Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

SPEECH & MUSIC DISCRIMINATION FOR ANALYSIS OF RADIO STATIONS

Ayesha Nishat, Antharapu Jyoshna B.tech Students,
Ms. J Stella Mary Associate Professor
Department of Electronics and Communications Engineering Bhoj Reddy Engineering College for Women, Vinaynagar, Santoshnagar X roads, Saidabad, Hyderabad - 500059

Abstract
A computationally efficient feature, called Minimum Energy Density (MED) was applied to discriminate audio signals between speech and music in the radio stations programs. The presented binary classifier is based on testing two features: energy distribution and differences between energy in channels. We analysed 240 hours of signals, from 10 Polish radio stations. Our analysis enables us to provide information about content of particular radio stations.
In this project, the speech and music discrimination is mainly focused on energy features. Since the algorithm is not meant to work in real time very long look ahead is not an issue. In order to use simple energy features because there is no need for compromising high accuracy by increasing classification resolution for radio stations analysis. In addition to the analysis of energy distribution in speech and music signals with searching for Minimum Energy Density(MED). It can analyse also the energy differences between channels. Superiority of MED over other energy based feature was shown in on benchmark database of recordings from radio and its usefulness was utilized as a part of VAD module.

I. Introduction

1.1 Introduction

Discrimination between speech and music has applications in different areas of speech processing, such as voice activity detection (VAD), automatic corpus creation and as part of modern hearing aids. For the purpose of this discrimination many features, in time as well as in frequency domain, have been proposed. The most common are 4 Hz modulation energy, entropy modulation, spectral centroid, spectral flux, zero-crossing rate. Recognition rate over 98%, has been reported for subsets of these features and their variations. Current research is focused on achieving high recognition ratewith aspect of minimizing required computations. In this project the focus is on speech/music discrimination based on energy features it analyses energy distribution in speech and music signals and upon this analysis it introduces a new feature called Minimum Energy Density (MED). We also analyze the energy differences between the audio signals and from this energy difference speech and music signals are discriminated.

1.2 Motivation

A human listener can discriminate easily between speech and music signals by listening to a short segment (i.e., few seconds) of an audio signal. An emerging multimedia application is a content-based audio and video retrieval. Audio classification is an important part of such systems. Automatic classification would remove the subjectivity inherent in the classification process and ultimately speed up the retrieval process. A major step in the design of a signal classification system is the selection of a "good" set of features that are capable of separating the signals in the feature space.
The focus is mainly on speech and music discrimination as our title name itself holds speech /music discrimination based on some energy features and can also analyse the energy difference between channels. Since the algorithm not meant to work in real time, for example WhatsApp (while we are watching a status which is of audio signal it may be music/speech and we have received a call, we answered the call, but the status which we are watching doesn't pause automatically, it generates disturbances for both listener and speaker). In addition to their analysis of energy distribution in speech and music signal is provided with starching for MED.

1.3 Objective

A computationally efficient feature, called Minimum Energy Density (MED) was applied to discriminate the audio signals speech and music in an audio clip. Thebinary classifier is based on testing two features energy distribution and differences between energy in channels. Simple energy features are used because there is no need for compromising high accuracy by increasing classification resolution for radio stations analysis.

1.4 Literature Survey

In India, private FM radio channels have developed tremendously in three phases and many more stations are yet to come. The broadcasting scenario has changed with the entrance of private FM radio stations. These are one of the most popular sources of entertainment. In this chapter, the scholar studied the various research papers, articles, blogs, books and magazines and reviewed the programming strategies, programme format and contentof the radio stations operated in India and internationally.

Scholar also reviewed the language and on-air behaviour of Radio jockeys of various radio channels.

According to the study of Pollard G. (1996), radio is a partner which is filling the silence and establishing contact with the outside world. The researcher also examined the importance of what broadcaster speaks, whyand how they speak it adds to satisfy the social and entertainment needs of listeners. Presenters always try to be spontaneous, genuine and comfortable while they are on-air. The pace of the morning programme establishes listeners' mood for the whole day and on the other hand, the latenight show pacified the tension of the day & keepsthe mood relaxed. According to the listeners, a preferred presenter is a friend whose company is pleasant and necessary.

