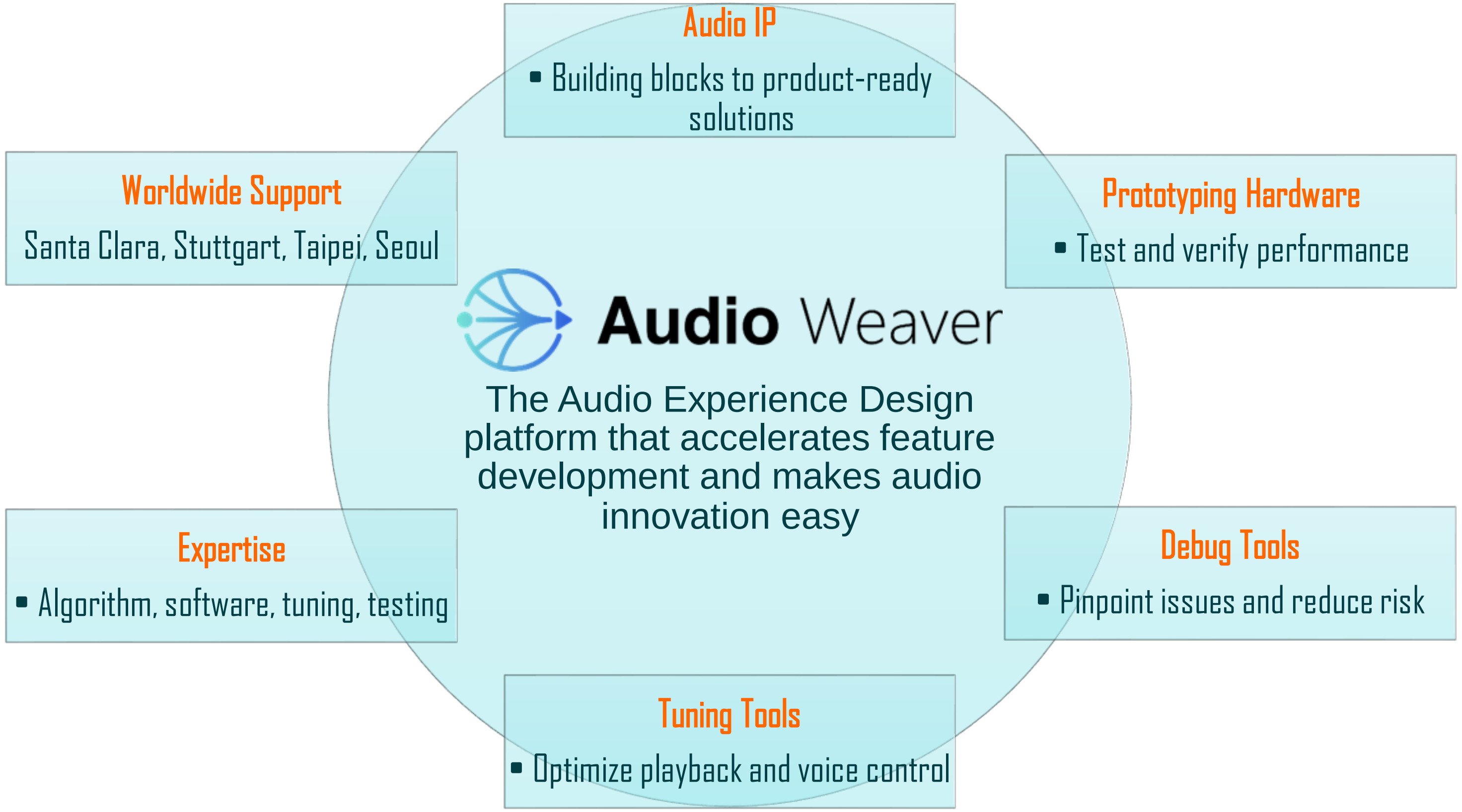
Comparing ML-Based Audio with ML-Based Vision: An Introduction to ML Audio for ML Vision Engineers

Josh Morris
Engineering Manager, Machine Learning
DSP Concepts

The Audio of Things Approach

DSP Concepts helps product makers deliver remarkable Audio Experience through a flexible and modular approach within a design platform environment. This system makes the entire workflow faster and easier across prototyping, design, debugging, tuning, production, and even over-the-air updates.



Audio IP

- Building blocks to product-ready solutions

Worldwide Support

Santa Clara, Stuttgart, Taipei, Seoul

Prototyping Hardware

- Test and verify performance

Audio Weaver

The Audio Experience Design platform that accelerates feature development and makes audio innovation easy

Expertise

- Algorithm, software, tuning, testing

Debug Tools

- Pinpoint issues and reduce risk

Tuning Tools

- Optimize playback and voice control

Motivation for Talk

Processing at the edge is getting more and more powerful making it possible to do things that were reserved for the cloud

Audio is becoming increasingly popular

- Standalone applications
  - Smart assistant
  - Voice control
  - Denoising
- Multi-modal applications
  - Industrial sensing
  - Anomaly detection

Agenda

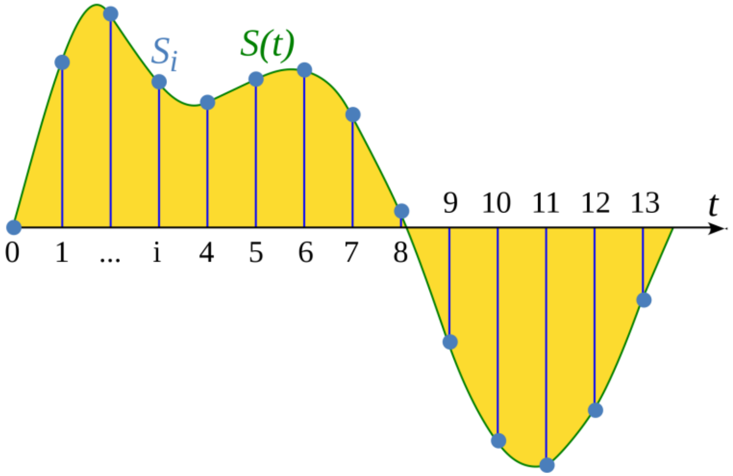
- **Feature engineering**
  - How is audio different from vision?
  - How is audio like vision?
- **Implementation**
  - Similarities
  - Differences

