Predominant Musical Instrument detection

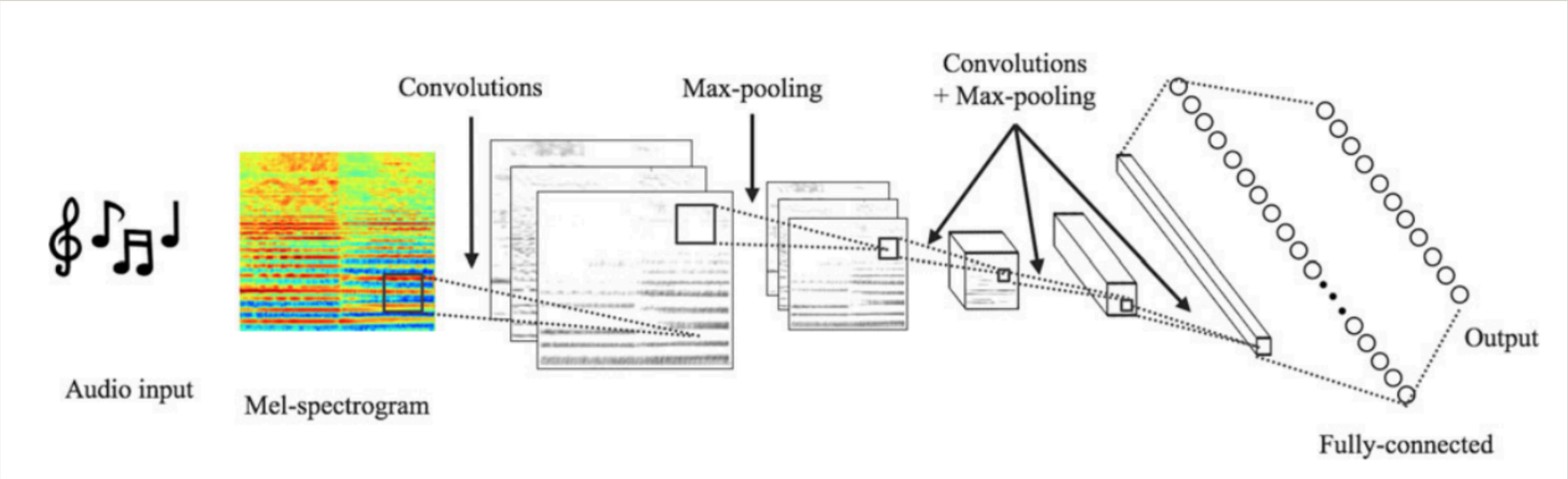
Igali Ayan, Barlykov Beket, Bekbolat Yernar

Team members

1. Igali Ayan: **Team leading, researching, coding**
2. Barlykov Beket: **Coding, audio preprocessing**
3. Bekbolat Yernar: **Coding, Dataset collection**

Problem

Identifying musical instruments in polyphonic music recordings for searching music, musical genres and transcription of music



Why this topic?

Showing **1-14** of **14** results for **Predominant Instrument Recognition** ×

Showing **1-25** of **26** results for **Predominant Instrument classification** ×

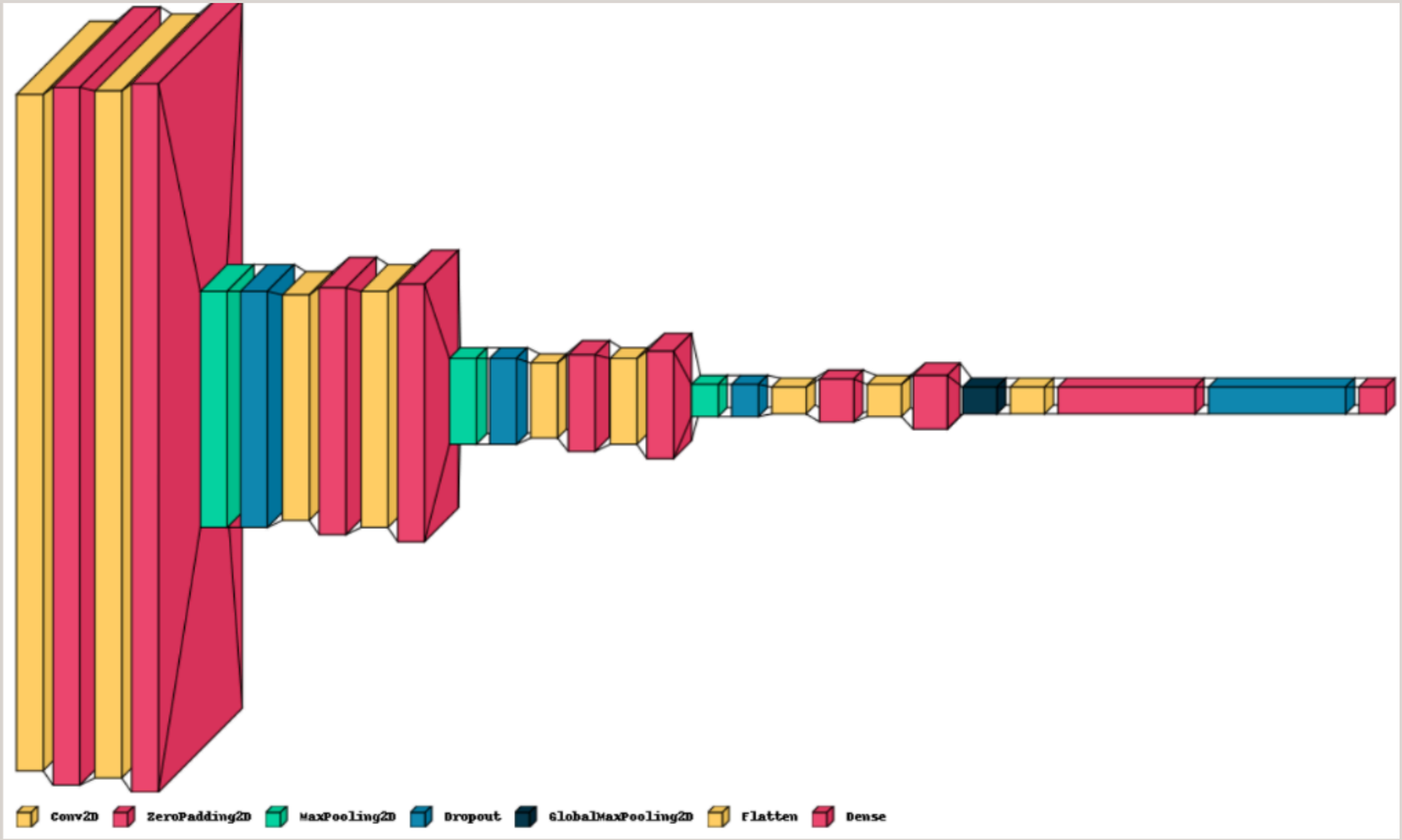
Showing **1-25** of **1,220** results for **genre classification** ×

Showing **1-25** of **569** results for **genre recognition** ×

screenshots from ieeexplore

Research on what we based

Han, Y., Kim, J., & Lee, K. (2016). Deep convolutional neural networks for predominant instrument recognition in polyphonic music.
IEEE/ACM Transactions on Audio, Speech, and Language Processing

