Music Speech Discrimination

Yash Malviya, Shiv Kaul, Kushaagra Goyal

ymalviya, shivkaul, goyalk@stanford.edu

Abstract

We present a classifier that classifies audio samples into music or speech. Our classifier uses MFCCs and chroma features which capture timbre and pitch information from an audio sample. We demonstrate the separability of music-speech samples by using PCA to reduce our feature space down to two dimensions. We implemented SVM, GDA, and Naive Bayes classifiers on our dataset. The Naive Bayes classifier gives 91% accuracy over the dataset with 4-fold cross-validation. When tested on completely unseen dataset this same classifier achieves an accuracy of 98.75%

1. Introduction

Music-Speech Discrimination has widespread application in the multimedia domain. It can help to skip over advertisements in audio broadcasts and allow for efficient encoding of audio samples since music and speech parts of an audio can be compressed more efficiently using different techniques. Figure 1 shows Power Spectral Density (PSD) plots for music and speech samples, clearly illustrating their differences. Several features such as zero-crossing rate, spectral centroid, spectral roll-off, MFCC's (Mel Frequency Cepstral Coefficients) and Chroma Features have been demonstrated to be useful in classifying and characterising audio signals. Since the number of features are huge in our task, we also tried feature reduction techniques like PCA on our data.

We used these features along with various classifiers including Support Vector Machines (SVMs), Naive Bayes and Gaussian Discriminative Learning for this task. Our final observation was that Naive Bayes Classifier worked best on our application and gave 91% accuracy over the dataset with 4-fold cross-validation.

The rest of this report is organized as follows - Section 2 discusses the relevant existing work in this field, Section 3 discusses our methodology to perform speech-music classification. It gives details about the dataset and the features used in our classifier. Section 4 gives the results where we compare the performance of different classifiers and give our evaluation results. Section 5 gives a conclusion and the proposed future work for this project.

2. Relevant Work

Music speech discrimination has been a problem of interest for a long time [1, 3, 4, 5, 6, 7, 8]. Researchers have tried a variety of techniques to enhance this classifier. Therefore considerable effort has been spent in identifying features and using the right classifier [3, 4] for this application. Recently a deep learning network has also been proposed to solve this problem [7].

MFCC features have been used for modeling audio signals for considerable time [2, 6, 8]. Apart from capturing information about timbre of the audio it also indicates information generation of vocal patterns by humans. Music speech discrimination is one among its many applications. Recently chroma features [1] which measure the energy distribution among different pitches has also proven to be a useful feature for speech music discrimination.

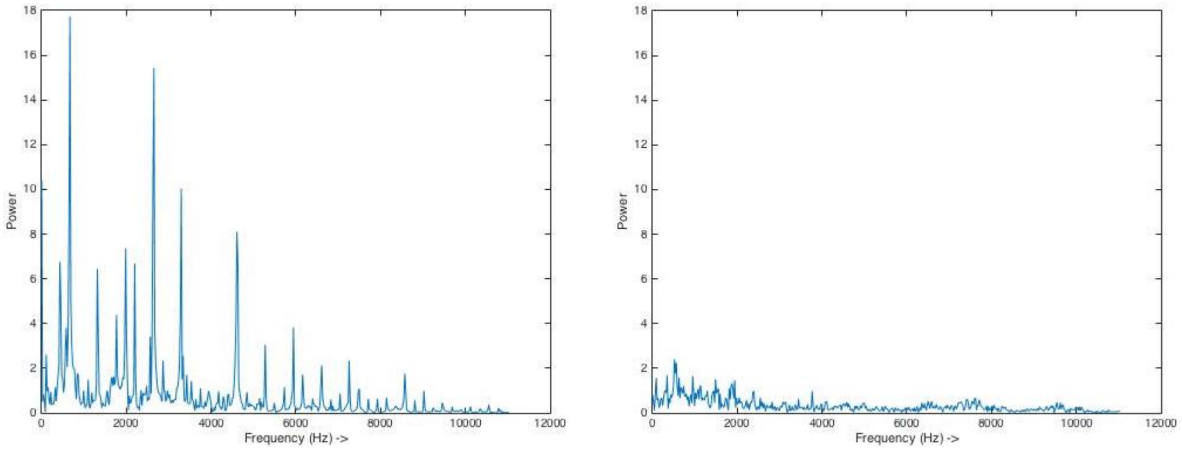
In this work we combine MFCC and chroma features for each audio sample and test its performance with a number of different classifiers to identify the best.