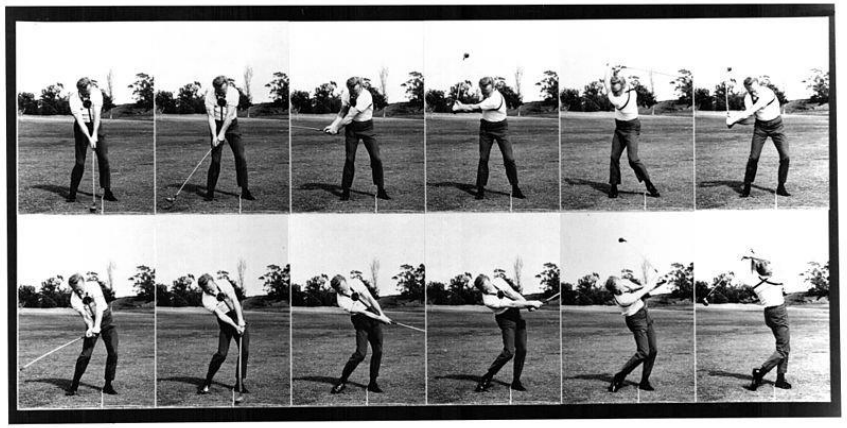
- **Common problems in vision and their audio analogues**
  - Classification
  - Sequence decoding
  - Restoration/denoising

Features

How Are Audio Features Different from Vision Features?

- **Audio is a sequence of samples**
  - Somewhere between video/images
  - Inherent left to right structure to data
  - Sample rate
  - Bit depth

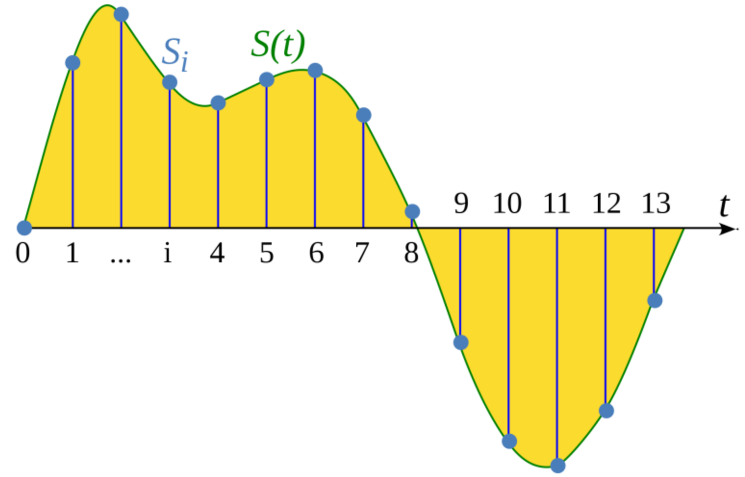


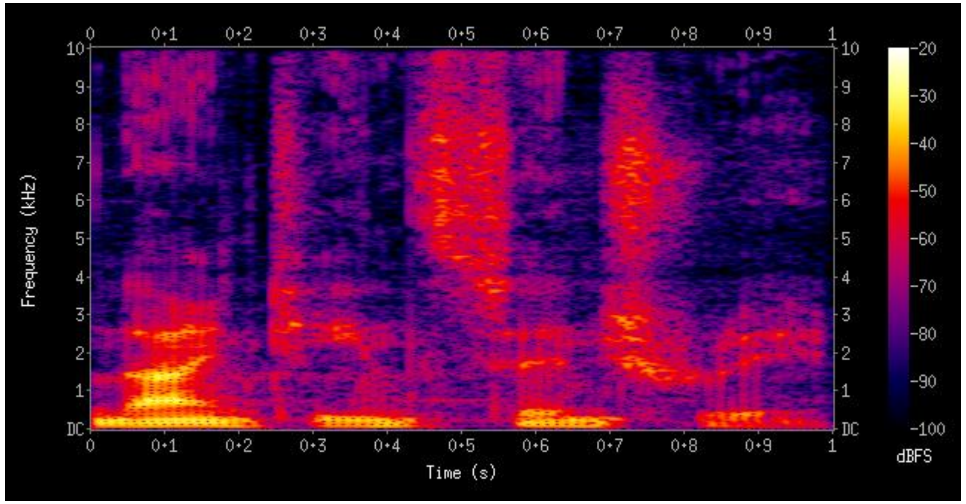


https://en.wikipedia.org/wiki/Sampling_(signal_processing)
https://en.wikipedia.org/wiki/File:Mike_Austin_Sequence.JPG

Manipulating Audio from Time Domain to Frequency Domain

A lot of techniques employed for ML audio-based solutions borrow techniques from vision. This means we need to take a 1D sequence of samples and make them look like an image.

