Identity Recognition from audio

Her Går Det Godt Podcast

Frederike Durow
Gian Paolo Currà
Guglielmo Borzone

Agenda

- Introduction
- Preprocessing the Dataset
- MLP vs LSTM vs KNN (Accuracy comparison)
- Future improvement

Introduction

Task: Identity recognition from audio

Dataset: “Her Går Det Godt” 300 5-sec .wav files

- 100 wavs where Esben is speaking
- 100 wavs where Peter is speaking
- 100 wavs where both are speaking

Her Går Det Godt

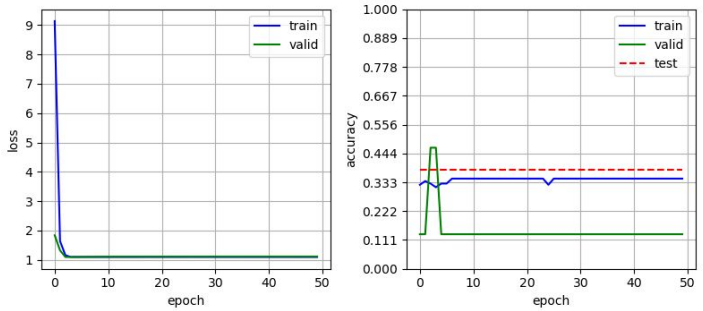
https://podcastguides.dk/podcast-her-gar-det-godt/

Dataset preprocessing

- **Dataset 1:**
  - 2 classes (Peter, Esben)
  - short (1000 by slicing 5-sec samples to 1-sec samples)
- **Dataset 2:**
  - 3 classes (Peter, Esben, both)
  - long (300 and not sliced)

Features & Normalization

- **Wavs:** wave amplitude
- **Mfcc:** Mel-Frequency Cepstral Coefficients
- **Chroma:** chromagram from amplitude
- **Spectral Contrast:** difference in amplitude between peaks and valleys in the spectrum



MLP for sliced dataset, normalized wavs

- All features have been tested normalized and not normalized

librosa

https://librosa.org/doc/latest/index.html

3 POSSIBLE SOLUTIONS

KNN, MLP, LSTM

Thought Process

1. Search for the best feature that can represent the data and decide whether or not to normalize (hyp random initialized)
2. Search for the best-tuned learning algorithm

K-Nearest Neighbours

Settings:

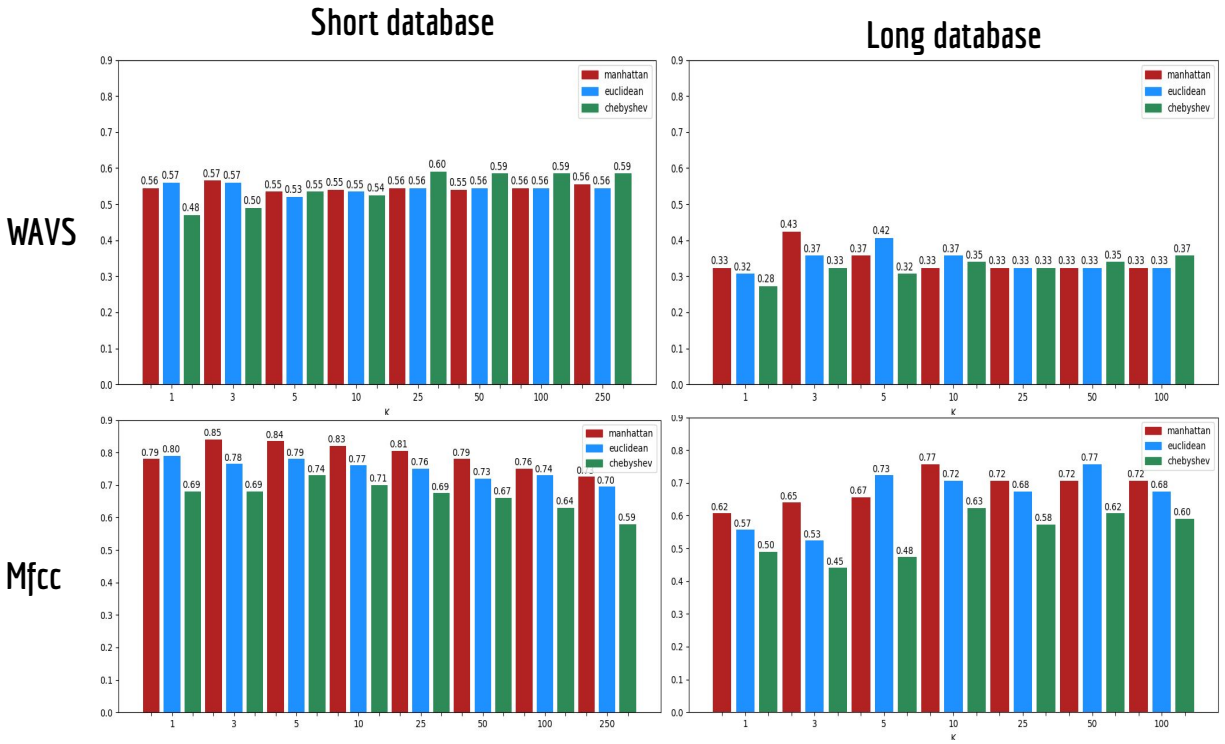
- K: 1, 3, 5, 10, 25, 50, 100, (250)
- Metrics: Euclidean, Manhattan,
- Chebyshev

Short with 2 classes:

- K: 3
- Metric: Manhattan

Long with 3 classes:

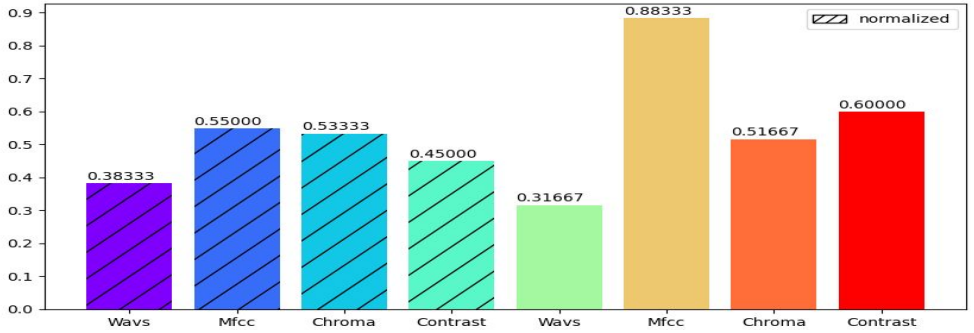
- K: 10 or 50
- Metric: Manhattan or Euclidean



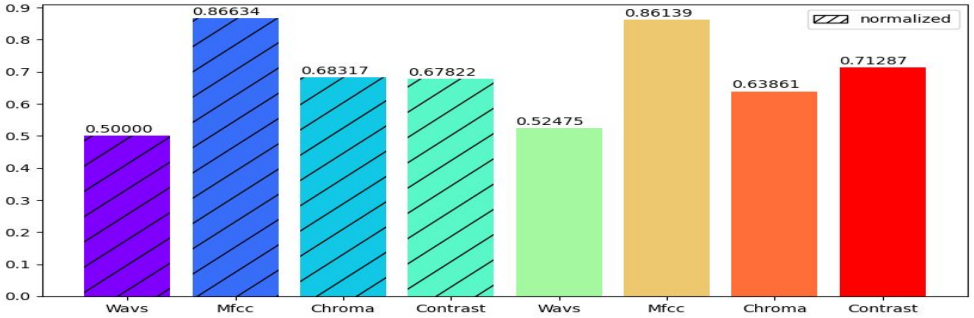
Multi Layer Perceptron

Search for the best feature (mfcc, not norm):