3. Methodology

3.1. Dataset

We combined two different datasets : GTZAN and Columbia Dataset. GTZAN contains 128 audio clips evenly distributed between music and speech, each 30 seconds in duration. Similarly the training set in Columbia contains 120 audio samples. After combining these two datasets, we have 248 audio clips of 15 seconds each (taking the first 15 sec from each of the GTZAN samples), out of which half are music and half are speech. For our evaluation we use 4-fold cross validation.

The Columbia dataset has a test set which has over 60 audio clips each 15 second long. This consists of twice the number of music samples as compared to speech. We keep this test set hidden from the classifier



(a) Music

(b) Speech

Figure 1: A plot of Power Spectral Densities of a music and speech audio sample. Evidently music has high energy concentrated in certain frequency values whereas speech is mostly evenly distributed.

until the end. After deciding our final classifier using cross validation on the training set we check its accuracy on this test set. We obtain a super impressive accuracy of 98.75 on this set.

Moreover, we generated more training data by permutating the audio samples. We divided our clip into 3 parts and formed 6 clips from 6 different permutations we can achieve by reordering these smaller clips. While testing using the permutated data-set, we first split the initial 248 samples into 62+186 samples. We then permutated these 186 samples and generate 186*6 samples and train the classifier using them and test on the remaining 62 samples. We repeated this process 4 times, picking 62 samples differently in each iteration.

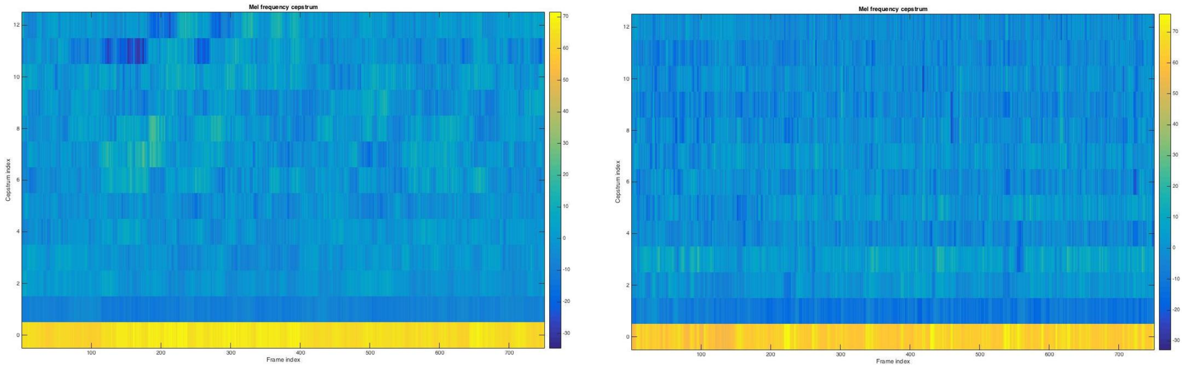
3.2. Feature Set

We tried various different features to train our model. The set of features are given below:

- Spectral Centroid : This feature is associated with musical timbre and is calculated by taking the weighted average of the power spectrum with respect to the frequencies.
- Spectral Flux : This feature indicates the rate at which the power spectrum of a signal is changing, by taking differences between the adjacent values in the spectrum.
- Zero Crossing Rate : As its name suggests, the zero crossing rate is the rate at which a given signal changes sign.
- MFCCs : Mel Frequency Cepstrum Coefficients (MFCC) are one of the widely used features in Music Information Retrieval. They are commonly derived by first taking the fourier transform, mapping it to the mel scale and then taking DCT of the mel log powers. We used an existing code library online for extracting the MFCC features from an audio signal [12]. We set the frame width to be 20ms for our purposes. For a 15 second sample, MFCC produces 9750 features. MFCCs are based on the timbre of sound signals, whereas Chroma features (explained later) are based on the pitch of a signal. We analyse the performance of our classifier using each of these features individually and as well as in combination with one another. Figure 2 shows a visualization of MFCC features for a music and speech sample.
- Chroma Features : These features encode the short-time energy distribution of music signal over 12 traditional pitch classes of equal-tempered scale. To compute these features we use an existing code base [10]. It decomposes a given audio signal into 88 frequency bands with center frequencies corresponding to the pitches A0 to C8. Then for each subband, the short-time mean-square power (STMSP) is computed. The features measure the local energy content of each pitch subband.

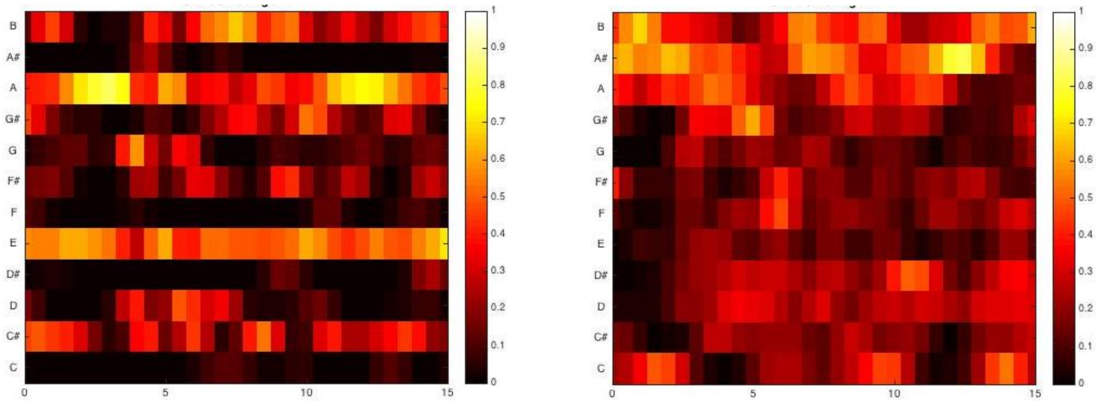
  Further we calculate short-time statistics over energy distributions within the chroma bands and obtain Chroma Energy Normalized Statistics (CENS). Since CENS features can be processed efficiently, have low temporal resolution, and are strongly correlated to the short-time harmonic content of the underlying audio signal, we use CENS features for our classification. Figure 3 shows a visualization of chroma features for a music and speech sample.

In the midterm report, when we were using only GTZAN dataset and using 30 second samples, we demonstrated that Chroma features and MFCC features together achieve high accuracy. Other additional features used in conjunction with MFCCs and chroma features do not increase classification accuracy. Figure



(a) Music (b) Speech

Figure 2: A visualization of the MFCC features



(a) Music (b) Speech

Figure 3: A visualization of CENS features used to classify music and speech. In (a) we can observe sharp variation between CENS parameters for neighboring pitches, characteristic of music. Whereas (b) shows less stark distribution of CENS parameters.

4 shows a distribution of the contribution to accuracy from each of the above features and their combinations. (Only using GTZAN dataset) Since other features don't add much to the combination we are using only MFCC and Chroma features in our final model.

3.3. Data Visualization

We applied principal component analysis (PCA) in order to better visualize the effectiveness of our features. For each 15 second training example, a total of 10110 features (9750 from MFCCs and 360 from CENS) were used for classification. We reduced the feature space down to two principal components and plotted the results to view the separation between music and speech examples (seen in 5). For this plot we chose to use the non-permutated dataset so as to not clutter our plot. Applying PCA and then doing classification gave poor results, hence we decided to use the complete set of features for our classifier.

4. Results

In this section we discuss a comparison of the different types of classifiers we tested for our application. In Figure 6 the training and testing accuracy of different classifiers can be seen, both with and without data permutation.