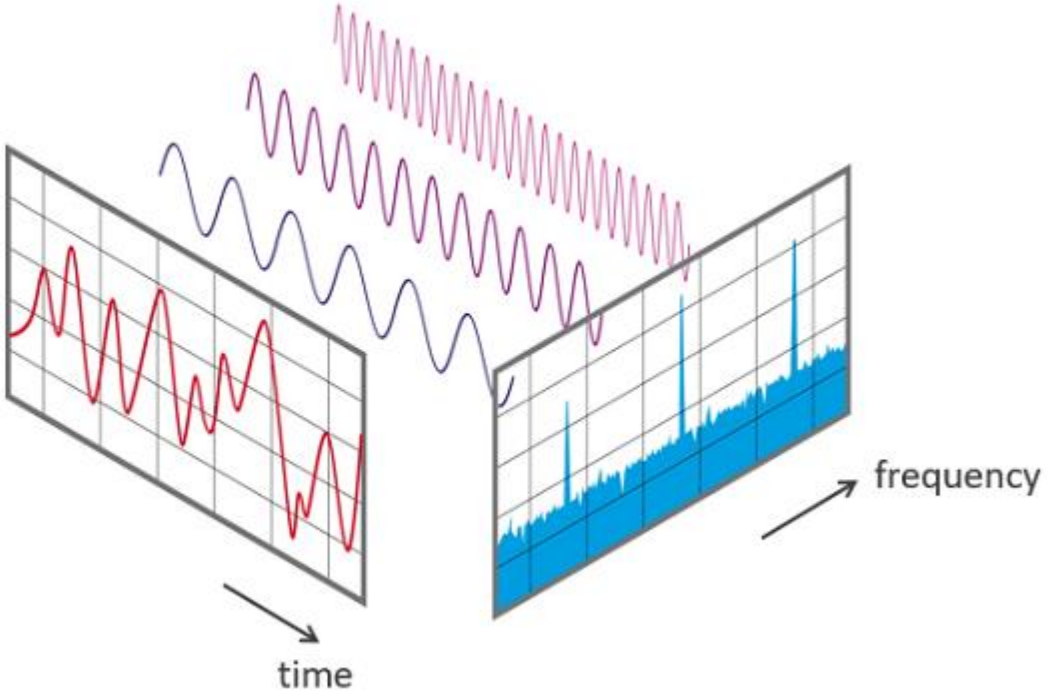




https://en.wikipedia.org/wiki/Sampling_(signal_processing)
https://upload.wikimedia.org/wikipedia/commons/c/c5/Spectrogram-19thC.png

Fast Fourier Transform

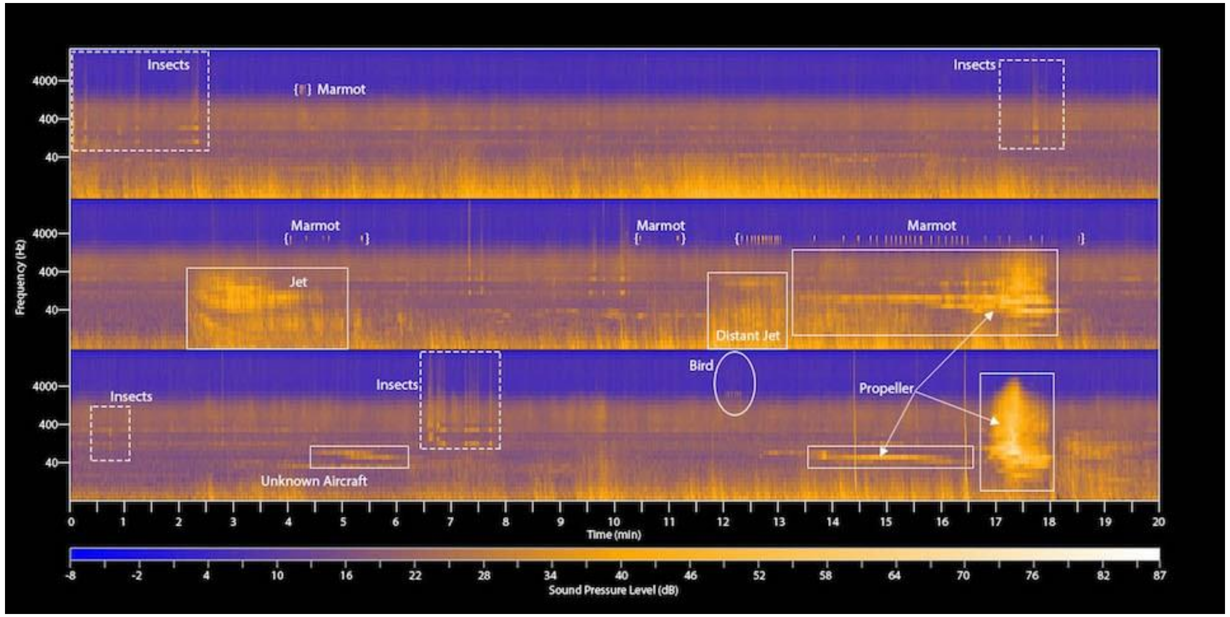
- **Fast Fourier transform (FFT)**
  - Shows frequency over time
  - Linearly spaced frequency bins
  - Can apply processing in the frequency domain and then use an inverse fast Fourier transform (IFFT) to get time domain audio



https://commons.wikimedia.org/wiki/File:FFT-Time-Frequency-View.png

Power Spectrogram

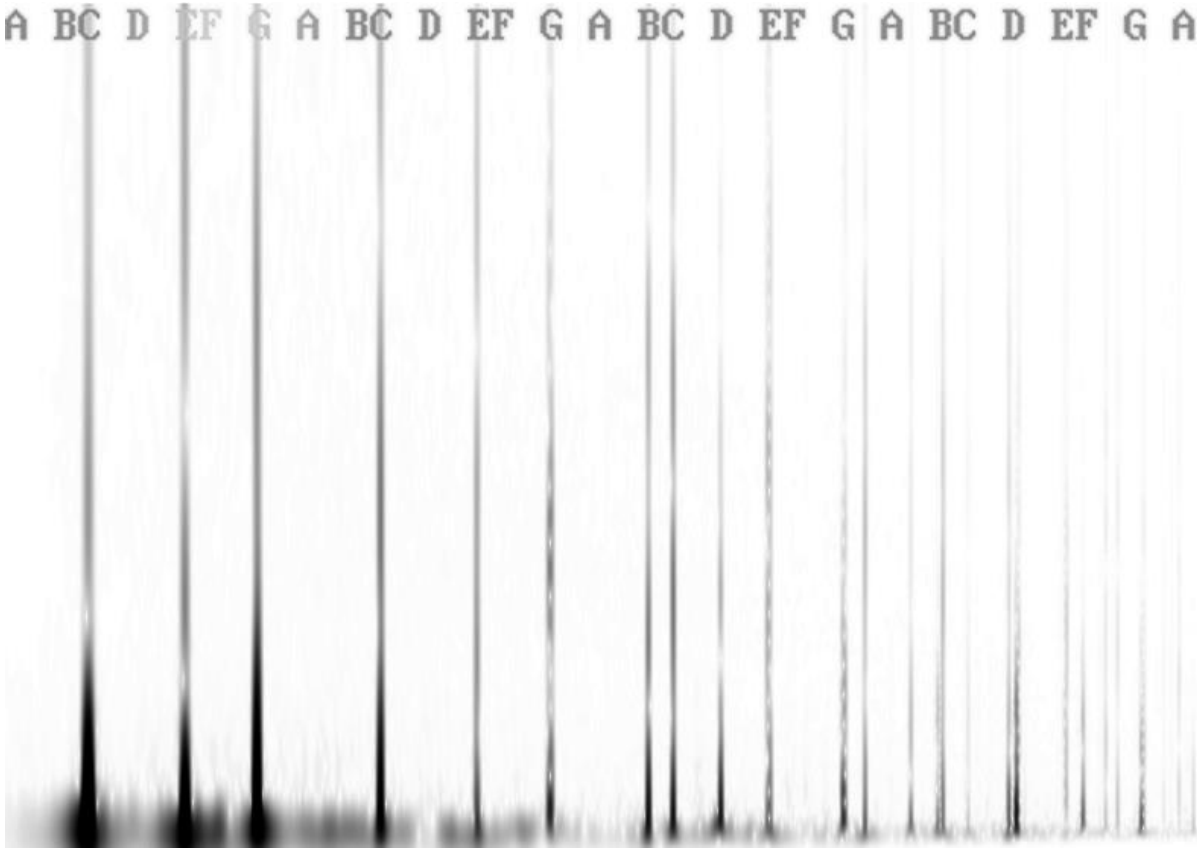
- **Built from short time Fourier transform**
  - Repeat Fourier transform with set window and hop size
    - short-time Fourier transform (STFT)
  - Take magnitude squared of frequency bins
  - Common for classification tasks