Timmermans et.al. (2003) studied that suitable use of voice and intelligible speaking for several hours a day is an athletic activity. Due to stress, radio professionals usually smoke and drink too much coffee. The job of the radio profession demands the punctuality, deadlines, proper preparation, presentation of the programs and voice care. Radio students are also suffering from vocal hoarseness and allergy due to soft drink, beer, late night meal and their lifestyle. The researcher recommended that in every profession, vocal hygiene program/workshops must be conducted.

Glevarec H. (2005) found that the radio stations of France mainly cater to the young listeners. From 9 pm to midnight, young listeners share their problems and experiences regarding career, relationships and so on withthe presenter and ask for the solution/advice as well. It was unable to find that the field experts give them advice and a team of presenters holding this conversation with callers.

Fitzgerald R. (2006) discussed the various format of radio programmes in this study. Researcher opined that the phone-in programme format provides a chance to the listeners to share their views without requiring any technical equipment. The researcher also found that the relationship of the language is associated with the topic and the target listeners. The formal and informal style of the presenter depends upon the content and the listeners. Being a friend, the voice on the radio often provides company, relieve and hope to the listeners.

II. Classification of Audio Signals

2.1 Introduction

Audio or speech signal is usually present in analog form but with tremendous growth in digital systems now most of applications are based on human computer interactions. Human voice is in analog form but real time applications demand digital version of this data because it offers easy analysis, demands less storage space and can be transmitted up to longer distances with more security using various encryption-decryption techniques. For analysing performance of any audio signal over digital systems, the first most requirements is to convert analog information into some frames that can be passed through window function in order to judge its various parameters. In digital signal processing there are various windows that have different advantages as well as disadvantages over signal analysis; for making decision about which window is more suitable for certain audio application, it is good to perform comparative analysis. Major requirement for audio processing is to convert analog signal into digital signal with minimum or negligible losses at transmitter end so that actual information can be easily recovered at receiving end.

For maximum extraction of signal at receiver it is important to ensure that transmitted signal contains enough power so that it can travel up to longer distanced through noisy channels and can deliver accurate information at receiver. Various Digital signal processing tools are used to maintain this power level of transmitted signal; the window techniques are most powerful solution to keep control over various parameters of signal so that its power level can be retained up to desired value. This report presents comparative analysis of Hamming, Hanning and Blackman window for audio signal processing. To judge their performance, some MATLAB based simulations are performed for calculation of power spectral density of speech signal using different windows and with the help of Fast Fourier Transform, total power contained by each signal is calculated.

Blackman window so that signal with more power can be transmitted towards destination.

2.2 Sampling

The principle of digital audio signals can be easily represented in the MATLAB programming environment by means of vectors of real numbers, as is also the case with any discrete-time signal. The term discrete time refers to the fact that although in nature time runs on a continuum in the digital world we can only manipulate samples of the real- world signal that have been drawn on discrete-time instances. This process is known as sampling and it is the first stage in the creation of a digitalsignal from its real-world counterpart. The second stage isthe quantization of the samples, but for the moment we will ignore it.

To give an example, the first sample may have been taken at time 0 (the time when measurement commenced), the second sample at 0.001 s, the third one at 0.002 s, and so on. In this case, the time instances are equidistant and if we compute the difference between any two consecutive time instances the result is Ts = 0.001 s, where Ts is known as the sampling period. In other words, one sampleis drawn every Ts seconds. The inverse of Ts is the

celebrated sampling frequency, i.e. Fs = 1 Ts Hz. The sampling frequency is measured in Hz and in this example it is equal to Fs = 1 0.001 = 1000 Hz i.e. 1000 samples of the real-world signal are taken every second.

