Sound Scene Classification: Comparing the Performance of Spectrum-based and Time-based Features

[1]Aliaa Mahgoub, [2]Iran Roman & [2]Juan P. Bello

[1]Brooklyn Technical High School

[2] New York University

Introduction

Recognizing sound scenes in realistic soundscapes would inform machines about their setting, creating context-aware machines.

To do this, we **train "listening machines"** using audio data, which they use to **find patterns** and learn.

Machines can process *these* **features extracted** from data:

- **Frequency-based features** help machines process audio like humans hear sounds at the level of the cochlea.
- **Time-based features** quantify characteristics of the signal's time-domain plot.

Problem Statement

Since we don't know what features and classification systems are best at classifying sound scenes, we can use those features to **train classification systems** and **compare their performance**.

Research Questions

1- What **combinations** of time-based and spectrum based features will result in the best **accuracy**?

2- How will the **K-Nearest Neighbors** classifier perform compared to a **Neural Network** classifier?

Hypothesis

Classifiers will perform best with frequency-based features because they help machines process audio like humans hear sounds.

A Neural Network will perform better than the K-Nearest Neighbors classifier because it learns from its errors to understand complicated relationships.

Methods: Acoustic Features

Root Mean Square (RMS)

$$rms = \sqrt{\frac{1}{N}\sum_{n}|x(n)|^2}$$

Zero-crossing Rate (ZCR)

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1}1_{\mathbb{R}_{<0}}(s_t s_{t-1})$$

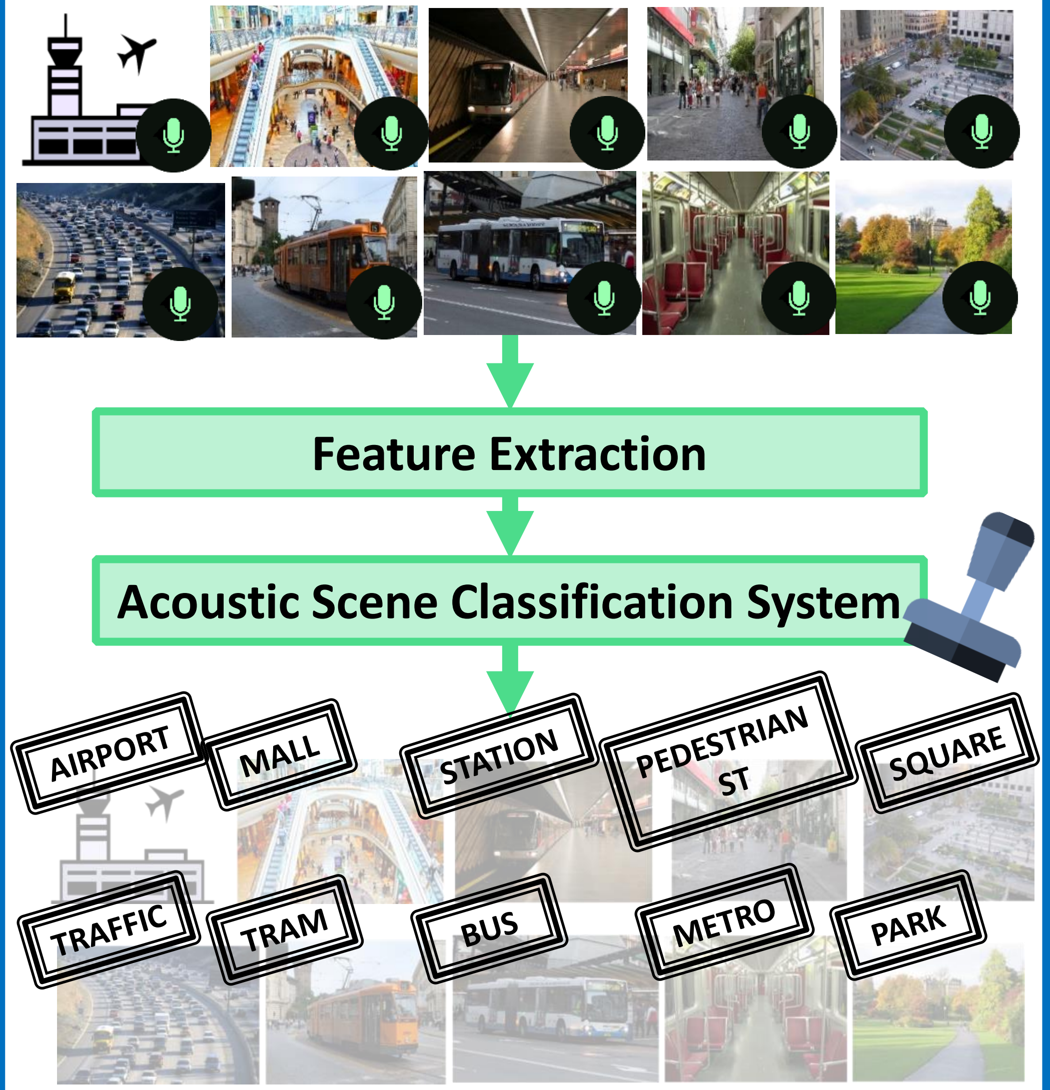
Spectral Centroid (SC)