https://upload.wikimedia.org/wikipedia/commons/9/99/Mount_Rainier_soundscape.jpg

Constant Q Transform

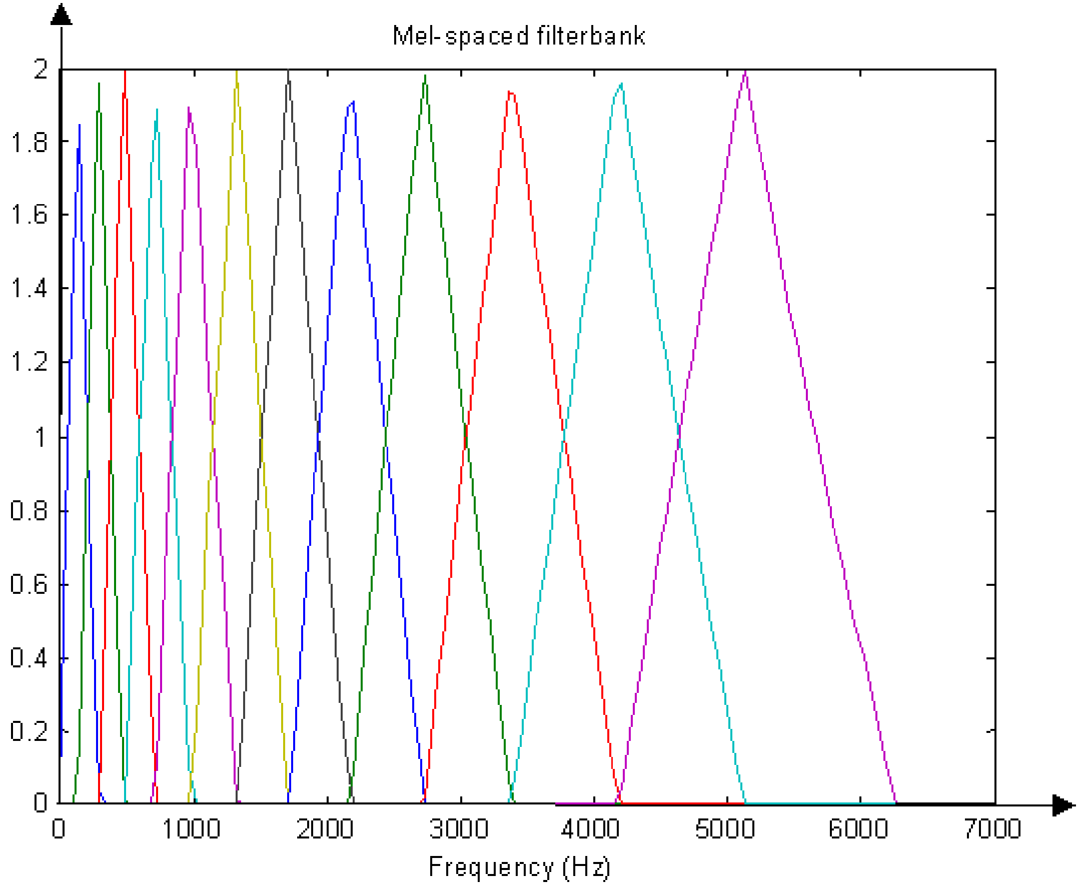
- **Constant Q transform (CQT)**
  - Logarithmically spaced frequency bins
  - Popular for musical applications
  - More computationally efficient since fewer bins are needed to cover a frequency range



https://en.wikipedia.org/wiki/Constant-Q_transform#/media/File:CQT-piano-chord.png

Mel Spectrogram

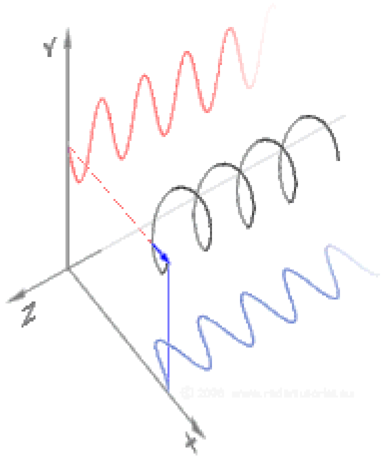
- **Mel spectrogram**
  - Triangular frequency windows
  - Filter banks that attempt to approximate human hearing
    - Humans struggle to hear frequencies that are close together. This anomaly is known as masking.

