Automatic Sound Recognition

CHANAKYA DEV
8 MAY 2019

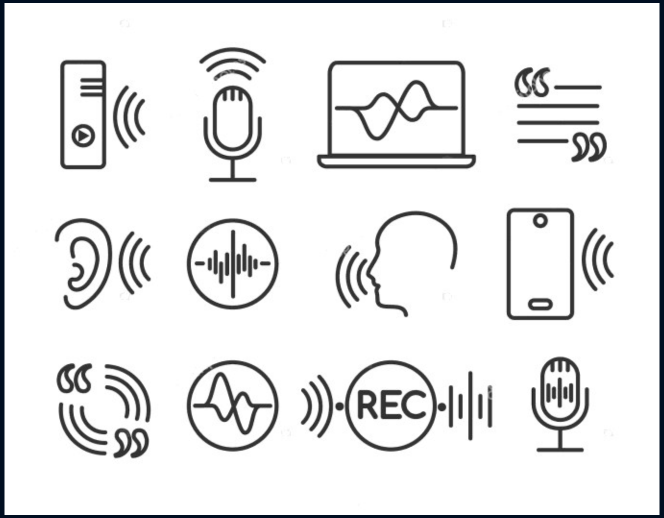
Channel Islands
CALIFORNIA STATE UNIVERSITY

The Challenge:

- Chimpanzee vocalizations in Issa Valley, Tanzania
- Raw sound files with chimp calls, birds, leaves, human generated sounds
- Implementation on Raspberry Pi 3
- Realtime classification
- Requires concepts of DSP, ML and embedded systems
- Localization, surveillance, Ecomonitoring

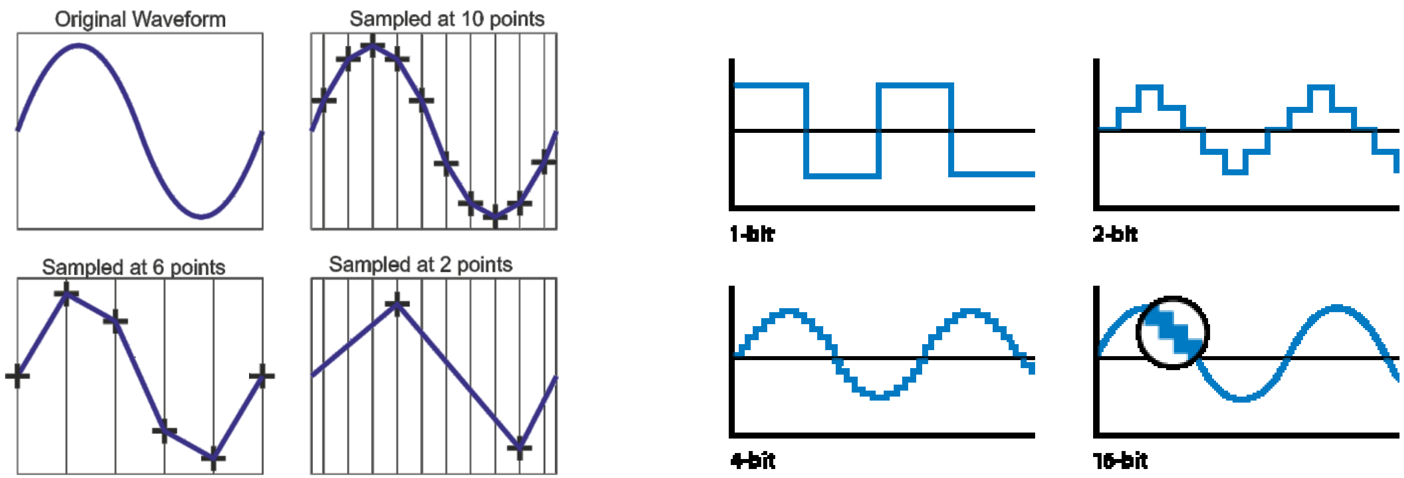
Key Steps

- Signal preprocessing
- Feature extraction
- Classification
- Realtime Inference



