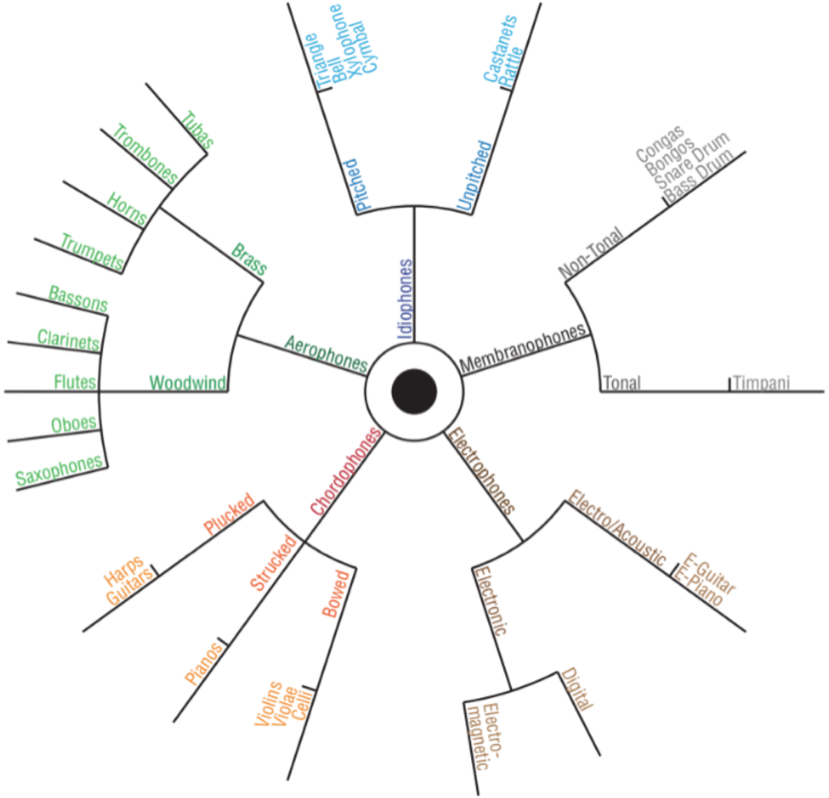
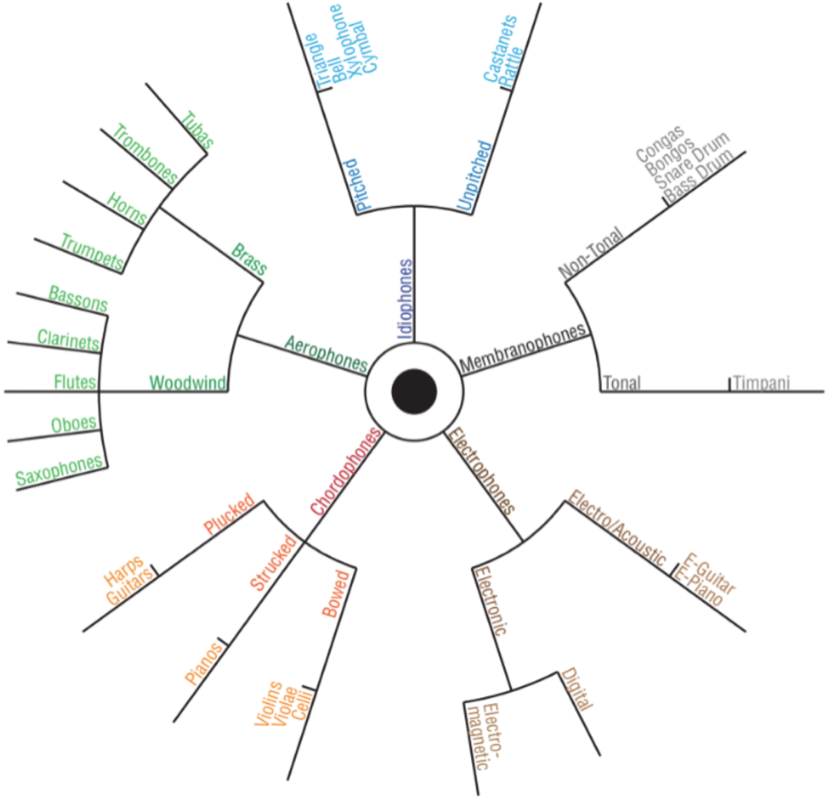
Musical Instrument Taxonomy Classification

Aidan Johnson and Deniz Alpay



How can we distinguish these instruments by ear?