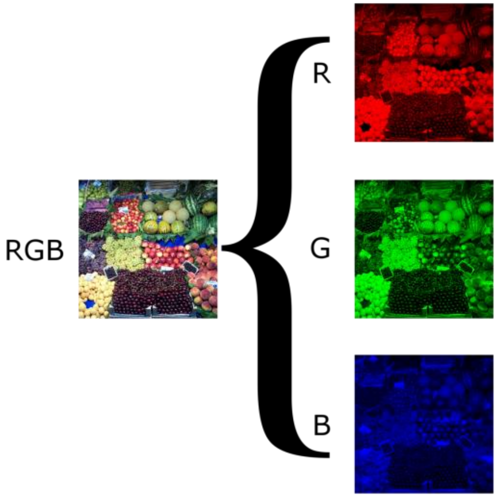


http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/speaker_recognition.html

How Are Audio Features the Same as Vision?

- **Feature normalization**
- **Multi-channel features for complex audio**
  - Color channels in audio
  - Real and imaginary components in frequency time



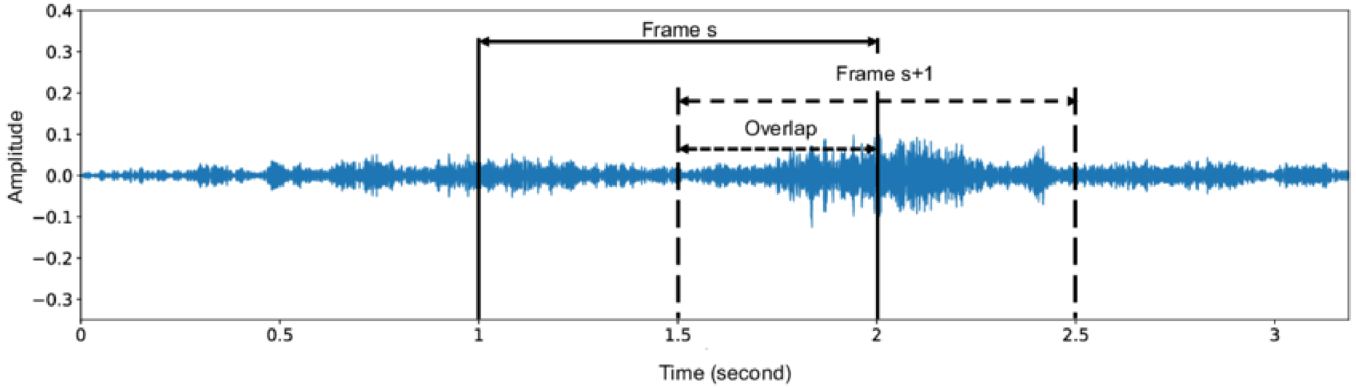


https://en.wikipedia.org/wiki/Fourier_transform#/media/File:Rising_circular.gif
https://en.wikipedia.org/wiki/Grayscale#/media/File:Beyoglu_4671_tricolor.png

Implementation

How Do Implementations of Audio Solutions Differ from Vision Solutions?

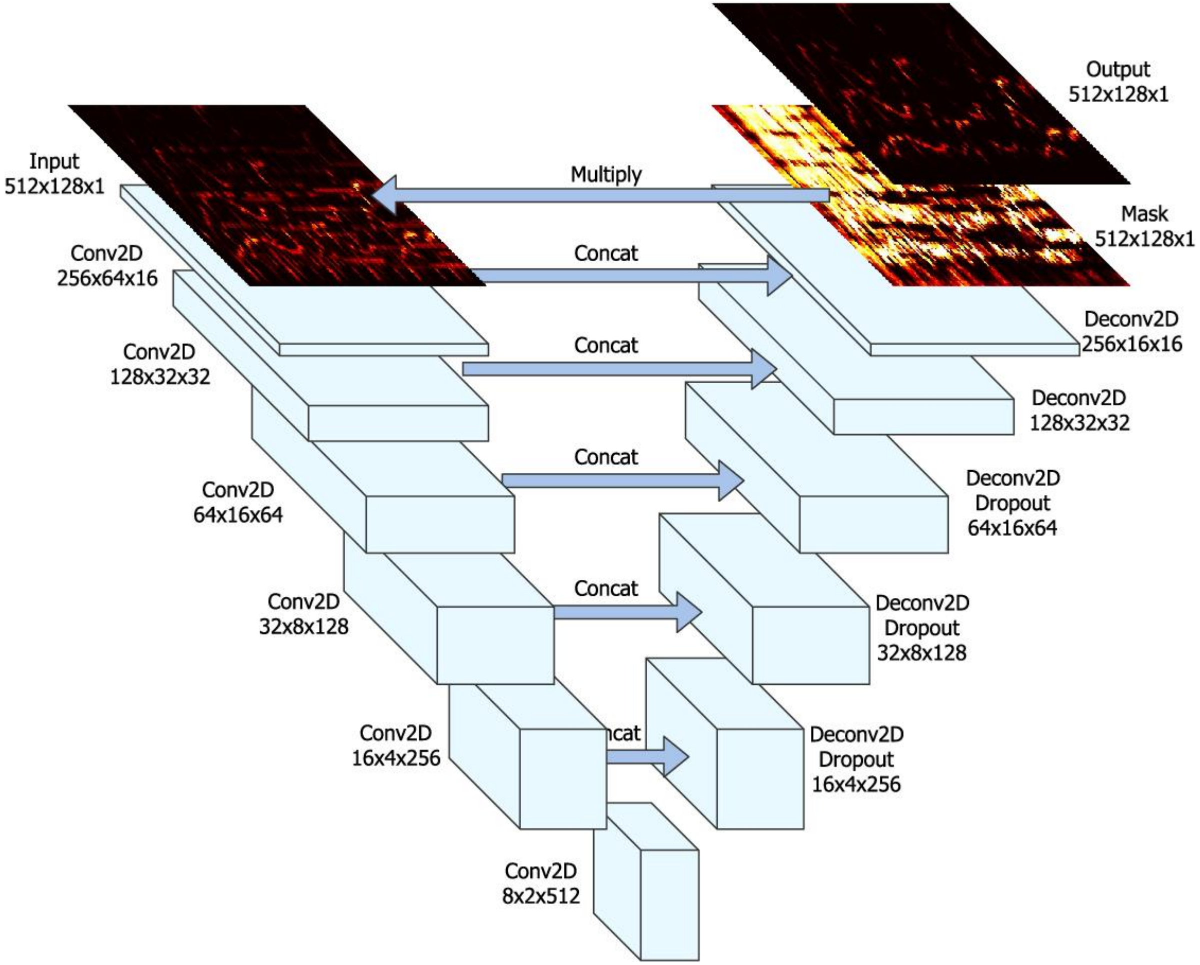
- **Take audio and create an image-like input with fixed dimensions.**
  - Take the input signal into the frequency domain
  - Take a window of feature vectors
  - Slide window with a hop, usually smaller than the input dimension of the model



