Speech-Music Classifier

Comprehensive evaluation of performance and investigation into mixed-signal behaviour

Author:
Aadam Osman
OSMAAD003

May 28, 2019

Contents

1 Abstract **3**

2 Introduction **3**

3 Literature Review **4**

4 Aim **5**

5 Apparatus **5**

5.1 Programming Tools 5

5.2 Data 6

6 Methodology **6**

6.1 Feature Selection 6

6.2 Feature Extraction 6

6.3 Classifier Model Selection 7

6.4 Training, Testing and Evaluation 8

6.4.1 Training and Testing 8

6.4.2 Evaluation Criteria 8

7 Results **9**

8 Discussions and Conclusion **11**

List of Figures

1 ROC curves of all models after aggregating the results after each fold of testing 10

2 kNN Response to Artificial Rap 10

3 SVM Response to Artificial Rap 11

4 RF Response to Artificial Rap 11

List of Tables

1 Top 5 most correlated values of MFCC coefficients against target variable in increasing order 7

2 Area Under Operating Receiving Characteristic Curve Values for each model aggregated over all validation folds 9

3 Skewness values for all models in “Artificial Rap” analysis 10

1 Abstract

The performance of classical machine learning algorithms were investigated for the purpose of binary music/speech classification using the AUROC metric. Three models were implemented and tested namely kNN, SVM and Random Forest. The Random Forest model performed the best out of the three but all three models performed fairly well for the task. The models behaviour to mixed signals (signals containing both music and speech) was also investigated by the use of "artificial rap" generated from "freestyle" audio snippets, containing speech signals, mixed with "instrumental" audio snippets, containing music signals. The freestyle-instrumental pairs were biased in such a way that the frequency distribution, plotted over the predicted probabilities, was always more skewed to the right for the classification of the instrumentals as music. Despite this the combined signal's skewness, of the frequency distribution, always gravitated towards 0 implying that the model was uncertain every time signals were mixed even if it was considerably more confident in the one class over the other. This insinuates that these models don't seem to have a propensity to classify one class over the other.

2 Introduction

The distinction between speech and music audio is quite a trivial task for humans with standard hearing. Their are numerous applications where this classification would be useful, however, most of these, in order for them to be considered feasible, would require the process to be automated - without the need of human supervision.

Its purpose in real-time radio receivers allow listeners of FM radio channels to automatically change stations when a commercial plays.[10] Other real-time applications for Automatic Speech Recognition (ASR) in broadcasting include the ability to determine how much air-play is dedicated to music or ads (seeing if stations are in regulatory compliance)[10] and disabling the speech-recognizer during non-speech segments of new's broadcasts.[9] Low bit-rate audio encoding is another particularly useful application. As a rule of thumb, speech encoders perform better on speech and audio encoders better on music.[2] Developing a universal encoder for both is a non-trivial task. An alternative approach is to use a multi-mode encoder and select an appropriate encoder module based on the output of a speech/music classifier.[9] This is already employed in the parametric encoder of the MPEG-4 standard.[6] Emerging applications include content-based audio and video retrieval where audio classification plays a pivotal role.[9]

The construction of these classifiers usually begins with carefully selecting features - either based on previous or newly motivated literature. The selected features are then extracted on a frame by frame basis. Once the extraction process is complete, the data is inputted to a model(s), of one's choice, in a process referred to as training. This model(s) is now able to make predictions - this refers to whether a frame is classified as a music or speech signal, in the context of this report.

3 Literature Review

A large amount of available literature tackles the problem of speech/music audio classification by first discussing and justifying chosen features and then moving onto explanations of the classification frameworks used. More often than not, the emphasis is drawn on the former. This literature review will similarly begin with a discussion of such popular features and how they relate to speech and music then it will discuss the different types of classification frameworks and current state of the art methods.

Some of the popular features, initially proposed by Scheirer and Slaney[1], are: 4Hz modulation energy, Percentage of ”Low-Energy” Frames, Spectral Roll-off point, Spectral Centroid and the Zero-Crossing Rate. The 4Hz modulation energy is the 4Hz Mel-Frequency energy obtained via use of the Mel-Frequency Cepstrum Coefficients (MFCC's). Speech has a characteristic energy modulation peak around the 4Hz syllabic rate[5]; thus, it tends to have more modulation energy at 4Hz than music does.[1] Percentage of ”Low-Energy” Frames refers to the proportion of frames with RMS power less than 50% of the mean RMS power within a certain window.[1] Speech's energy distribution is more left skewed than music - due to more quiet frames.[1] Spectral Roll-off measures the amount of energy in the higher frequency range of the spectrum past a chosen percentile. It helps distinguish between ”voiced” and ”unvoiced” speech which has a lower Spectral Roll-off.[1] Spectral Centroid can be thought of as the “balancing point” of the spectral power distribution.[1] It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weight.[4] Music with percussive sounds, which is often present in music, includes high frequency components which push this spectral mean higher.[1] Zero-crossing rate is the number of time-domain zero crossings within a frame.[10] It is shown to be an extremely effective discriminator and Saunders[10] built a real time classifier, using 16ms frame data, solely based on this feature.

