Using Artificial Neural Networks (ANNs) with Acoustic Reverberation to Classify Room Size

Tyrell Martens, Angela De Sousa Costa, Brayden Carlson - Departments of Neuroscience and Computer Science
Dr. Matthew Tata - Department of Neuroscience
TataLab, Department of Neuroscience, University of Lethbridge

TataLab
Neuroscience, Robotics, and AI

Abstract

Acoustic reverberation occurs when sound waves reflect off surfaces in an environment. Useful information about the size and configuration of an acoustic space, and the locations of sounds in that space, is often encoded in these echoes. The human auditory brain uses these cues to understand the auditory scene, suggesting that artificial neural networks might also be able to extract this information. Here we show that a simple convolutional neural network trained on simulated realistic acoustic environments can learn to classify the size of a space using only reverberation as a cue. The use of an ANN to classify room size based on sound alone could aid in assistive technologies for the visually impaired, and in localization and mapping algorithms for autonomous navigation (e.g. robotics, self-driving cars).

Introduction

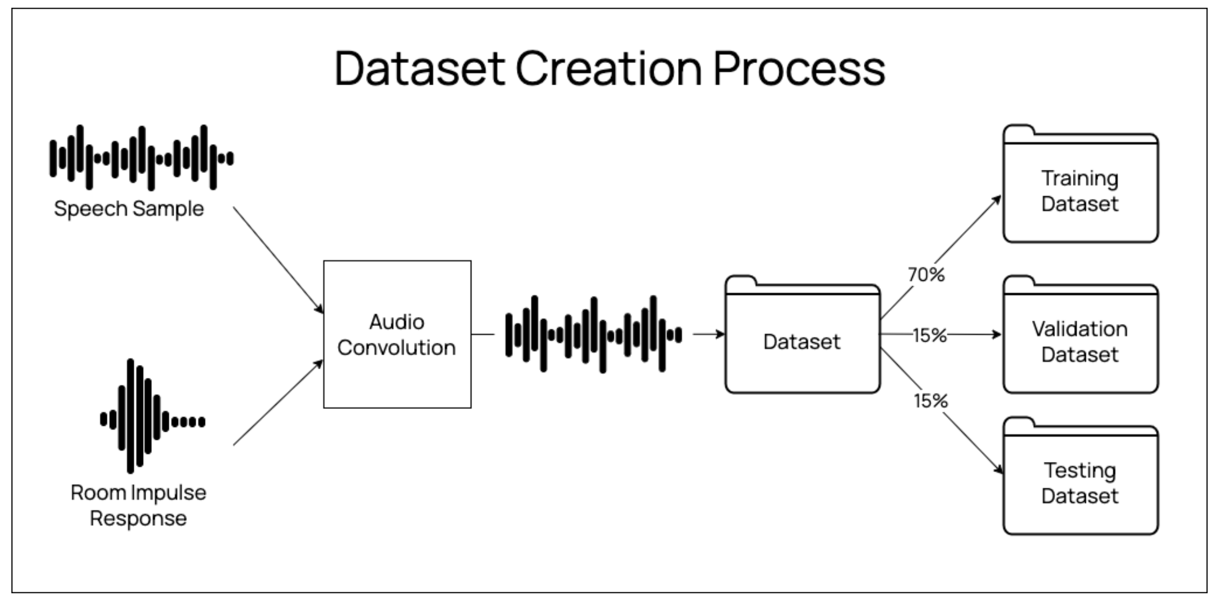
The complexity of reverberation makes it difficult for computers to understand the information cues contained in reverberant sounds. However, with the use of a ANN this issue may be addressed as they have strong feature extraction.

Artificial neural networks

- are a specialized, advanced topic in artificial intelligence that mimic the learning present in organisms.
- demonstrate high accuracy when given classification tasks.

Convolutional neural networks (CNNs)

- are a type of artificial neural network
- have shown exceptional performance when working with grid-structured inputs, such as 2D images.
- have strong spatial dependencies within local regions of the grid making them a great choice for image classification tasks [1].