https://www.jonnor.com/2021/12/audio-classification-with-machine-learning-europython-2019/

How Are Implementations of Audio Solutions Like Vision Solutions?

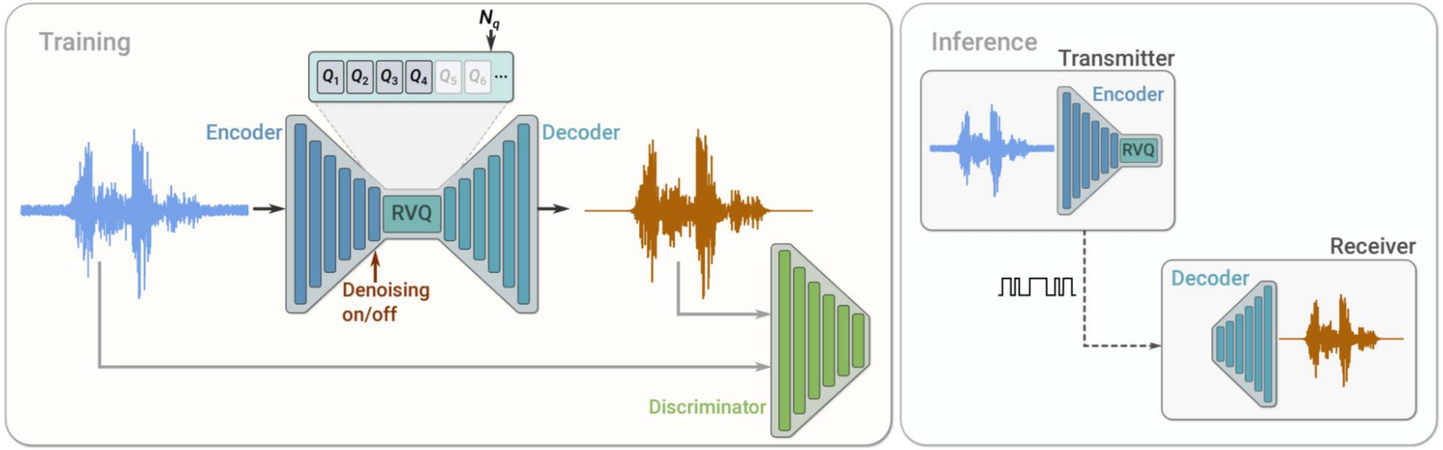
- **Use the same layers**
  - Convolutional layers
    - Conventional
    - Depth-wise separable
  - Residual blocks
  - RNN
- **Use the same training tricks**



https://www.semanticscholar.org/paper/Singing-Voice-Separation-with-Deep-U-Net-Networks-Jansson-Humphrey/83ea11b45cba0fc7ee5d60f608edae9c1443861d

How Are Implementations of Audio Solutions Like Vision Solutions?

- **Some popular vision model architectures show up**
  - EfficientNet
- **Similar concepts**
  - Encoder-decoder network
  - Transfer learning

