The Andrew and Erna Viterbi Faculty of
ELECTRICAL & COMPUTER ENGINEERING

SIPL
Signal and Image Processing Lab

TECHNION
Israel Institute of Technology

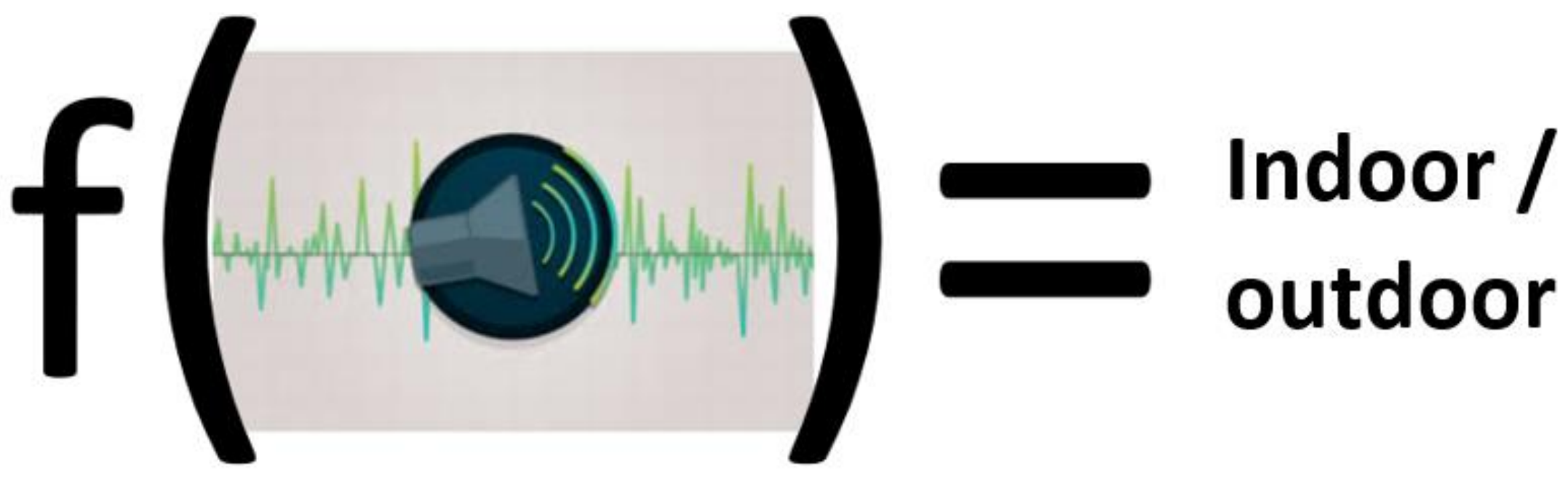
Indoor/Outdoor Classification of Voice for Mobile Devices

Gabriel Mannes, Odelia Longini and Ori Bryt

In Collaboration with RAFAEL ADVANCED DEFENSE SYSTEMS LTD.

Introduction

- The acoustic detection and classification area of research is now developing at a rapid pace, and special sessions on the topic are commonly encountered at international signal processing conferences
- Intelligence gathering often includes voice recordings, and the ability to detect and classify them can be important for security needs



