MUSIC AND SPEECH DISCRIMINATION

Afagh Farhadi, Thomas Culeton, Jiayi Ren

Department of Electrical and Computer Engineering
University of Rochester

ABSTRACT

Our project on source classification of music and speech required us to extract and manipulate key audio features from individual samples. We then fed the data from these features through a classifier to train our model. Once trained, we fed testing data to our model, which gave us our confusion matrices. Our trained model had a 97% accuracy with classifying music, and a 100% accuracy with classifying speech. Our finished project included a user-friendly interface, and the capability of performing both live tests and tests from our dataset.

Index Terms**—** Music and speech classification

1. INTRODUCTION

Source classification is a challenging concept in audio signal processing, which involves the extraction and manipulation of certain key features. Some approaches in the recent literatures [1], [2], [3] are constructed. In this work, we explored the discrimination between music and speech samples through classification of audio features and machine learning. A robust system with a simple user interface, which is capable of both live-testing and testing from collected data sets, is our result.

2. METHOD

To classify music and speech, an audio signal was analyzed in both time domain and frequency domain. For instance, signal intensity or amplitude in time domain can be used for analysis. On the other hand, frequency spectrum can be obtained by applying Fast Fourier Transform. Hamming window was chosen to divide audio signal into smaller frames. A general classification flow chart of our training phase and testing phase is shown in Fig.1.

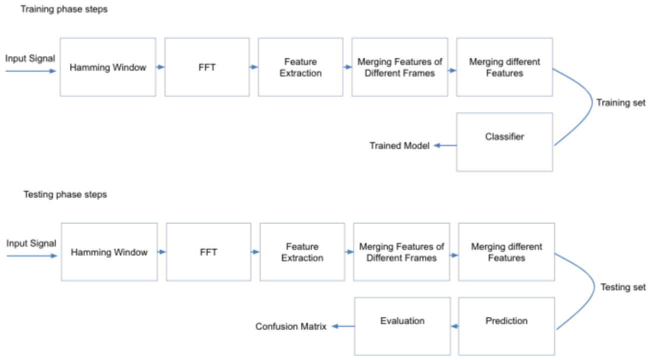


Fig.1. Classification Method Flow Chart

2.1. Feature Extraction

In our project, we focused on six features to extract. These features consisted of: the root mean square (RMS) of the audio signal; the spectrum roll-off frequency, which is the frequency at which 85% of the signal power is below; the zero-crossing rate (ZCR), or how often the plotted signal crosses the time-axis; the spectral centroid, or "center of frequency" where the power is concentrated; the spectral flux, which measures the spectral differences between audio frames; and finally, the short term energy (STE), which measures the voiced and unvoiced segments of a signal.

1. Root Mean Square

This feature relates to the amplitude of audio signal is calculated by (1). Typical RMS plots of music and speech are shown in Fig.2.

$$\text{RMS} \triangleq \sqrt{\sum_{n=1}^{N} x^2(n)} \tag{1}$$

It is important to notice that RMS calculation is volume dependent meaning louder signal would have much bigger average RMS value than quiet signal. To make this feature volume invariant, normalized RMS variance is calculated by the ratio between RMS variance and the square of RMS mean.

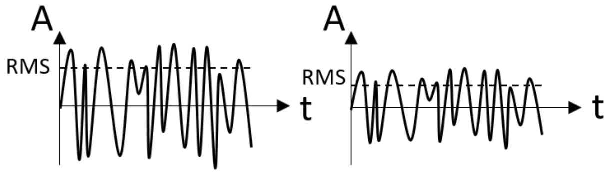


Fig.2. Typical RMS plots of music(right) and speech(left)

2. Zero Crossing Rate

This feature calculate the number of zero-crossings along time axis of a signal. In Fig.3, music signals tend to have much bigger zero-crossing rate comparing to speech signals.