Signal Preprocessing

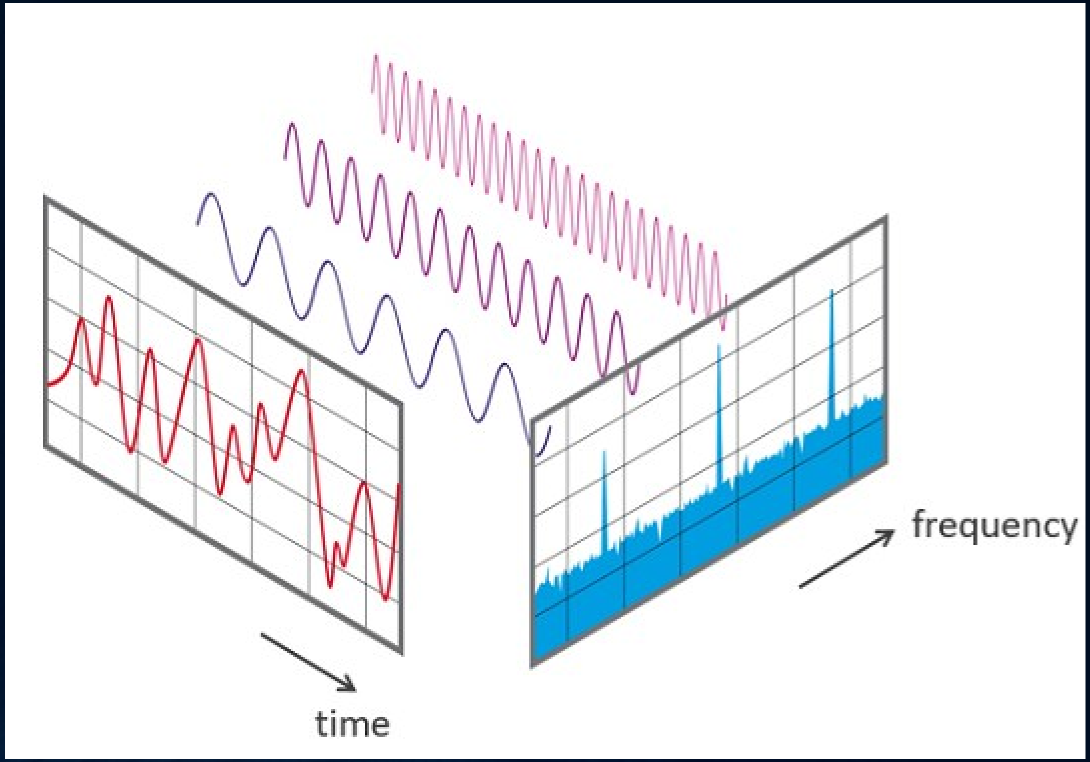
- Sound attributes like Sampling rate, n-channels, bit depth, length
- Depends on existing data and recording devices
- Normalization to remove bias during classification
- 11025 Hz, mono channel, 16 bit depth, 4 second chunks



