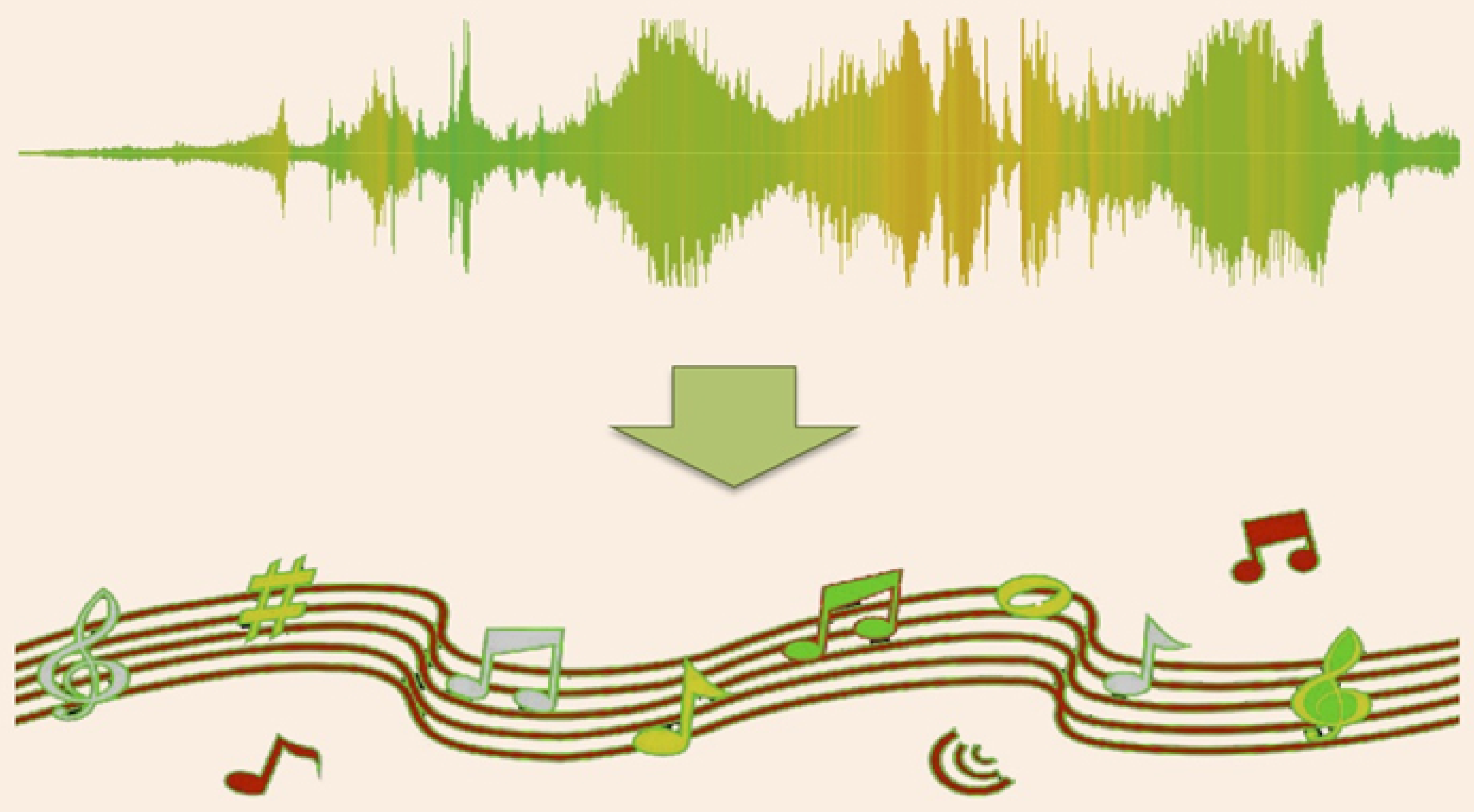
A Computational Model for Automatic Music Transcription