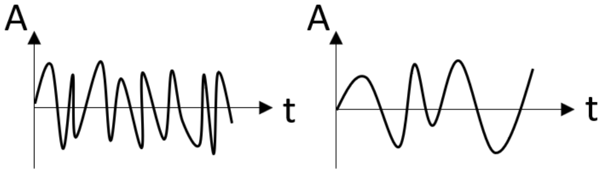


Fig.3. Zero-Crossing of music(left) and speech(right)

3. Spectrum Roll-off

This feature shows the frequency that most (c=80-95%) of the signal power is under that frequency. The power distribution of speech and noise are different across different frequencies. For music power in concentrated across higher frequencies but for speech is mostly in lower frequencies illustrated in Fig.4.

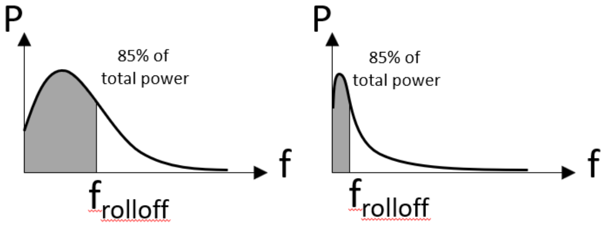


Fig.4. Spectrum Roll-off of music(left) and speech(right)

4. Short-Time Energy (STE)

This feature measures the short-time energy within a window frame. An equation is shown in (2), where x is the signal and w is the window function.

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])^2 \tag{2}$$

In speech there are voiced and unvoiced segments of the signal in which the amplitude changes to lower values for unvoiced segments and increases in voiced segments. However, for music there is no unvoiced segment, so the STE is larger in music signal than speech.

5. Spectral Flux

This feature is the rate of change from power spectrum represented as the 2nd norm of the frame to frame spectral amplitude difference vector (3).

$$\text{SF}=\|\ |X(k)| - |X(k-1)|\ \| \tag{3}$$

Music has bigger rate of change comparing to speech.

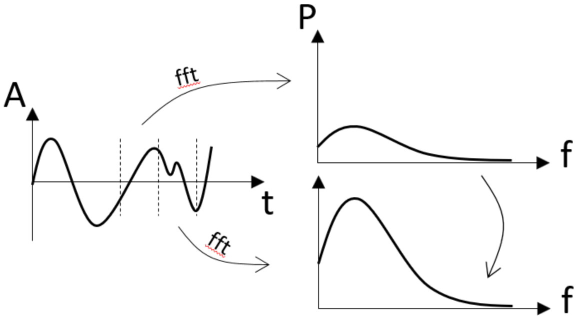


Fig.5. Spectral Flux of a signal

6. Spectral Centroid

This feature shows the center of frequency at with most of the power of signal is concentrated. Music signals have the tendency of bigger spectral centroid than speech signals illustrated in Fig.6.

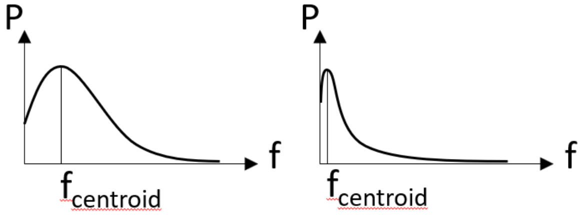


Fig.6. Spectral Centroid of music(left) and speech(right)

2.2. Classification Process

To create a working model, the model first had to be trained via machine learning. Once the model was properly trained, it could then be tested for its accuracy. These processes were similar, yet they differed towards the end of their individual processes. For both processes, the input signal was first split up into data chunks using a Hamming Window, and then a Fast Fourier Transform was applied to these chunks. The next step of both processes was to extract the features described above, and analyze these features. Once analyzed, both processes merged the individual features of different frames, so that each feature had one set of data for the input signal. These features were then concatenated in both processes. In the training process, this concatenated data was fed into a classifier, which would interpret the data and learn whether it was music or speech, based on the training set of data. In the testing process, however, the trained model would evaluate the data it had collected, and predict whether it was music or speech. These predictions were evaluated for accuracy, and placed into a confusion matrix.