2.3 Playback Audio Signal

A useful MATLAB command, which sends a vector (stream) of samples, x,to the sound card device for playback purposes is the sound(x, fs, nbits) command, where fs is the sampling frequency based on which x will be reproduced and nbits is the number of bits that are used to represent each sample

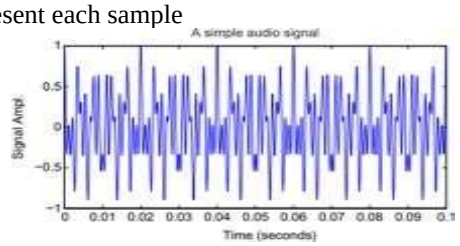

Figure :1. A synthetic audio signal.

If the sampling frequency is not provided then a default value is assumed by MATLAB (around 8 KHz). Similarly, depending on the operating environment,the default bit representation may utilize 8 or 16 bits. In order to avoid distortion duringthe playback operation, the values in x have to be in the range [−1,1].

2.4 Mono and Stereo Signals

In MATLAB, a column vector represents a single-channel (monophonic— MONO) audio signal. Similarly, a matrix with two columns refers to a two channel signal (stereophonic—STEREO), where the first column represents the left channel and the second column represents the right channel. The following code creates a STEREO signal. The left channel contains a 250 Hz tone (cosine signal) and the right channel a 450 Hz tone. Figure 2. provides a separate plot of each channel over time.

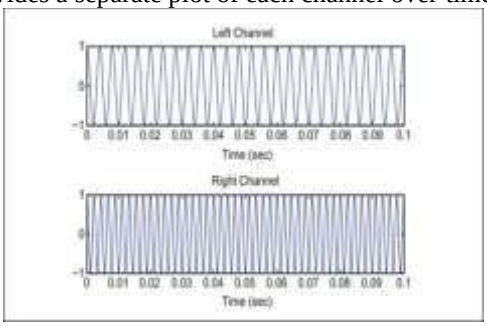

Figure :2. A STEREO audio signal.

III. Features of Audio Signals

3.1 Introduction

Feature extraction is an important audio analysis stage. In general, feature extraction is an essential processing step in pattern recognition and machine learning tasks. The goal is to extract a set of features from the dataset of interest. These features must be informative with respect to the desired properties of the original data. Feature extraction can also be viewed as a data rate reduction procedure because we want our analysis algorithms to be based on a relatively small number of features. In this case, the original data, i.e. the audio signal, is voluminousand as such, it is hard to process directly in any analysis task. So there is need to transform the initial data representation to a more suitable one, by extracting audio features that represent the properties of the original signals while reducing the volume of data. In order to achieve this goal, it is important to have a good knowledge of the application domain, so that decision of best features can be analyzed.

3.2 Short term and mid term processing

In most applications, the audio signal is analyzed by short-term processing technique according to which the audio signal is broken into possibly overlapping short- term windows (frames) and the analysis is carried out on aframe basis. The main reason why this windowing technique is usually adopted is that the audio signals are non- stationary by nature, i.e. their properties vary (usually rapidly) over time [5. More specifically, consider an audio recording consisting of a short conversation (1 s long) between two individuals, that is followed by the shout of a third person (also 1 s long). Therefore, this audio recording consists of two main events: the conversation (normal intensity signal) and the shout (high-intensity signal). It is obvious that the signal changes abruptly from the state of the conversation to the state of the shout. From a very simplified perspective, this can be considered as a change of stationarity, i.e. theproperties of the signal shift from one state to another. In such situations, it would not really make sense to compute, for example, the average intensity of the samples of the whole recording because the resulting value would be dominated by the more intense samples that were recorded during the shout of the third person. Instead, it would be more useful to break the recording into short segments and compute one value of (average) intensity per segment. This is also the main idea behind short-term processing.