$$c = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

Spectral Bandwidth (SB)

$$b = \sum_{n} S[n](f[n]-c[t])^{p^{\frac{1}{p}}}$$

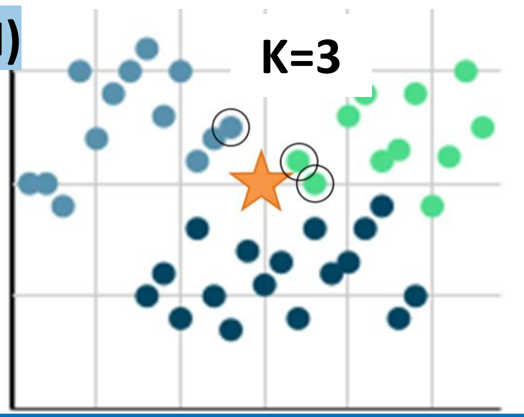
Methods

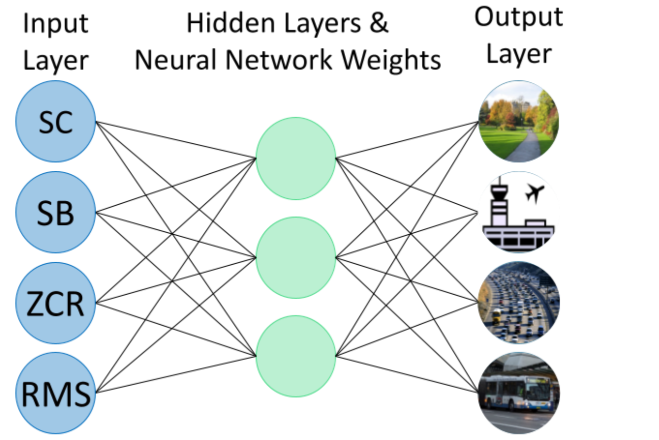


Methods: Classifiers & Cross-validation

K-Nearest Neighbors (KNN)

1. Find the k nearest neighbors.
2. Vote for classes.
3. Object is assigned to class with **majority vote**





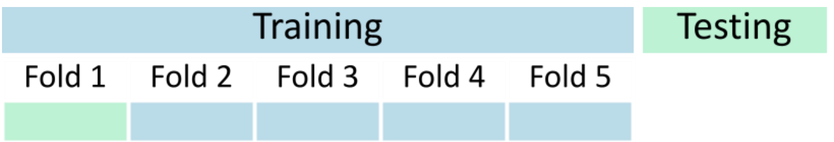
Neural Network (NN)