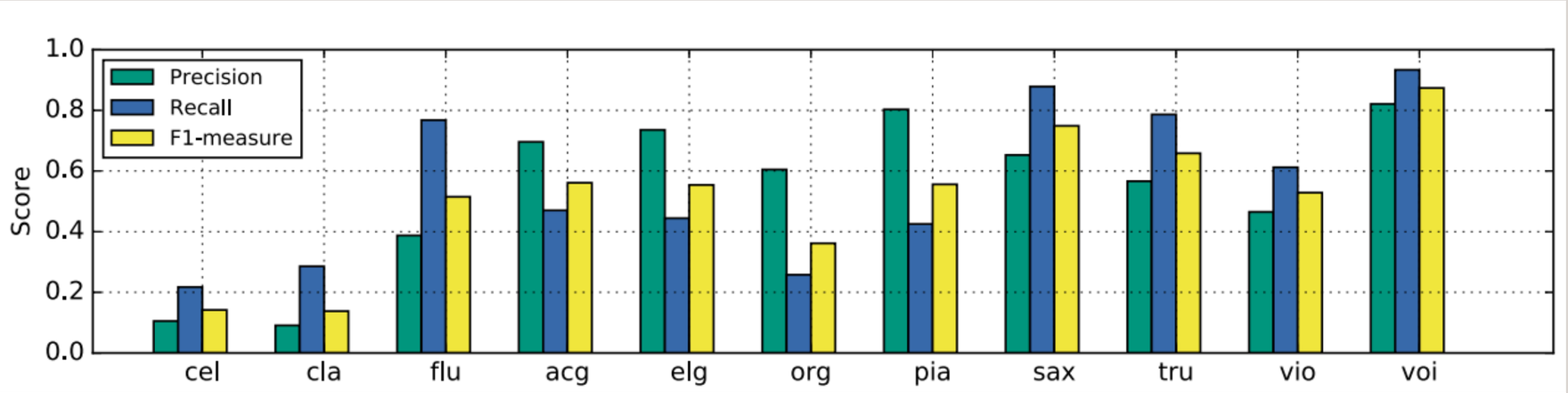


Their proposed model

Their results on multi-label classification

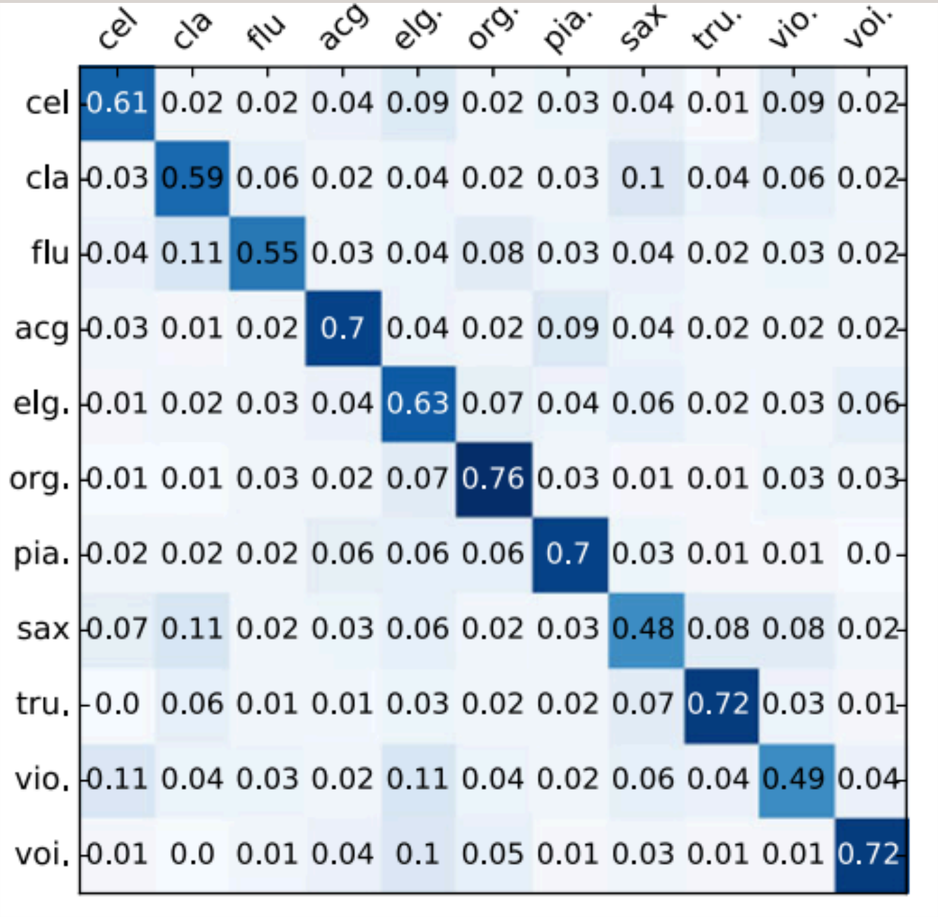
F1-score	Precision	Recall
0.629	0.657	0.603

Their results on multi-label classification



Their results on single-label classification

Accuracy on single instrument classification: 0.633

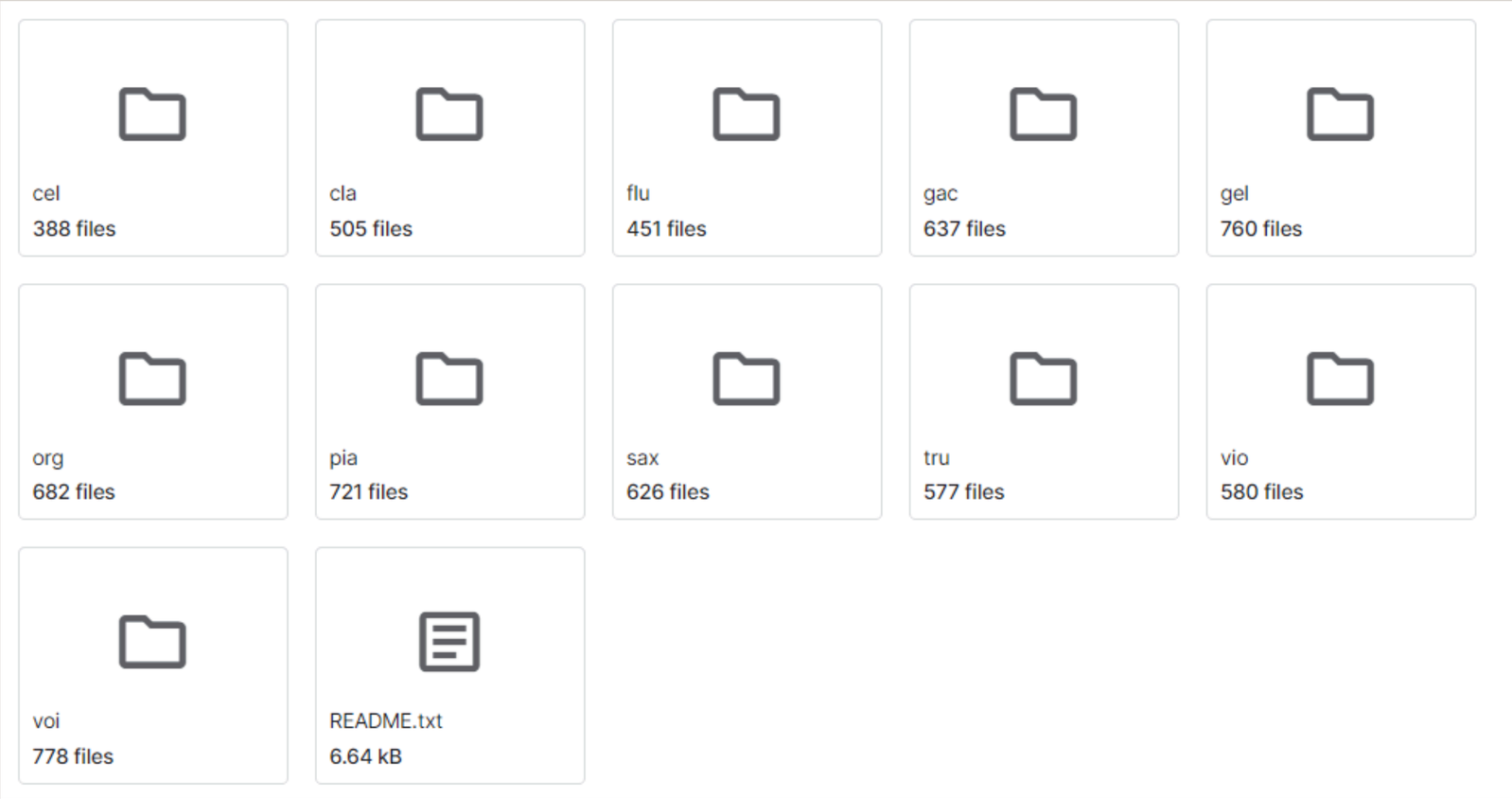


Dataset

IRMAS: 6705 audio files of ‘wav’ format of differents like:

a. cello (cel) - 388
b. clarinet (cla) - 505
c. flute (flu) - 451
d. acoustic guitar (gac) - 637
e. electric guitar (gel) - 760
f. organ (org) - 682
g. piano (pia) - 721
h. saxophone (sax) - 626
i. trumpet (tru) - 577
j. violin (vio) - 580
k. human singing voice (voi) - 778

Stucture of IRMAS dataset folder



Kazakh national instruments

1. **Dombyra** (dom) - 587

2. **Kobyz** (kob) - 575

Audio preprocessing

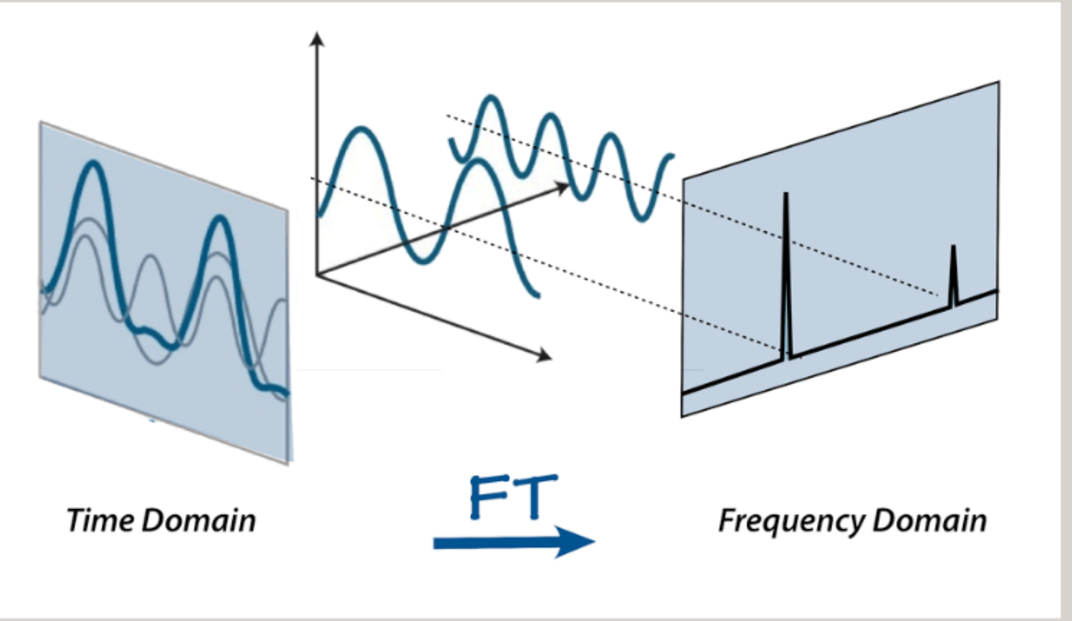
Fourier Transform

FT transfers a signal from real-valued function in the **time domain** to a complex-valued function in **frequency domain**:

$$X(f) : \mathbb{R} \to \mathbb{C}$$

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} dt$$

original signal



$$X(f) = x + iy = \rho e^{i\phi}, \rho = \sqrt{x^2 + y^2}, \phi = \arctan\left(\frac{y}{x}\right)$$

