Musical Genre Classification using Melodic Features and Source Separated Audio from Polyphonic Signals

Dhaivat Shah
ds3267@columbia.edu

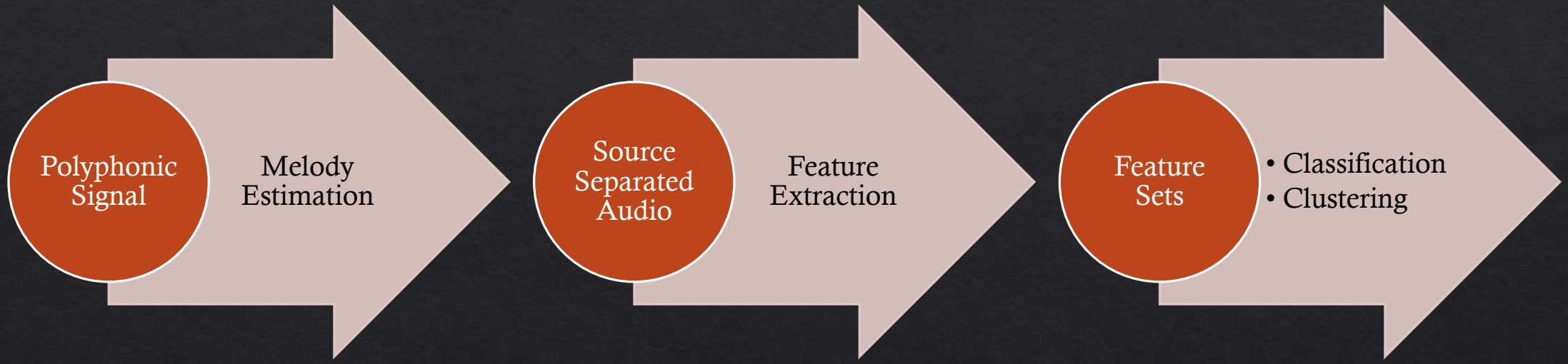
Introduction

- **Musical Genre**
  - The term used to describe music that has similar properties, in those aspects of music that differ from the others
- **Melody**
  - The single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music.
- **Pitch**
  - Perceptual property that allows the ordering of sounds on a frequency-related scale.
- **Polyphany**
  - Texture consisting of two or more simultaneous lines of independent melody, as opposed to music with just one voice (monophony) or music with one dominant melodic voice accompanied by chords (homophony).
- **Source Separation**
  - Estimation of main melody from a polyphonic signal.

Problem Definition

- Given an audio signal, classify it into previously known categories.
- Given many audio signals, cluster them into similar genres.
- **Applications**
  - Casual Users: Listening to similar kinds of music
  - Academic/Professional users: Finding related sequences to analyze and study
  - Music industry: Efficient retrieval and storage
- **Issues**
  - While this comes naturally to people, it is quite difficult programmatically, due to polyphony.
- **Why not let people do it?**
  - With the volume generated, it is simply impossible.