Zhengshan Shi, Tony Yang, Huijie Yu

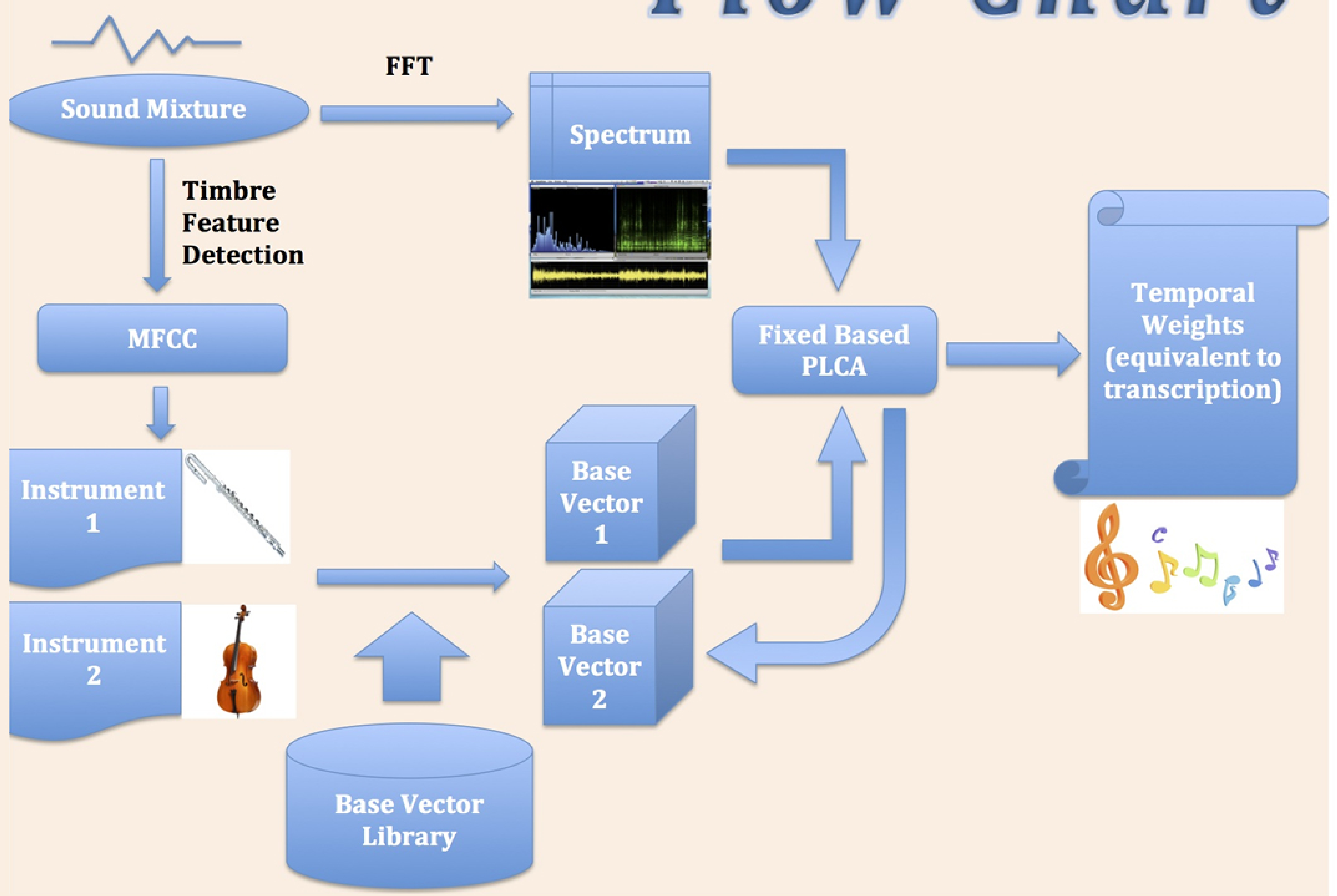


Introduction

The aim of our project is to build a computational model for an automatic music transcription system. Given a sound mixture of a couple of monophonic instruments (for example, a flute-cello duet), our system is expected to generate the music transcriptions for each instrument voice respectively.