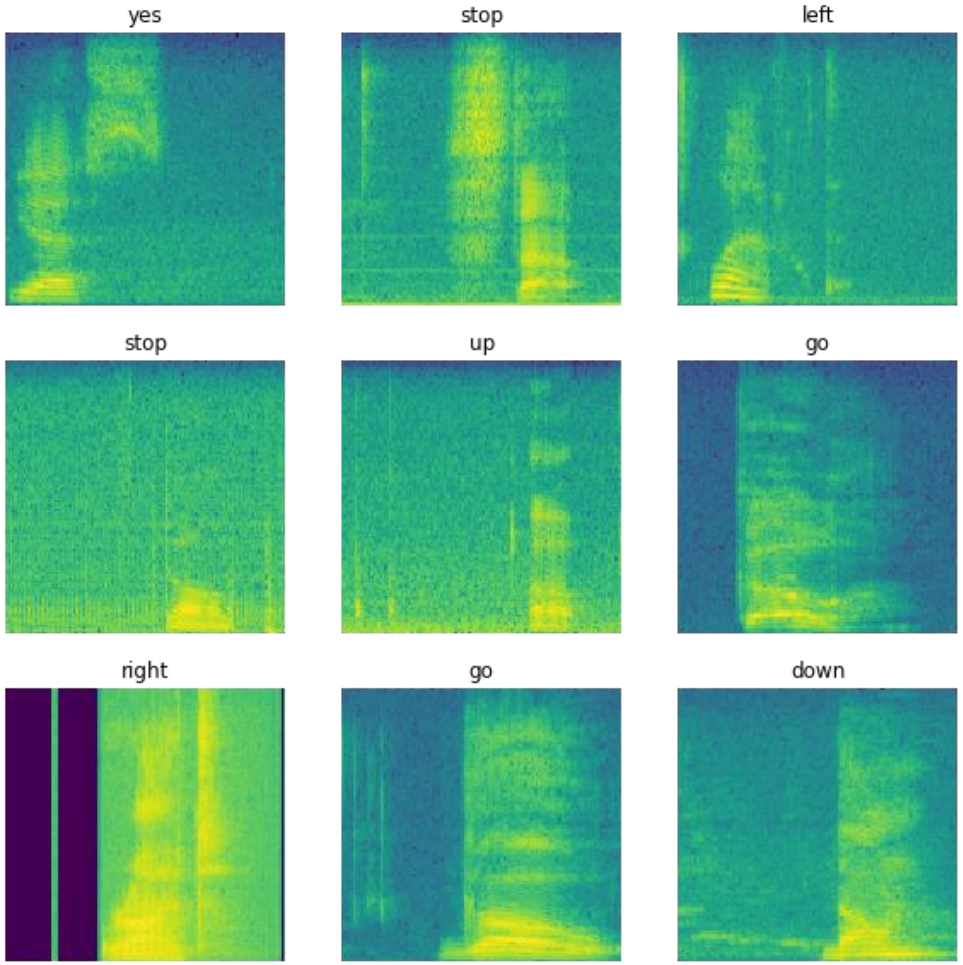


https://ai.googleblog.com/2021/08/soundstream-end-to-end-neural-audio.html

Common Problems in Vision and Their Audio Analogues

Classification

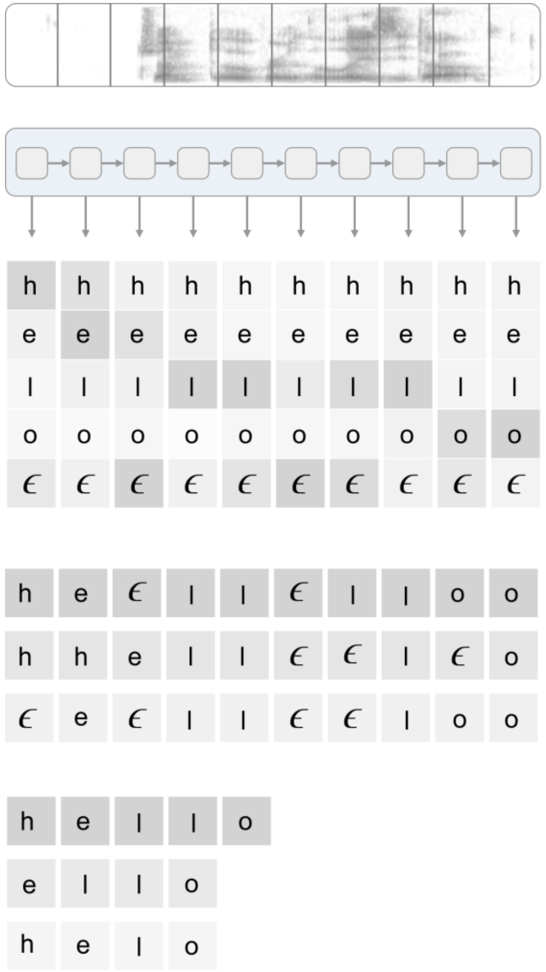
- **Very similar**
- **Convolution layers pull out structural information**
- **Trained on frequency domain features**
- **In audio, we usually have a sliding window**
  - Like working with a camera stream
- **Can be important for energy savings in complex systems**
  - Motion detection
  - Voice activity detection



https://www.tensorflow.org/tutorials/audio/simple_audo

Sequence Decoding

- **Vision**
  - Optical character recognition
- **Audio**
  - Automatic speech recognition
- **Both look for structural information in their inputs and decode them to a character sequence**
- **Similar architectures**
  - RCNN
  - Transformer



We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

h	h	h	h	h	h	h	h	h	h
e	e	e	e	e	e	e	e	e	e
l	l	l	l	l	l	l	l	l	l
o	o	o	o	o	o	o	o	o	o
ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ

The network gives $p_t(a \mid X)$, a distribution over the outputs {h, e, l, o, ϵ} for each input step.

h	e	ϵ	l	l	ϵ	l	l	o	o
h	h	e	l	l	ϵ	ϵ	l	ϵ	o
ϵ	e	ϵ	l	l	ϵ	ϵ	l	o	o

With the per time-step output distribution, we compute the probability of different sequences

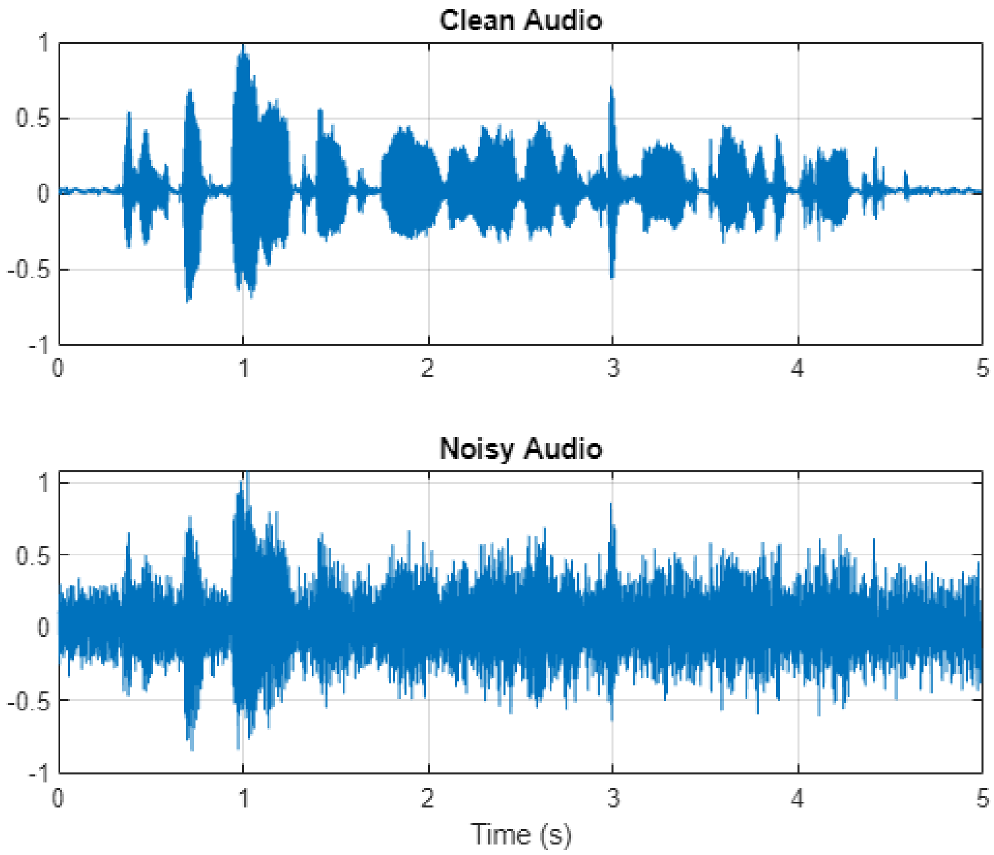
h	e	l	l	o
e	l	l	o	
h	e	l	o	

By marginalizing over alignments, we get a distribution over outputs.