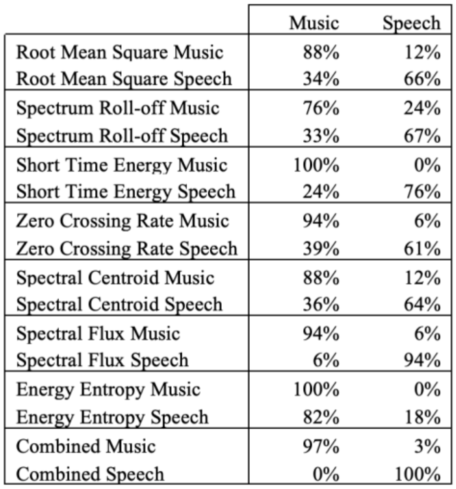
2.2.1 Classifiers

The classifiers used in this project were based off the Gaussian Mixed Model classifier created during Assignment 5. A classifier works by indexing the data chunks that are fed into it, as either music or speech. While only one classifier is required to make this project work, it can be enhanced by using multiple classifiers, each trained using different sized data chunks. These multiple classifiers were all trained off the same data, so they are all given the same start. The testing data was fed through all classifiers simultaneously, and then a majority vote of the classifiers led to the prediction and evaluation of the sample. One potential idea for the different classifiers was to introduce biasing to some of the classifiers, where the weights given to the different indexes would be different. For example, one classifier would be biased towards giving speech a higher weight, while another would be biased towards giving it a lower weight. This would work if we were indexing into more than two classes, but with two classes, the combination of indexes would return the same weighted average in both cases.

3. RESULTS

The dataset we used to train and test was from GTZAN music/speech collection. Thirty different tracks for each class (music/speech) were chosen; all tracks were in Mono 16-bit wav. format sampled at 22050 Hz. We divided the selected tracks into two separate groups so that one could be used for training and the other could be for testing.

The confusion matrices generated by our trained model were analyzed shown in Table 1, and Fig.7 was created for an easier visual comparison.



	Music	Speech
Root Mean Square Music	88%	12%
Root Mean Square Speech	34%	66%
Spectrum Roll-off Music	76%	24%
Spectrum Roll-off Speech	33%	67%
Short Time Energy Music	100%	0%
Short Time Energy Speech	24%	76%
Zero Crossing Rate Music	94%	6%
Zero Crossing Rate Speech	39%	61%
Spectral Centroid Music	88%	12%
Spectral Centroid Speech	36%	64%
Spectral Flux Music	94%	6%
Spectral Flux Speech	6%	94%
Energy Entropy Music	100%	0%
Energy Entropy Speech	82%	18%
Combined Music	97%	3%
Combined Speech	0%	100%

Table 1. Confusion Matrix Collection

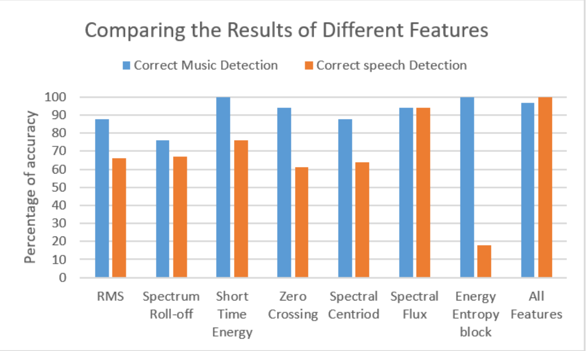


Fig.7. Visualization of Confusion Matrix

As one can see by the data, certain features were better at individual classification than others, for example Spectral Flux compared to Spectrum Roll-off. It should also be noted that the individual features were almost always more accurate with classifying music than with speech. Despite this, the concatenation of all the features was more accurately able to classify speech. With a classification accuracy of 100% for speech, and 97% for music, one can say with confidence that the trained model excels at doing its job.