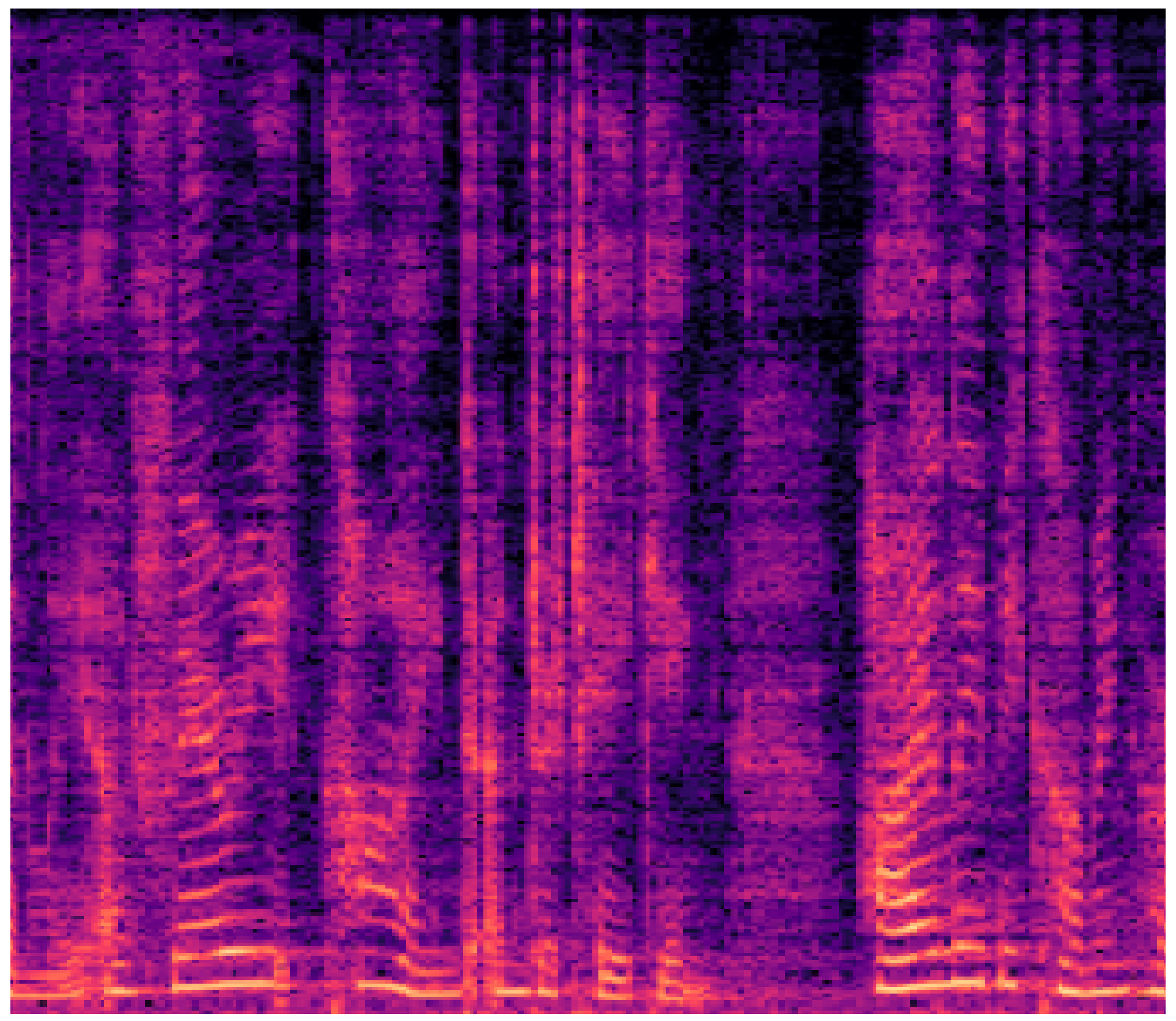
Simulated realistic acoustic environments generated using audio convolution with speech samples and room impulse responses.

