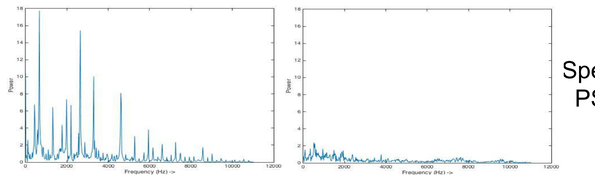
Music Speech Discrimination

Shiv Kaul, Kushaagra Goyal, Yash Malviya
Department of Electrical Engineering, Stanford University

Introduction

- Speech music discrimination has widespread applications in multimedia domain.
- For e.g., adaptive audio encoding for music/speech audio samples
- Music requires higher bandwidth and resources to capture it with good quality
- Speech on the other hand has lower frequency as well as quality requirements

Music PSD

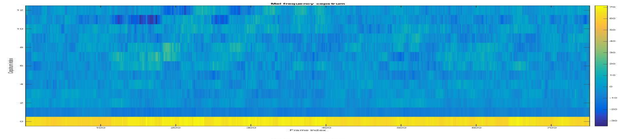


Speech PSD

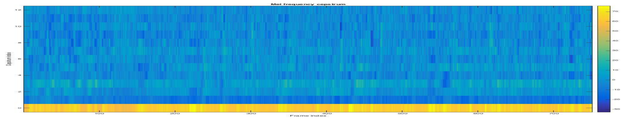
Methodology

- Dataset -> GTZAN + Columbia -> 248 speech+music audio samples, each 15 sec (124 - speech, 124 - music samples)
- We use two types of features (360 Chroma, 9750 MFCC features)
  - MFCC - Mel-Frequency Cepstrum Coefficients
    - Frame width - 20 msec, 13 coeff. per frame
    - MFCC's vary with the timbre of audio signal
  - Chroma Features (CENS - Chroma Energy Normalized Statistics)
    - Encode short-time energy dist. over 12 pitch classes
    - Short-time stats over energy dist. in chroma bands gives CENS

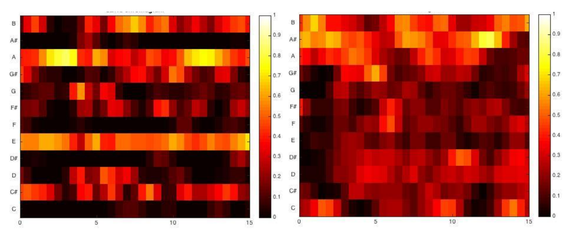
Music MFCC



Speech MFCC



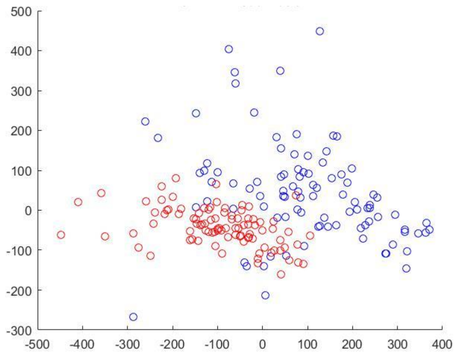
Music CENS



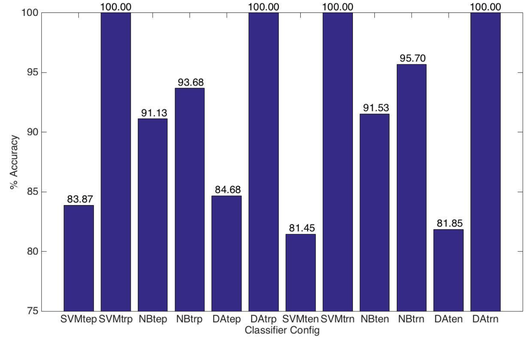
Speech CENS

Processing Steps

- Data Permutation - inspired from image classification
  - To expand our small dataset we permute audio samples to generate new samples
  - This approach increases our datasize 6-fold, reduces overfitting
- Principal Component Analysis
  - MFCC+Chroma features creates a large set of features
  - To visualize we applied PCA to convert the features space into two dimensions. Music/Speech clusters can be seen distinctly



- Classifier
  - Choice of classifier is essential to ensuring low generalization error
  - Strong classifiers like SVM overfit the training set and produce high test error
  - Whereas weaker classifier like Naive Bayes generalise better

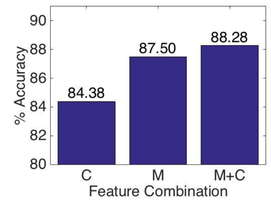


te - test, tr - train, p - permuted, n - not permuted

Results

- Chroma and MFCC features combined give best accuracy
- Best classifier: Naive Baye's with permuted data set, using 4-fold cross validation. Achieved 91.13% accuracy

	Speech (predicted)	Music (predicted)
Speech (actual)	30.75	0.25
Music (actual)	5.25	25.75



- From the above matrix it is observed that we have a higher recall for speech samples than for music samples

	Music	Speech
F-Score	0.8987	.9208

- Testing on unseen test data from Columbia dataset has 98.75% accuracy
- Demo!

Future Work

- Implement a CNN to perform speech/music classification and compare it with a feature-based approach
- Analyze why there is higher recall for speech samples in our model.
- Compare classification accuracy of music samples with and without vocals

References

1. George Tzanetakis, GTZAN Music/Speech Collection, University of Victoria.
2. Dan Ellis, The Music-Speech Corpus, Columbia University, 2006.
3. Kamil Wojcicki, HTK MFCC MATLAB, MathWorks, 2011.
4. Meinard Muller et al., Chroma Toolbox: MATLAB Implementations for Extracting Variants of chroma based Audio Features, International Society for Music Information Retrieval, 2011.