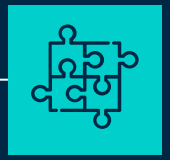
DEEP LEARNING FOR ACOUSTIC SIGNAL PROCESSING:

Speaker Recognition through One-Shot Learning and Siamese Neural Networks

- Jaime Pérez -

TABLE OF CONTENTS



1 Deep Learning in Acoustics

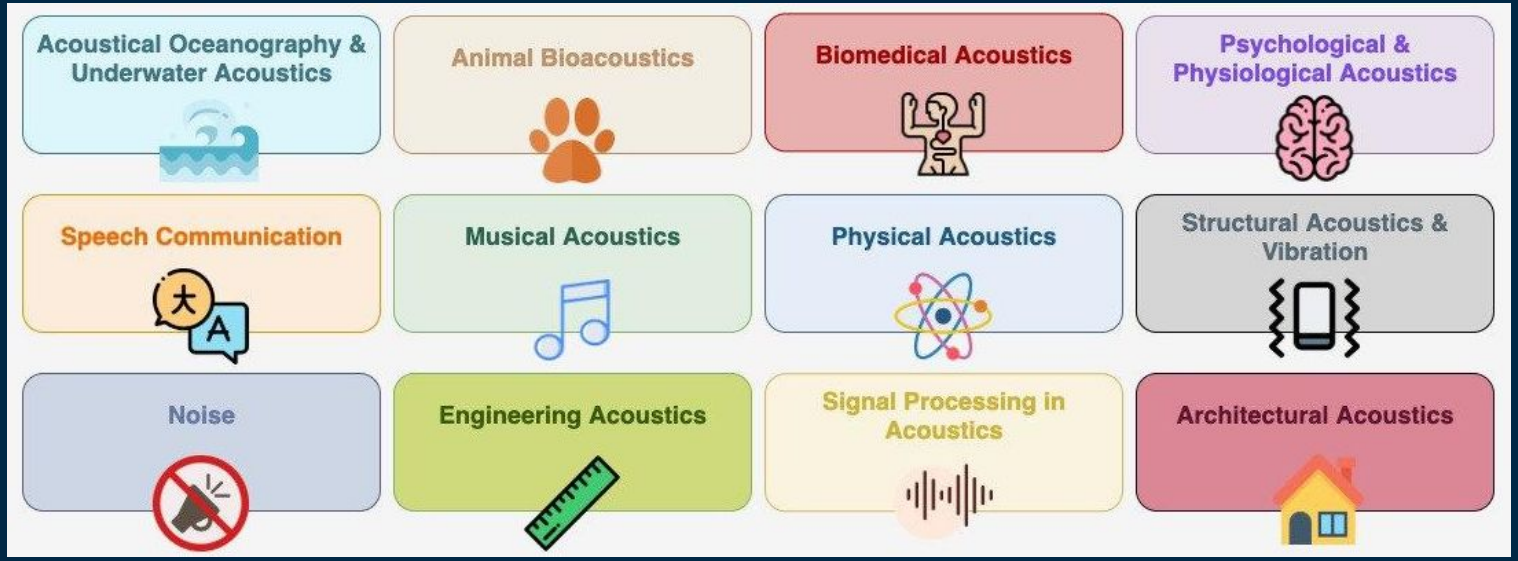
2 Case Study

3 Experiments

4 Conclusions

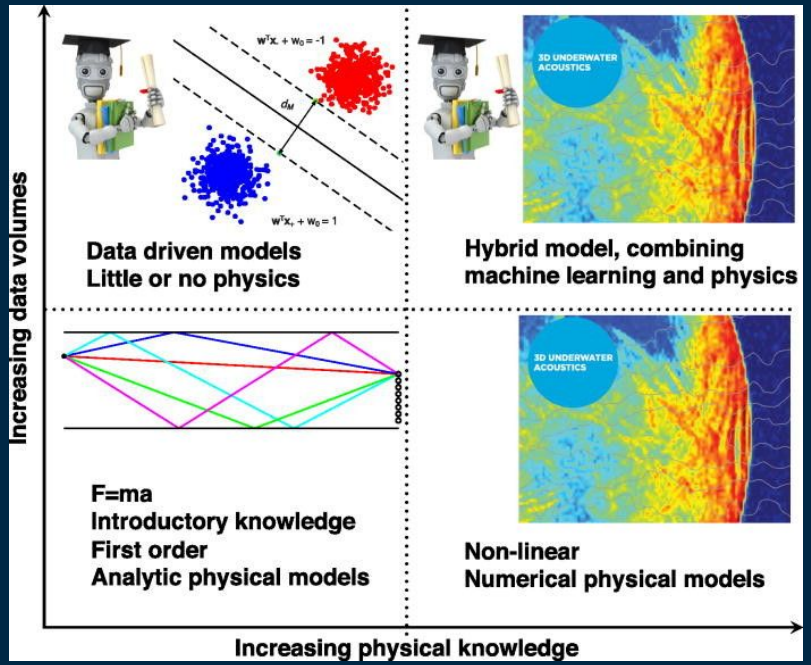
1 Deep Learning in Acoustics

Map of Acoustics



Source: Agustín de los Riscos Mayorga

Deep Learning Applications for Acoustics



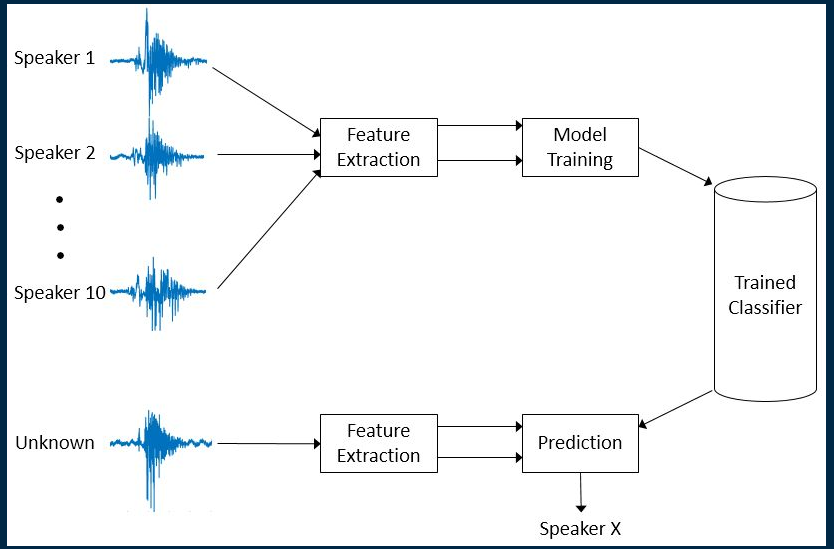
2

Case Study:

Speaker Recognition through One-Shot Learning and Siamese Neural Networks

Motivation

- What is speaker recognition?
  Identification of a person and distinguishes from others, based on its voice characteristics
- Speaker Verification (1:1)
  vs.
  Speaker Identification (1:N)
- Few-shot learning & deep learning?
  Siamese Neural Networks

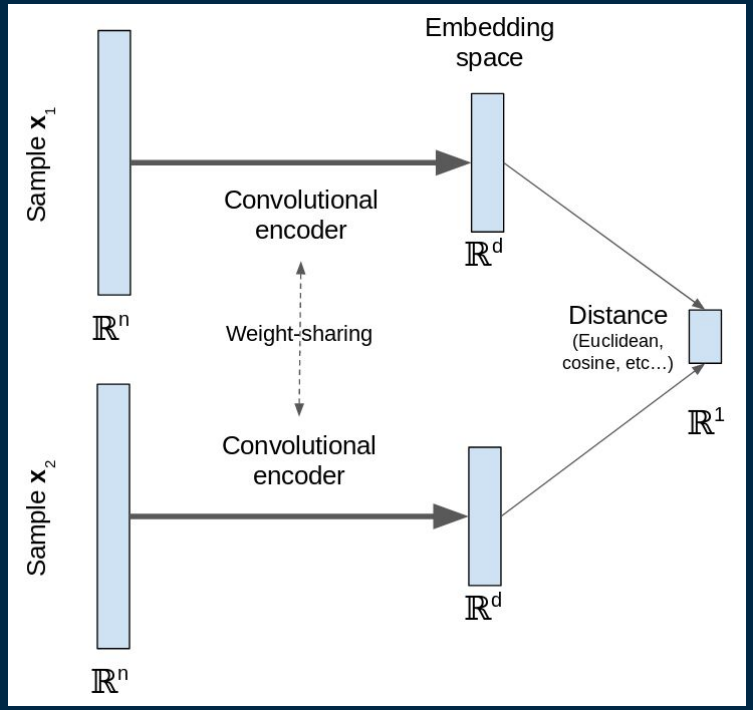


Use Cases

- Simplify translating speech tasks
- Improved and personalized services (Alexa, Google Home, Customer service bots, etc.)
- Complement biometric verification methods in security systems
- Criminal investigations