Long database

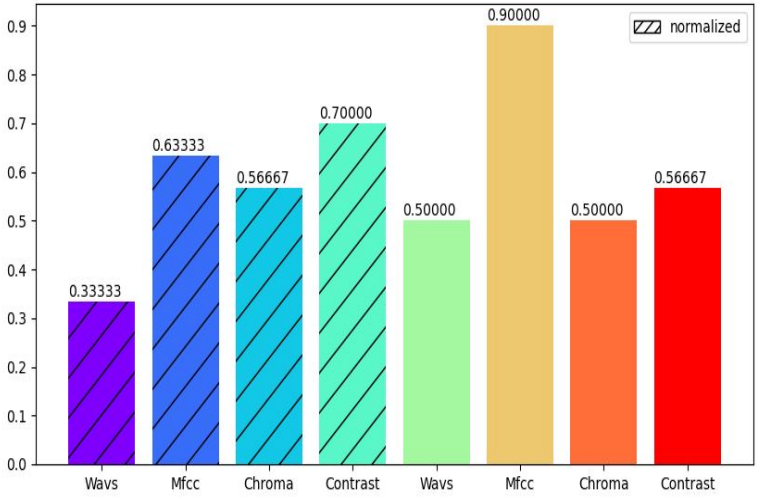


Short database

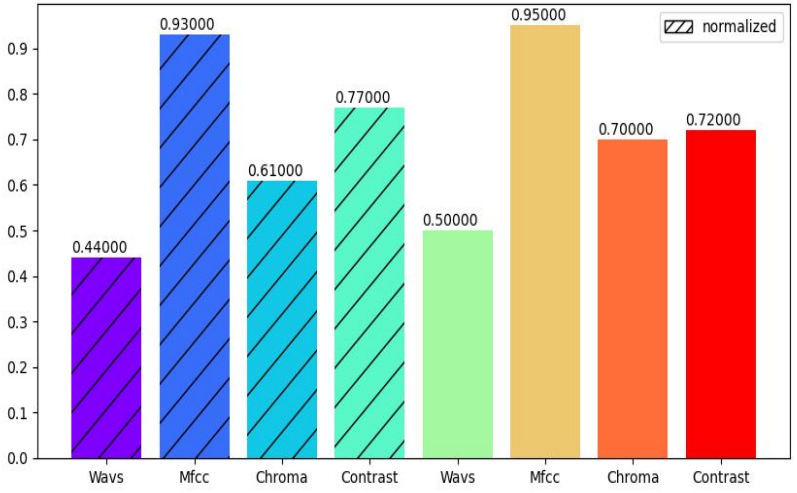


Recurrent NN with LSTM

Long & 3 Classes



Short & 2 Classes



Hyper parameters optimization

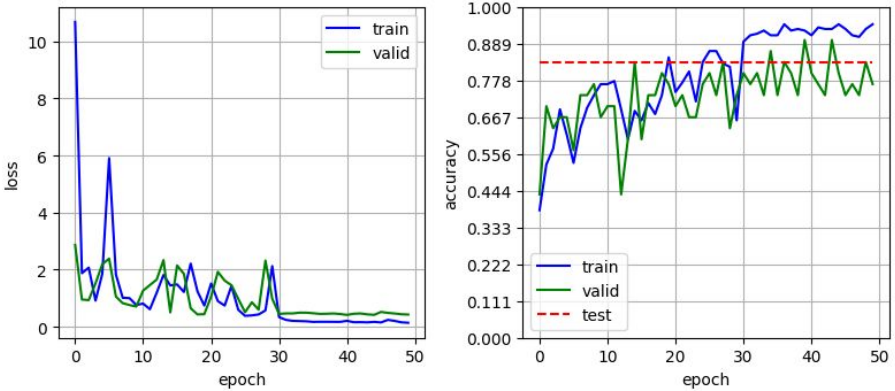
Hyper-parameters considered (mfcc, not norm):