The above features, apart from 4Hz energy, rely on features extracted in the ”standard” frequency domain but other features extracted from other domains have also been used namely Line Spectral Frequencies[9] and Mel-Frequency (such as MFCC's).[11] MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz and frequencies are perceived to be evenly spaced by the human ear.[7] The rough idea is that it captures certain pitch elements that, similar to how we use our ears, help discriminate important phonetics in speech.[7] Line spectral frequencies (LSFs) have been used in classifiers built for multimedia applications.[9] They are a transformation of linear prediction (LP) coefficients which is widely used for linear predictive coding, in audio and speech signal processing, for efficient storage.[3]

Apart from frequency domain features some information theory based features have also been shown to be good discriminators. Entropy Modulation, defined as the signal's entropy, is one such feature as music tends to be more "ordered" than speech.[8]

The classifiers used in previous literature tend to be "supervised learning" algorithms, where the models are trained with labelled data. Some commonly used models include Gaussian Mixture Model(GMM), k-Nearest Neighbours(kNN), Support Vector Machines and Hidden Markov Models.[11] With the rise in GPU's and the increase in popularity of Deep Learning (DL), many new approaches utilize DL algorithms such as Artificial Neural Networks (ANN's)[12] and Convolutional Neural Networks(CNN's).[13]

4 Aim

To investigate the performance of classical machine learning algorithms for binary music/speech classification and to investigate the behaviour of the chosen models to mixed signals (containing both music and speech).

5 Apparatus

5.1 Programming Tools

All processing was done in the **Python** programming language and was written and run on Google's **Colaborotory (Colab)**. Colab is a free Jupyter notebook environment that requires no setup and runs entirely on the cloud. It was used to allow easy collaboration (as two people can edit a single notebook at a time) and because of the free access to GPU compute time - greatly reducing model training time.

Librosa, a popular audio processing package, was used to load audio files and extract the chosen features for training. **Scikit-learn (sklearn)** was used to build and train the classification algorithms. **Pandas** was used for some basic Exploratory Data Analysis with **SciPy** and **NumPy** being used for basic data manipulation. Visualization was done via the **matplotlib** package.

5.2 Data

The “GTZAN music/speech collection” dataset was used for this project and was collected from Marsyas - a “Music Analysis, Retrieval and Synthesis for Audio Signals” open source project.[14]. Each audio segment is a single channel 16 bit track sampled at a sampling rate of 22050Hz. A total of 128 audio samples were used (64 each for both music and speech respectively).

The “freestyle” audio files, used in the “artificial rap” analysis explained further on, was downloaded from Soundcloud and is also sampled at 22050Hz. A total of 16 such files were used.

6 Methodology

6.1 Feature Selection

The choice of features for this report was decided based on two simple criteria- ease of extraction and proven efficacy in past literature. Given these it was decided on the following features:

- Zero-Crossing Rate
- Spectral Roll-Off
- Spectral Centroid
- Percentage of Low Energy Frames
- Mel-Frequency Cepstrum Coefficients (MFCC)

Most of the above, except “percentage of low energy frames”, have simple built-in functions in Librosa making it fairly simple to extract.

6.2 Feature Extraction

In most previously done work a window frame between 16-25ms was used, when extracting features, as this was small enough to be used in real-time applications[10] yet large enough to capture intrinsic characteristics of the signal.[8] A 23ms overlapping window frame is used in this project. It is generally ideal to have infinite overlap but to reduce computational cost an overlap of 50% was used.

The percentage of low energy frames was implemented as per Slaney's definition [1] - by computing the mean RMS within a 1s window and counting how many frames have RMS values less than fifty percent of that value (then normalizing over the total no. of frames in the window). The same frame size and overlap was used as all the other features, along with the RMS calculations being performed with the Librosa library.

The number of MFCC's that the Librosa function computes is defaulted to 20 coefficients. Using all 20 would make the dimensions of our input vectors (for the classifiers) too large thus making training slow and could affect the models ability to classify. A simple Pearson correlation was computed, using Pandas, with the MFCC's against the target variable - defined as 1 for music and 0 for speech.