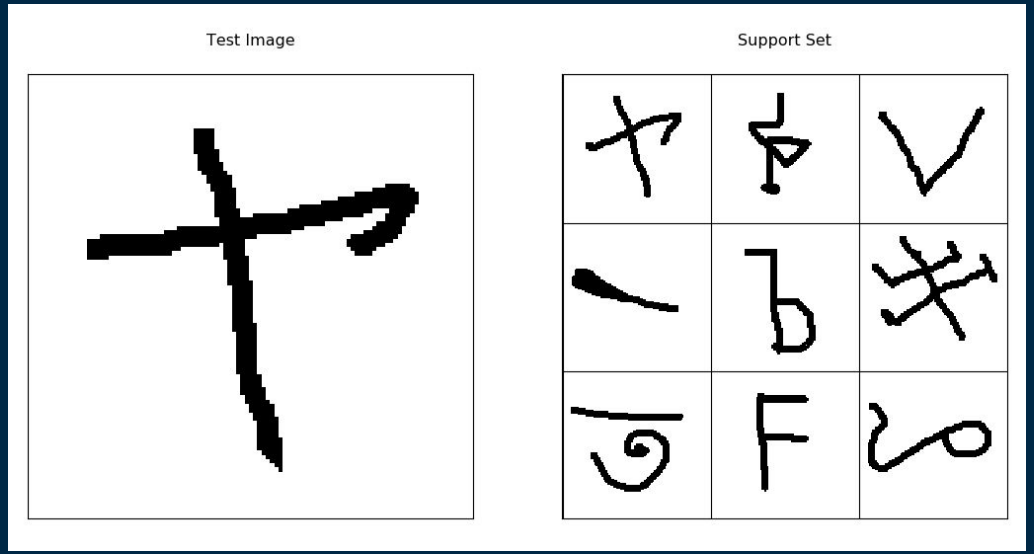
Methodology

- Siamese Neural Network



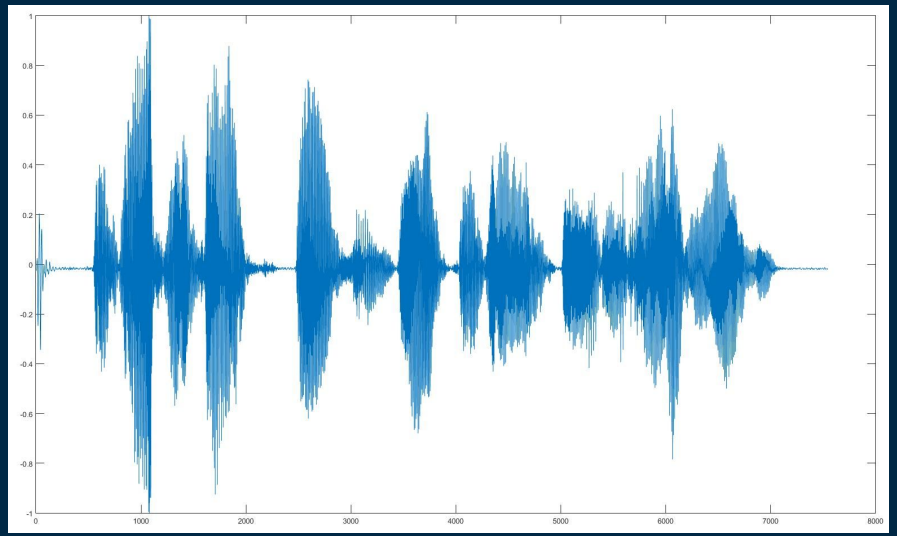
Methodology

- Siamese Neural Network
- Validation: n-shot k-way classification task



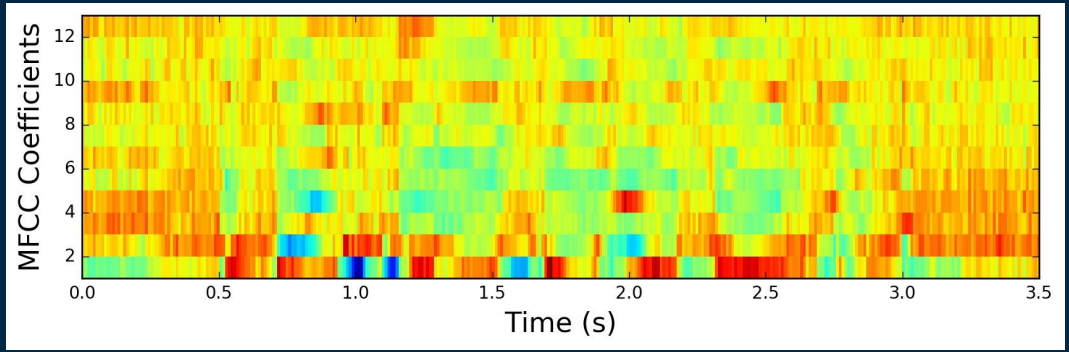
Methodology

- Siamese Neural Network
- Validation: n-shot k-way classification task
- Data Representations:
  - Raw Audio