- Number of hidden neurons [10, 100, 500, 1000]
- Starting learning rate [0.01, 0.001, 0.0001]
- Batch Sizes [1, 8, 32, 128]
- Number of hidden layers [1, 3, 5]

From all this nets we selected the one with the best test accuracy

Multi Layer Perceptron

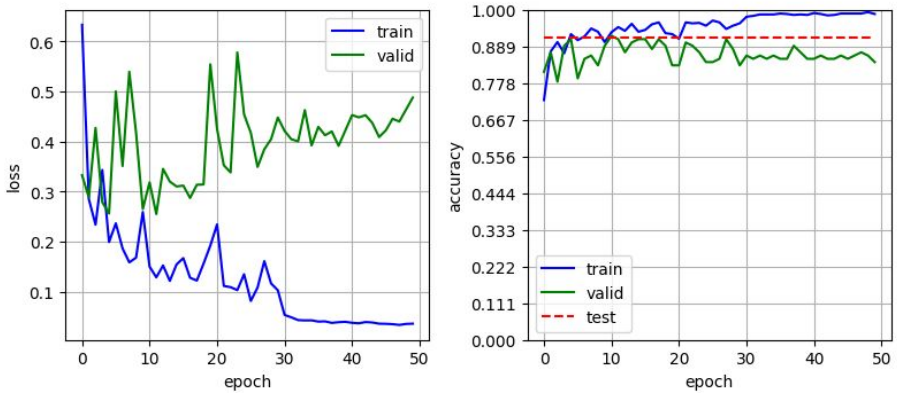
Long database



Test Accuracy: 0.817

N_layers: 1; N_hidden_neurons: 1000; LR: 0.0001; Batch_size: 8

Short database



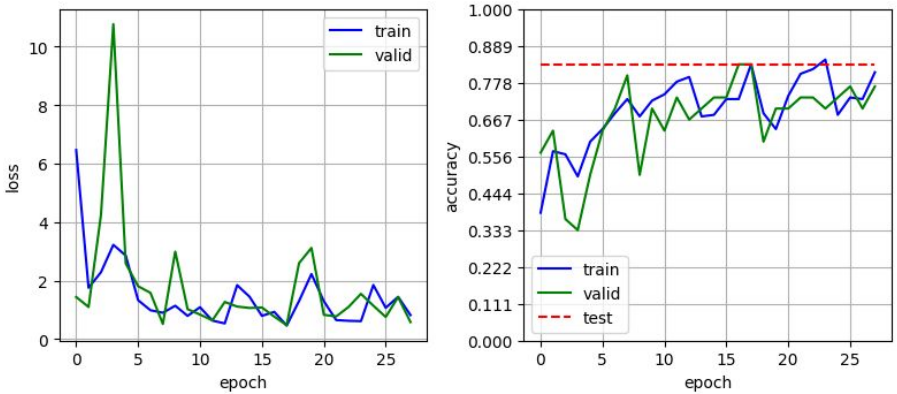
Test Accuracy: 0.935

N_layers: 1; N_hidden_neurons: 100; LR: 0.0001; Batch_size: 8

Multi Layer Perceptron

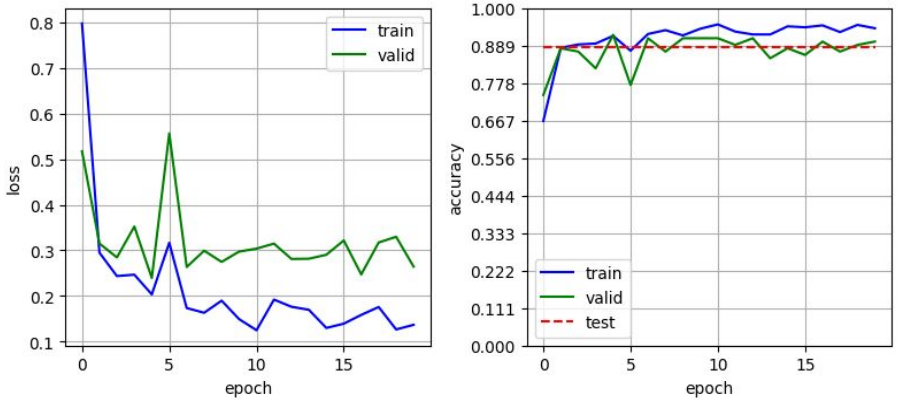
Introducing a patience value of p for early stopping in the epoch