Sampling Rate   Bit Depth

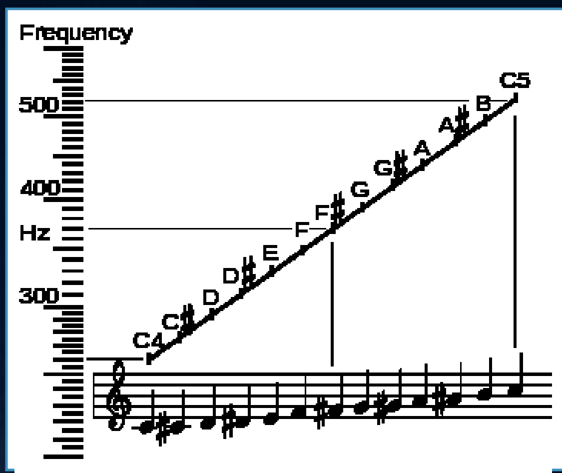
Feature Extraction

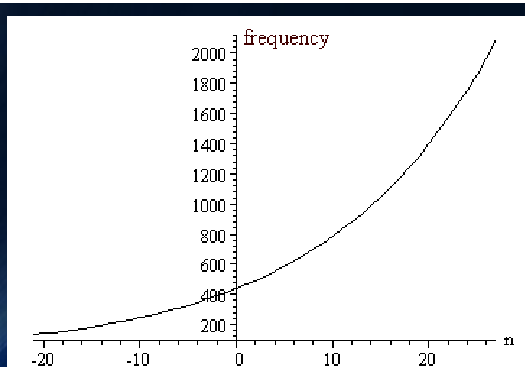
- Time domain vs Frequency Domain
- Fourier transform
- Psychoacoustic properties

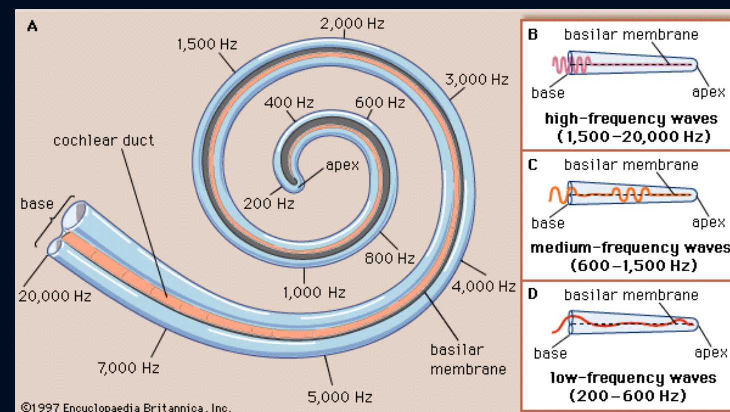


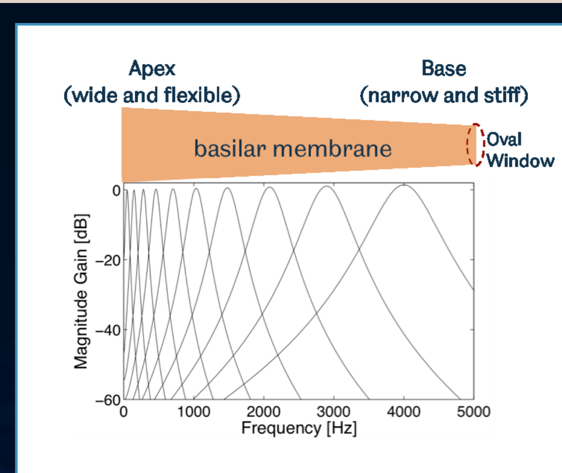
Feature Extraction part.2

- MFCC's : Frequencies equally spaced on the Mel scale
- GTCC's : Frequencies represented as how humans perceive sound