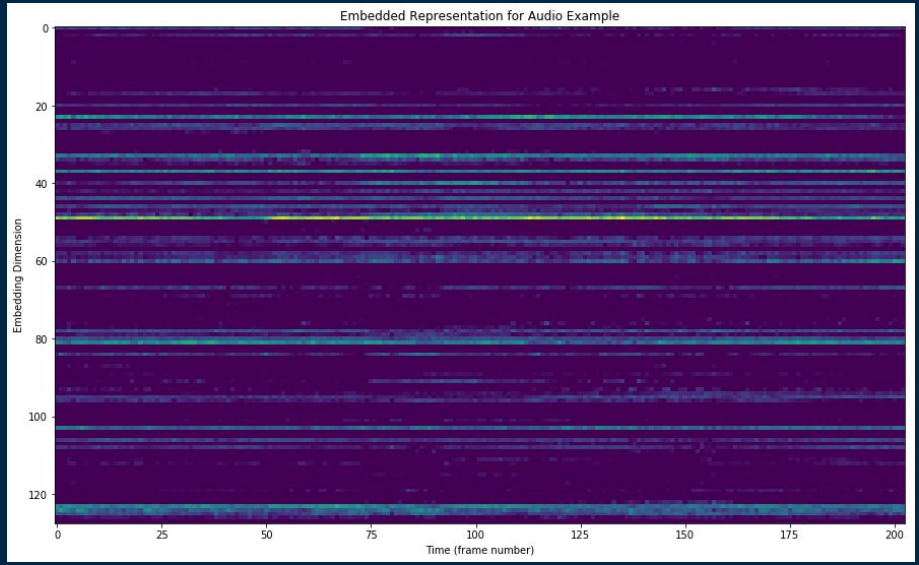
Methodology

- Siamese Neural Network
- Validation: n-shot k-way classification task
- Data Representations:
  - Raw Audio
  - MFCCs



Methodology

- Siamese Neural Network
- Validation: n-shot k-way classification task
- Data Representations:
  - Raw Audio
  - MFCCs
  - VGGish Embeddings



Experiments 3

Best Results

Neural Network Structure:

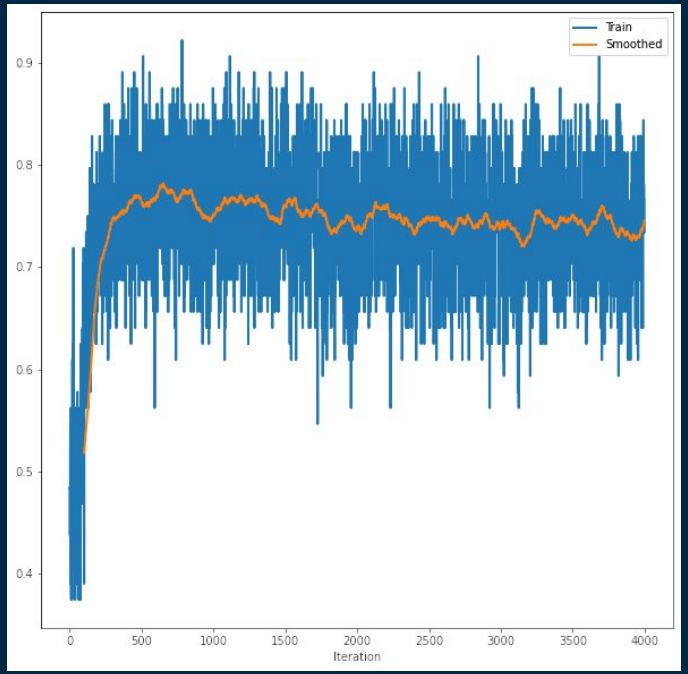
- Input: MFCC
- 3 x CNN Blocks
  - Filters: 128 | 256 | 384
  - Stride: 3 x 3
  - Batch Normalization
  - Dropout: 0.2
  - Max Pooling
  - Activation Function: ReLU
- Global Max Pooling
- Fully Connected Layer
  - Units: 1024
  - Dropout: 0.2
- Euclidean Distance