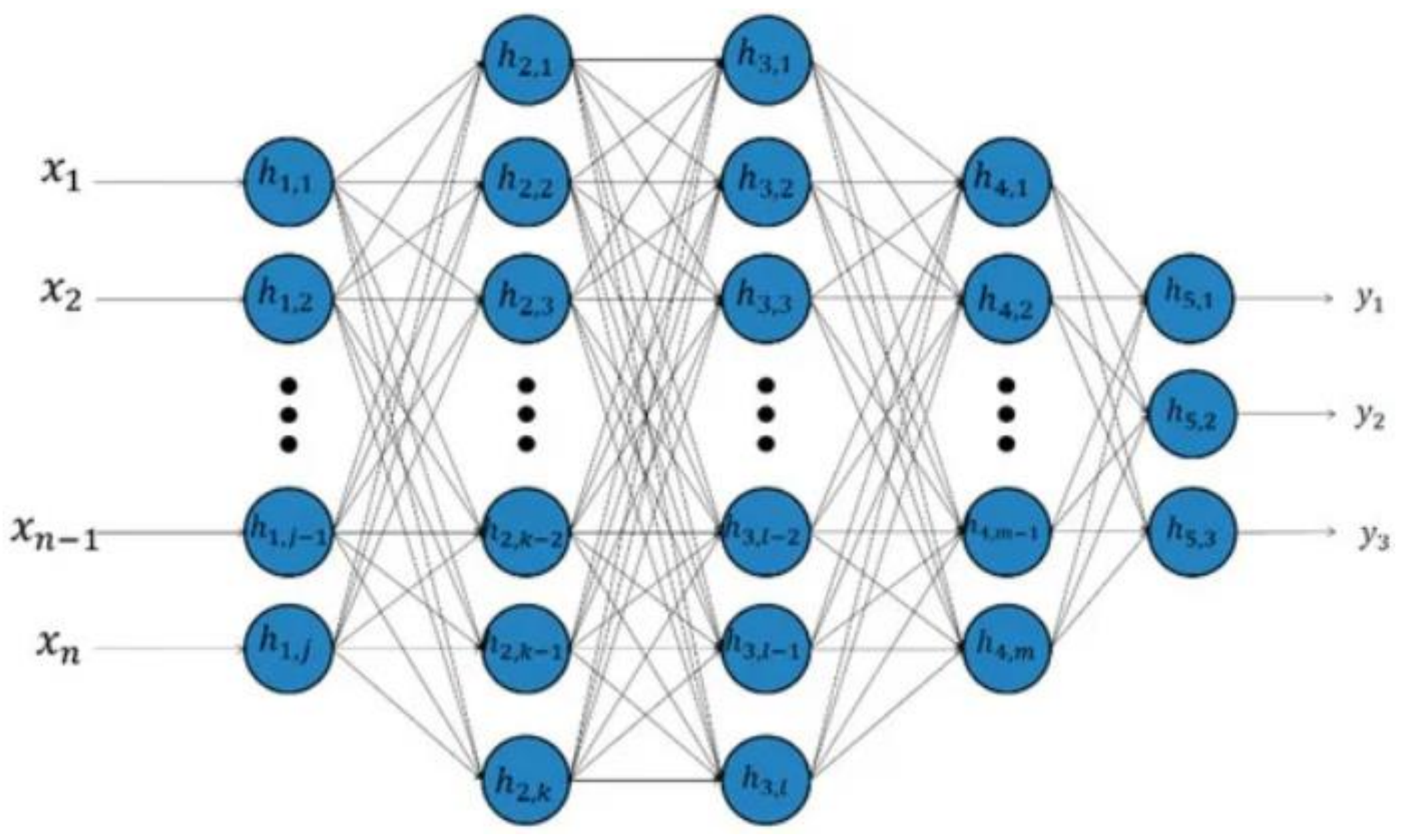
Goals

- Classify two-way radio recordings to indoor/outdoor classes.
  - The project goal will be achieved with Deep learning techniques
  - Simulate Rafael's Database

Challenges

- Lack of data for deep learning network training
- How could we use and maybe adjust a different wide database?