Approach

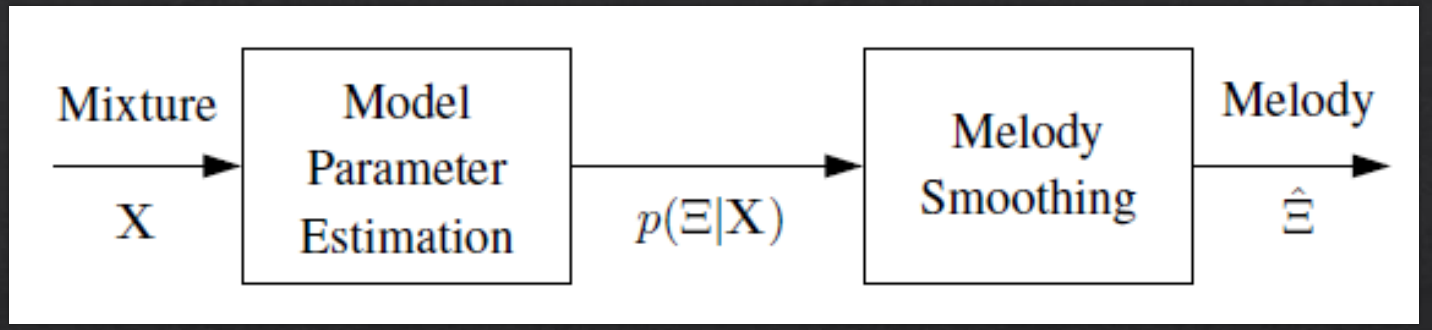


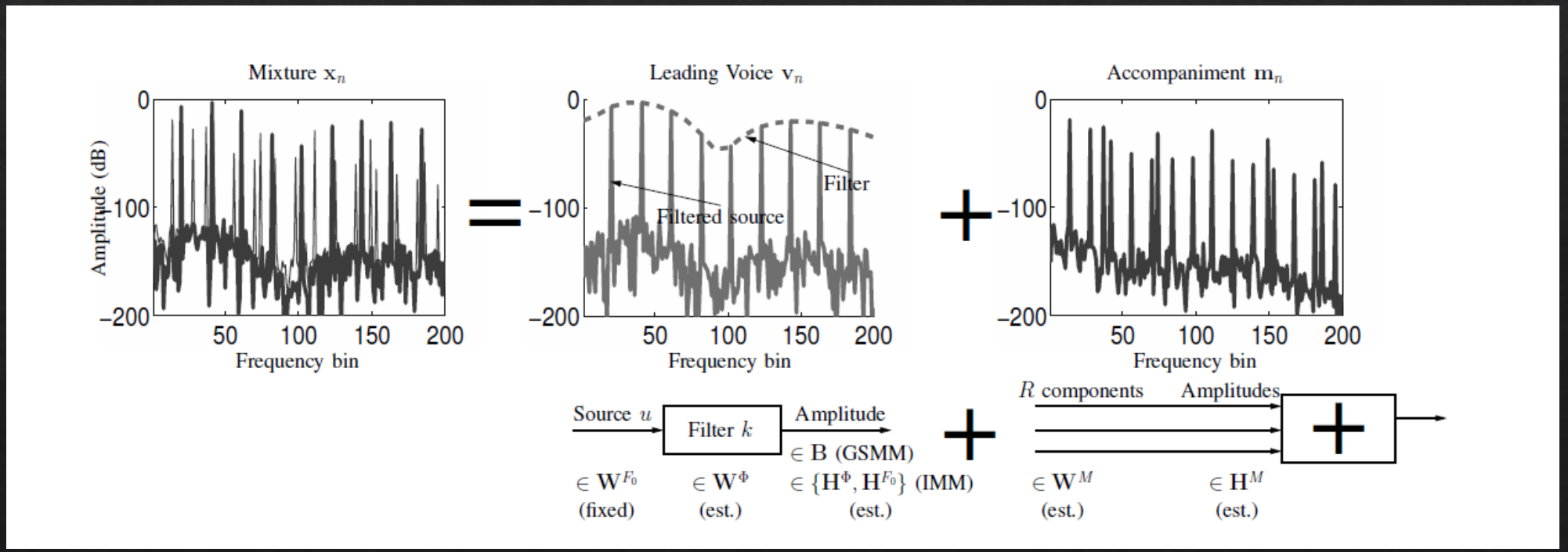
- Melody Estimation
- Feature Extraction
- Classification