Short-Term Feature Extraction

In most audio analysis and processing methods, the signal is first divided into short- term frames (windows). This approach is also employed during the feature extraction stage; the audio signal is broken into possibly overlapping frames and a set of features is computed per frame. This type of processing generates a sequence, F, of feature vectors per audio signal. The dimensionality of the feature vector depends on the nature of the adopted features. It is not uncommon to use one-dimensional features, like the

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

energy of a signal, however, in most sophisticated audio analysis applications several features are extracted and combined to form feature vectors of increased dimensionality. The extracted sequence(s) of feature vectors can then be used for subsequent processing/analysis of the audio data.

Mid-Term Windowing in Audio Feature Extraction

According to this type of processing, the audio signal is first divided into mid-term segments (windows) and then, for each segment, the short-term processing stage is carried out.1 At a next step, the feature sequence, F, which has been extracted from a mid-term segment, is used for computing feature statistics, e.g., the average value of the zero-crossing rate. In the end, each mid-term segment is represented by a set of statistics which correspond to the respective short-term feature sequences. During mid-term processing, we assume that the mid-term segments exhibit homogeneous behaviour with respect to audio type and it therefore makes sense to proceed with the extraction of statistics on a segment basis. In practice, the duration of mid-term windows typically lies in the range 1–10 s, depending on the application domain.

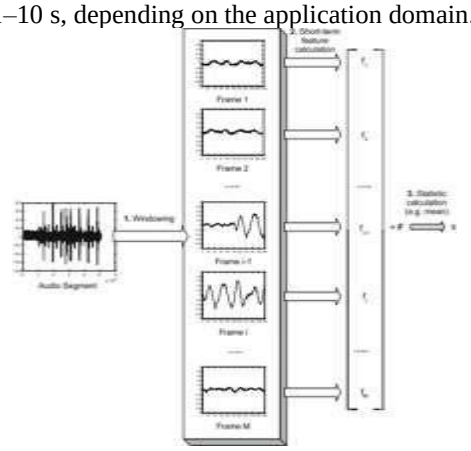

Figure :3. Mid term feature extraction

3.3 Extracting Features from an Audio File

The Function stFeatureExtraction() generates short-term feature sequences from an audio signal, and function mtFeatureExtraction() extracts sequences of mid-term statistics based on the previously extracted short-term features. We will now show how to break a large audio file (or audio stream) into mid-term windows andgenerate the respective mid- term audio statistics. If the audio file to be analysed has a very long duration, then loading all its content in one step can be prohibitive due to memory issues. This is why function readWavFile() demonstrated how to read.

IV. Block Diagram and Software Description

4.1 Introduction

This chapter includes Block diagram description and description of software. In this project the MATLAB tool is used in order to show discrimination of Music and Speech signal in terms of figures. Programming assignments in this project will almost exclusively performed in MATLAB, a widely used environment for technical computing with a focus on matrix operations

4.2 Block Diagram

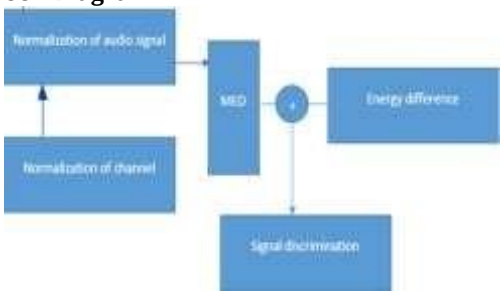

Figure :4. Block Diagram

An Audio signal is normalized which means changing its overall volume by fixed amount to reach a target level then it is provided as input to Minimum Energy Density (MED) which analysis energy distribution of Speech and Music Signals. The energy difference between channels is analysed therefore MED and Energy difference both collaborates and provides Signal discrimination.

Normalization of channel

The channel normalization operation normalizes each channel of convolutional network individually. This approach works by dividing the channels into groups and computes within each

group the mean and variance for normalization i.e. normalising the features within each group.