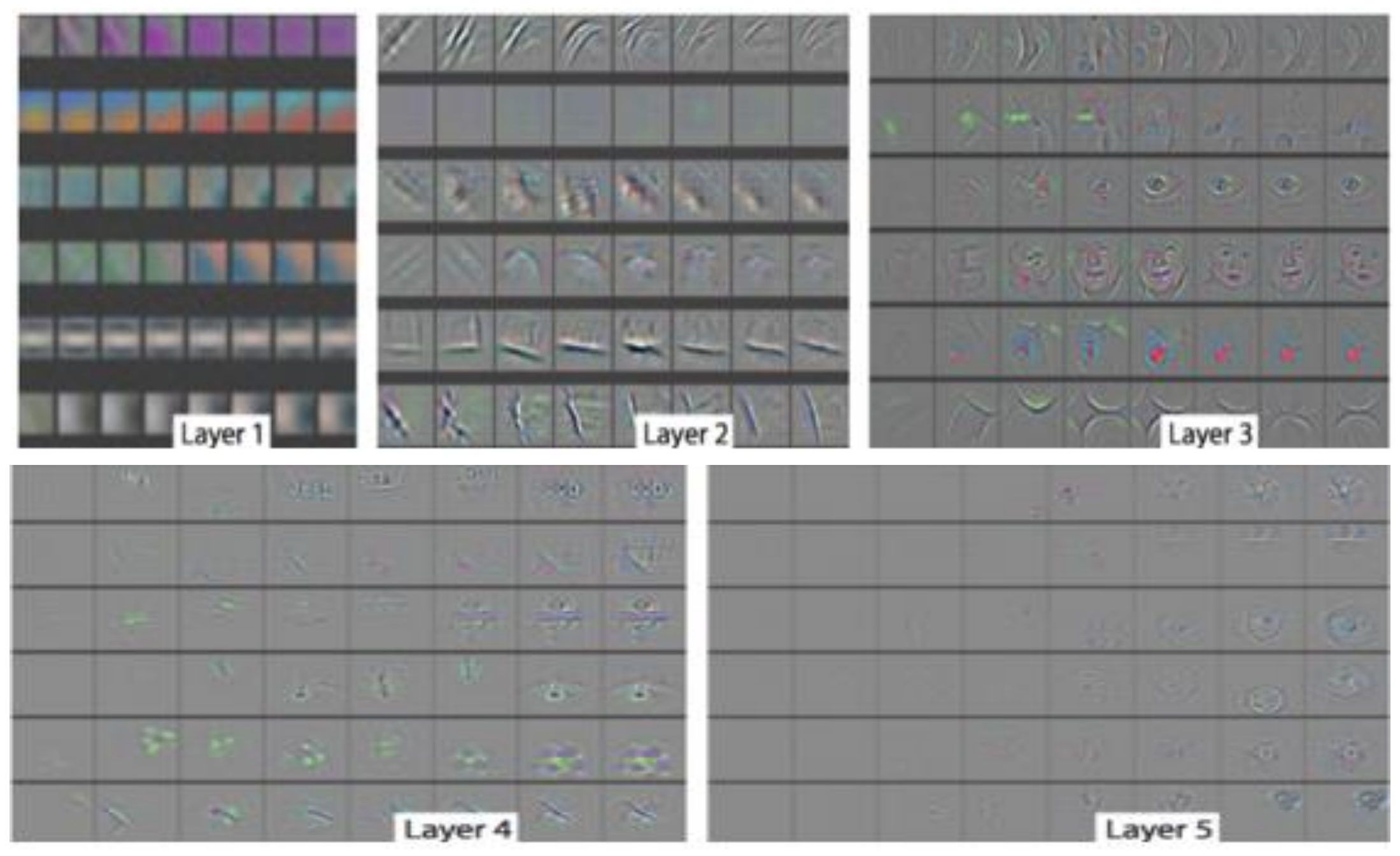
Deep Learning



- A subdomain of Machine learning
  - Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed
  - Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions
- Characterized by having many hidden layers
- An Artificial neural network is a model based on a collection of connected units or nodes called "artificial neurons", which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit information, a "signal", from one artificial neuron to another