magnitude phase

From Fourier Series to Fourier Integral

Fourier Series

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i \frac{n}{T} t}$$

$T \to \infty$

Fourier Integral

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{2\pi i f t} df$$

Fourier Coefficient

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-2\pi i \frac{n}{T} t} dt$$

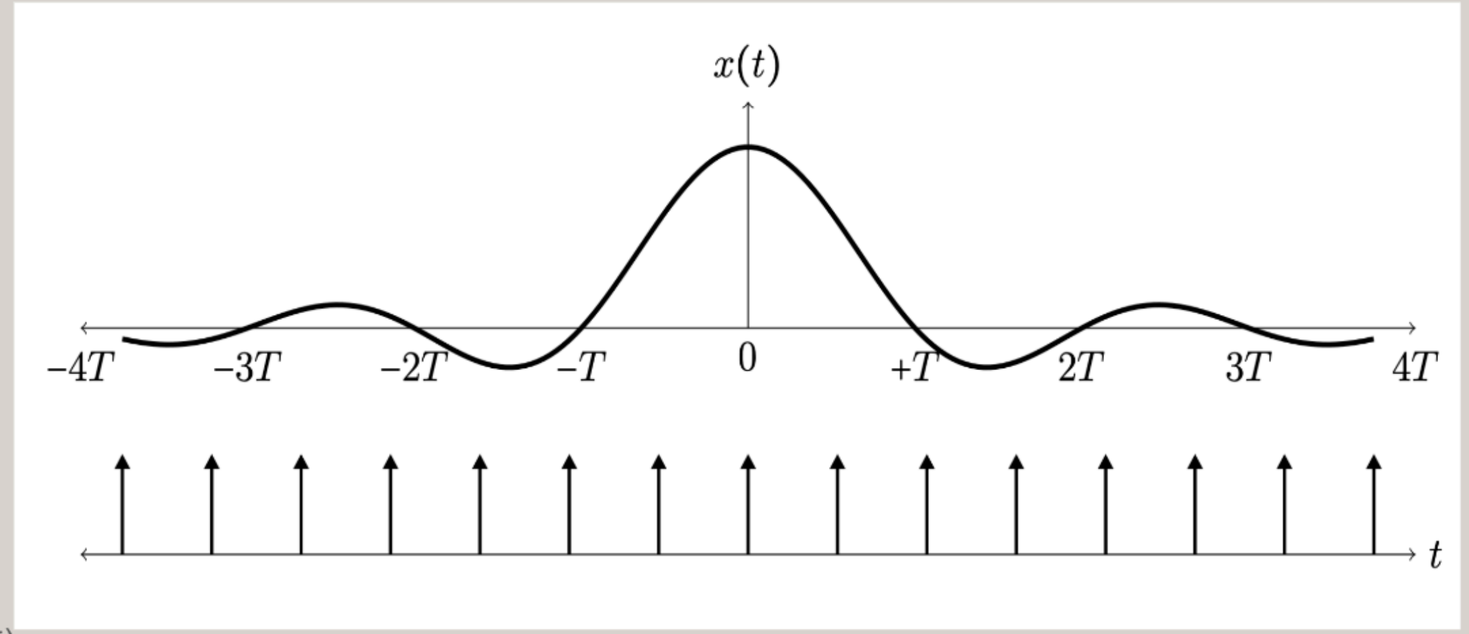
Fourier Transform

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} dt$$

Sampling

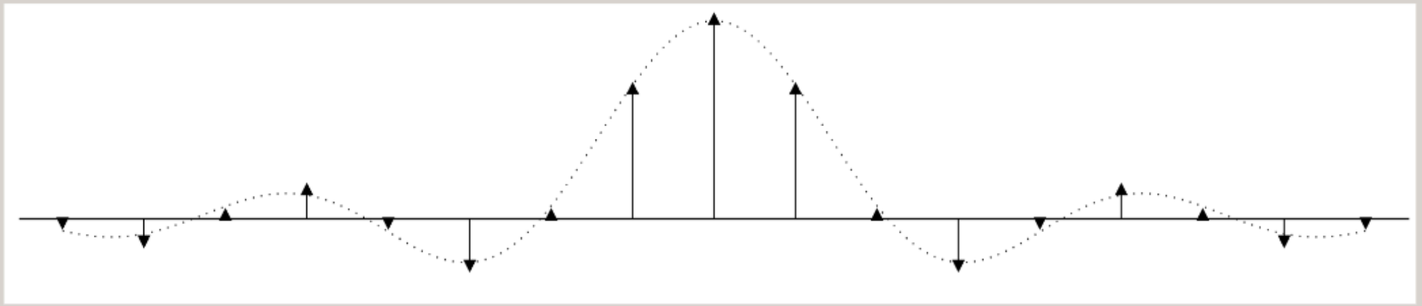
Multiplication by an impulse corresponds to sampling:

$$f(t)\delta(t-a) = f(a)\delta(t-a)$$



Thus the sampled waveform is[*]

$$\sum_n x(t)\delta(t-nT)$$



Apply Fourier Transform to Sampled signal

$$\tilde{X}(f) = \int_{\mathbb{R}} \sum_n x(nT)\delta(t - nT)e^{-2\pi i f t}dt =$$

$$= \sum_n x(nT) \int_{\mathbb{R}} \delta(t - nT)e^{-2\pi i f t}dt =$$

$$= \sum_n x(nT)e^{-2\pi i f nT}$$

Shannon-Nyquist Sampling Theorem (Kotelnikov)

Theorem: *A function $x(t)$ containing no frequency higher than f Hz, is completely determined by sampling at $f_s = 2f$ Hz.*

Equivalently: To resolve all frequencies in a function, it must be sampled twice the highest frequency present.

