01

A Neural Network Model for Music Instrument Recognition

Ismael Martínez Ferrer

Spiced Academy

Costmary Function

Introduction

- Ismael Martínez Ferrer
- Civil Engineer
- From Valencia, Spain
- Living in Berlin, since 2004

Audio in Machine Learning

MIR: MUSIC INFORMATION RETRIEVAL

Broad and interdisciplinary: signal processing, music theory, psychology, machine learning

Other tasks: Recommender systems, Music Generation, Music Transcription

MER: MUSIC EMOTION RECOGNITION

Valence, energy, tension, anger,
fear, happy, sad, tender

MIDDLE LEVEL FEATURES

Provide explainability:
rhythm, melody, harmony

MUSIC GENRE RECOGNITION

Genre Classifiers:
Jazz, Blues, Country, Rock

MUSIC INSTRUMENT RECOGNITION

Isolated instruments
Recognition in a song context
Timbre, harmonics, frequencies, attack
Main approaches:
Unsupervised k-NN and Neural Networks

The Task: Instrument Recognition Model

Music Instrument Recognition scaling from two to four Instruments:

2 Instruments: distorsion electric guitar + singer

3 instruments: add clarinet

4 instruments: add piano

Samples: 3-second snipets with one predominant instrument (soloing) extracted out of actual songs

2916 3-seconds
.wav audio files

train set:
1468 samples
validation set:
1446 samples

DATA

librosas library:
audio file to signal
melspectrograms

Processing

CNN:
2 layers/ 3 layers
16/ 25 filters
w/ batch norm.
w/ dropout
stride: 2x2, no padd.

Model

monitor accuracy and
loss functions
hyper paramter study
loss function:
cross-entropy