Error Function

$$J = \sum -\frac{1}{N}log(\sum y \cdot \hat{y})$$

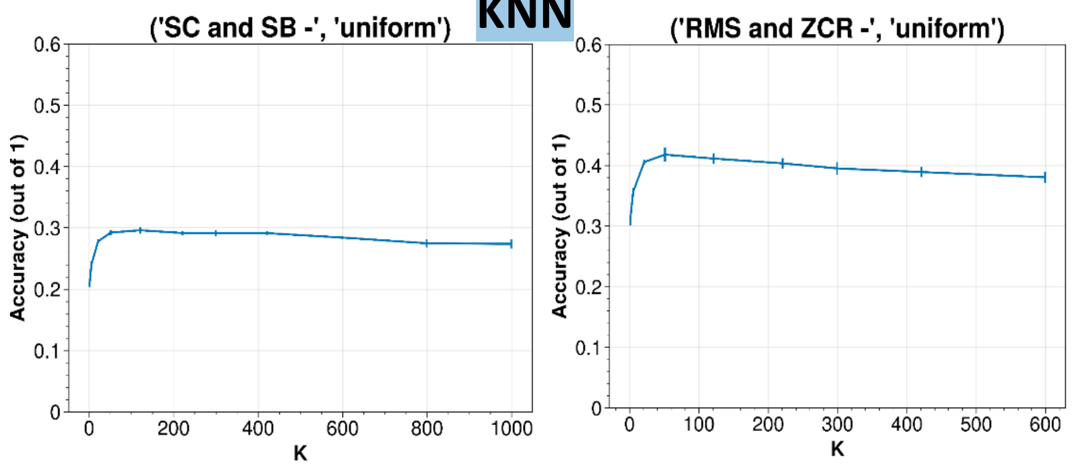
$$\hat{y} = softmax(xW)$$

Cross-validation

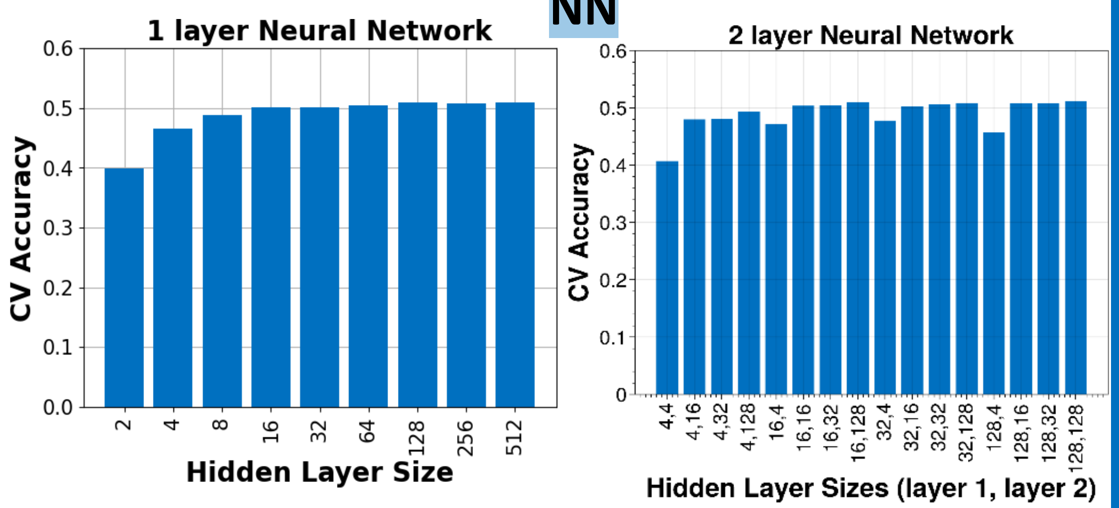


Results: Cross-validated Training

KNN

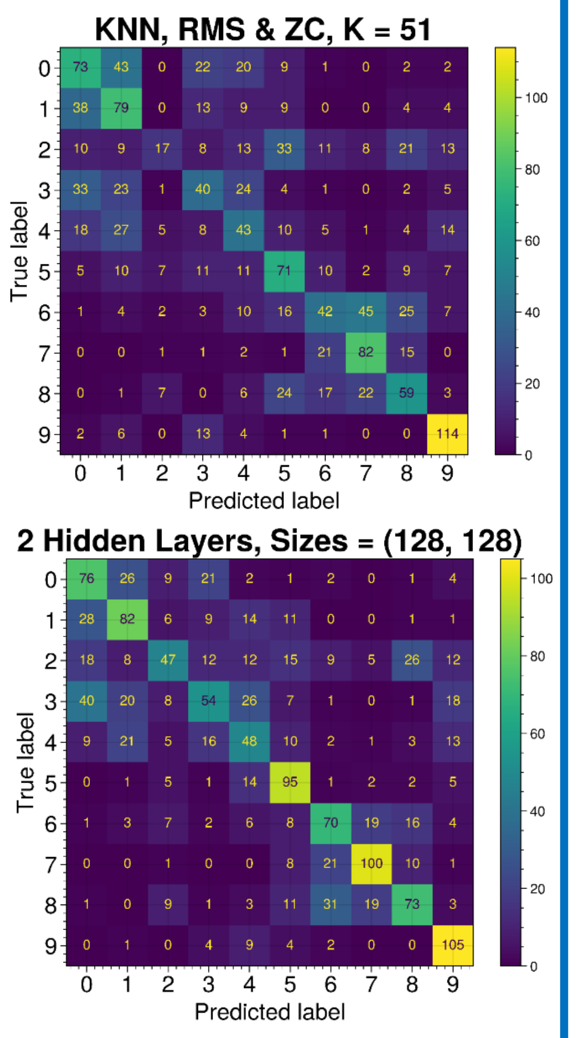


NN



Results: Performance

Model	Validation Accuracy	Test Accuracy
NN (2 layers, Sizes = (128,128))	51.4%	**51.6%**
NN (2 layers, Sizes = (16,128))	**51.8%**	51.5%
NN (1 layer, Sizes = 512)	**51.0%**	50.9%
NN (1 layer, Sizes = 2)	40.3%	39.9%
KNN (K=51, RMS & ZCR, Uniform)	**41.5%**	**41.3%**
KNN (K=221, RMS & ZCR, Distance)	40.9%	**41.3%**
KNN (K=221, SC & SB, Distance)	29.7%	29.8%
KNN (K= 221, All features, Uniform)	29.7%	29.5%
KNN (K=121, SC, SB, & ZC, Uniform)	29.7%	29.3%



Conclusion

- Overall, the performance of the Neural network increased as the sum of the sizes of the layers increased but **plateaued** quickly at around **50%**.
- The 2-hidden-layer Neural Networks with hidden layer sizes **(16,128) and (128,128)** performed the best overall at over **40% more than chance**.
- The K-Nearest Neighbors classifier performed best with a combination of the two **time-based features**, with an accuracy of **over 40%**

References

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 9–13. November 2018. URL: https://arxiv.org/abs/1807.09840.

Cartwright M, Mendez AE, Cramer J, Lostanlen V, Dove G, Wu HH, Salamon J, Nov O, Bello J. SONYC Urban Sound Tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network.

Heittola T, Mesaros A, Virtanen T. TAU Urban Acoustic Scenes 2019, Development dataset.

Takahashi G, Yamada T, Makino S, Ono N. Acoustic scene classification using deep neural network and frame-concatenated acoustic feature. Detection and Classification of Acoustic Scenes and Events. 2016 Sep 3.

Project Repository

https://github.com/AliaaMahgoub/sound-scenes

Acknowledgements

Center for K12 STEM Education NYU Tandon School of Engineering

The Pinkerton Foundation