Unlike batch normalization, group normalization is independent of batch sizes, and also its accuracy is stable in a wide range of batch sizes. By using the channel normalization the each channel will get individually distributed over the signals after the channel normalization the each signal will be also divided into the channel in different variations where it undergoes the minimum energy density where the each signal will be differentiated with their particular frequency ranges by using the energy difference method then the discrimination of speech and music will takes place accordingly.

Normalization of audio signal

The normalize audio is to change its overall volume by a fixed amount to reach a target level. It is different from compression the changes volume over time in varying amounts. It doesn't effect dynamics like compression and ideally doesn't change. By using the normalize audio we can get the maximum volumes as well as matching volumes. Hence, normalize audio signal is provided as input to minimum energy density which analysis energy distribution of speech and music signals.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

Audio normalization is the application of a constant amount of gain to an audio recording to bring the amplitude to a target level (the norm). Because the same amount of gain is applied across the entire recording, the signal-to-noise ratio and relative dynamics are unchanged working principal types of audio normalization exist. Peak normalization adjusts the recording based on thehighest signal level present in the recording. Loudness normalization adjusts the recording based on perceived loudness.

Normalization differs from dynamic range compression, which applies varying levels of gain over a recording to fit the level within a minimum and maximum range. Normalization adjusts the gain by a constant value across the entire recording.

Minimum Energy Density(MED)

From energy distribution we can say speech has more low energy frames than music. We also know that speech has 4 Hz energy modulation, which implies four energy minima in 1 s window. These facts allow us to suspect that the presence of the frame with energy below some calculated threshold is sufficient to distinguish between speech and music. The disadvantage of this approach is inability to rely on some fixed threshold value, due to differences in signals power. To overcome that, we calculate distribution of short time frame energy inside some time window, which we refer to as normalization window. Normalization window has to be long enough to capture the nature of the signal. For example for 1 swindow seems a bad idea, since in case of window containing breathe pause we would get distribution close to uniform and information about low energy of that window would be lost. We define normalized short time frame energy as

$\bar{E}(n) = E(n)$ PN $k=1$ $E(k)$. (eq 4.1)

Energy Difference

The energy difference is the process where the music and speech are divided apart individually which is given as audio input by differentiating their density ranges as well as the frequency ranges of both the speech and music. As compared to both speech and music both the frequency ranges are different from each other so the each signal which has it's own frequency in the channel with the help of normalization.

The energy difference is distributed with the help of MED the is minimum energy density. Because the channel will carry the signals with respect to their own frequency ranges as it undergoes to the energy density process and to the energy difference depending on their density and frequency ranges then finally the signal discriminationwill takes place where the music and speech will be provided separately in the output.

Signal Discrimination

The discrimination of signals in the channel with the help of the features of audio signals, like the different signals have different frequencies as well as the feature extraction from the audio file with the help of their time domain features of both speech and music and the energy will be compared between both the signals, where pure audio contains a large amount of data but in it's original shape it is not very conductive to determining the content. It is also highly redundant and noisy in the sense that contains a high percentage of information that provides noevidence, for this reason we extract so called audio features from the audio.

V. Advantages, Disadvantages and Applications

5.1 Introduction

In this chapter it includes the advantages, disadvantages and applications of project Speech/Music Discrimination for Analysis of Radio Stations.

5.2 Advantages

- High Discrimination
- Good Communication
- Less attack of noise
- Clear Transmission
- High Signal Density
- Can separate voice and music in any application

5.3 Disadvantages

- Frequency Collaboration
- Signal Corruption

5.4 Applications

Radio broadcast monitoring
Multimedia indexing
Audio coding
Automatic speech recognition
Voice recording

VI. Results and Discussion

This Chapter includes the input signal which is an Audio Signal, Energy differences, Spectrum of input signal, Speech and Music signals.