How can we distinguish these instruments by ear?

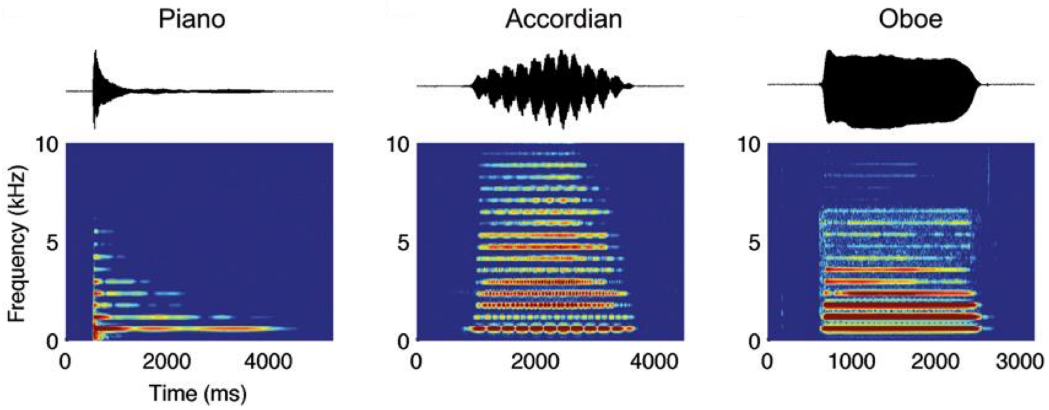
Timbre

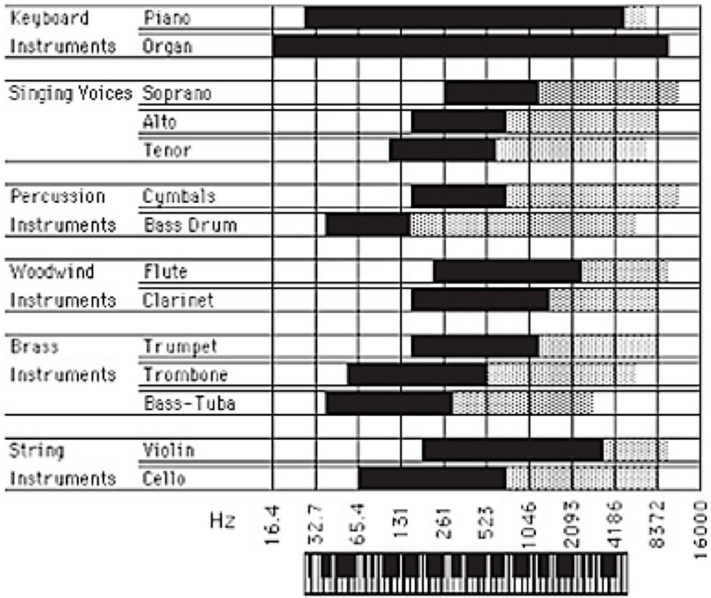
What is Timbre?

- Humans can perceive difference in the sound produced by different instruments at same pitch and loudness
- In music, this distinguishing characteristic is referred to as timbre—the distinguish quality/characteristic of a sound
- Measurable definition of timbre: uncertain
- Timbre is multidimensional but abstract
  - Spectral (MFCCs)
  - Temporal (ADSR envelope)

Introduction

Goal: Classify instruments in audio of a single instrument.





Feature Extraction

- Mel-frequency cepstral coefficients (MFCCs)
- Spectral shape statistics

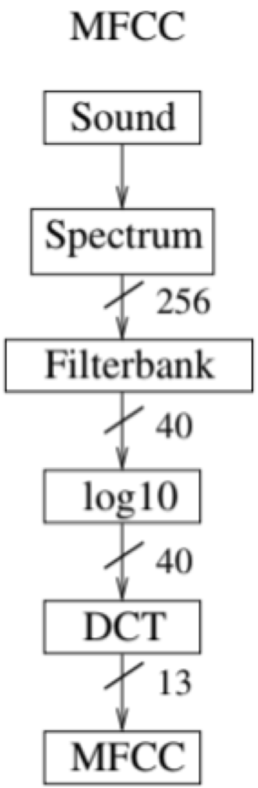
$$\mu_i = \frac{\sum_{n=1}^{N} f_k^i * a_k}{\sum_{n=1}^{N} a_k}$$ *Raw moments*

$$centroid = \mu_1$$ *Mean*

$$spread = \sqrt{\mu_2 - \mu_1^2}$$ *Variance*

$$skewness = \frac{2\mu_1^3 - 3\mu_1\mu_2 + \mu_3}{spread^3}$$ *Asymmetry (at mean)*

$$kurtosis = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{spread^4} - 3$$ *Flatness (at mean)*