State of the Art

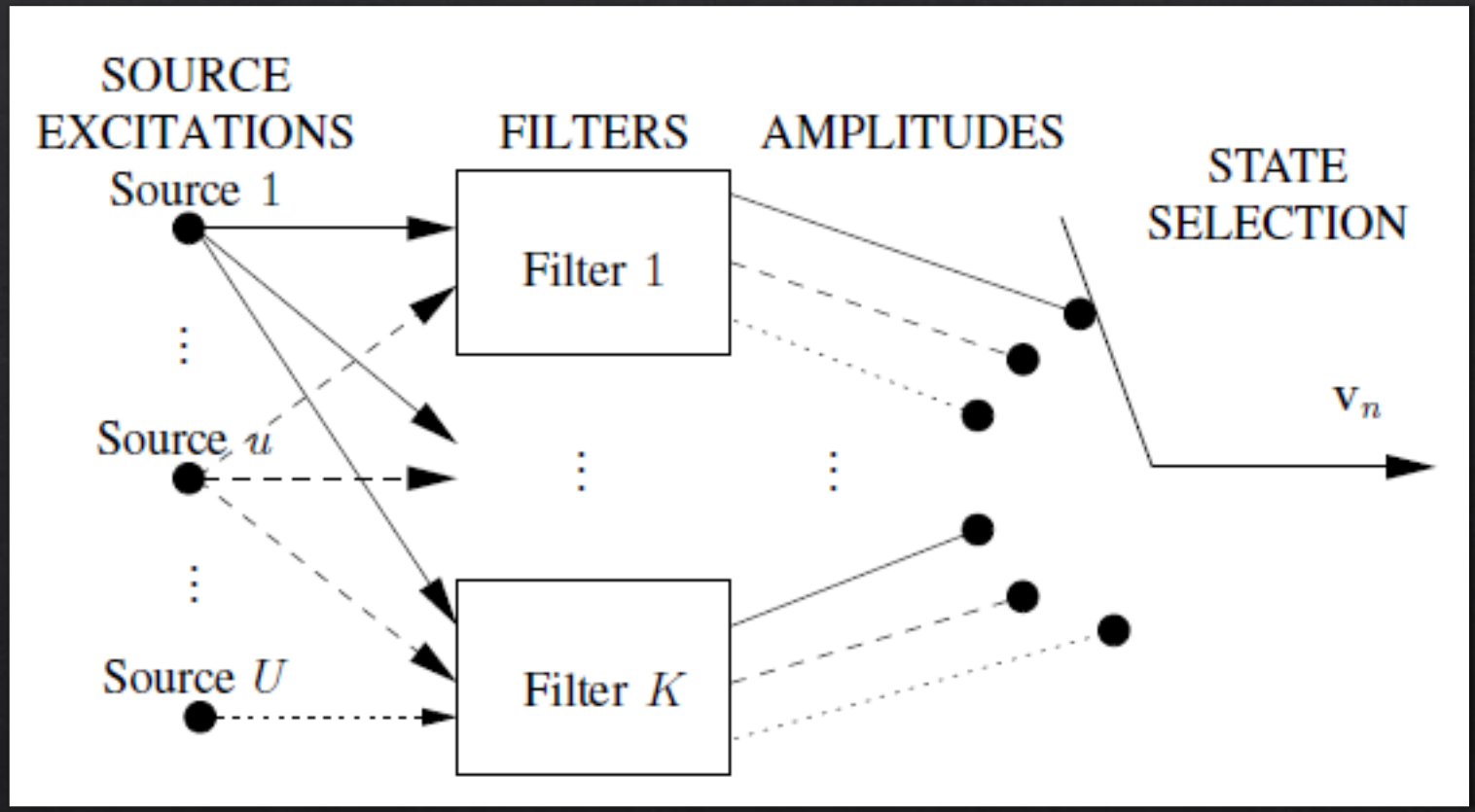
- Proposed Techniques:
  - Temporal
  - Energy
  - Spectral Shape
  - Perceptual features
- Slow Temporal modulations of music recordings and the sparse representation based classifiers.
  - Accuracy: 91%
  - Previous best: 82.5%
- Pitch Based
  - 50-60%
  - Synthetic Dataset: 90-95%