Table 1: Top 5 most correlated values of MFCC coefficients against target variable in increasing order

MFCC coefficient no.	5	6	17	1	2
correlation	-0.117233	0.098009	0.142868	0.254073	0.348580

From these results it was decided to use only MFCC coefficients 5, 17, 1 and 2 (6 was below 0.1 so it was ignored).

6.3 Classifier Model Selection

The models were selected on a similar criteria to that of the features - simplicity and and proven efficacy in past literature. Simplicity referring to ease of implementation. Based on this criteria the following models were selected:

- k-Nearest Neighbours (kNN)
- Support Vector Machines (SVM)
- Random Forest (RF)

All the above classifiers are "classical" machine learning algorithms and can be implemented simply using sklearn. The Random Forest is the only model that was not used in any of the reviewed literature and is a "relatively" new machine learning model first proposed in 1995.[15] In most current machine learning applications it is a very popular model especially for classification problems. It tends to produce very good results and as a result was included in our model choices.

6.4 Training, Testing and Evaluation

6.4.1 Training and Testing

In order to create a reliable evaluation of the performance and to ensure that the models are tested thoroughly it was decided to use a 5-fold cross validation. Here the data-set is divided into 5 partitions and for 5 permutations, trained on 4 of them and then tested on the 5th. The models are stored at each fold, then discarded and recomputed at the next fold. The final presented results will be on the aggregated performance of the models on each fold. This reduces the variance on the performance (than if just a single hold out was used) and prevents "lucky" data segments, when splitting into testing and training data, from skewing the results.

The features are then normalized using Librosa's normalize function and fed to the model for training (frame by frame). The target variable was set to 1 if its music and 0 if its speech.

The k parameter for the kNN model was set to 5 - to search for the 5 nearest neighbours when classifying. SVM was implemented with a non-linear radial basis kernel and using the NuSVC mathematical formulation. RF was implemented using 100 trees and the "Gini impurity" criteria to measure the quality of a split. All other parameters were left to sklearn's default settings.

6.4.2 Evaluation Criteria

Most reviewed literature use accuracy and error as measures of a model's performance. Other common metrics include precision and recall. These metrics are associated with a single probability threshold, whereby if the models predicted probability is higher than the threshold it predicts a certain class. A more accurate method of determining the performance is by computing the Area Under the Receiver Operating Charecteristic curve (AUROC). The Receiver Operating Characteristic curve (ROC) is a plot of the true positive rate (TPR), on the vertical axis, vs the false positive rate (FPR), on the horizontal axis, for varying threshold probabilities. Each accuracy, precision and recall metric,commonly seen in other papers, corresponds to a single point on the ROC graph. The AUROC is a better metric than the others, mentioned earlier, as it takes into account unbalanced data and measures the trade-off between TPR and FPR as one varies the threshold probability.

The next part of our evaluation is a form of heuristic analysis whereby we analyze the response of our models to "artificial rap." Audio snippets of "free-style" rap music, which just consists of speech, and audio snippets of "instrumentals," which come from the music data set that was used to train the models, were tested. The test consists of first testing each model with the free-style segments and instrumental segments individually, then testing each model with the combined signals. A frequency distribution of the predicted probabilities were found, along with their associated skewness. Where a positive skewness corresponds to left skewed data and a negative skewness corresponds to right skewed data. This was done for:

- freestyle audio - predicted for speech
- instrumental audio - predicted for music
- mixed "rap" audio - predicted for both speech and music

The purpose of this is to see if there are any interesting observations for this type of classification that could potentially be investigated in later works. One would assume one of two things are likely to show up. Firstly that because the mixed signal is essentially in a class that belongs to neither music nor speech, and since there is no explicit class for this, the classifiers prediction will gravitate towards a 50/50 prediction as it is uncertain of the outcome. This would imply that the skewness should always gravitate towards 0. Another result that could appear is that for some reason if the propensity for classification is higher for one class ie the skewness more negative for one of the classes than the other, for a given freestyle and instrumental pair, the combined signal's skewness would be skewed closer to that class. So the classifier is consistent in some form, with its own predictions, under superposition. Of course neither of these could happen but our heuristic assumption is that it will be one or the other. So that it would be easier to visualize the results the freestyle audio files were "cherry picked" such that none of the models predicted them very well. This ensured that instrumentals as music was always the more "confident" prediction.

7 Results

Table 2: Area Under Operating Receiving Characteristic Curve Values for each model aggregated over all validation folds

Classifier	AUROC Values
kNN	0.9230
SVM	0.8874
RF	0.9729

The dotted reference line in Figure 1 is the performance of a model that just randomly guesses. It is used as a reference to see how much better a model is doing than if it were to just randomly guess and not making any predictions.