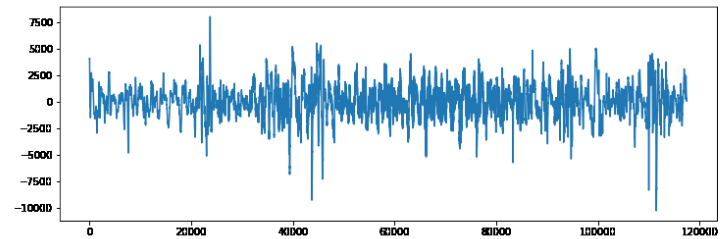


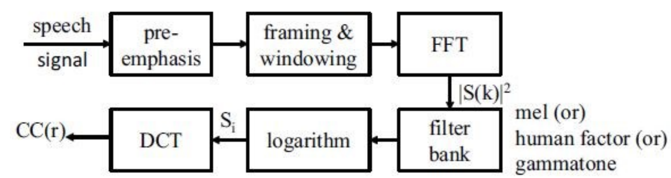


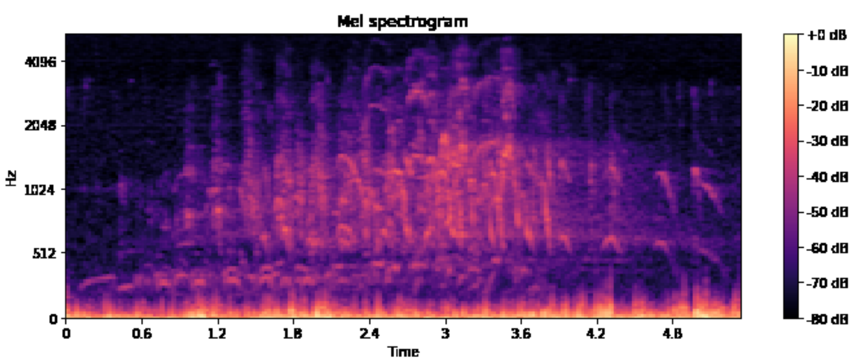


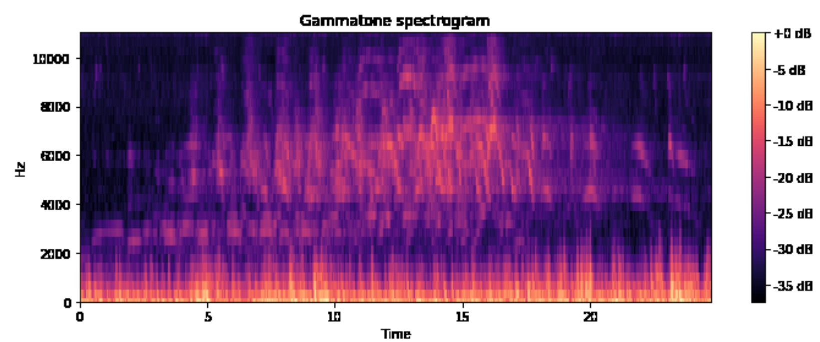
Building the classifier: Input vectors

- MFCCs and GTCCs extracted from the filterbanks
- 2D vector depending on number of CC features and length