Melody Extraction Process





Source/Filter Model

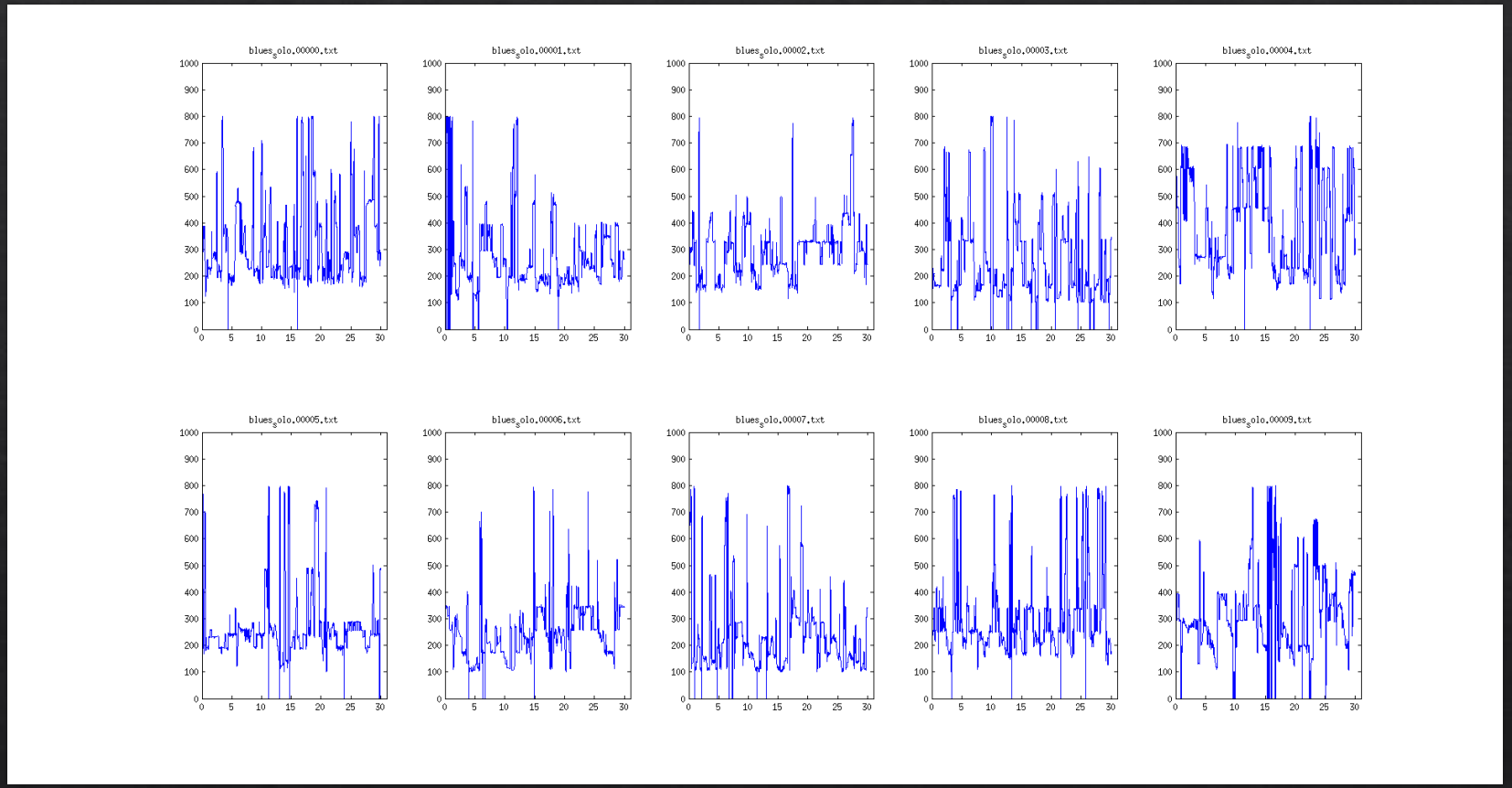


Source Separation

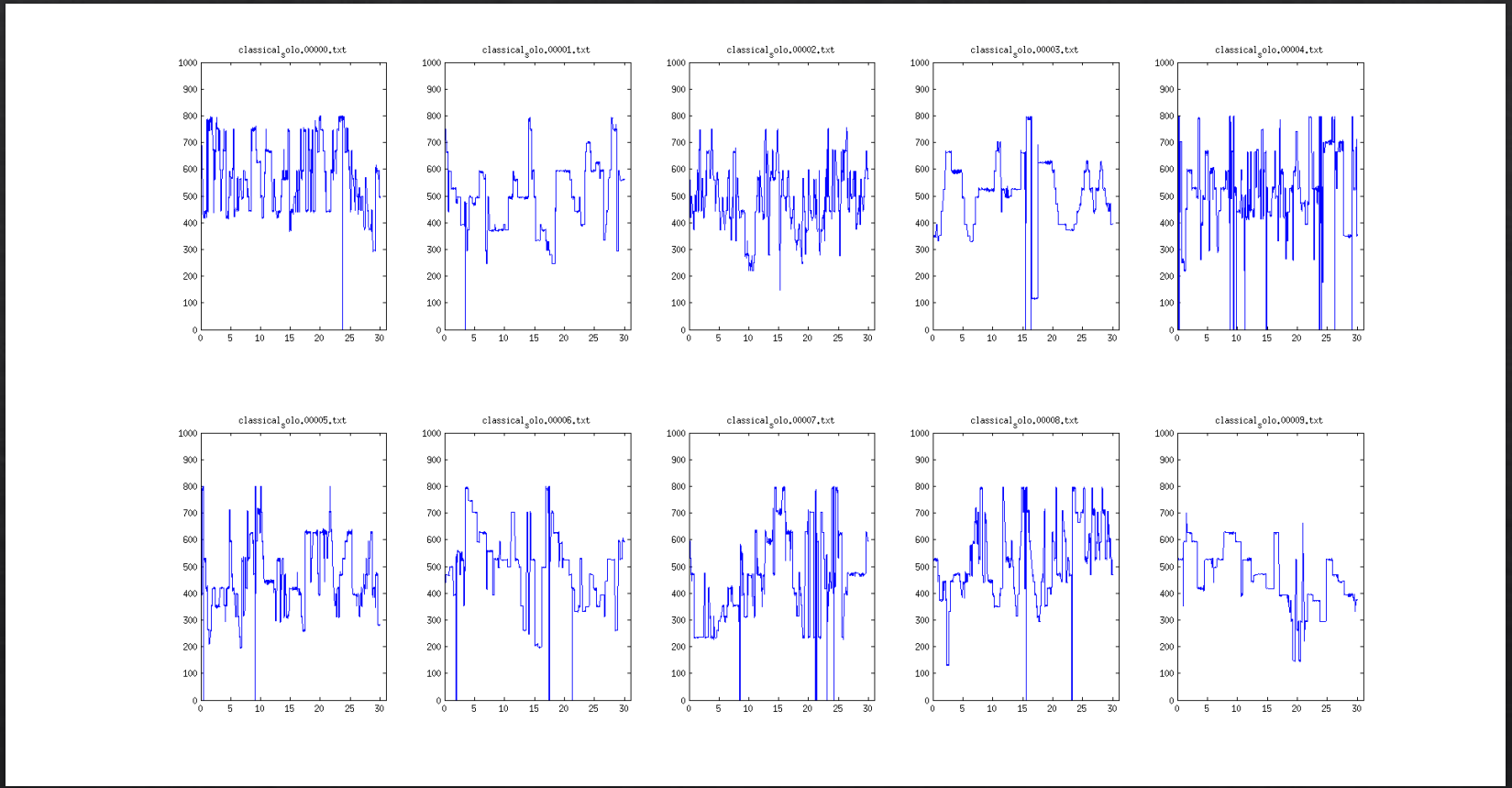
Original

Solo

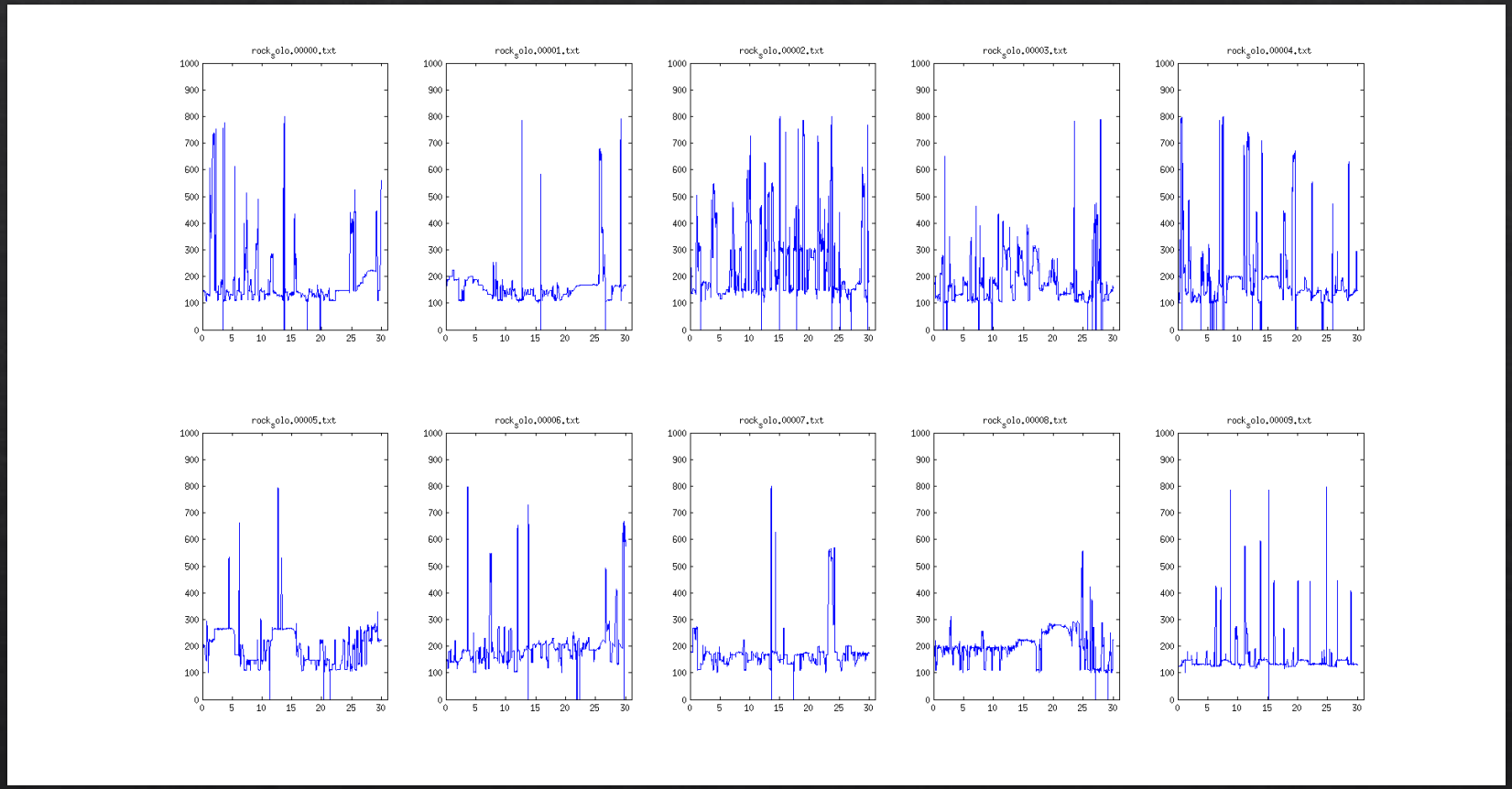
Pitch (F0) - blues



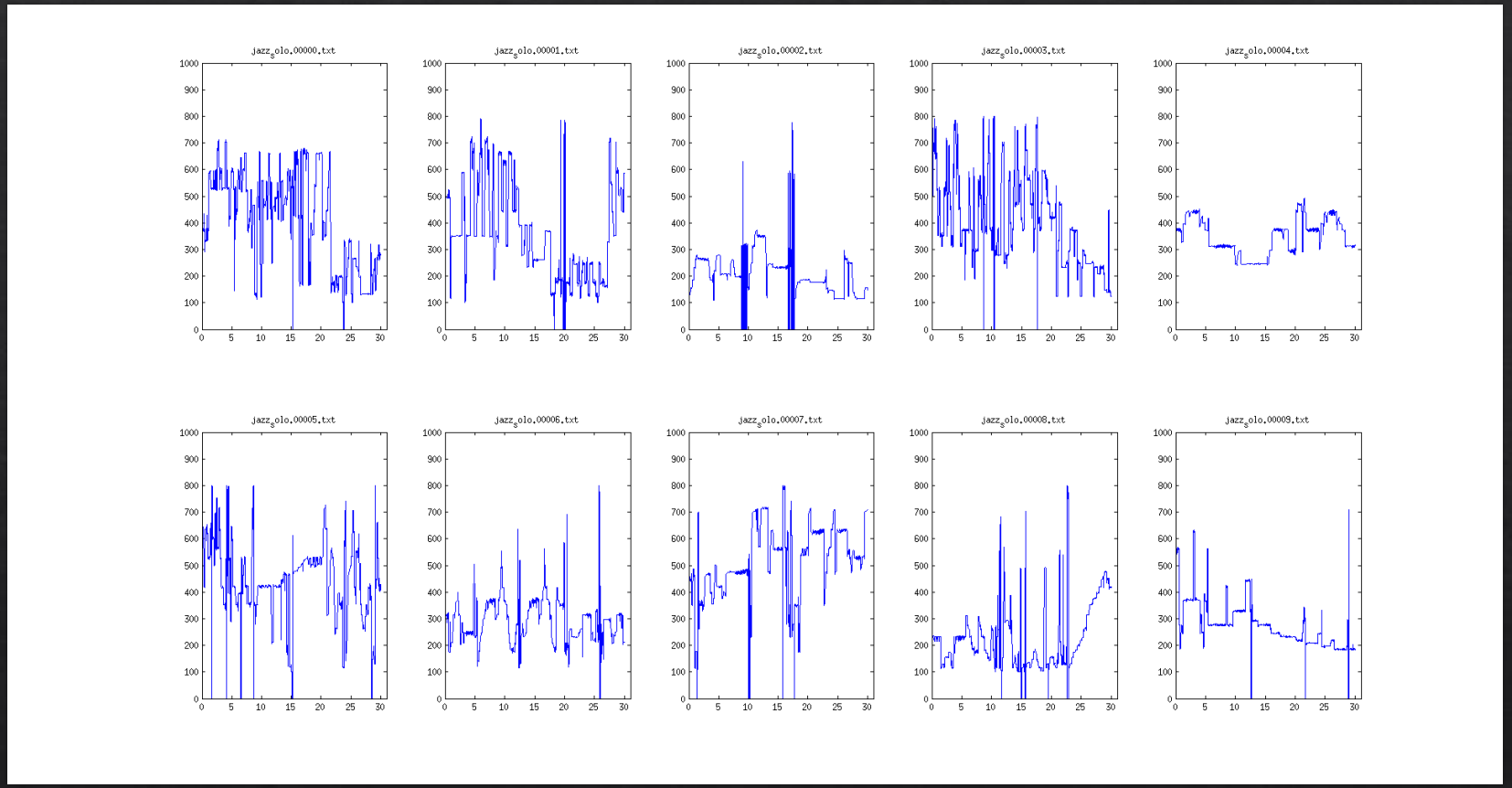
Pitch (F0) - classical



Pitch (F0) - rock



Pitch (F0) - Jazz



Feature Set

- **Pitch Based**
  - Mean
  - Variance
  - Max
  - Min
  - Skewness
  - Kurtosis
- **Vibration**
  - Mean
  - Variance