https://distill.pub/2017/ctc/

Restoration/Denoising

- **Vision**
  - Directly regress the image
- **Audio**
  - Regress a gain mask which is applied to audio stream
  - Applied in a streaming fashion
    - Window and hop



https://www.mathworks.com/help/audio/ug/denoise-speech-using-deep-learning-networks.html

Conclusion

- **Feature engineering**
  - After some preprocessing things are more similar than not
- **Implementation**
  - Windowing with a set stride and hop allows us to deal with streams of data

- **Common problems in vision and their audio analogues**
  - Classification
  - Sequence decoding
  - Restoration/denoising

Resources

Getting Started with Audio

Audio Classification using Transfer Learning

https://www.tensorflow.org/tutorials/audio/transfer_learning_audio

Speech Command Recognition

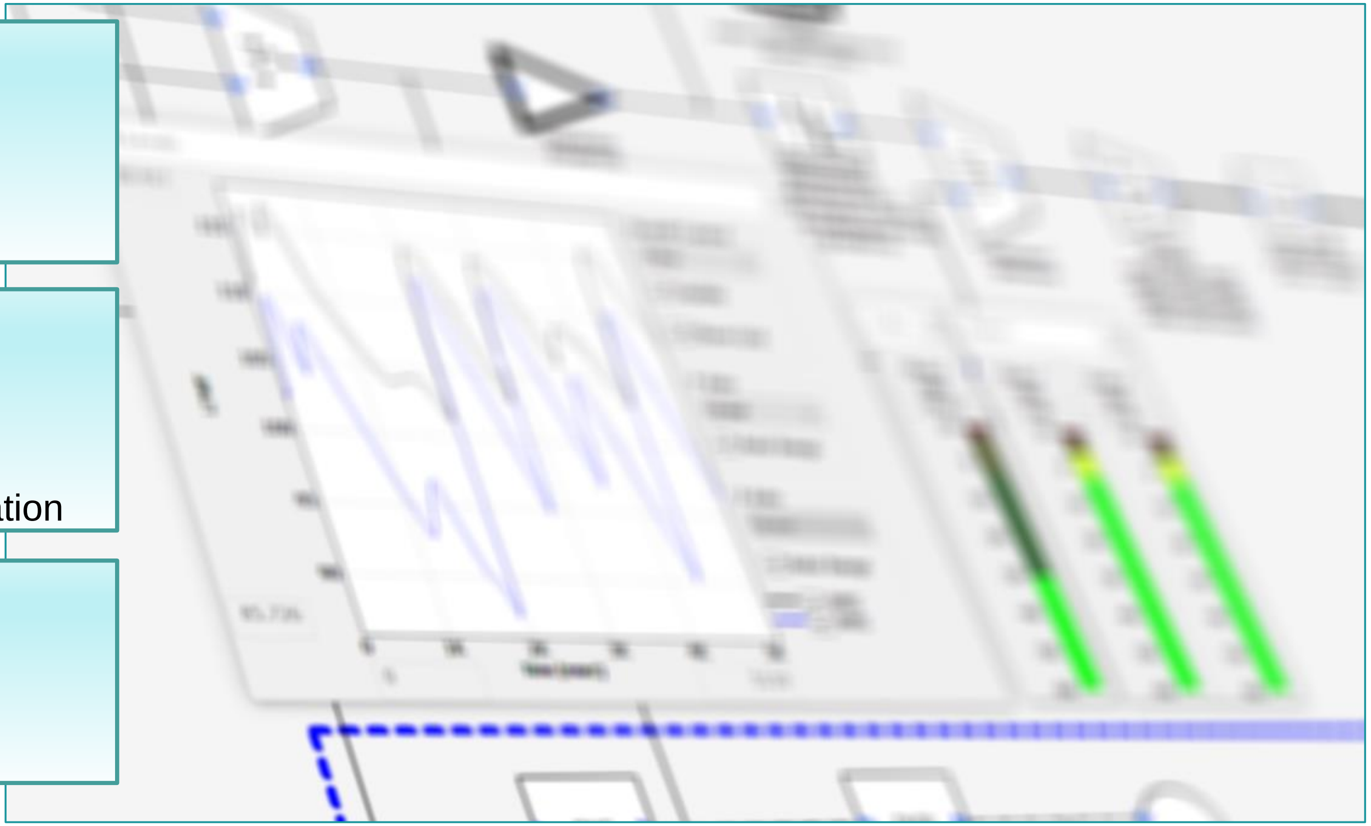
https://www.tensorflow.org/tutorials/audio/simple_audio

Get a 30 Day Trial of Audio Weaver

https://w.dspconcepts.com/audio-weaver

The Audio Weaver Framework: Overview

Audio Weaver **accelerates** audio feature development and **enables** collaboration across product teams. With over 550 optimized processing modules, audio designs can be developed and implemented on hardware **without writing any DSP code**.



AWE Designer

Windows-based graphical design environment

- ✓ Standard Edition: Design GUI
- ✓ Pro Edition: Works with MathWorks® MATLAB® platform

AWE Core

The embedded processing engine

- ✓ Optimized target-specific libraries
- ✓ Available for multiple processors
- ✓ Supports multicore and multi-instance implementation

Audio IP Modules

Building blocks for product developers

- ✓ From low-level primitives to complete designs
- ✓ From DSP Concepts and our third-party partners