Approach

With the use of Python, PyTorch and PyAudio we

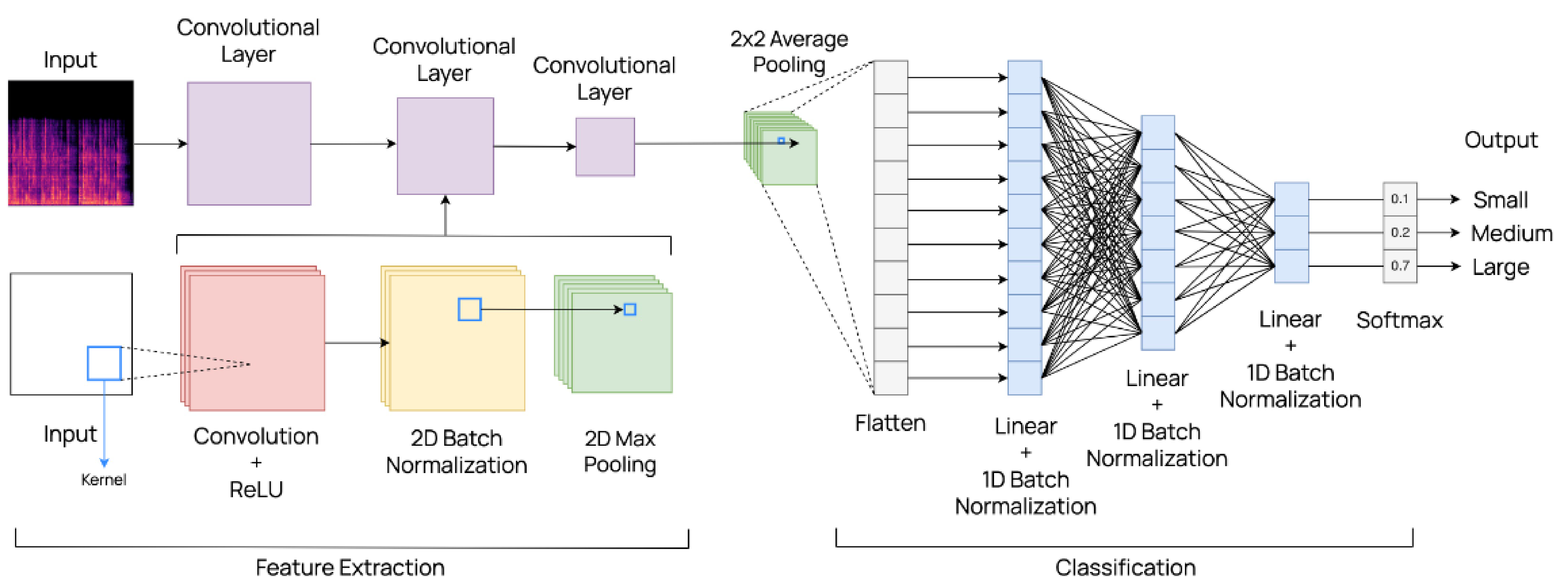
1. Created a dataset for training the neural network.
   - Used acoustic convolution to create speech audio samples at small, medium, or large room reverberant intensities [2].
   - Our dataset consisted of 20 203 room impulse responses from 200 rooms for each reverberate-intensity [3], and 28 539 speech audio samples totaling 100 hours [4].
   - The final dataset consisted of 28 539 audio files for each reverberate-intensity, for a total of 85 617 convolved audio files.
2. Transformed the audio samples into relevant data for a convolution neural network.
   - Normalize
   - Resample
   - Mixdown
   - Cut
   - Pad
   - Mel Spectrogram
3. Created a CNN that uses the Mel Spectrograms as inputs.
4. Trained the network using the dataset resulting in a network that can infer room-size from any audio sample.