- **Others**
- **MFCC**
- **Combined**

Features

- **Skewness**
  - Skewness is a measure of the asymmetry of the data around the sample $s = \frac{E(x-\mu)^3}{\sigma^3}$
- **Kurtosis**
  - Kurtosis is a measure of how outlier-prone a distribution is. $k = \frac{E(x-\mu)^4}{\sigma^4}$
- **MFCC**
  - Take the Fourier transform of (a windowed excerpt of) a signal.
  - Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
  - Take the logs of the powers at each of the mel frequencies.
  - Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
  - The MFCCs are the amplitudes of the resulting spectrum.

Classification

- K Nearest Neighbors (KNN) Classifier
  - An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors
  - Value of k: depends on number of genres.
- Gaussian Naive Bayes
- Support Vector Machine Classifier (SVC) - ‘rbf’ kernel
- Linear SVC - ‘linear’ kernel

Clustering

- **Technique:**
  - K Nearest Neighbors (KNN)
- **Scores:**
  - **Homogeneity**: each cluster contains only members of a single class.
  - **Completeness**: all members of a given class are assigned to the same cluster.
  - **V-measure**: Harmonic mean of the previous two measures.
  - **Adjusted Rand index**: Function that measures the similarity of the two assignments, ignoring permutations and with chance normalization.
  - **Adjusted Mutual Information:** is a function that measures the agreement of the two assignments, ignoring permutations.