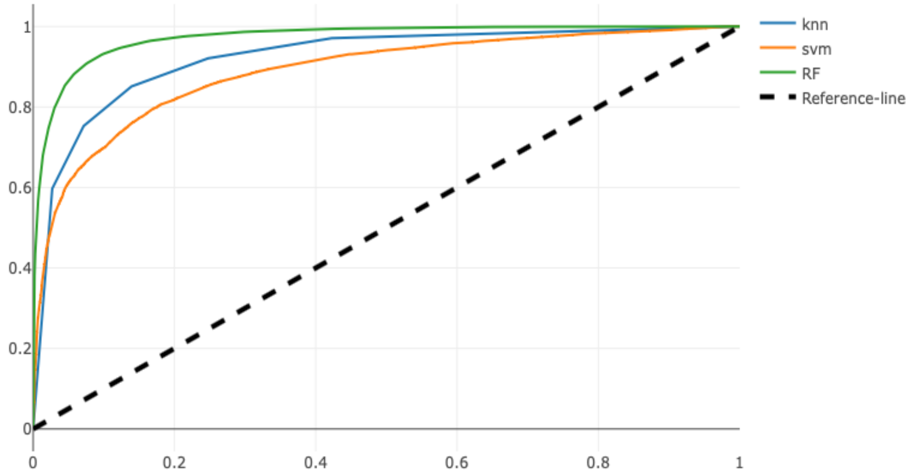

Figure 1: ROC curves of all models after aggregating the results after each fold of testing

Table 3: Skewness values for all models in “Artificial Rap” analysis

Classifier	Freestyle(as speech)	Instrumental(as music)	Combined(as speech)	Combined(as music)
KNN	-0.9835	-5.1923	0.0768	-0.0768
SVM	-0.4802	-1.1042	0.0701	-0.0701
RF	-0.8606	-3.7480	0.0420	-0.0420