$$f = \frac{f_s}{2}$$ is called **Nyquist frequency**

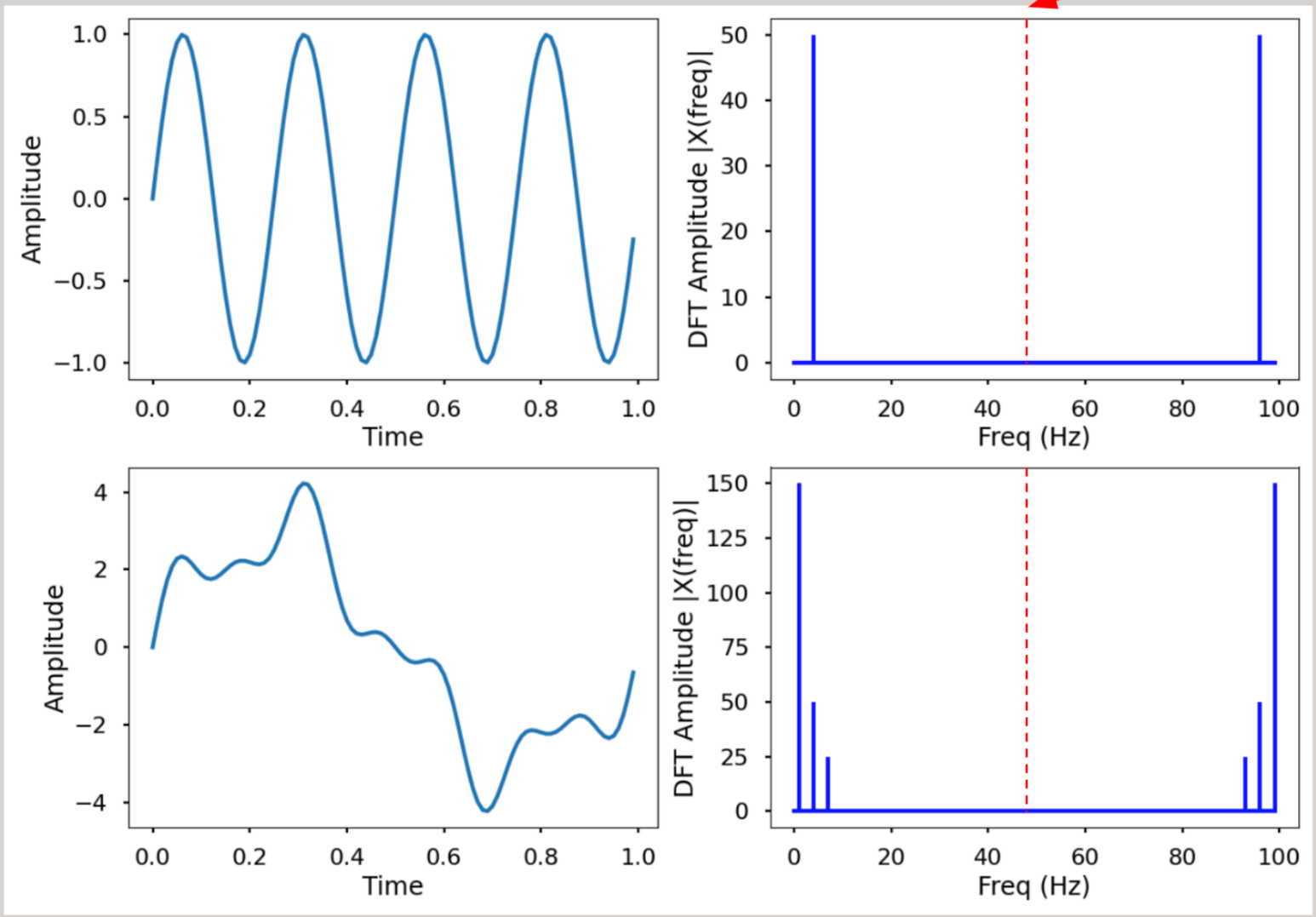
For Nyquist frequency

$$f = k\frac{f_s}{N} \quad \Rightarrow \quad k = \frac{N}{2}$$

Nyquist frequency is always in the middle of a spectrum

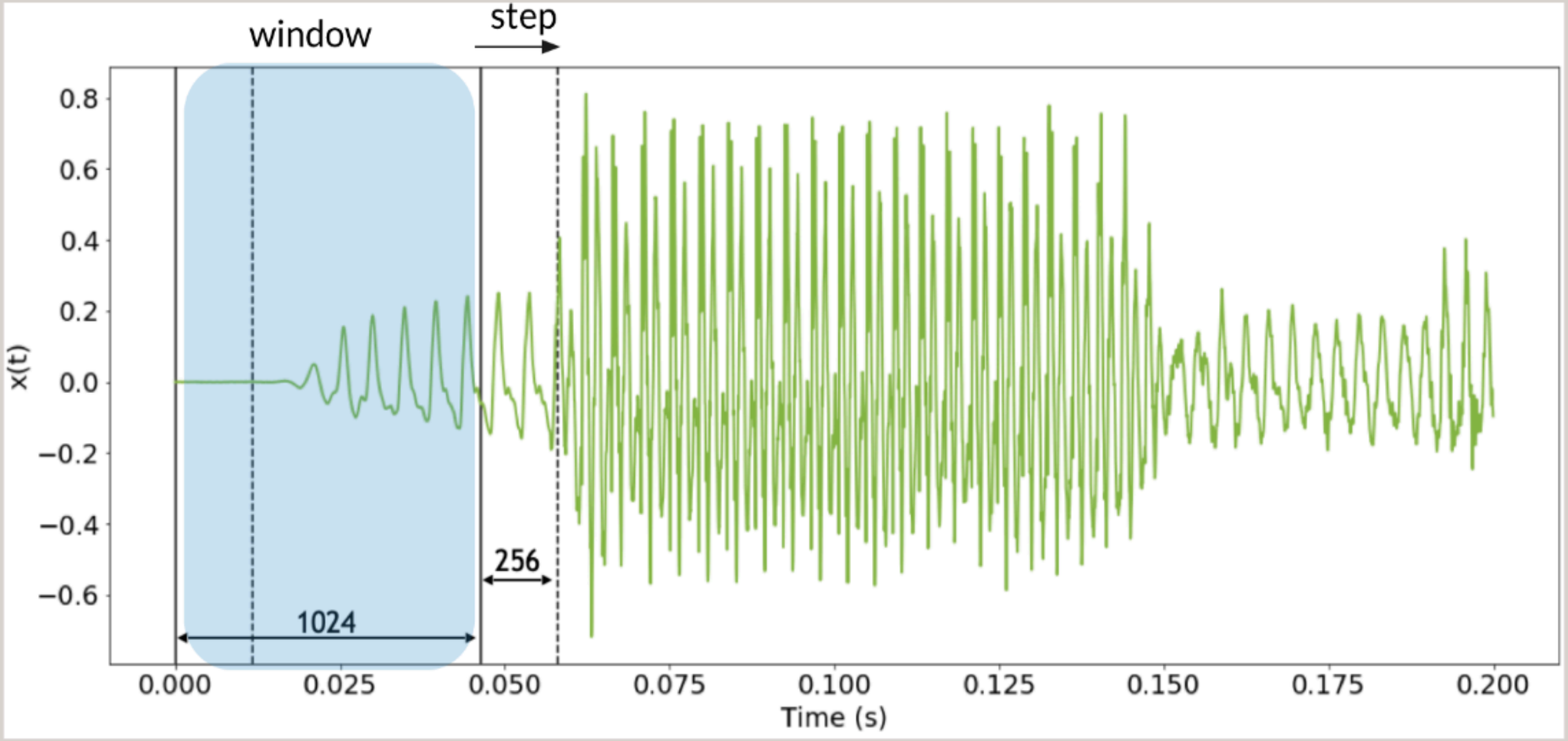
https://www.youtube.com/watch?v=FcXZ28BX-xE&t=277s

DFT: Symmetry of coefficients



Short-Time Fourier Transform (STFT)

Need to examine local spectrum to see how it changes!

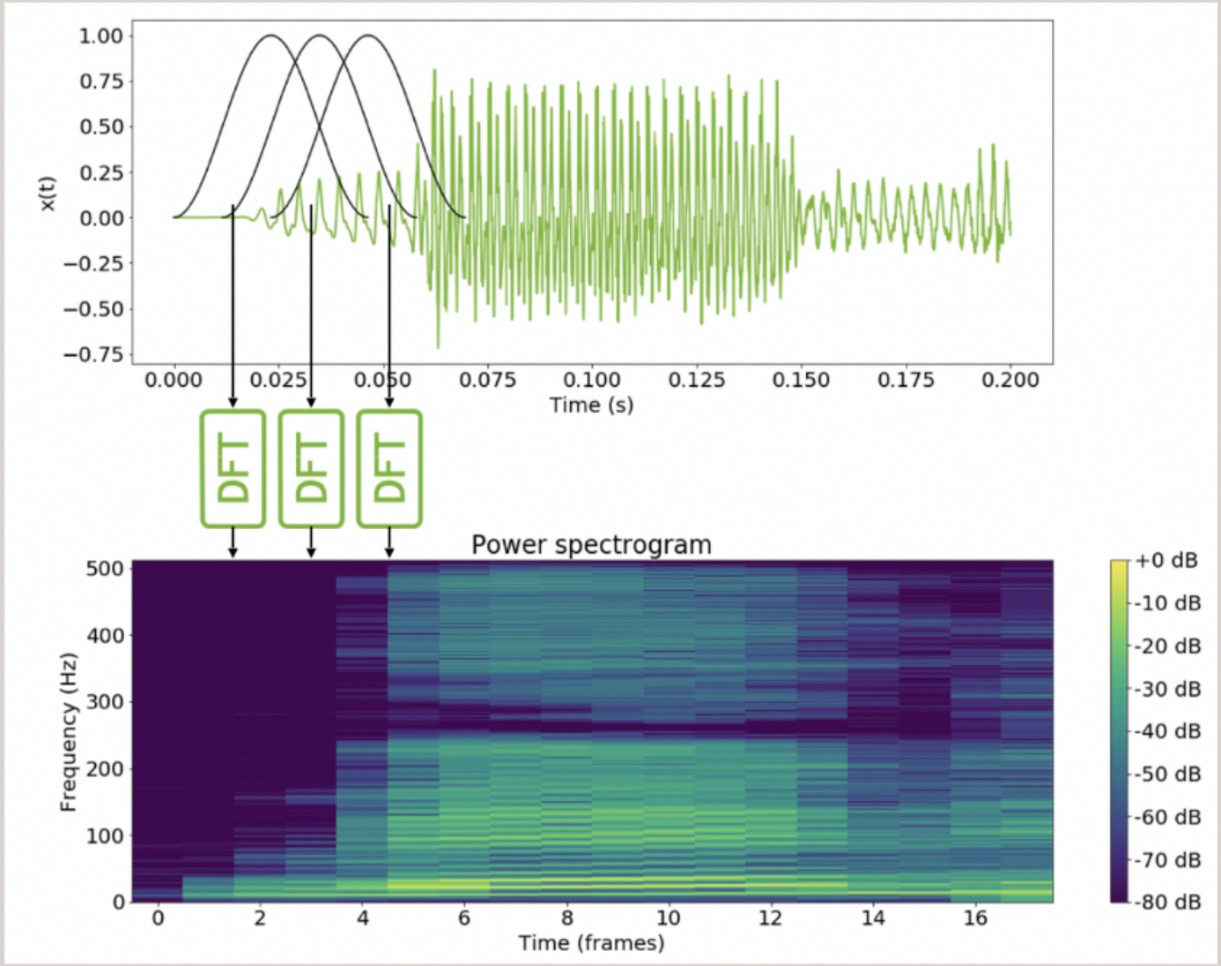


Spectrogram

- Stack STFTs of a window sliding over the signal
- Each window is called **acoustic frame**

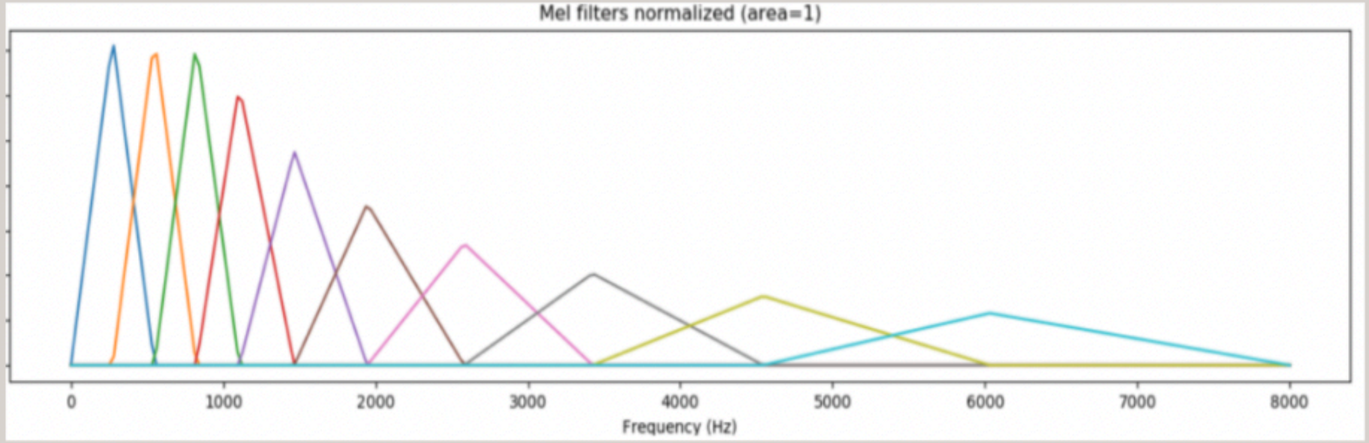
Important hyperparameters:

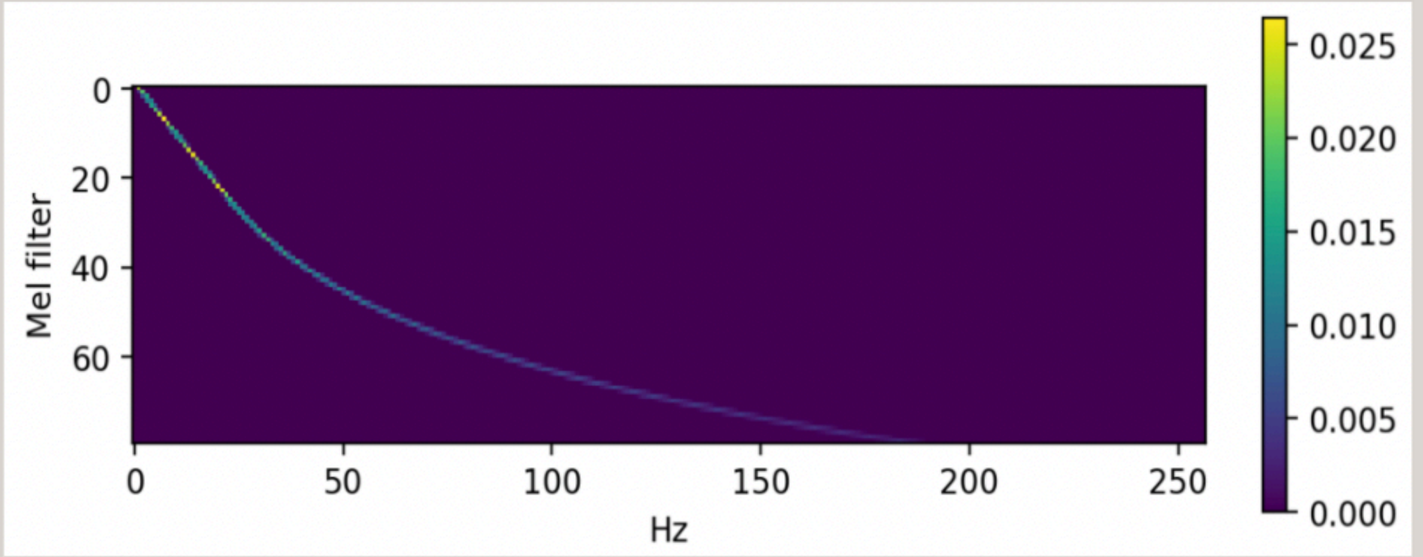
- **Window length** and **shape**
- **Hop length**: the length of overlap between windows