6.1 Results

Results were obtained in the form of figures which consists of five figures, input audio signal is obtained as figure 1, energy differences is obtained as figure 2, spectrum of input signal is obtained as figure 3, Speech signal obtained as figure 4 and Music signal as figure 5.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

Input Audio Signal

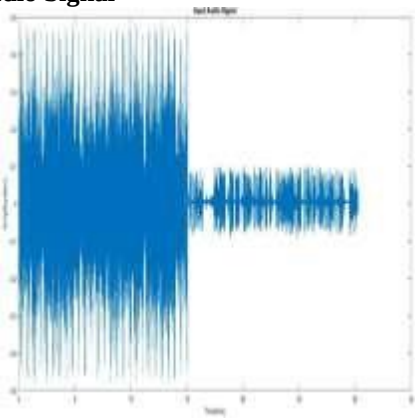

Figure :5. Input Audio Signal

Audio signals are the representation of sound, which is in the form of digital and analog signals. Their frequencies range between 20 to 20,000 Hz, and this is the lower and upper limit of our ears. Analog signals occur in electrical signals, while digital signals occur in binary representations. Here an audio signal is given as input which consist of both music and speech signals and the file name is given as test2mono.

Energy Differences

The energy difference iss the process where the music and speech are divided apart individually which is given as audio input by differentiating their density ranges as well as the frequency ranges of both the speech and music. As compared to both speech and music. As compared to both speech and music both the frequency ranges are different from each other so the each signal which has it's own frequency in the channel with the help of Normalization.

From energy distribution speech has more low energy frames than music is analysed. speech has 4 Hz energy modulation also, which implies four energy minima in 1 window. These fact allow us to suspect that the presence of the frame with energy below some calculated thresholdis sufficient to distinguish between speech and music.

Calculation of MED involves normalization of audio signal by the normalization window. This normalization window has to be long enough to capture the nature of thesignal.

We define Minimum Energy Density (MED) for k-th classification window as $MED(k) = \min\{\bar{E}(n) : (k-1) \bullet M + 1 \leq n \leq k \bullet M\}$,

Signal which is the difference between MED and ED produces a distributed function based on the ouput speech and music signal can be separated.

To find Energy density Peter D. Welch scientist created a function in spectral centroid , in our MATLAB code we are using

[Pxx, f] = Pwelch(centroid,1885, 300, 1885, fs, 'centroid', 'power')

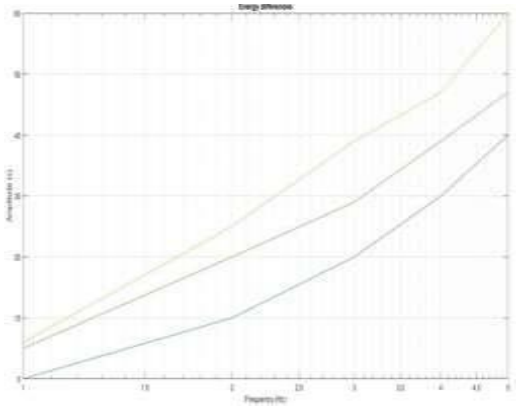

Figure:6. Energy Differences

This figure represents the energy difference of an input signal in which it consists of minimum energy density which is represented by yellow colour, Energy density which is represented by blue colour and the difference between minimum energy density and energy difference which is represented by red colour.

Spectrum of Input Signal

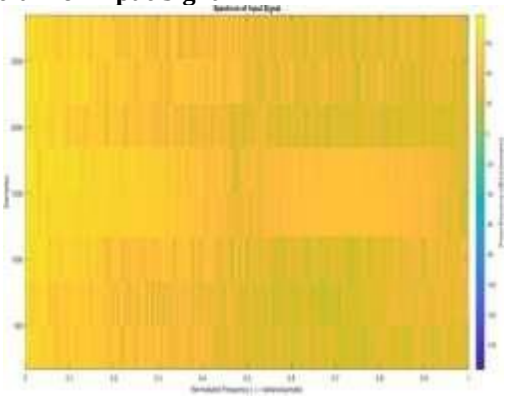

Figure:7. Spectrum of Input Signal

This figure represents spectrum of input signal in which there is combination of both the signals. In this project we have taken only one audio clip so that the spectrum is not visible but from figure we can say that the dark colour represents high frequency signals and light colour represents low frequency signals . From this spectrum we can discriminate speech signal and music signal separately.