Long database (p=15)



Test Accuracy: 0.833

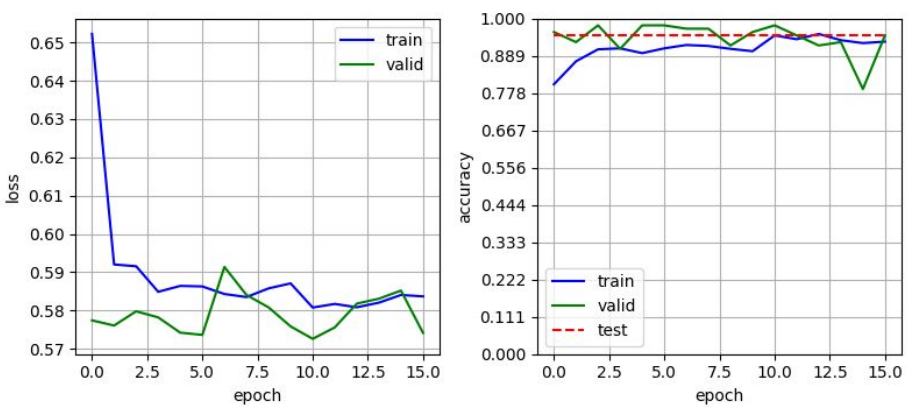
Short database (p=10)



Test Accuracy: 0.886

Recurrent NN with LSTM

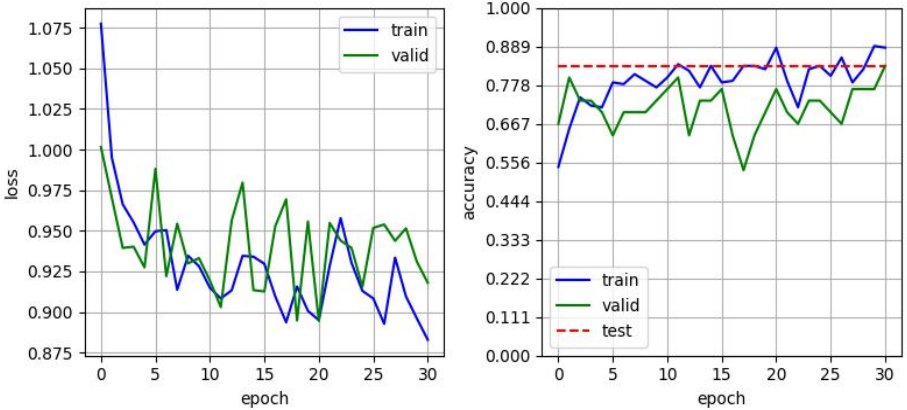
Long database



Test Accuracy: 0.95

N_layers: 4; N_hidden_neurons: 1000; LR: 0.0001; Batch_size: 8

Short database



Test Accuracy: 0.83

N_layers: 2; N_hidden_neurons: 500; LR: 0.001; Batch_size: 8

Comparison

	KNN	MLP	LSTM
Short with 2 C	85%	93	95%
Long with 3 C	77%	83%	83%

Improvements

- Data augmentation
- Introducing patience or other early stop methods during the hyper-parameters phase
- Transfer learning through auto encoder