Training Parameters:

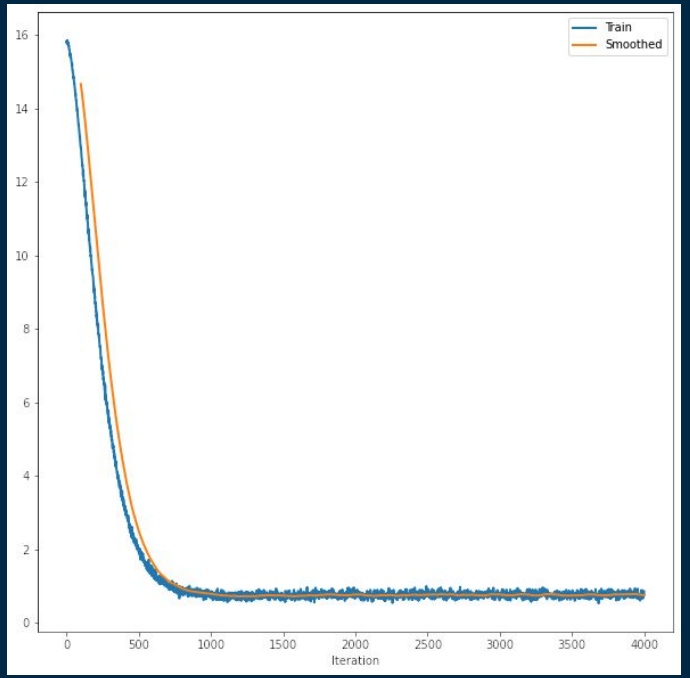
- Optimizer: RAdam
- Loss Function:
  Binary Cross-Entropy
- Batch Size: 64