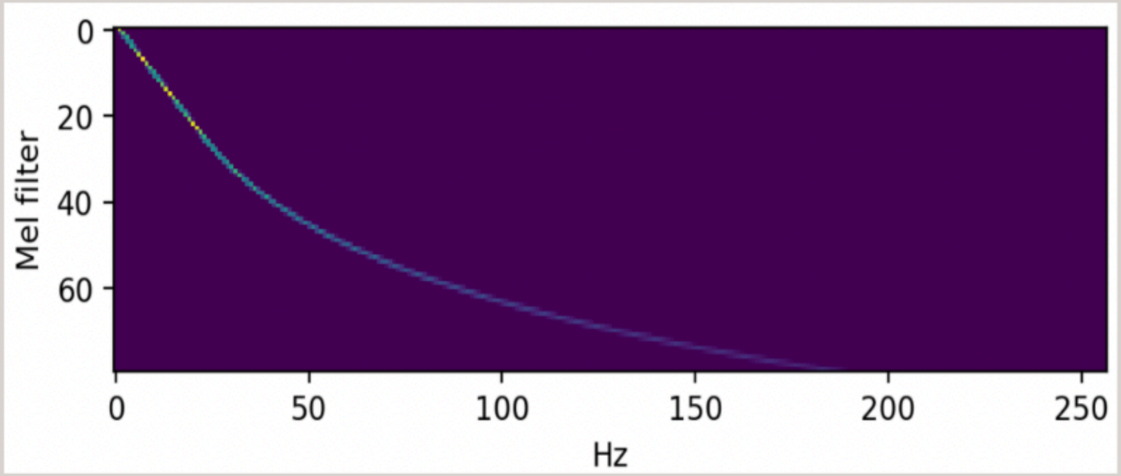
Mel filter bank

- Mel spectrum is produced by linear projection with Mel filter bank **F**.
- Given a spectrogram **S**, the Mel-Spectrogram **M** is produced as **M = F S**
- Typically, a logarithm is applied to the result in the end.





Mel-spectrogram

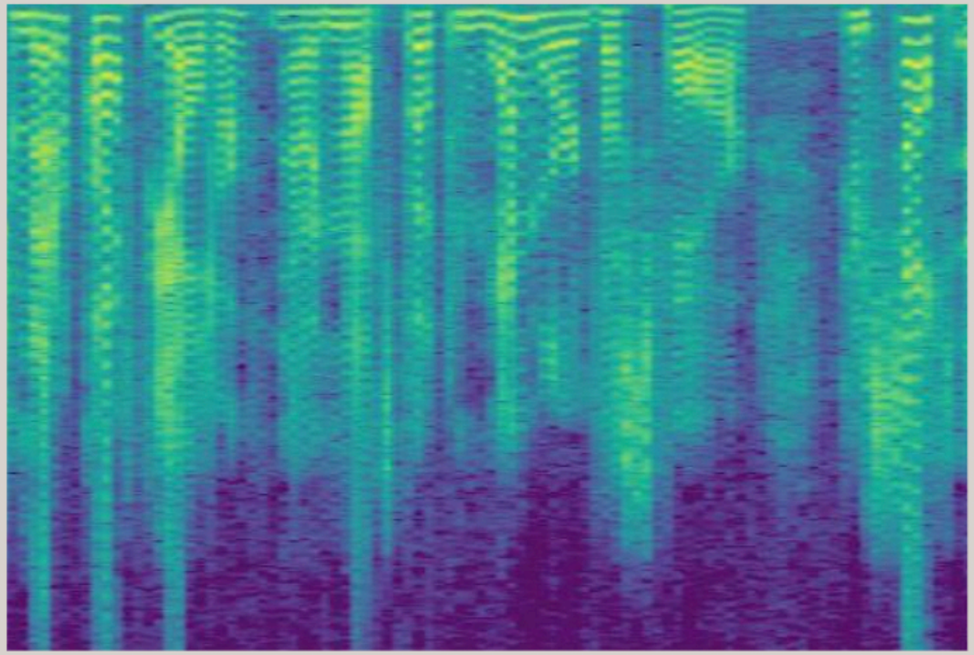


Filterbank F

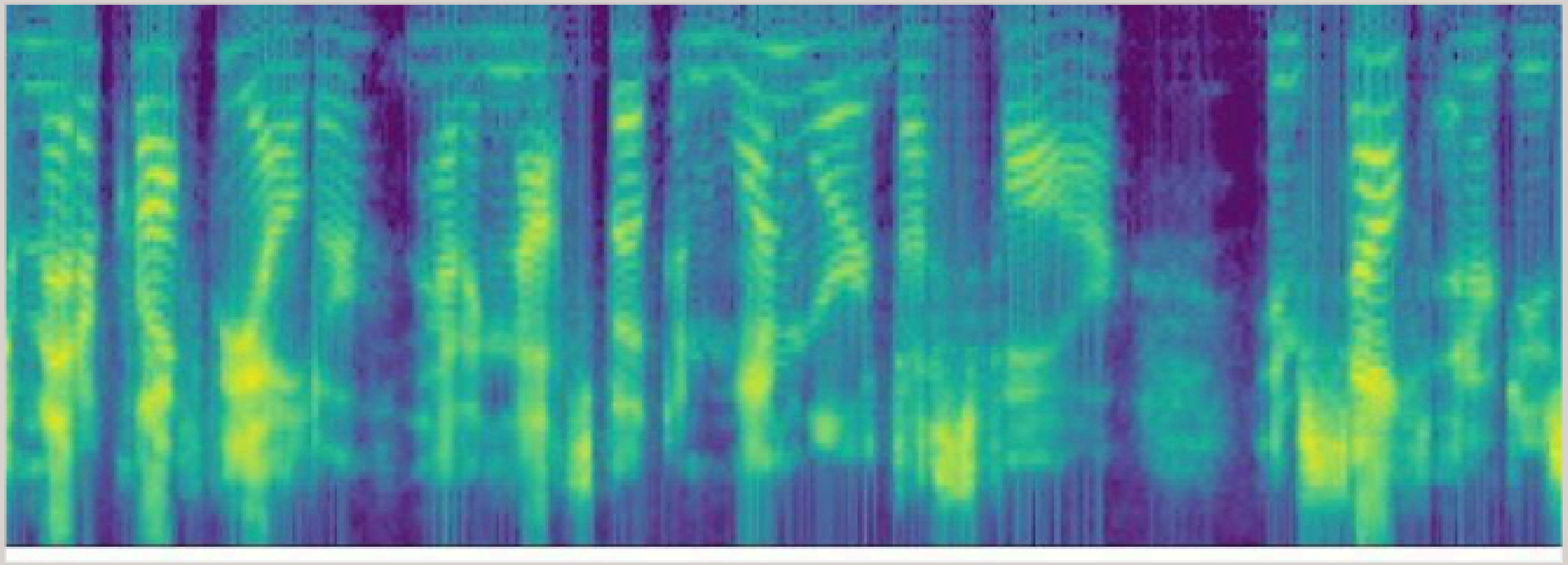
×

0

8000



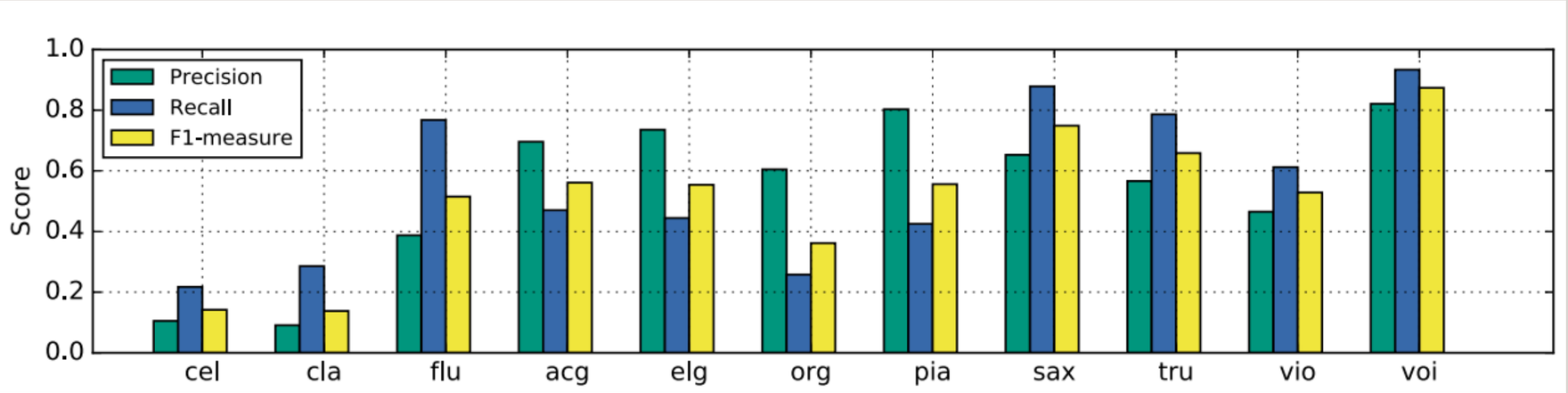
Spectrogram S



Mel-Spectrogram = F * S

Problems that we faced

1. Poor description of proposed model and inability to repeate it
2. Dataset inconsistency: cello has less data than other instruments