Outcome

The Data Set

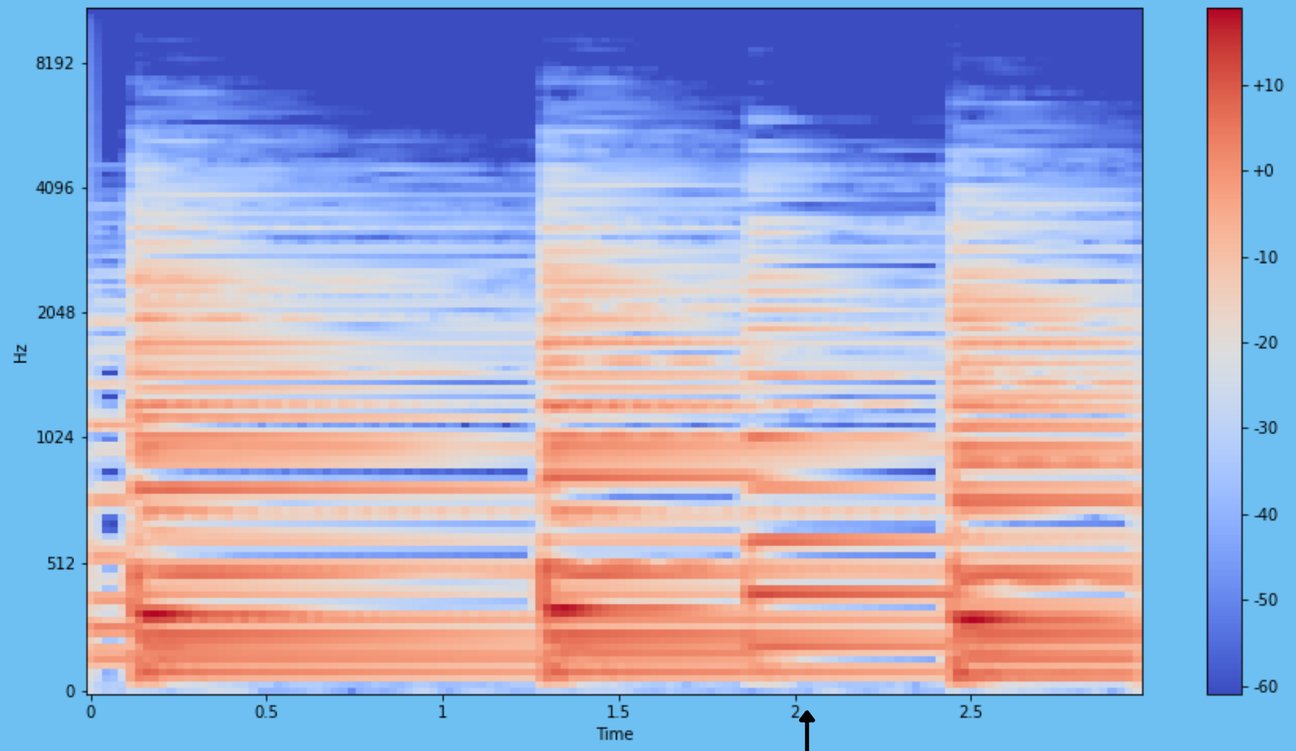
Medley-solos-DB

- 21571 audio clips as WAV files
- sample rate = 44100 kHz, bit depth = 32 bits with a single channel (mono)
- 8 instruments: clarinet, distorted electric guitar, female singer, flute, tenor saxophone, trumpet and violin
- duration: 2972 ms, 65536 discrete-time samples
- train, validation, test: 5841, 3494, 12236
- description: one predominant instrument in a song context

https://zenodo.org/record/3464194#.YnrhsFDP2Ul

Audio Processing

- time domain: discrete time intervals
- frequency domain: freq [Hz] to mel fr [Hz]
- Amplitude: power to dB



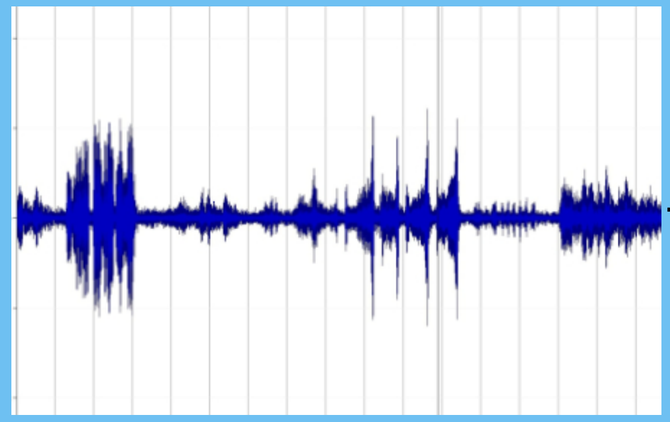
Spectrogram:

- 3s at 22050 Hz: 129 time intervals
- 129 mel freq bands
- db scale

(129, 129 , 1)