The whole project is divided into two parts. First, we implement the instrument classification. We use k-nearest neighboring and logistic regression in this part. Second, based on the result from the instrument classification, we will create basis vectors for the corresponding instrument we predict and then implement the source separation. The algorithm we used in this part is PLCA.

Flow Chart



Sound Mixture

FFT

Spectrum

Timbre Feature Detection

MFCC

Instrument 1

Instrument 2

Base Vector 1

Base Vector 2

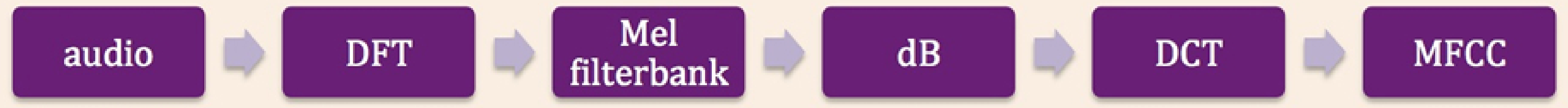
Base Vector Library

Fixed Based PLCA

Temporal Weights (equivalent to transcription)

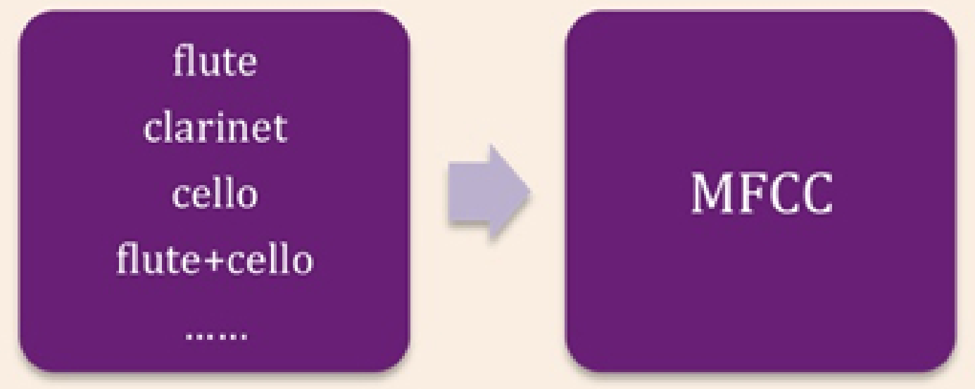
INSTRUMENT CLASSIFICATION

- Feature Extraction: MFCC



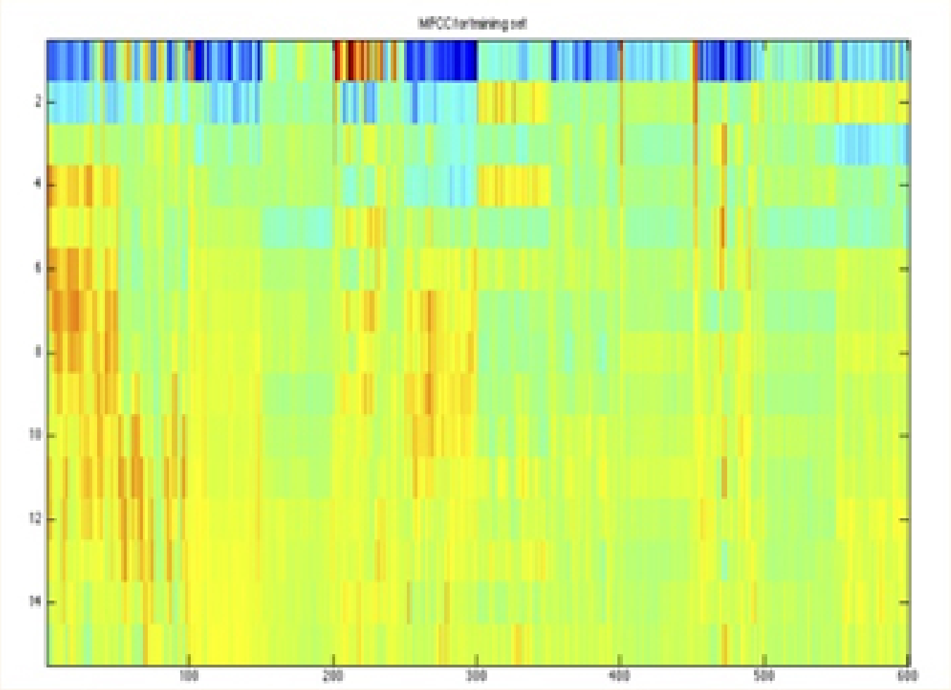
audio → DFT → Mel filterbank → dB → DCT → MFCC