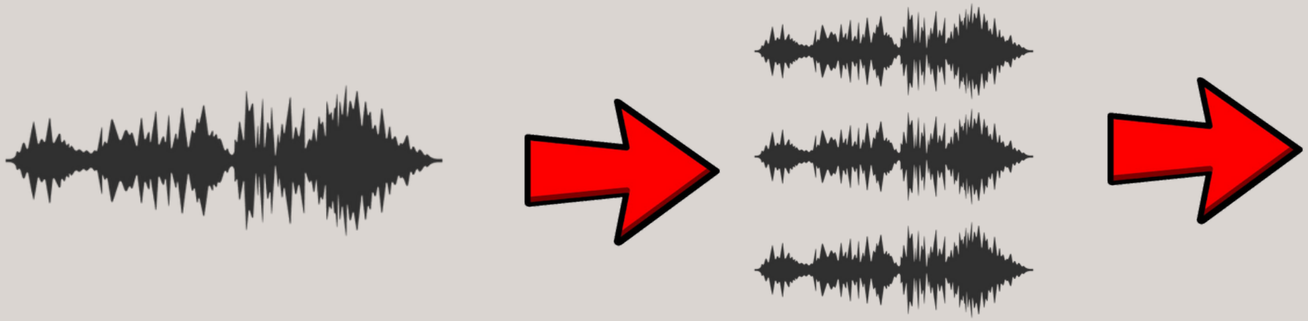


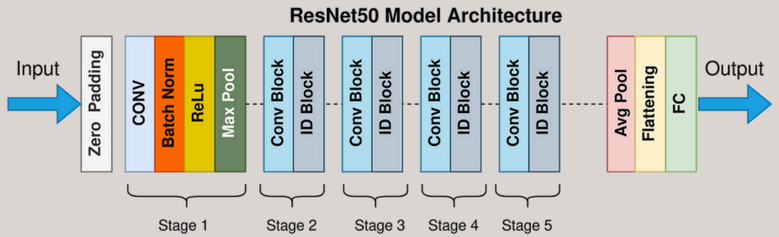
How we solved problem

precisely speaking first problem

We took other way

Interpret audio spectrogram as “image”



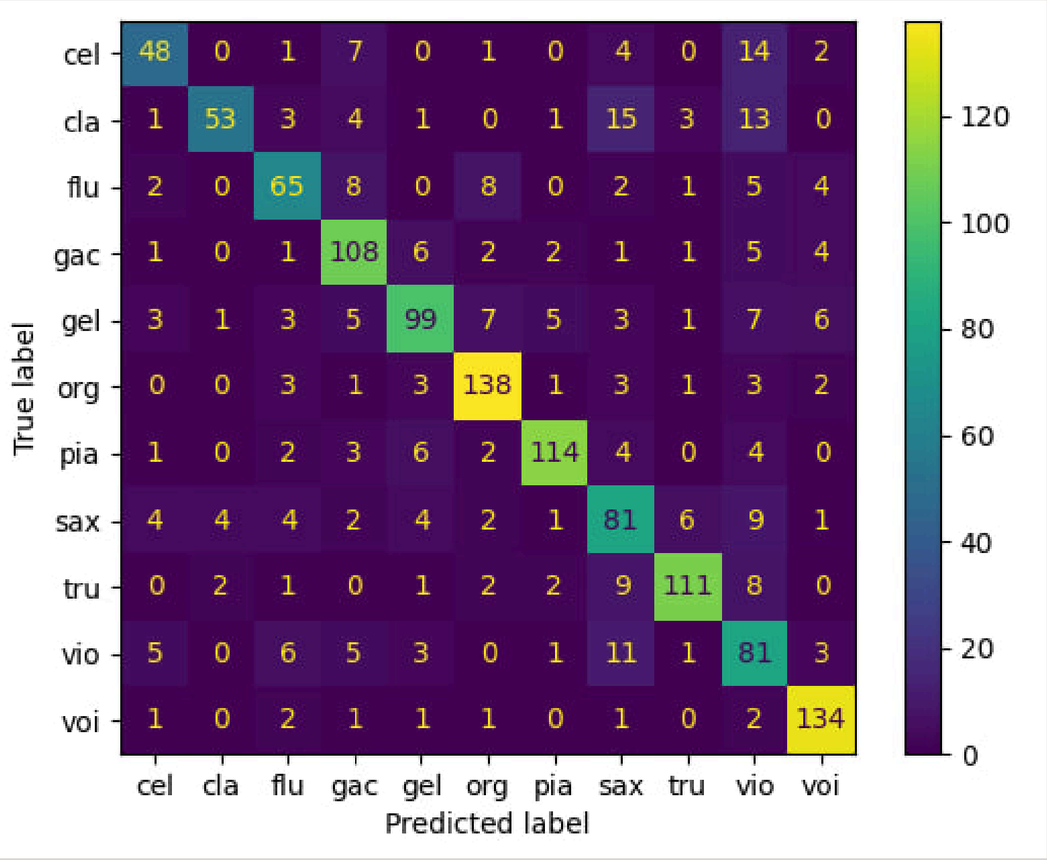


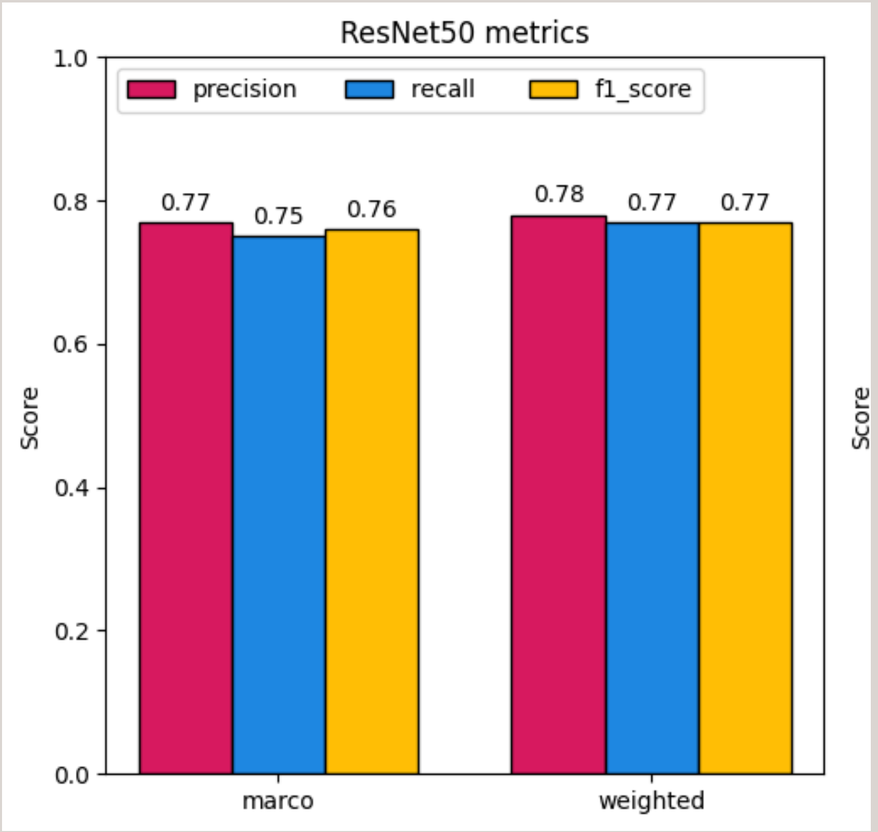
Model parameters

1. **Epochs** - 20
2. **Batch size** - 32
3. **Learning rate** - 0.0001
4. **Optimizer** - Adam
5. **Size**:
    a. Training size - 70%
    b. Validation - 10%
    c. Test - 20%

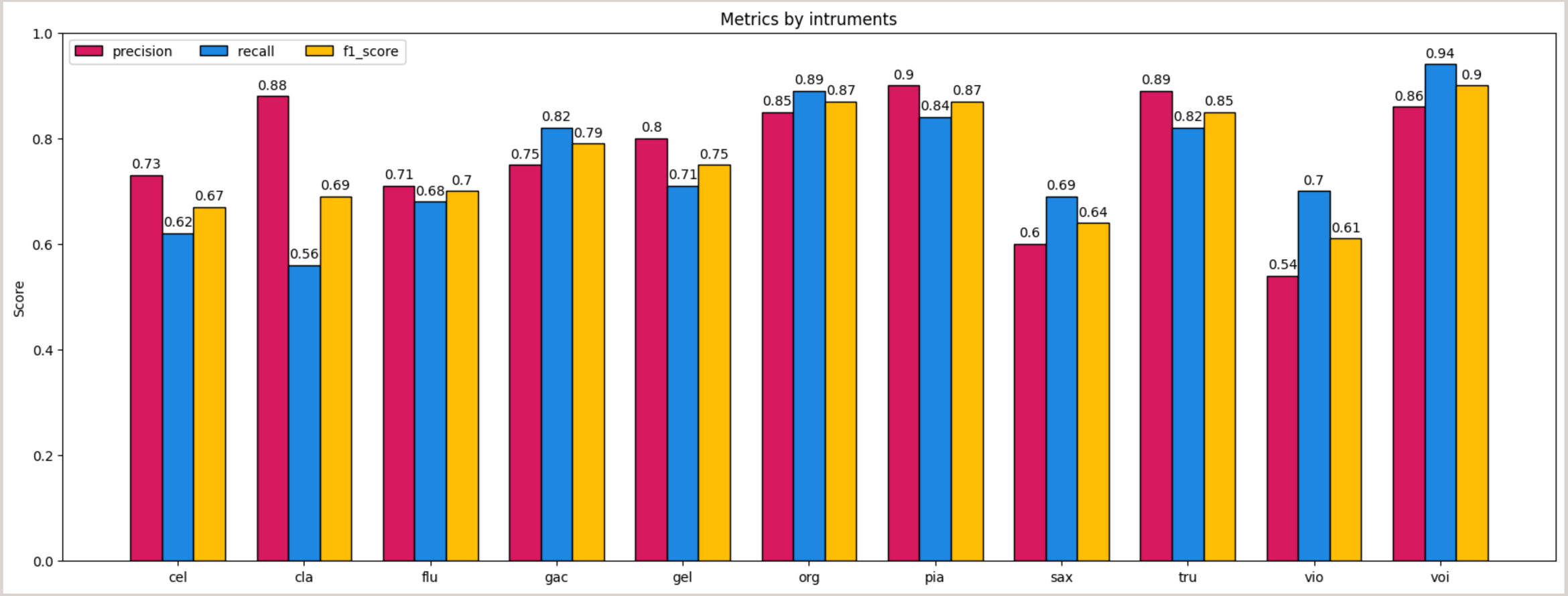
Resnet-50 only on IMRAS metrics instrument