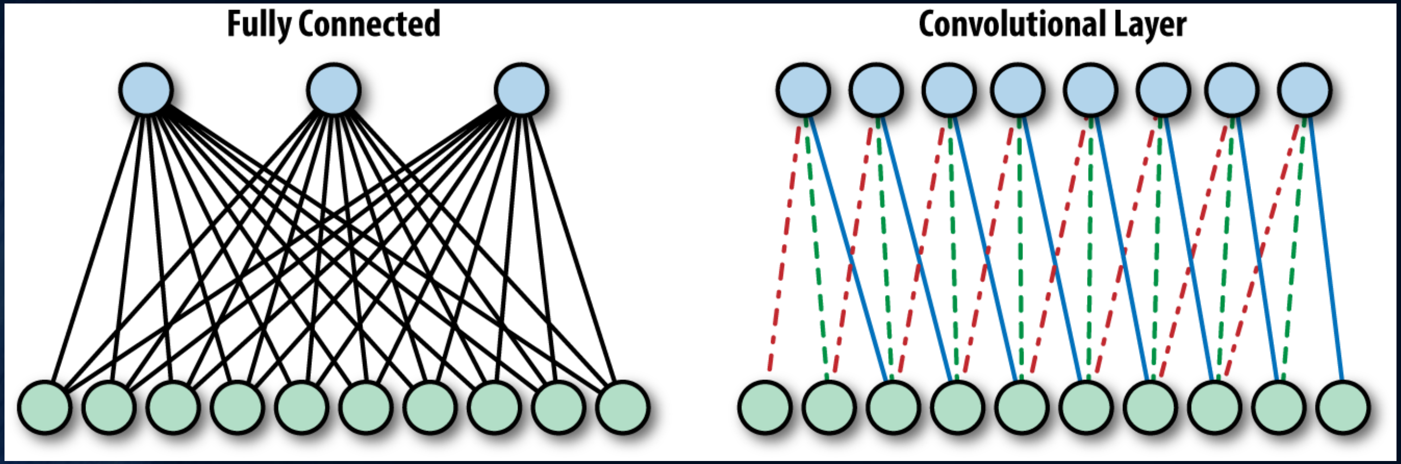


Building the classifier: Sound corpus

- Increases classifier robustness
- Urbansound8K dataset
- This dataset contains 8732 labeled sound excerpts (<=4s) of urban sounds from 10 classes
- 0 = air_conditioner
  1 = car_horn
  2 = children_playing
  3 = dog_bark
  4 = drilling
  5 = engine_idling
  6 = gun_shot
  7 = jackhammer
  8 = siren
  9 = street_music

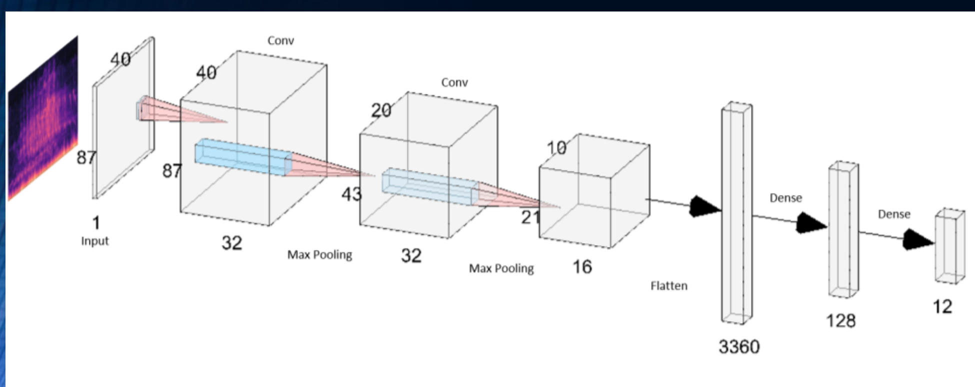
Building the classifier: Network Architecture

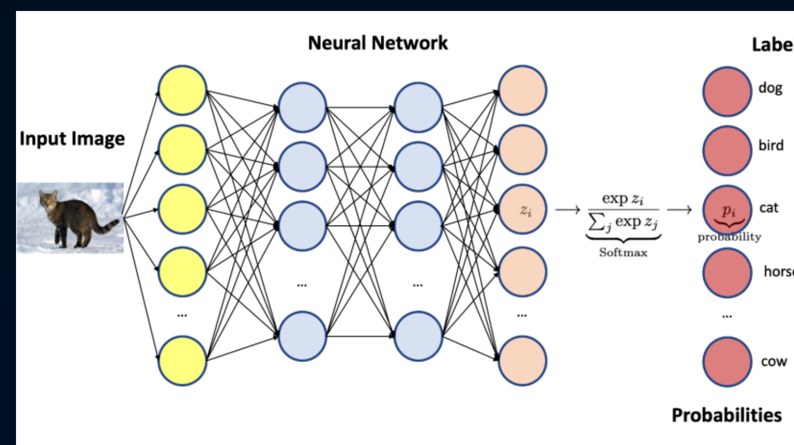
- Experimented with 3 classifiers
- SVM's: Preferred choice pre deep learning era
- Neural Network with fully connected layers: expensive and robust
- Neural Network with Convolutional Layers: lightweight and relevant