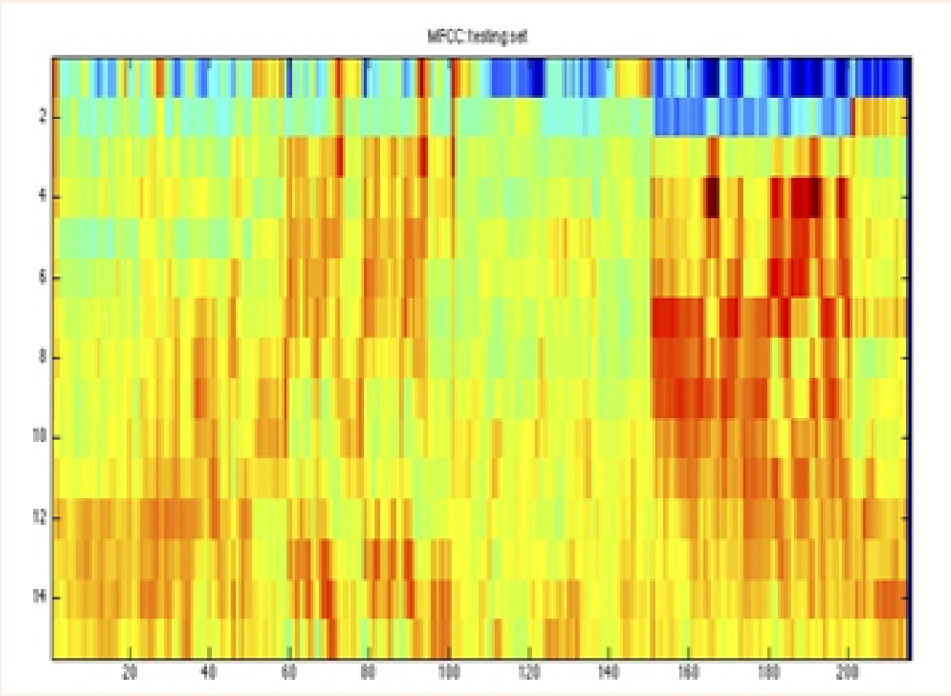
- Training & Testing



flute
clarinet
cello
flute+cello
……

MFCC





- Classification Algorithm
  - ❖ K-nearest Neighboring
  - ❖ Logistic Regression

- Result

Test Case I:

Hit Ratio	flute	clarinet	trombone	cello	piano	violin
K-NN	0.98	0.70	0.96	1.00	0.80	1.00
Softmax	0.86	0.86	1.00	1.00	0.66	0.96

Test Case II:

Hit Ratio	flute	clarinet	cello	flute+cello	clarinet+piano
K-NN	0.96	0.70	1.00	0.98	0.56
Softmax	0.86	0.86	1.00	1.00	0.58