Training Phase

Accuracy



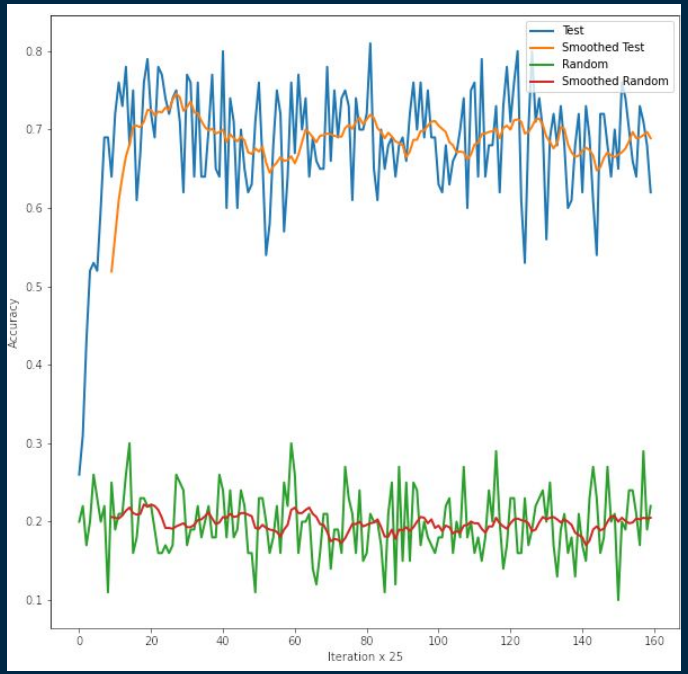
Loss



Validation Phase

- 1-shot 4-way classification, evaluated every 25 batches over 100 tasks

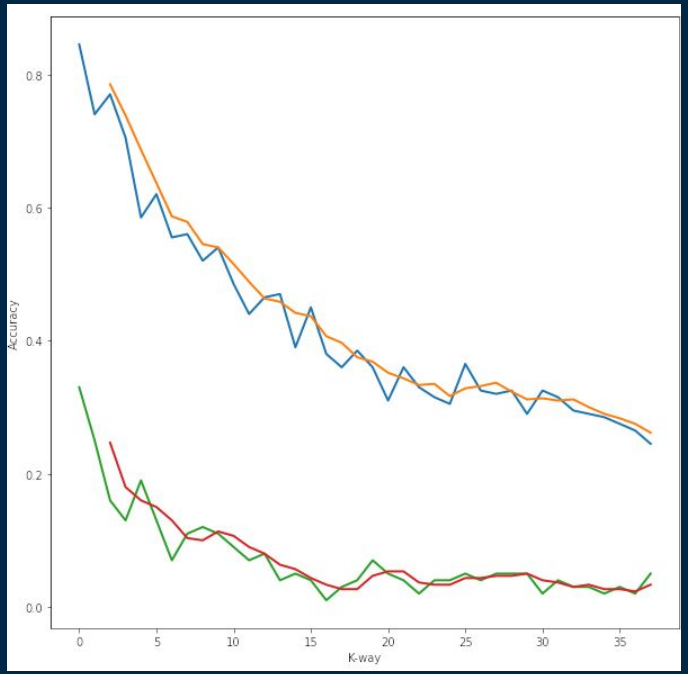
Accuracy



Testing Phase

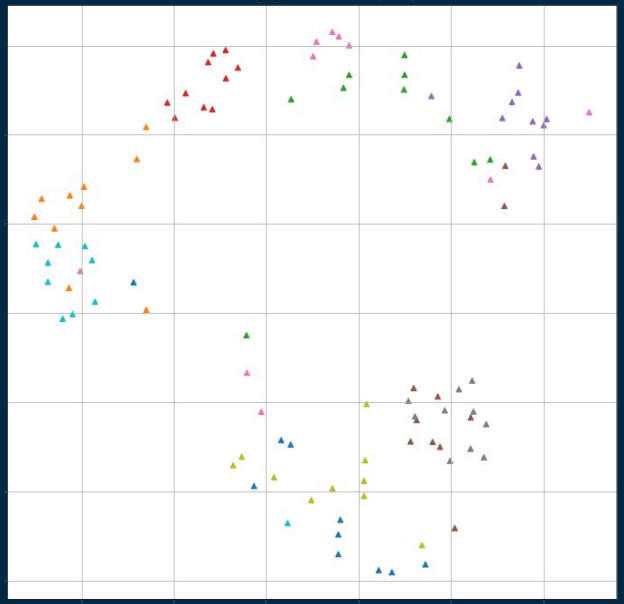
- 1-shot k-way classification, evaluated over 100 tasks per k

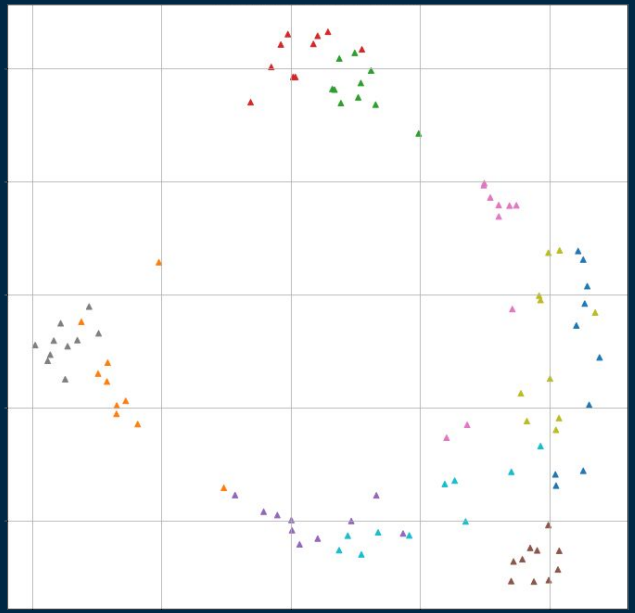
Accuracy



Embedding Space Visualization

- Vector embeddings obtained from 2 sets of 10 random selected speakers (10 random audios fragments each one)
- Dimensionality reduction through t-SNE algorithm





Conclusions

4

Result Discussion

- The three types of data representations have obtained satisfactory results
  - Small difference in performances, but huge in dimensionality
  - Raw audio is a very inefficient representation for transmitting information or extract patterns
- Best results obtained with MFCCs
  - Promising results for the join of Deep Learning and specific domain knowledge
- Siamese Network architecture has perform the task quite successfully
  - Very time and computational consuming approach
  - Number of possible pair combinations grows exponentially

Future Directions

- Intensive search of hyperparameters (length audio fragments, subsampling rates, neural structures, extended training times, different distance metrics, etc.)
- Use of CNN and LSTM neurons
- Implement validation tasks n-shot k-way, for n > 1
- Better baseline model (e.g. k-NN)
- Compare performance of test with new samples against classes seen during training phase

THANKS!

jaime.perez.sanchez@gmail.com

linkedin.com/in/jaime-perez-sanchez

github.com/jaimeperezsanchez