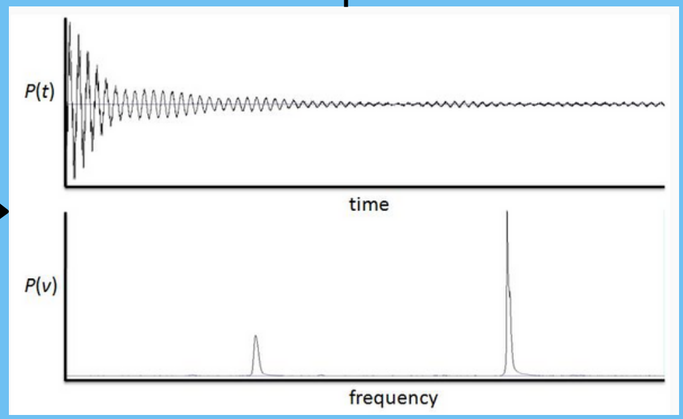
optional: feed other features

for ex.: sound convolution

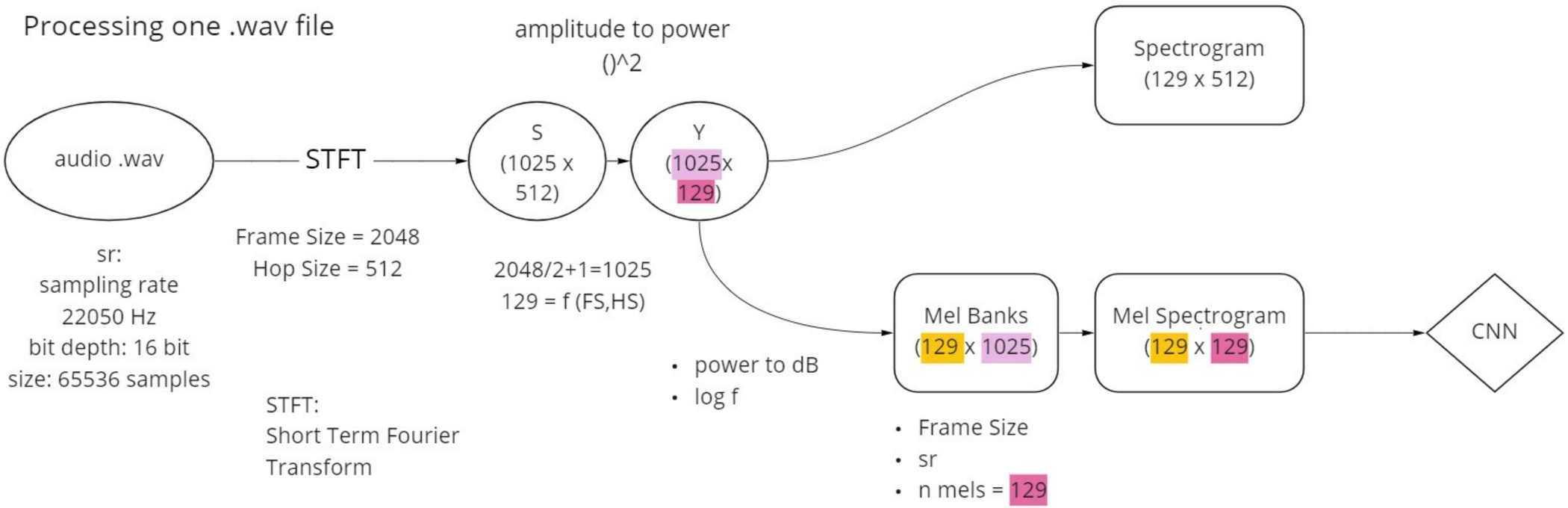


sound wave

STFT



Data Processing



The Model: table

Task		Neural Network	activation funct.	batch norm. /dr
2 instr.	0a	2 CNN (16) f=3	relu / tanh	
	1	2 CNN (16) f=3	relu / tanh	BN
	2	3 CNN (16) f=3	relu / tanh	
	2B	3 CNN (16) f=3	relu / tanh	BN, dr
	3	2 CNN (25) f=3	relu / tanh	
3 instr.	0a	2 CNN (16) f=3	relu / tanh	
4 instr.	0a	2 CNN (16) f=3	relu / tanh	
	2	3 CNN (16) f=3	relu / tanh	