Mel Spectrogram



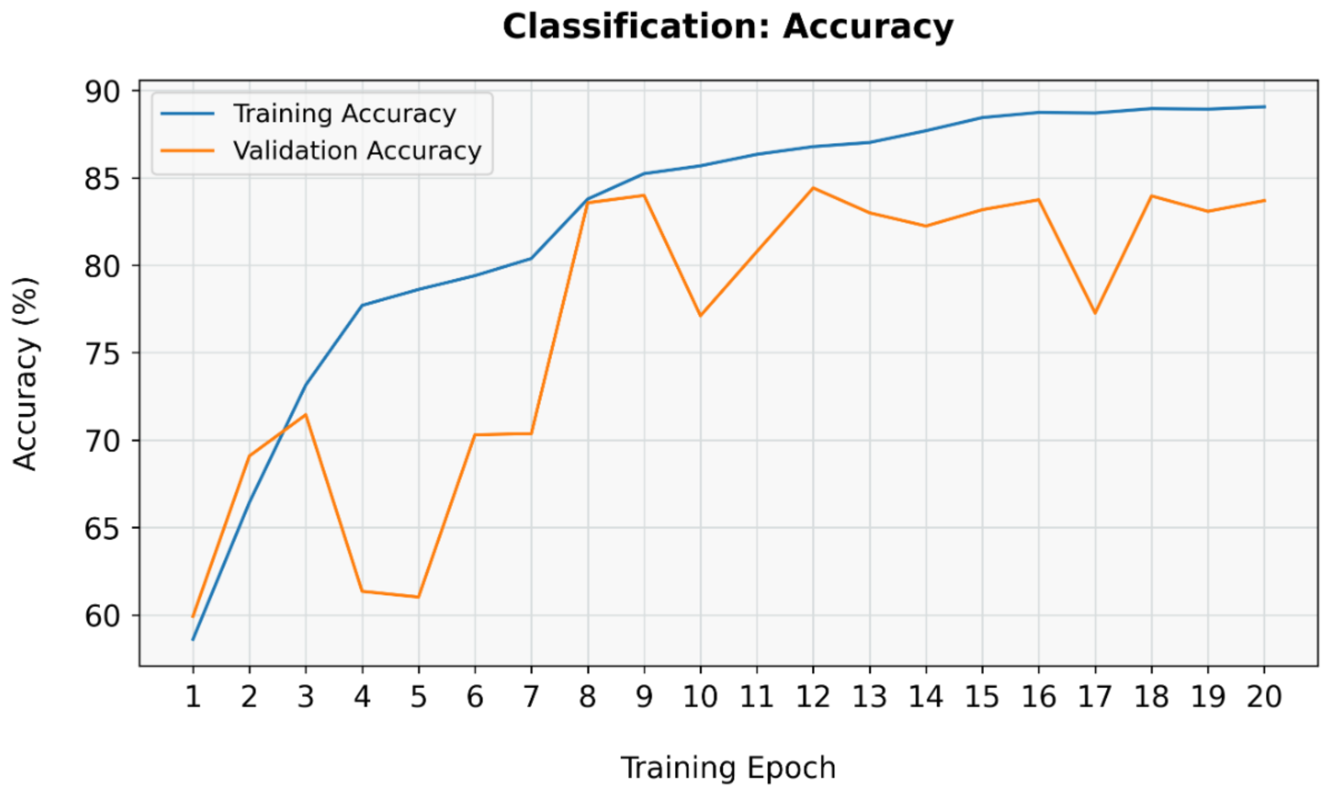
Mel Spectrogram of a speech audio signal that was convolved with a large-room room impulse response. The spectrogram decomposes the audio sample into its frequencies, producing an image. The y-axis represents frequency, the x-axis represents time, and the colours represent audio intensity.

Reverb Classification Neural Network Architecture



The architecture of the convolution neural network created to solve audio reverberation classification. Feature extraction aspects analyze the images and discover patterns present. The fully-connected classification layers interpret and classify the data from the feature extraction into the three different labels.

Results



Accuracy of the audio reverb classification neural network during training. A training epoch is one full cycle of training over the entire dataset. Training accuracy accounts for 70% of the dataset, validation accuracy accounts for 15% of the dataset.

The final accuracy of our neural network was

- 89.0% for training.
- 83.7% for validation.

Accuracy is measured throughout training by making inferences and evaluating if they are correct. Chance performance is 33% accuracy.

Discussion / Conclusion

Neural networks excel at solving classification problems, which was demonstrated in our network that was able to classify audio reverb into three different room sizes at an accuracy of 89%. This provides the foundation for advancing research in auditory machine learning. Future aspirations are to use an ANN with reverberation to aid in speech recognition.

Acknowledgements / References

Thank you to Autonomous Systems Initiative for providing the funds to supply this research.

[1] Aggarwal, Charu C., Neural Networks and Deep Learning: A Textbook. Springer International Publishing, 2023.

[2] Hass. J, Convolution: a form of cross-synthesis, Indiana University, 2021, https://cmtext.indiana.edu/synthesis/chapter4_convolution.php

[3] Panayotov. V, Chen. G, Povey. D, Khudanpur. S, LibriSpeech ASR Corpus Dataset, ICASSP, 2015, Retrieved from https://www.openslr.org/12

[4] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, T. Yoshioka; "REVERB challenge" EURASIP Journal on Advances in Signal Processing, 2016, Retrieved from https://github.com/RoyJames/room-impulse-responses

[5] Martens. T, De Sousa Costa. A, Carlson. B, Reverb Speech, 2023 https://github.com/Hotrod1220/reverb_speech