It is evident from Figure 6 that strong classifiers like SVM (linear kernel) and gaussian discriminant analysis overfit the training data, achieving a training accuracy of 100%. However, accuracy on test set is 83% and 84% respectively, highlighting the poor generalization accuracy of these classifiers.

Naive Bayes, on the other hand, with a training accuracy of 93% and testing accuracy of 91% generalizes best. The structure of MFCCs and Chroma features gives some guarantees on their statistical independence [9, 11]. This offers some support to the intrinsic assumption of the Naive Bayes model, justifying its impressive performance.

Table 1 shows a confusion matrix and F-score values for the Naive Bayes. As can be seen in confusion matrix, in a 62 sample test set (averaged over 4-fold cross validation), music samples are miss classified more often than speech. This is verified by the slightly higher F-score for speech than music. Such a difference is

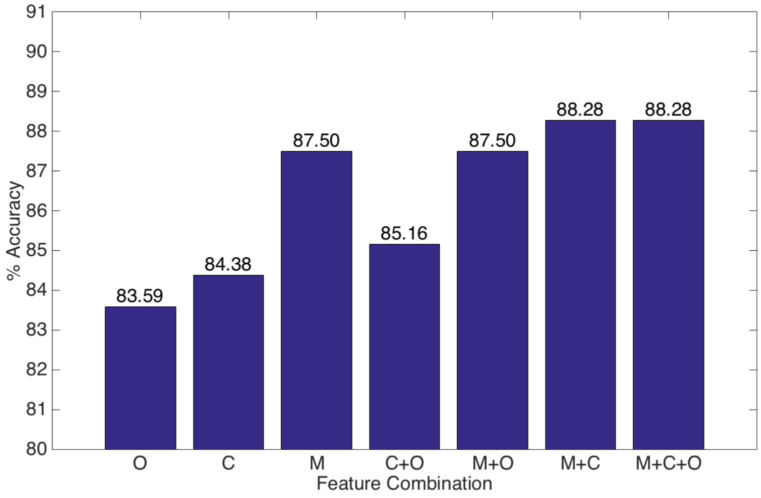


Figure 4: Dist. of accuracy contributed by different combinations of classifying features. M-MFCC, C-Chroma features, O-Spectral Flux, Spectral Centroid and Zero Crossing Rate

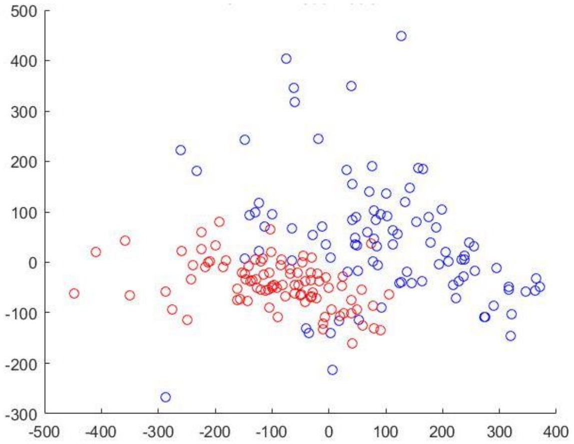


Figure 5: PCA applied to our Naive Baye's classifier. The red data points correspond to speech samples while the blue data points correspond to music samples.

expected since the classifier distinguishes music from speech on the basis of certain high frequency content or pitch information. If it cannot detect these types of information it is bound to incorrectly classify speech as music. This causes the slight bias towards speech.