The Model: conclusions

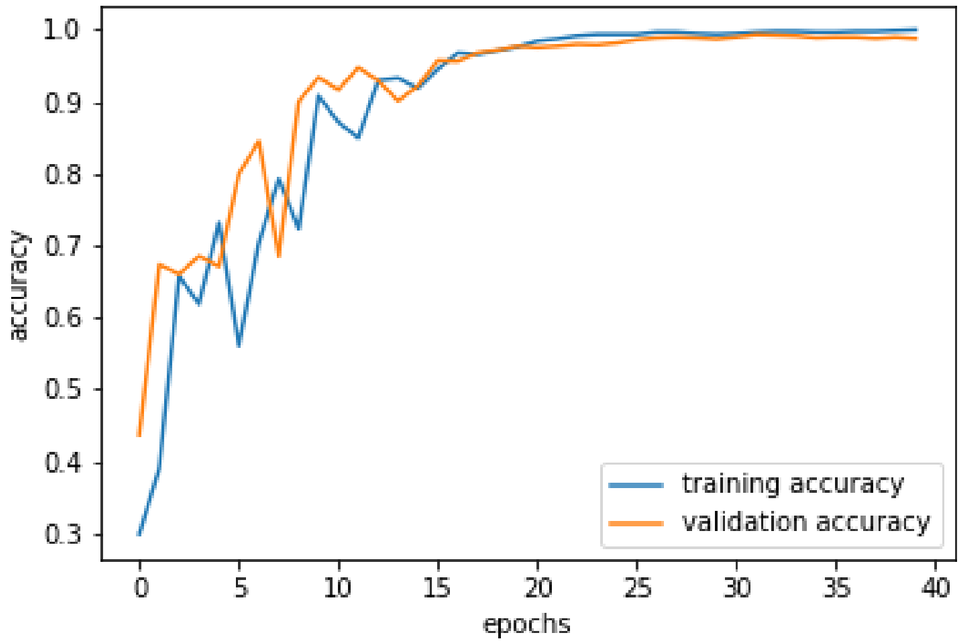
Task	id	Neural Network	train acc.	val. acc.	comments
2 instr.	0a	2 CNN (16) f=3	99.7 %	98.4 %	adjusting act. funct. allowed improvement on loss
	2	3 CNN (16) f=3	100%	100%	3 layers: turbo learning overkill
	3	3 CNN (16) f=5	100%	99.4%	slower learning and computation
3 instr.	0a	2 CNN (16) f=3	100%	98.8%	very good performance with a simple model
4 instr.	0a	2 CNN (16) f=3	99.7 %	89.6 %	good performance with simple model
	2	3 CNN (16) f=3	100%	94.5 %	very good performance adding 3rd layer

Outcome

3 Instruments: guitar, singer, clarinet

Simple model: 2 CNN

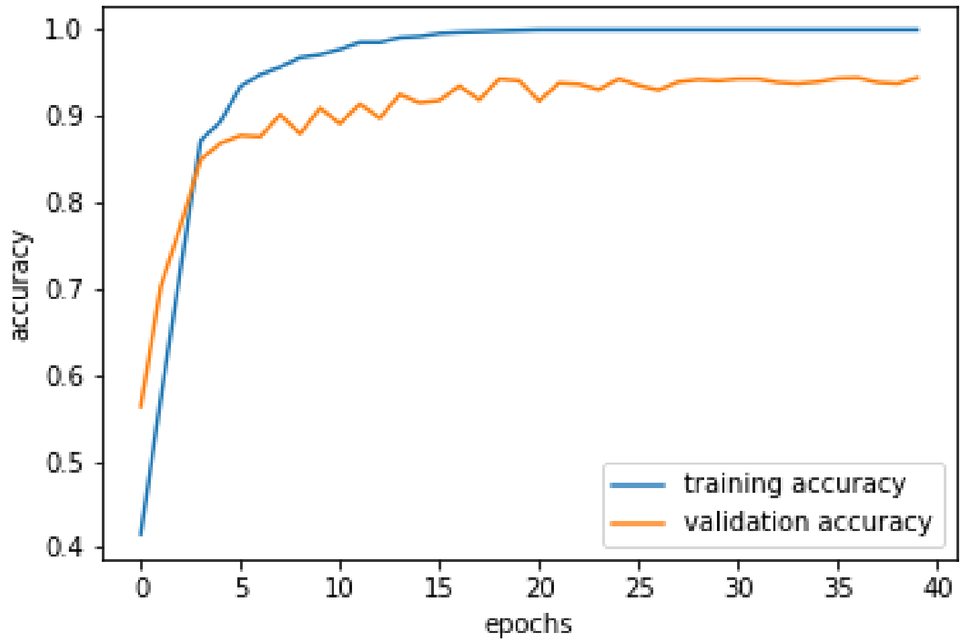
test accuracy: 98.8 %



4 Instruments: guitar, singer, clarinet, piano

Deep Network: 3

test accuracy: 94.5 %



Music Box: an application with musical functionalities

- Instrument Recognition
- Song Recommender
- Playlist genre classification
- Music Generation according to playlist

Thank

You

for the great experience to all my cohort students and teachers

Spiced Academy
Costmary Function.