Building the classifier: CNN

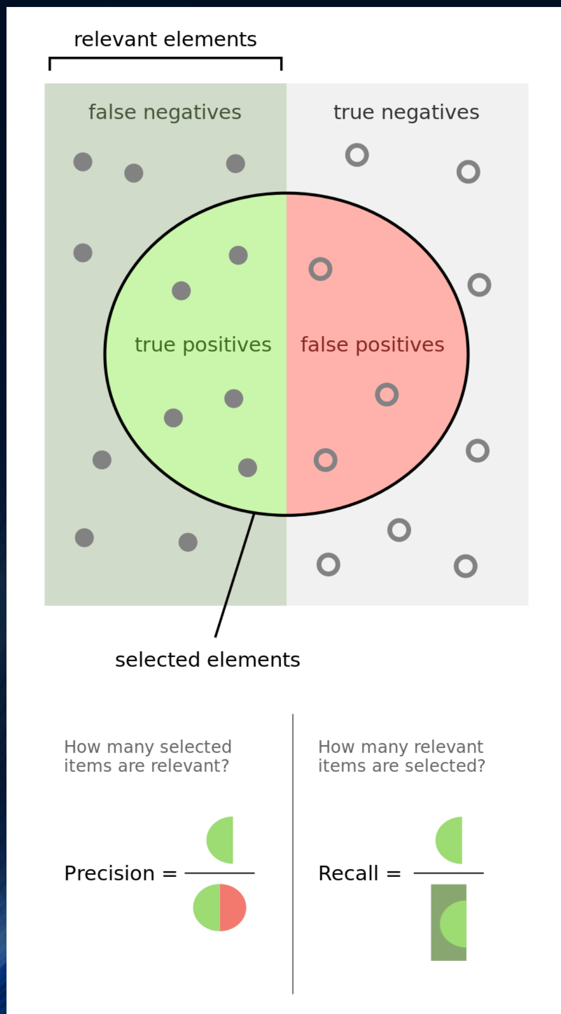
- Experiment with layer depth and number of hidden nodes
- Follows commonly used pattern of CONV-POOL-CONV-POOL layers
- Penultimate layer is FC with the last layer as a dense Softmax layer
- Grid search for hyperparameter tuning





Building the classifier: Results

- Accuracy Metrics : Precision-Recall



	precision	recall	f1-score	support
0	0.82	0.97	0.89	805
1	1.00	0.89	0.94	168
2	0.89	0.91	0.90	786
3	1.00	0.80	0.89	537
4	0.93	0.92	0.92	648
5	0.97	0.94	0.95	759
6	1.00	1.00	1.00	14
7	0.91	0.95	0.93	637
8	0.99	0.96	0.98	719
9	0.93	0.92	0.93	816
10	1.00	0.99	1.00	1154
11	1.00	1.00	1.00	1199
avg / total	0.95	0.94	0.94	8242

Realtime Inference: Raspberry Pi

- Raspberry Pi 3b+
- SoC: Broadcom BCM2837 (roughly 50% faster than the Pi 2)
- CPU: 1.2 GHZ quad-core ARM Cortex A53 (ARMv8 Instruction Set)
- GPU: Broadcom VideoCore IV @ 400 MHz.
- Memory: 1 GB LPDDR2-900 SDRAM.
- USB ports: 4.
- Network: 10/100 MBPS Ethernet, 802.11n Wireless LAN, Bluetooth 4.0.

Realtime Inference: Audio Chunks

- The ASR system takes 4 secs of audio as input
- Continuous audio stream is chunked
- For inference at t=0-1 seconds, we take the audio frame from t'=t-3 to the current frame.
- Inference is performed every second
- For inference on Block 1:

Realtime Inference: Classifier Characteristics

- Low memory footprint : model <30 mb
- Classification power : >90% accuracy
- Prediction speed : Inference time <1 second
- Evaluate tradeoffs

Future Improvements

- Embedded systems with specialized hardware (Nvidia Jetson Nano)
- Usage of LSTMs for context and long term dependencies
- Optimal spectral characteristics
- Traingulation and localization

Thank You