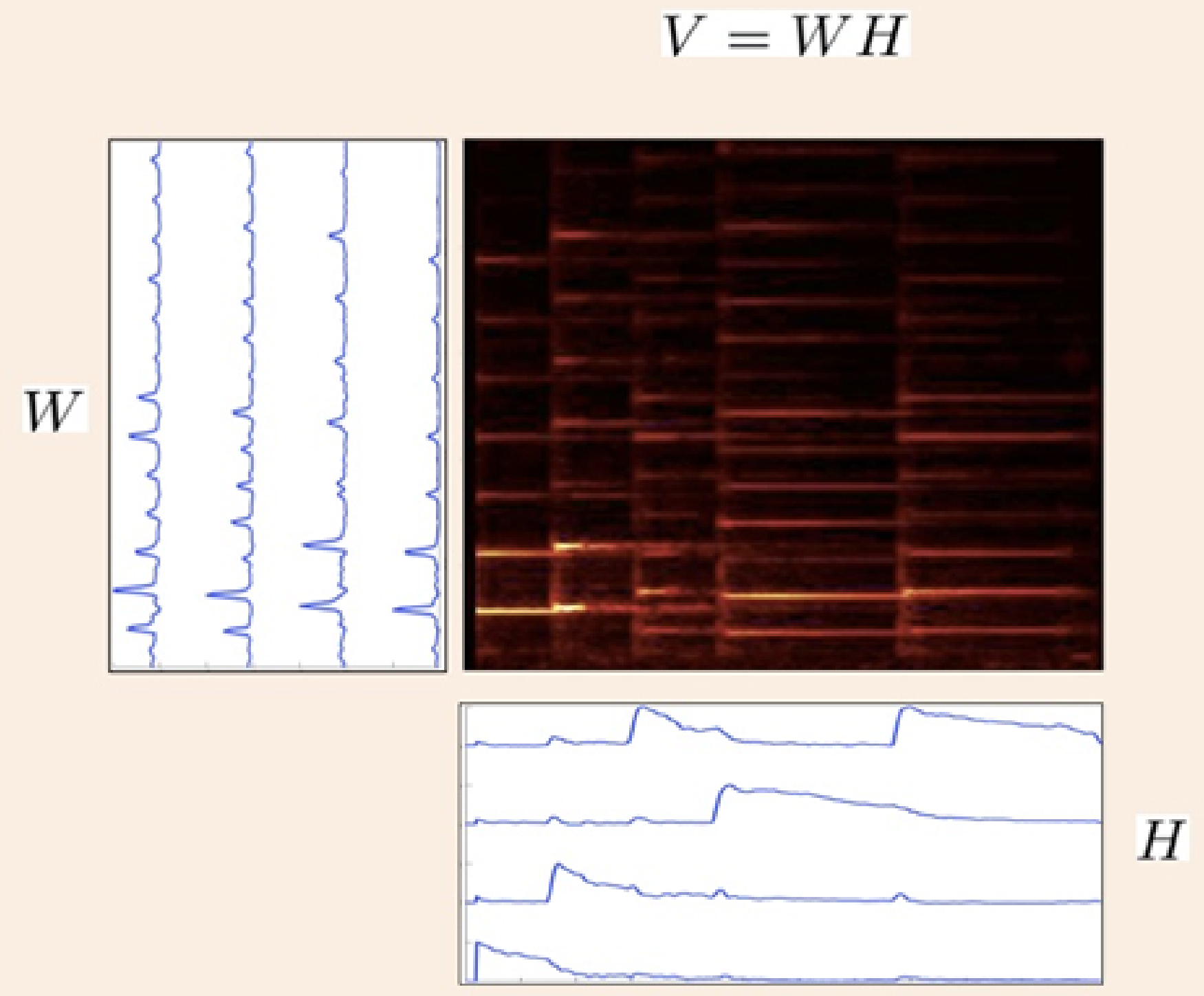
PLCA: Probablistic Latent Component Analysis

$$\mathbf{E-Step:}\quad Pt(z|f) = \frac{Pt(f|z)P(z)}{\sum_z Pt(f|z)P(z)}$$

$$\mathbf{M-Step:}\quad Pt(z) = \frac{\sum_f \mathrm{Pt(f)Pt(z|f)}}{\sum_z \sum_f Pt(f)P(z|f)}$$