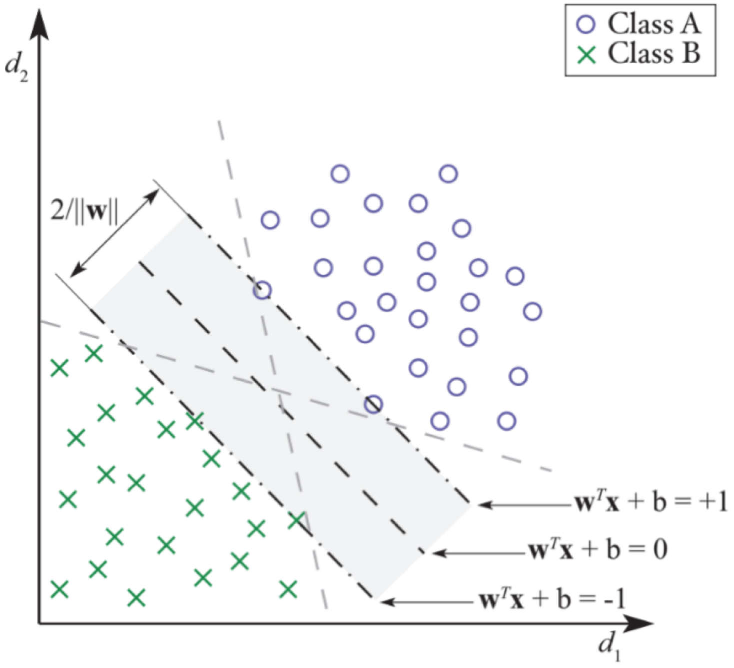
Data set

- IRMAS: https://www.upf.edu/web/mtg/irmas
  - 6 instruments
  - Labeled by primary instrument present and genre
  - 6705 training files, 2874 testing files
- Data decisions
  - Only used audio with a single instrument
  - Manually cleaned some of the testing data
  - Narrowed instruments to cello, saxophone, and violin
  - 395 training files, 75 testing files

Support Vector Machine (SVM)

Training:

- min $(½)||\mathbf{w}||^2$ such that:
  - $\mathbf{y}_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) \geq 0$ for all points
- N total training points
- Point $\mathbf{x}_i$
  - D-dimensions
  - Has binary class $\mathbf{y}_i$ (either +1 or -1)
- w: normal vector of hyperplane
- b: offset
- b/||**w**||: perpendicular distance hyperplane to origin



Classification: SVM

- Support vector machine with radial basis function kernel

	Cello	Saxophone	Violin
Accuracy	76.2%	82.2%	63.7%
Type I error (false positives)	11.2%	17.8%	3.36%
Type II error (false negatives)	12.6	0%	33.0%

Classification: SVM

- Support vector machine with radial basis function kernel

	Cello	Saxophone	Violin
Misclassifications	23.8%	17.8%	36.3%
Type I error (false positives)	11.2%	17.8%	3.36%
Type II error (false negatives)	12.6	0%	33.0%
Number of support vectors	25,469	20,799	31,416

- One vs all

Gaussian Mixture Model (GMM)

x: d-dimensional random vector

M: no. mixture components

c_m: component weight

$b_m(x)$: Gaussian density function with mean mu

$$p(x|\lambda) = \sum_{m=1}^{M} c_m b_m(x)$$

$$b_m(x) = \frac{1}{2\pi^{D/2}|\Sigma^{1/2}|} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$$

Classification: GMM

- GMM advantages:
  - Parametric
  - Interpretable
  - Computationally cheaper

Conclusion

- SVM:
  - Computationally expensive for complex data and more classes, but is much more effective than GMM
  - Classifies with reasonable and comparable—with respect to the literature—performance
- Broader implication of studying/quantifying timbre:
  - Human perception of sound (music, speech, etc.) using spectral and temporal acoustic features
  - Neural representation processing in human brain (auditory cortex)