Tools

- **Soxexam**
  - au to wav format conversion
- **separateLeadStereo**
  - Separate into solo, music parts – both GSMM and IMM
  - Estimate pitch values
- **python/matlab**
- **numpy/scipy**
- **Scikit Learn**
  - Classification and Clustering
- **MFCC**
  - Baseline calculation (Mean and Variance)

Dataset

- **GTZAN**
  - The dataset consists of 1000 audio tracks each 30 seconds long.
  - It contains 10 genres, each represented by 100 tracks.
  - The tracks are all 22050Hz Mono 16-bit audio files in .au format.
  - Genres – Blues, Classical, Country, Disco, Hiphop
    Jazz, Metal, Pop, Reggae, Rock
- **Processed**
  - First 10 from each of the genres.
  - Total of 100 tracks.
- **ReducedSubset**
  - Classical, Pop, Rock genres

Results – All Genres

Classification	KNN	Gaussian NB	Linear SVC	SVC
MFCC	**0.32**	0.52	**0.6**	**0.44**
Pitch Solo	0.28	0.36	0.22	0.22
Pitch Orig	0.22	0.3	0.28	0.2
Combined Solo	0.28	**0.54**	0.16	0.26
Combined Orig	0.22	0.3	0.1	0.2

Clustering	Homogeneity	Completeness	V Measure	Adjusted Rand Score	Adjusted Mutual Info
MFCC	**0.411**	**0.437**	**0.424**	**0.173**	**0.262**
Pitch Solo	0.379	0.418	0.397	0.160	0.234
Pitch Orig	0.343	0.412	0.374	0.122	0.203
Combined Solo	0.372	0.416	0.393	0.134	0.228
Combined Orig	0.350	0.401	0.374	0.128	0.207

Results – Specific Genres

Classification	KNN	Gaussian NB	Linear SVC	SVC
MFCC	0.46	0.93	0.73	0.6
Pitch Solo	0.46	0.93	**1.00**	**0.8**
Pitch Orig	0.46	**1.00**	0.93	0.6
Combined Solo	0.46	0.93	0.93	**0.8**
Combined Orig	0.46	**1.00**	0.86	0.6

Clustering	Homogeneity	Completeness	V Measure	Adjusted Rand Score	Adjusted Mutual Info
MFCC	0.389	0.441	0.413	0.340	0.342
Pitch Solo	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Pitch Orig	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Combined Solo	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Combined Orig	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Conclusion

- The method proposed does not work great for all genres in general.
- However, when used to classify or cluster genres suitable for the particular method, it performs way better than other approaches.
- Extracting pitch from the source separated audio as compared to the original signal, provides better results in most cases.
- Combining the features extracted with the standard MFCC features increases accuracy in most cases.
- In a classification task, augmenting pitch features to the other suited features should provide promising results.

References

- pyFASST – Jean-Louis Durrieu
  - Python implementation of the Flexible Audio Source Separation Toolbox
- Slaney, Malcolm. "Auditory toolbox." *Interval Research Corporation, Tech. Rep*10 (1998): 1998.
- http://marsyas.info/download/data_sets/
- Salamon, Justin, Bruno Rocha, and Emilia Gómez. "Musical genre classification using melody features extracted from polyphonic music signals."*Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012.
- Durrieu, J-L., et al. "Source/filter model for unsupervised main melody extraction from polyphonic audio signals." *Audio, Speech, and Language Processing, IEEE Transactions on* 18.3 (2010): 564-575.
- Sturm, Bob L. "An analysis of the GTZAN music genre dataset." *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, 2012.
- Panagakis, Yannis, Constantine Kotropoulos, and Gonzalo R. Arce. "Music genre classification via sparse representations of auditory temporal modulations." *Proc. European Signal Process. Conf., Glasgow, Scotland*. 2009.

Thank You