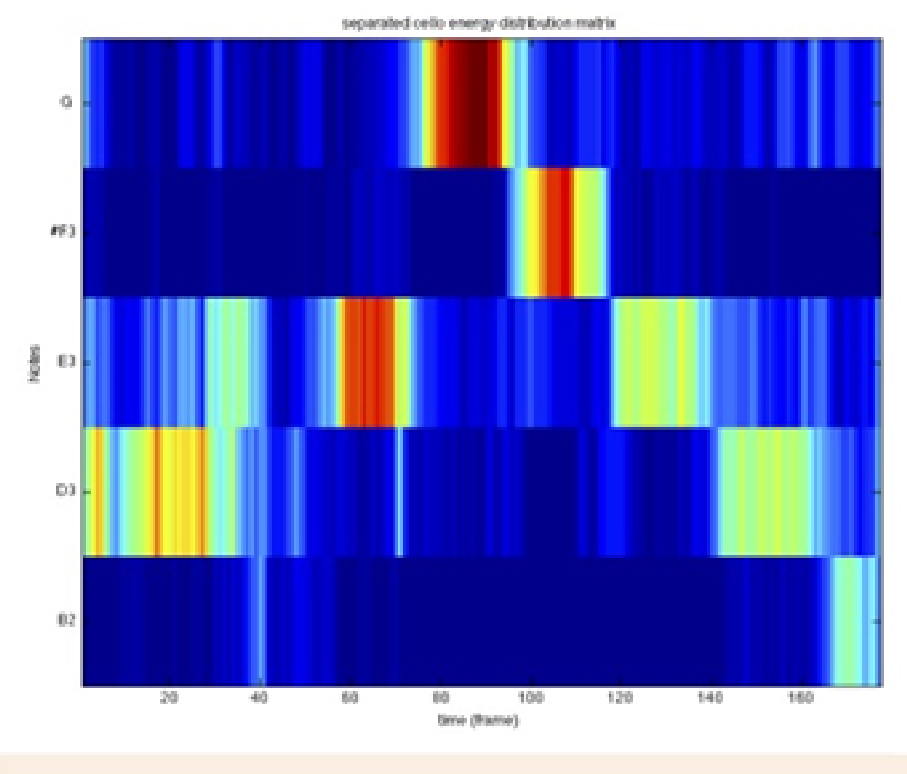
$$P(f|z) = \frac{\sum_t \mathrm{Pt(f)Pt(z|f)}}{\sum_t \sum_f Pt(f)Pt(z|f)}$$

- Decompose the audio spectrum (v) into a spectral basis matrix (w) and a temporal weight matrix (H).
- Set the initial guess of the spectral bases as single notes played by a particular instrument
- Update the temporal weight matrix and the spectral basis matrix using the EM algorithm.