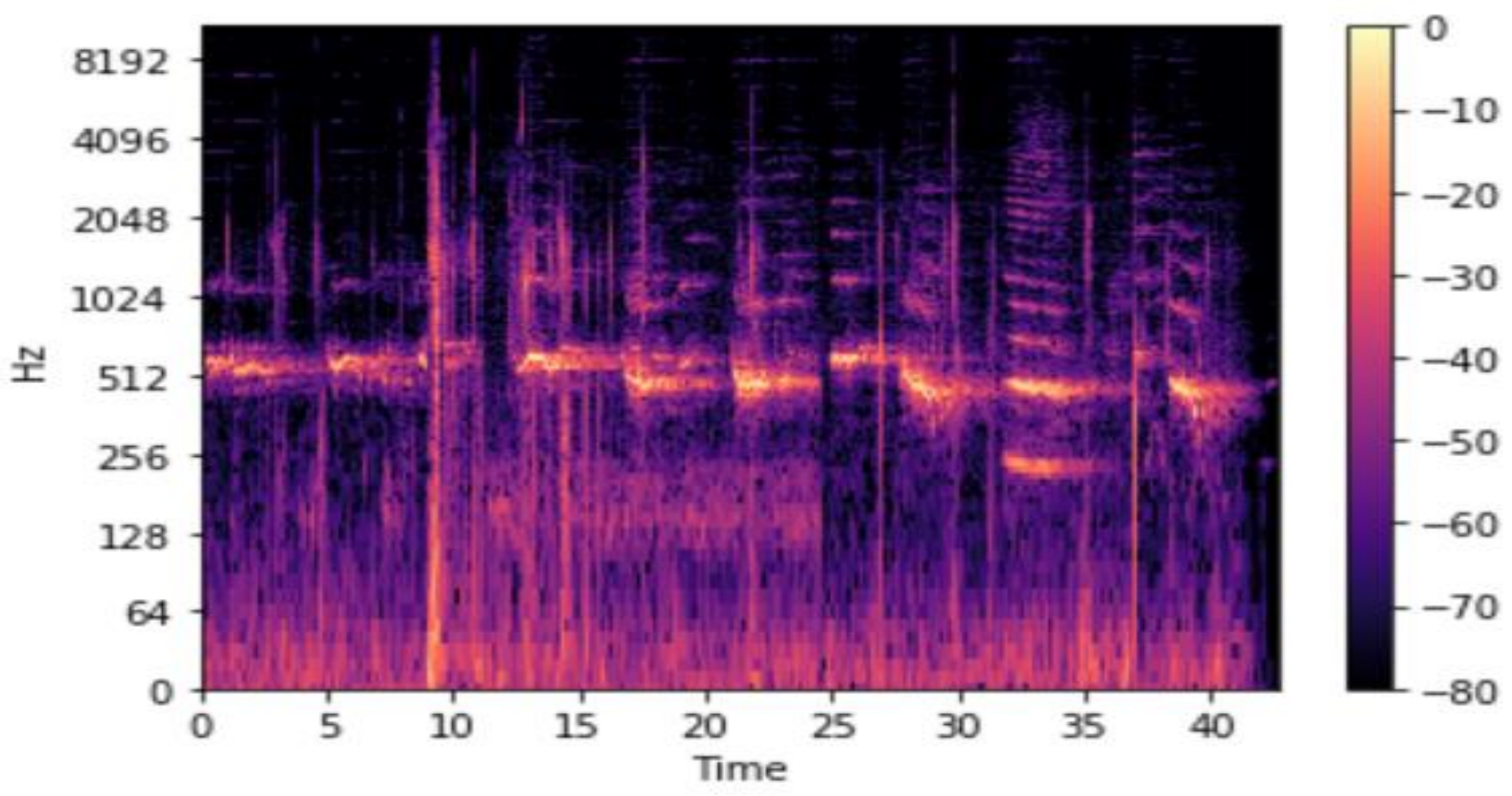
Transfer Learning

- Transfer learning and semi-supervised learning are a way to enable models to work better with limited amounts of data
- Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a different yet similar task
  - Different features are extracted in each of the layers



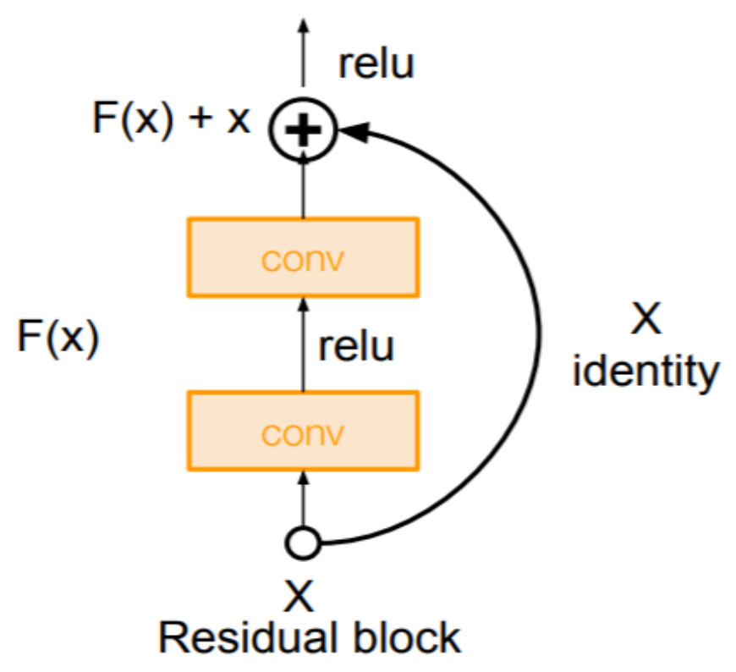
The Network

- Performing machine learning involves creating a model, which is trained on some training data and then can process additional test data to make predictions
- The model is based on McDonnel's work on DCASE challenge 2019
  - DCASE (detection and classification of acoustic scene and events) is a technology challenge that takes place every year and strengthens the understanding and importance of developing methods for detecting and classifying acoustic signals
- The models input is a log-mel spectrogram



Log-mel spectrogram

- The model architecture is based on ResNet - residual neural network
  - ResNet utilizes skip connections, or shortcuts to jump over some layers
    - skip connections prevent the problem of vanishing gradients



