy to perform visually

Neural networks can detect such EVEs when using spectra as input

pandora®

THE FUTURE

Investigate soft-attention further?
Publish some the dataset?
Use Screaminator's screams as training data?
Extend to Tuvan Singing?
Automatically classify different EVEs?

Publish some of this work (looking for collaborators!)

Use deeper architectures similar to state-of-the-art music autotagging

pandora®

References

Balke, S., Dorfer, M., Carvalho, L., Arzt, A., Widmer, G., Learning Soft-Attention Models for Tempo-Invariant Audio-Sheet Music Retrieval, ISMIR, Delft, The Netherlands, 2019

Bonada, J., Blaauw M., Generation of Growl-Type Voice Qualities by Spectral Morphing, In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, Vancouver, Canada, 2013.

Carr, C. J., Zukowski, Z., Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands. 6th International Workshop on Musical Metacreation (MUME), Spain, 2018.

Loscos, A., Bonada, J., Emulating Rough and Growl Voice in Spectral Domain, In Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx'04), Naples, Italy, 2004.

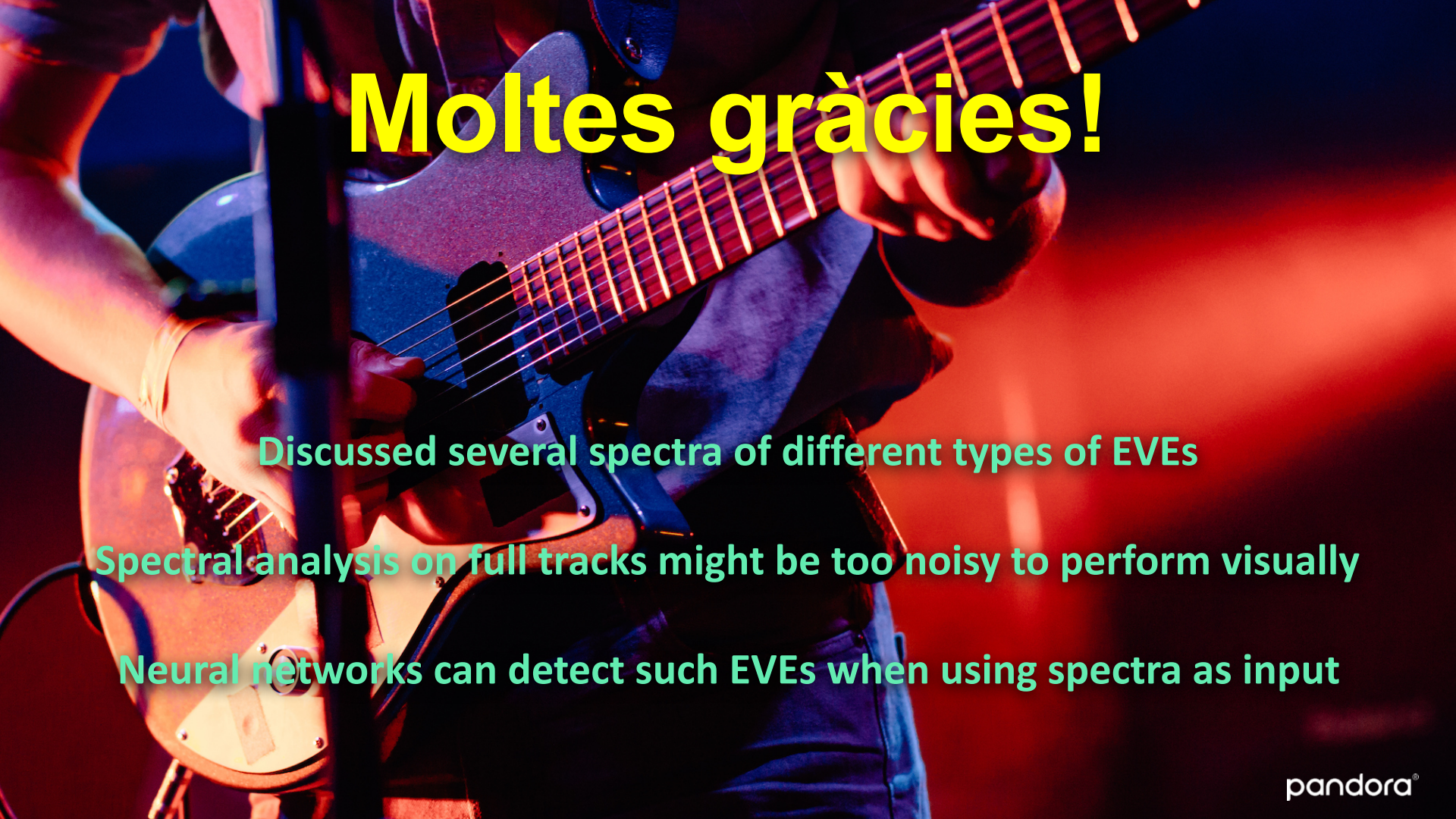
Mayor, O. Bonada, J. Janer, J., "KaleiVoiceKids: Interactive Real-Time Voice Transformation for Children" Proceedings of 9th International Conference on Interaction Design and Children; Barcelona, Spain, 2010.

Nieto, O., Unsupervised Clustering of Extreme Vocal Effects. Proc. of the 10th International Conference Advances in Quantitative Laryngology, Voice and Speech Research (AQL), pages 115-116. Cincinnati, OH, USA, 2013.

Nieto, O., Voice Transformations for Extreme Vocal Effects. Master's Thesis. Pompeu Fabra University. 2008

Pons, J., Serra, X., Designing efficient archi-tectures for modeling temporal features with convo-lutional neural networks. InIEEE International Con-ference on Acoustics, Speech and Signal Processing(ICASSP), 2017

pandora®


Moltes gràcies!
Discussed several spectra of different types of EVEs
Spectral analysis on full tracks might be too noisy to perform visually
Neural networks can detect such EVEs when using spectra as input
pandora®