Accuracy: 0.77



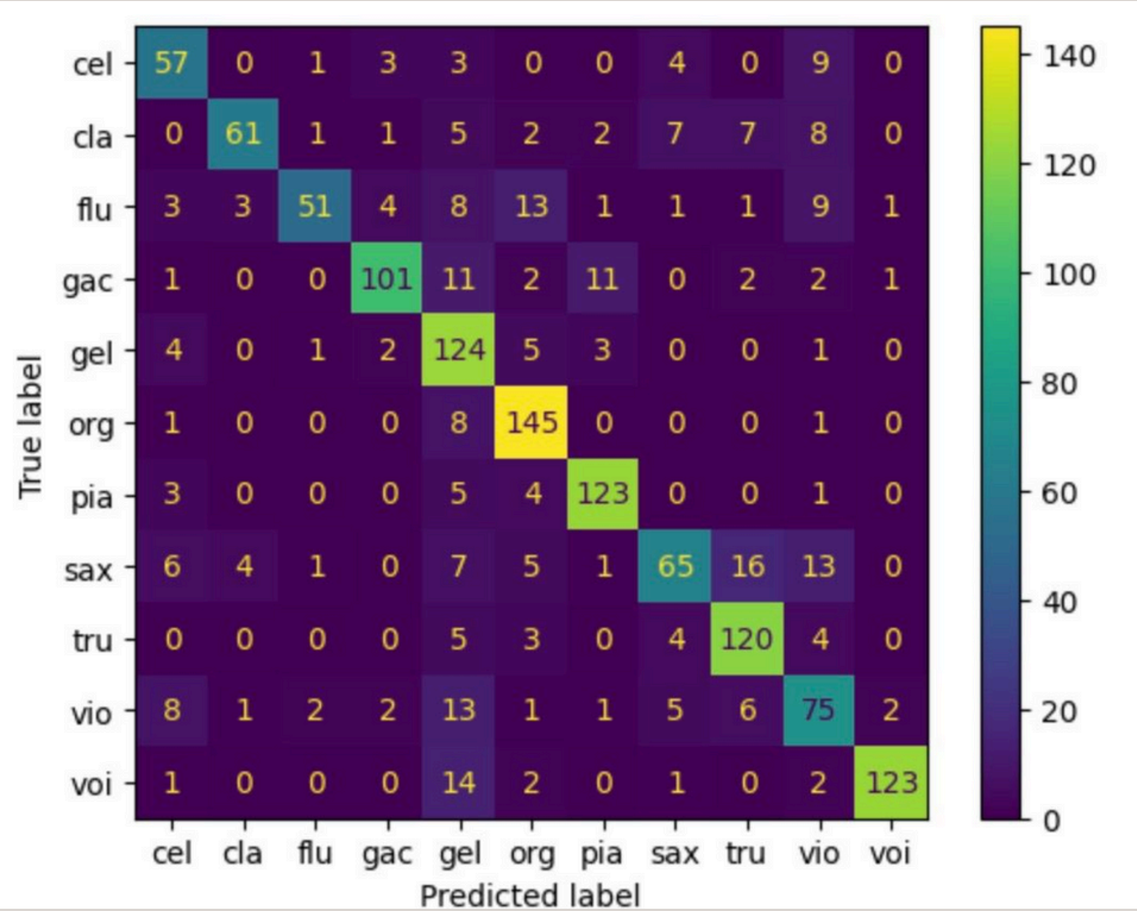


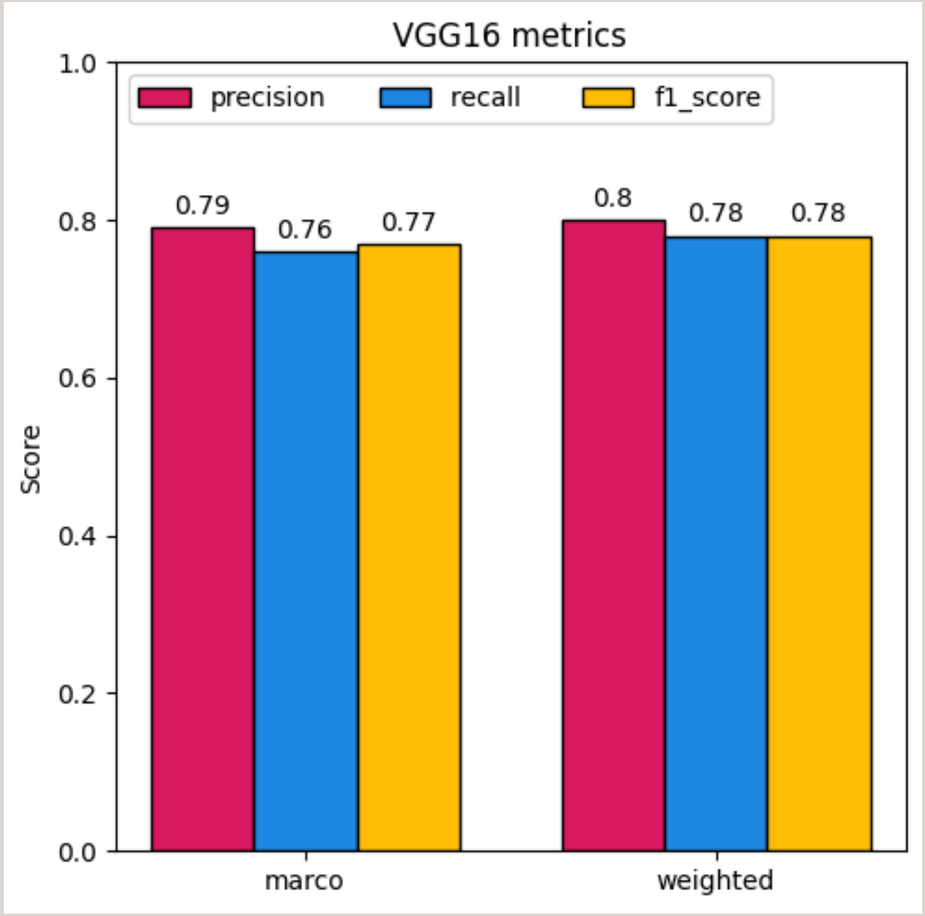
Resnet-50 only on IMRAS metrics instrument