Resnet layer

Database

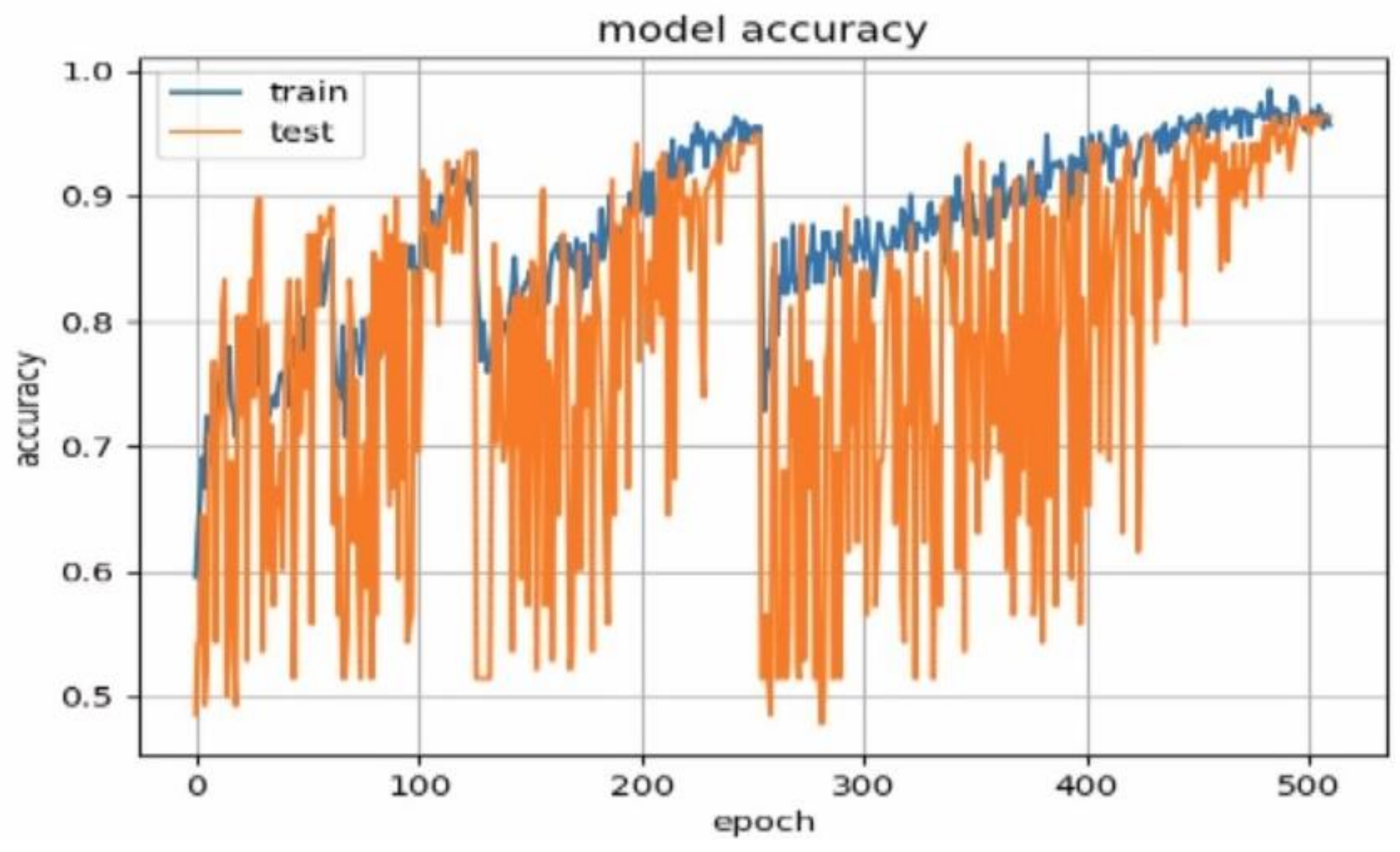
- Rafael's database
  - 668 samples - 388 outdoor, 280 indoor
  - 10 different rooms, 7 different outdoor areas
  - Various audio clip lengths
- DCASE's database
  - 10 acoustic scenes (park, airport, street traffic, ...)
  - 14,400 samples of 10 seconds
- Manipulating data – preprocessing:
  - Modeling the recording device by a frequency domain filter
  - Down-sampling the frequency
  - Adding main speaker in background scene
  - Audio clip length adjustment
  - Stereo to mono
  - Scene selection

Work Done

- Network architecture for audio classification
- Preprocessing - Manipulate a large database to fit Rafael's database
- Transfer learning from manipulated data
- Training network on 80% data as training and testing with 20% data, without intersection

Results

- Understanding the inability to inference two different acoustic scenes
- Classification ability of 96.3% on Rafael's database, without using DCASE



Final network results

Conclusions

- Artificially reducing databases gap – didn't work
- Importance of segment's length
- Further testing – Larger Database, RNN network

June 2021