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

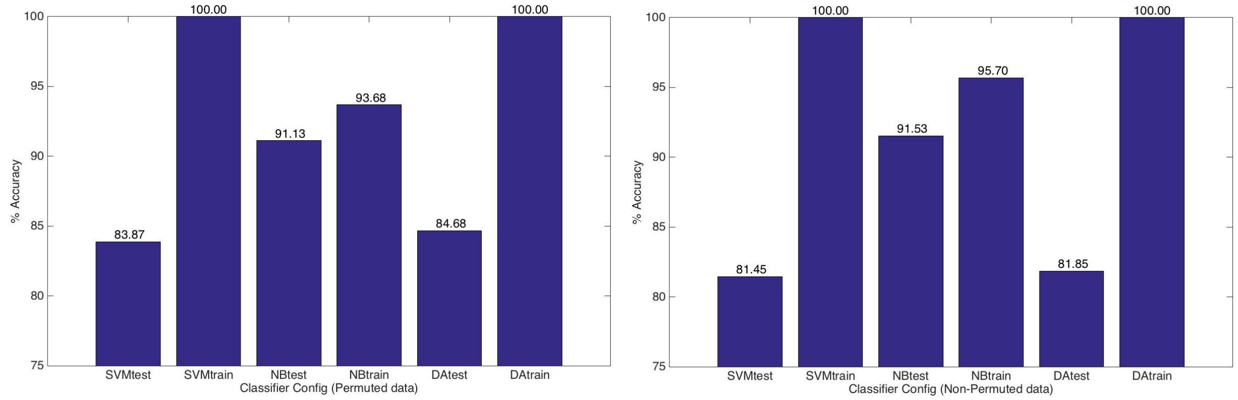
	Speech	**Music**
F-score	0.899	0.921
Precision	0.854	0.990
Recall	0.992	0.831

Table 1: Confusion matrix for our classifier. Underneath there is f-score, precision, and recall for music and speech samples

Our classifier achieves an accuracy of 91.13% on the training set with 4-fold cross validation. Additionally, when tested on an unseen dataset (test set from Columbia dataset) we get an accuracy of 98.75%. Such a high accuracy validates our model, but could have been enhanced by ease of classification of the dataset being tested. For any arbitrary dataset accuracy might decrease slightly.

5. Conclusion & Future Work

We would like to seek explanations for peculiarities in our results, such as why our best classier has a much greater recall for speech samples than it does for music samples. In addition, we would like to apply convolutional neural networks (CNNs) to automate the feature selection, training, and classification processes.



(a)

(b)

Figure 6: Classifiers like SVM, Gaussian discriminant analysis over fit the training set and perform bad with test set. Naive Bayes generalizes best, since independent feature assumption for Naive Bayes is true for this feature set.

We expect that CNNs will yield better performance compared to using standard features to classify audio samples as either speech or music. We will also investigate how music samples with and without speech influence our training and testing accuracies.

6. References

[1] Sell, Gregory, and Pascal Clark. *Music tonality features for speech/music discrimination.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[2] Logan, Beth. "Mel Frequency Cepstral Coefficients for Music Modeling." *ISMIR.* 2000.

[3] A. Masoumeh Velayatipour et al., *A Review on Speech-Music Discrimination Methods*, International journal of Computer Science and Network Solutions, 2014.

[4] Kim, Kibeom, et al., *Speech Music Discrimination Using an Ensemble of Biased Classifiers*, Audio Engineering Society Convention 139. Audio Engineering Society, 2015.

[5] K. El-Maleh, et al., *Speech/music discrimination for multimedia applications*, Acoustics, Speech, and Signal Processing, 2000.

[6] Rajesh, B. & Bhalke, D.G.,*Automatic genre classification of Indian Tamil and western music using fractional MFCC*, International Journal of Speech Technology, 2016.

[7] A. Pikrakis et.al.,*Speech-music discrimination: A deep learning perspective*, Signal Processing Conference (EUSIPCO), 2014.

[8] D. Hosseinzadeh et.al.,*Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs*, Multimedia Signal Processing, 2007.

[9] M. Bartsch, G. Wakefield,*To Catch A Chrous: Using Chroma-based Representations for Audio Thumbnailing*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2001.

[10] Meinard Muller et al., *Chroma Toolbox: MATLAB Implementations for Extracting Variants of chroma based Audio Features*, International Society for Music Information Retrieval, 2011.

[11] Taylor, Paul, *Text-to-Speech Synthesis*, Cambridge university press, 2009.

[12] Kamil Wojcicki, *HTK MFCC MATLAB*, MathWorks, 2011.

[13] George Tzanetakis, GTZAN Music/Speech Collection, University of Victoria.

[14] Dan Ellis, The Music-Speech Corpus, Columbia University, 2006.