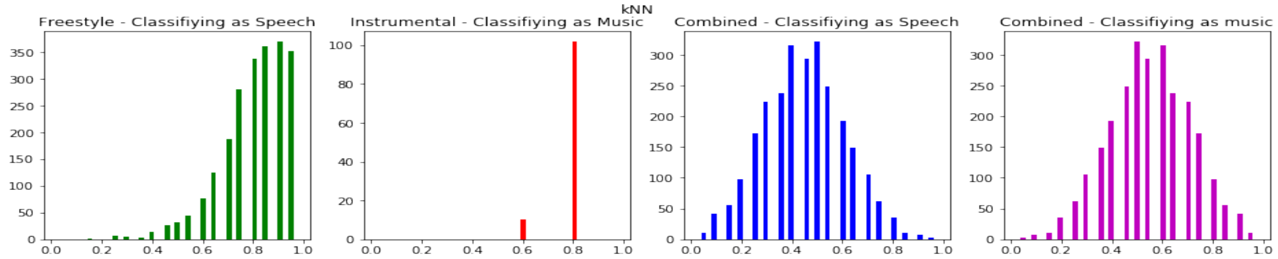

Figure 2: kNN Response to Artificial Rap

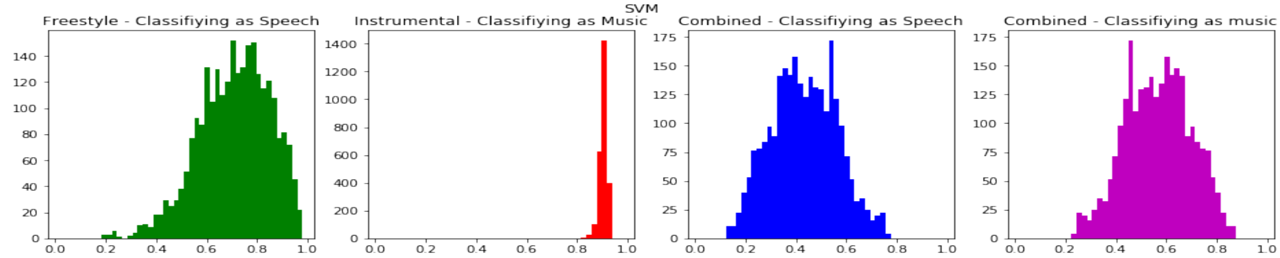

Figure 3: SVM Response to Artificial Rap

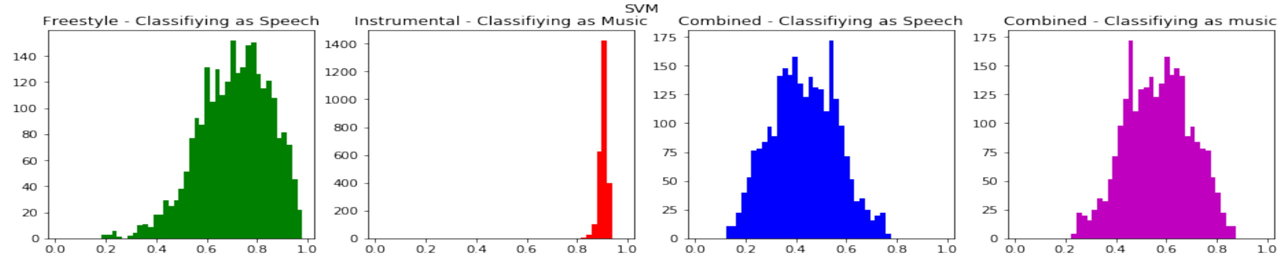

Figure 4: RF Response to Artificial Rap

8 Discussions and Conclusion

In terms of performance, the AUROC values in Table 2 show that all models performed quite well with Random Forest performing the best with an AUROC of 0.9729. The high performance of all the models is indicative of the discriminatory ability of the features proposed in past literature as the models were trained on a fairly modest dataset of 128 audio files.

The mixed signals results all tend to 0 skewness even with a clearly greater skewness for the instrumentals, classifying as music, in the freestyle-instrumental pairs. The average absolute value of the skewness for the combined signals is 0.06230; insinuating a shift towards a a normal distribution and a random classification. This is in agreement with the first proposed outcome and implies that the models is in essence are unsure which class the signal belongs to because they are neither music nor speech. In a strict theoretical sense it could be said that this makes the models robust as the objective of a binary classifier is to only classify into 2 fixed classes anything apart from those 2 classes it “should” be uncertain of.

References

[1] Scheirer, E. &amp; Slaney, M. n.d. Construction and evaluation of a robust multifeature speech/music discriminator1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. DOI: 10.1109/icassp.1997.596192.

[2] Ramprashad, S. n.d. A multimode transform predictive coder (MTPC) for speech and audio1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No.99EX351). DOI: 10.1109/scft.1999.781467.

[3] "Linear Predictive Coding." 2010. Principles of Speech Coding165–184. DOI: 10.1201/b15821-8.

[4] Gajic, B. &amp; Paliwal, K. 2006. Robust speech recognition in noisy environments based on subband spectral centroid histogramsIEEE Transactions on Audio, Speech and Language Processing14(2):600–608. DOI: 10.1109/tsa.2005.855834.

[5] Houtgast, T. &amp; Steeneken, H.J.M. 1973. The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility The Journal of the Acoustical Society of America54(2):557–557. DOI: 10.1121/1.1913632.

[6] Li, W. n.d. Overview of Streaming Video Profile Amendment in MPEG-4 video standard Proceedings of Workshop and Exhibition on MPEG-4 (Cat. No.01EX511). DOI: 10.1109/mpeg.2001.996452.

[7] H.mansour, A., Salh, G.Z.A. &amp; Mohammed, K.A. 2015. Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients AlgorithmsInternational Journal of Computer Applications116(2):34–41. DOI: 10.5120/20312-2362.

[8] Pinquier, Senac &amp; Andre-Obrecht. 2002. Speech and music classification in audio documentsIIEEE International Conference on Acoustics Speech and Signal Processing. DOI: 10.1109/icassp.2002.1004854.

[9] El-Maleh, K., Klein, M., Petrucci, G. &amp; Kabal, P. n.d. Speech/music discrimination for multimedia applications2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100). DOI: 10.1109/icassp.2000.859336.

[10] Saunders, J. n.d. Real-time discrimination of broadcast speech/music1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. DOI: 10.1109/icassp.1996.543290.

[11] Ajmera, J., Mccowan, I. &amp; Bourlard, H. 2003. Speech/music segmentation using entropy and dynamism features in a HMM classification frameworkSpeech Communication40(3):351–363. DOI: 10.1016/s0167-6393(02)00087-0.

[12] Harb, H. &amp; Chen, L. 2003. Robust speech music discrimination using spectrum's first order statistics and neural networksSeventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings. DOI: 10.1109/isspa.2003.1224831.

[13] Choi, K., Fazekas, G., Sandler, M. &amp; Cho, K. 2017. Convolutional recurrent neural networks for music classification2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/icassp.2017.7952585.

[14] About. n.d. Available: http://marsyas.info/.

[15] Ho, T.K. n.d. Random decision forestsProceedings of 3rd International Conference on Document Analysis and Recognition. DOI: 10.1109/icdar.1995.598994.