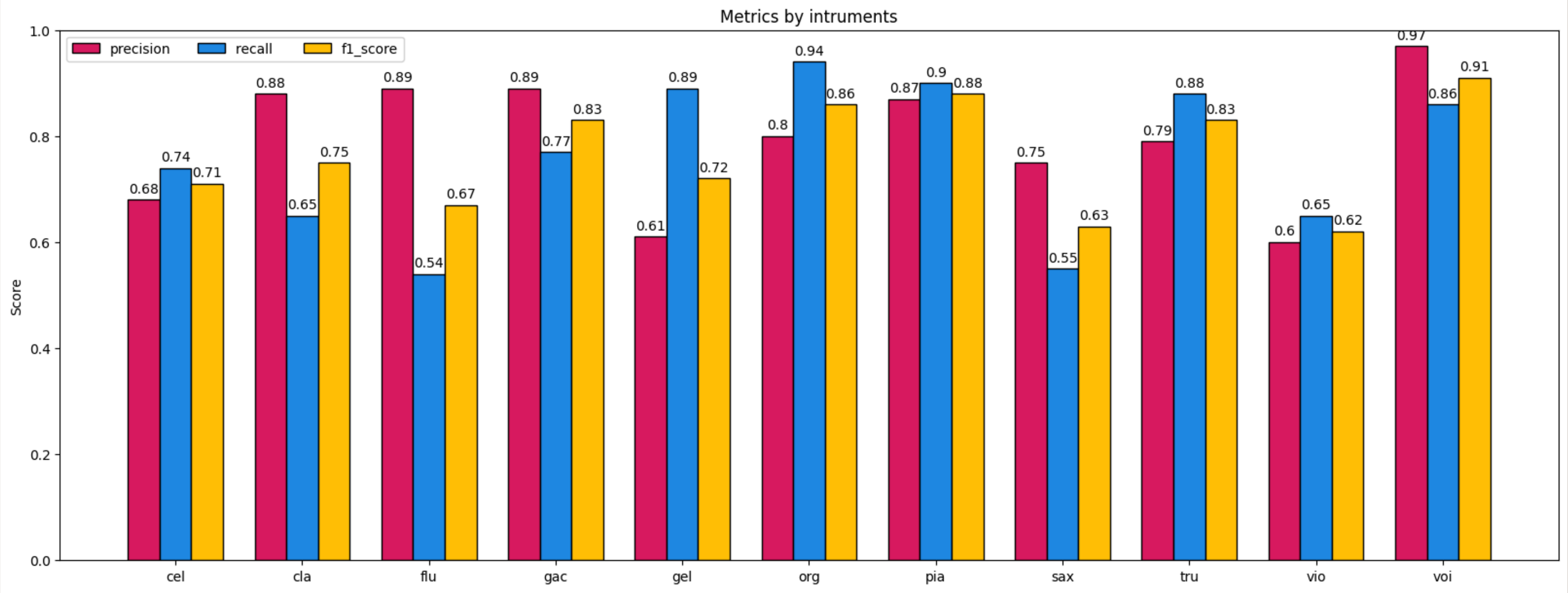
VGG-16 only on IMRAS metrics instrument

Accuracy: 0.78





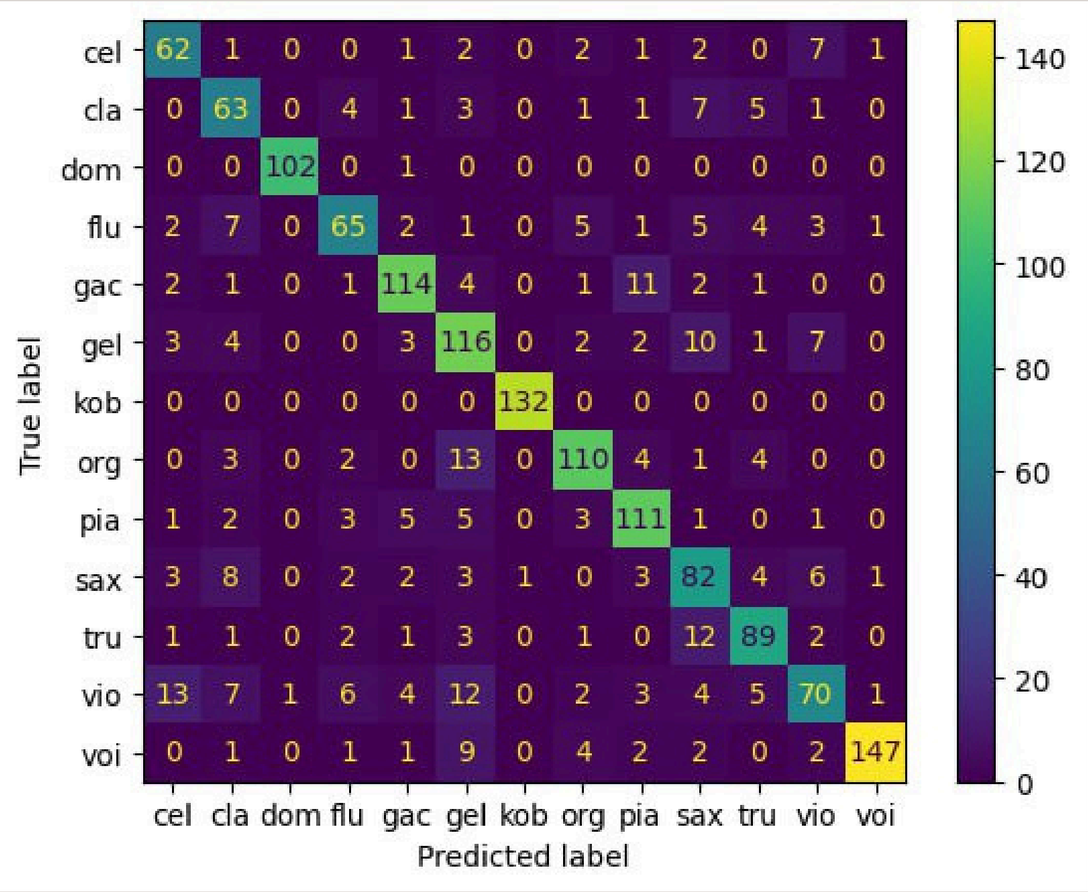
VGG16 only on IMRAS metrics instrument

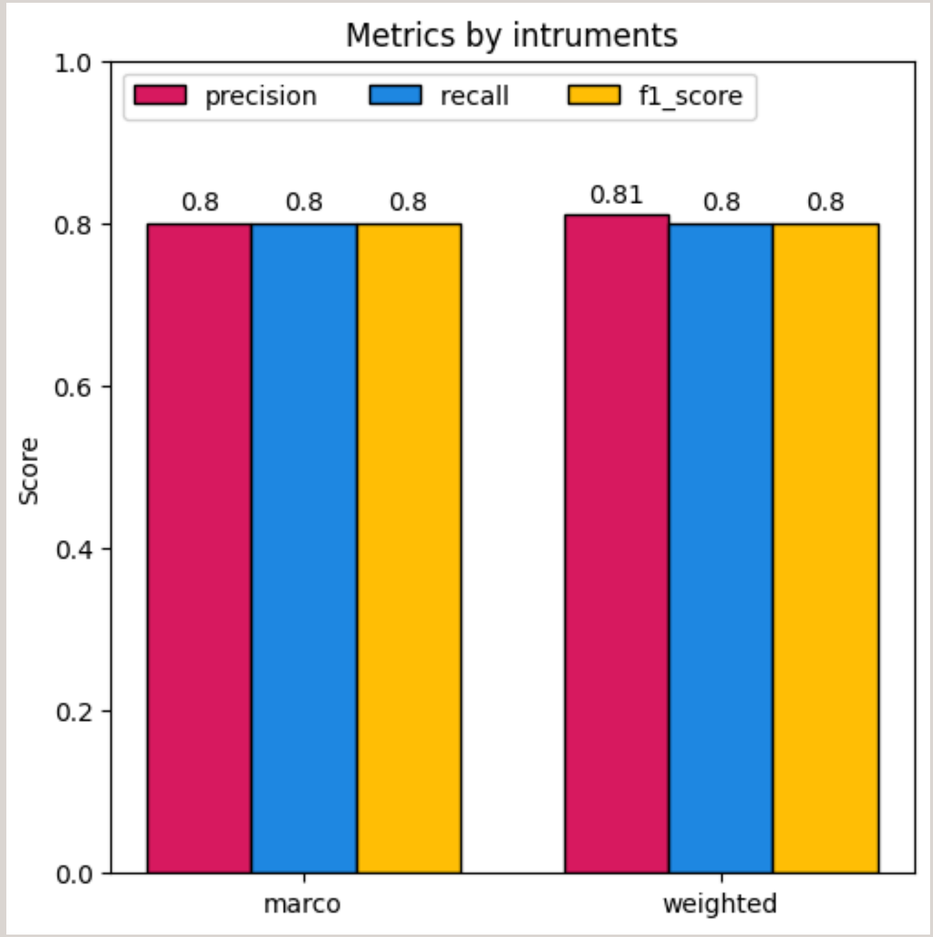


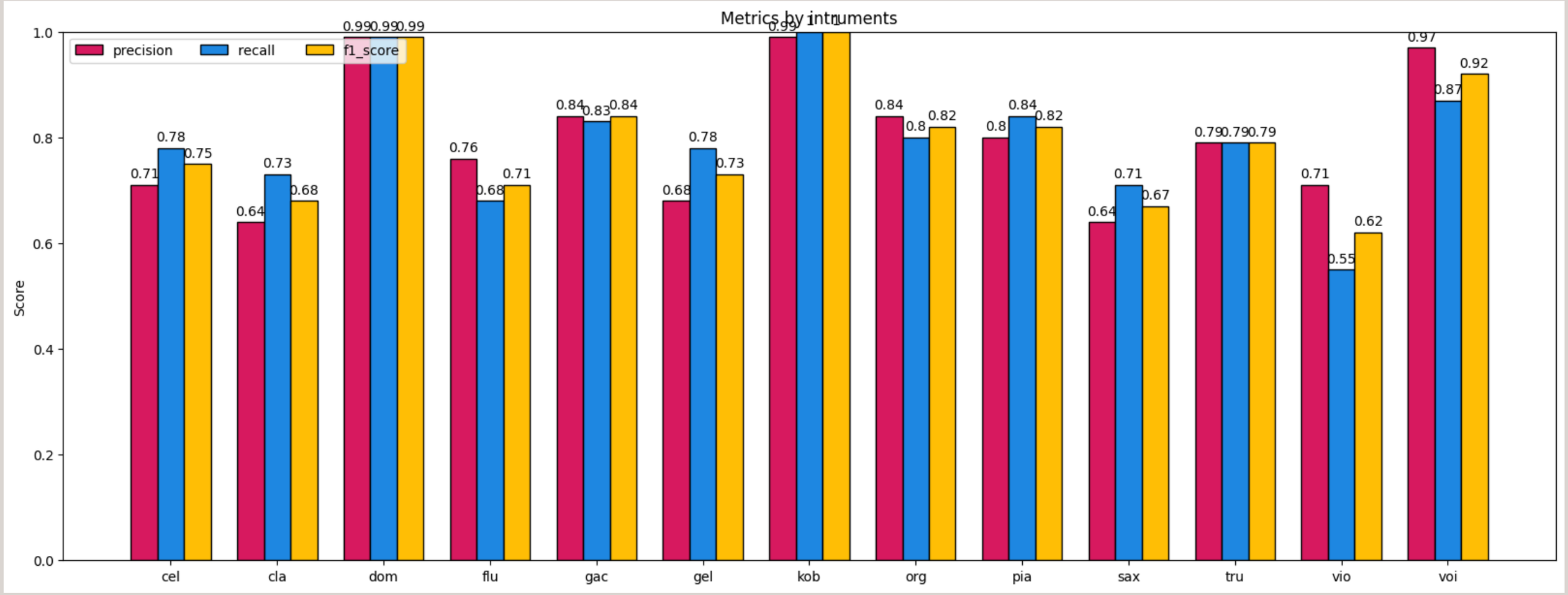
Resnet-50 on IMRAS metrics instrument with dombra and kobyz

We changed number epochs on this model ***from 20 to 30***

Accuracy: 0.802






Metrics by instruments
precision
recall
f1_score
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.71
0.78
0.75
0.64
0.73
0.68
0.99
0.99
0.99
0.76
0.68
0.71
0.84
0.83
0.84
0.68
0.78
0.73
0.99
1
1
0.84
0.8
0.82
0.8
0.84
0.82
0.64
0.71
0.67
0.79
0.79
0.79
0.71
0.55
0.62
0.97
0.87
0.92
cel
cla
dom
flu
gac
gel
kob
org
pia
sax
tru
vio
voi

Future steps

1. Add data to classes where there is little data - cello, flute, clarniet
2. Make dombyra and kobyz dataset more balanced
3. Add multilable classification