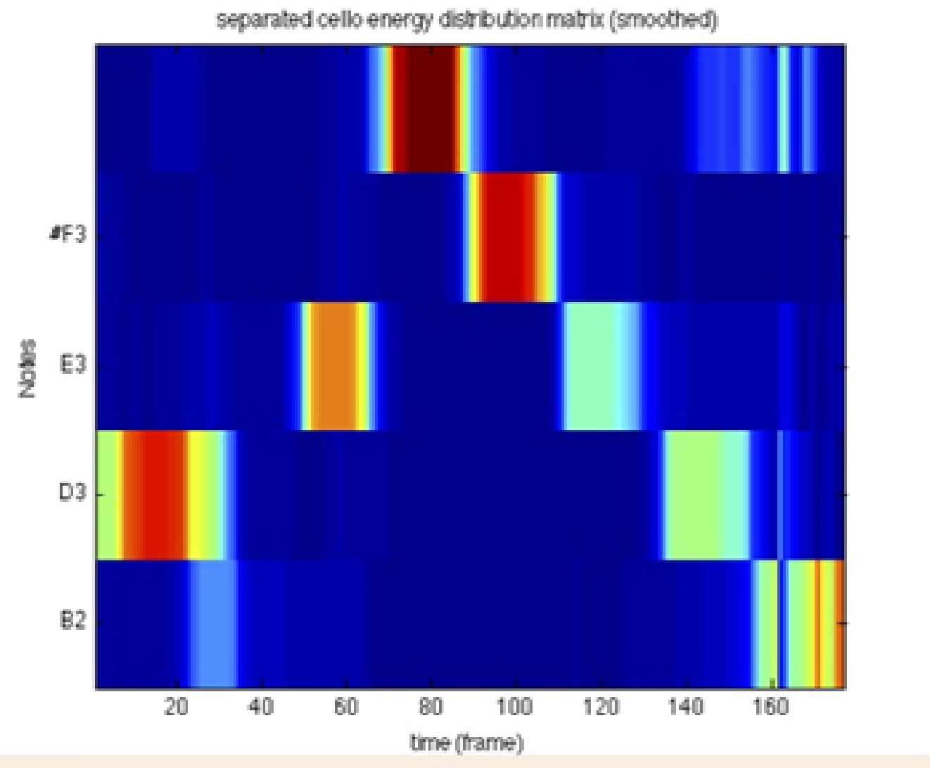


Evaluation

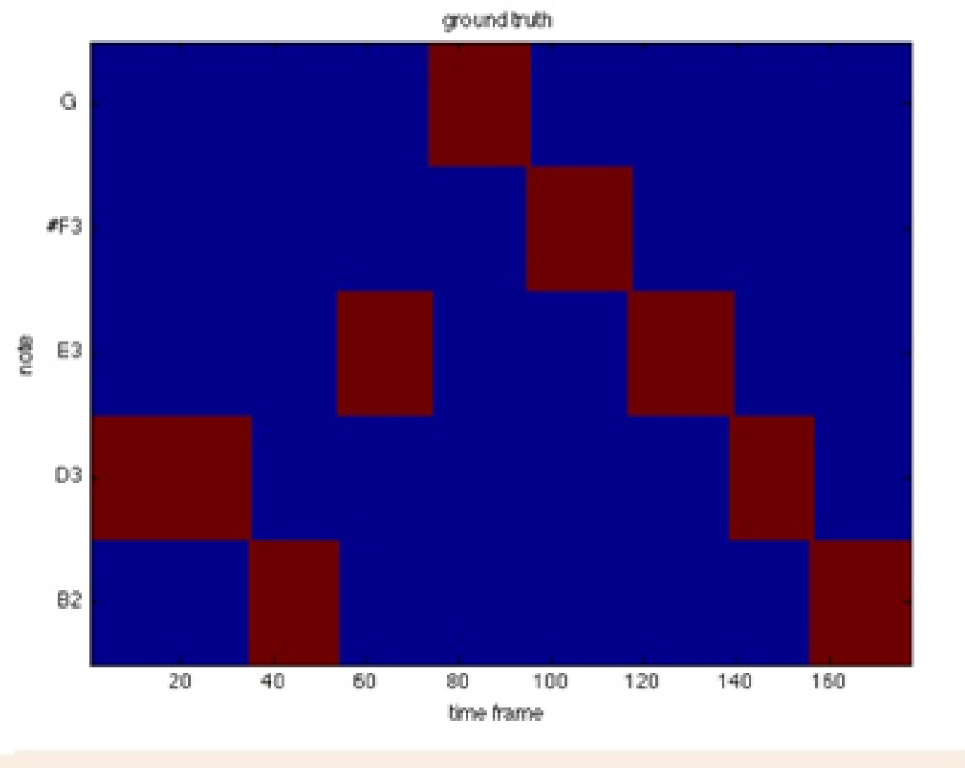
Audio source: J.S. Bach Suite en si mineur – Polonaise et Double 0-4.2s



Separated cello notes



Smoothed cello notes



Ground truth

- Separated temporal matrix into two parts, each for one instrument (cello and flute).
- Smoothed the results by moving average filter.
- Compared with musician-annotated ground truth and found 87% accuracy.