3.1. Live Test

Another part of this project is to perform a live test in which the trained model will recognize a music or speech signal at the same time as it is playing. This is one important step toward the final applications of the project. There are many challenges for this goal. The first one is that we do not have all the frames of the input when we start feature extraction, so we cannot use the same code as for offline test as it uses function such as STFT and feature extraction for all these frames at same time. The other problem is that in a very noisy situation such as the poster presentation for this course some of the feature face problem making an accurate decision. Another challenge for this task is to use a computational efficient feature extraction so that this process be fast enough to be performed at the same time of the signal playing.

3.2. Graphical Interface Unit

To make an efficient environment for using our project a user-friendly interface is designed. As we can see in figure, there are two main sections in this project. The first one is Offline test in which you can add the name of your desired input signal in designed part and click on the button. It will calculate the features of the input signal and used the imported trained model to classify the input signal. The other part of this interface is the live test in which it classifies the type of input signal at the same time as the signal is playing.

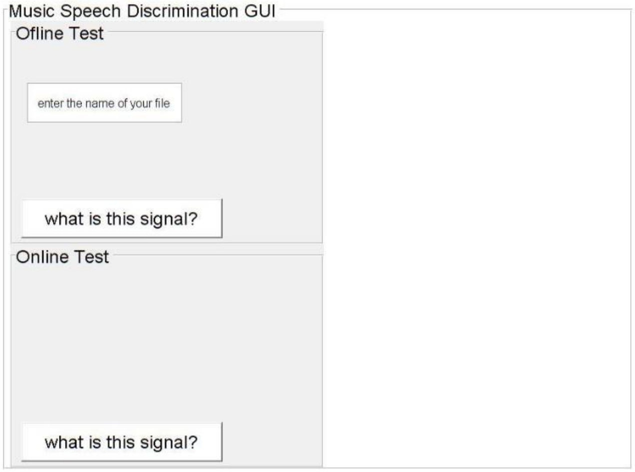


Fig.8. Graphical Interface Unit

4. CONCLUSIONS

In conclusion, our project accurately classifies between speech and music samples. However, there are several improvements possible to further increase the accuracy of discrimination. For example, this model could be improved by training it with samples containing a higher noise floor. In addition, adding more features to train would be beneficial. Furthermore, one could take this project in several different directions. One could further develop this model to discriminate between additional classes of samples, such as a Capella singing, or one could also enhance this model to separate classes in a multi-class sample, such as singing with an instrumental backing. The other step after this project is to use audio signal processing techniques to separate speech and music in a mixed input signal. In this case a pure speech or music file, be extracted from the input signal which could be a combination of speech, music and noise. This could be used as a hearing aid device to be used in noisy space or with background of noise to attenuate the undesired classes of signal and amplify the desired one. This could also be used in a live music show to delete the noise of people speaking during the show.

5. REFERENCES

[1] Al-Shoshan, Abdullah I. "Speech and music classification and separation: a review." Journal of King Saud University-Engineering Sciences 19.1 (2006): 95-132.

[2] Carey, Michael J, et al. A Comparison of Features for Speech, Music Discrimination. Ensigma Ltd., 1999, pp. 149–152.

[3] El-Maleh, Khaled, et al. "Speech/music discrimination for multimedia applications." 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). Vol. 4. IEEE, 2000.

[4] Giannakopoulos, Theodoros. "Some Basic Audio Features." MathWorks. 2014

[5] “GTZAN Music/Speech Collection.” Https://www.kaggle.com/Lnicalo/Gtzan-Musicspeech-Collection, 24 Oct. 2017.

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

[7] Scheirer, Eric, and Malcolm Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator." 1997 IEEE international conference on acoustics, speech, and signal processing. Vol. 2. IEEE, 1997.

.