Speech Signal

Speech signal has low energy frames than music signal. Speech signal compounded of single talker in specifictime period so it's amplitude remains same. In MATLAB code binomial probability distribution function is used.

y = binopdf(x , n , p) computes the binomial probability density function at each of the values in x using the corresponding number of trials in n and probability of success for each trial in p . x , n , and p can be vectors, matrices, or multidimensional arrays of the same size. In order to produce a speech signal it is representing x as

length of signal , n as samples and p as vector. (y = binopdf(0:1885,fs,1).

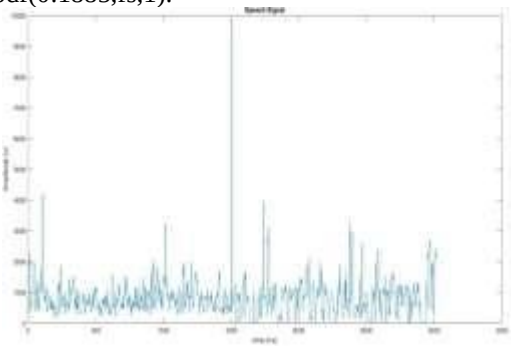

Figure:8. Speech Signal

Music Signal

Music signal has high energy frame than speech signal. A music signal is characterized by a rms voltage of 2 V and a bandwidth B of 15 kHz. A noise process adds a rms noise voltage of 4 mV to the music. In figure it has observed that there are many ups and downs of frequencies. For music signal we are using a binomial probability density function same as speech signal, in order to discriminate both music signal represented asy1=binopdf(0:1885, fs, 1).

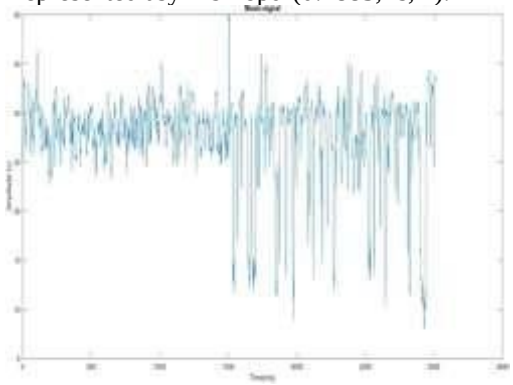

Figure:9. Music Signal

VII. Conclusion and future scope

7.1 Conclusion

The speech/music classifier which is a combination of two effective methods (MED feature and energy appearedmore precise then other tested methods. A computationally efficient feature, called Minimum EnergyDensity (MED) was applied to discriminate audio signals between speech and music in an audio clip. The binary classifier is used which is based on testing two features energy distribution and differences between energy in channels. Signal is analyzed and this analysis provides theinformation about content of particular audio clip.

The classifier is used to analyze signals of radio stations. This analysis gives insights into radio stations airplay schedule. Based on the presented classifier it is possible to identify similar types of radio stations. The classifier rating strongly depends on which recording is tested. The classifier achieves 100% accuracy if there is no acoustic signals which are the simultaneous combination of speech and music. These types of signals, appearing in commercial advertisements are difficult to classify. Our automatic classifier, based on the analysis of the normalized power variability, usually considers them as speech.

7.2 Future Scope

Speech music discrimination based on energy bands is good idea for separating the signals. But in the process,we may face energy loss which may creates data loss, proper preprocessing before classification is highly important which can be used as a future work.

Constructing a low cost real-time discriminator that could be inserted in car radio receivers could be another kind of project. In our project determination of signal is done, in future work we can determine and can also transmit the signallike removal of noise in videos and can also provide a clear video signal.

References

[1] John G. Proakis, Dimitris K. Manolakis, Digital Signal Processing, fourth ed., Pearson Education, 2019.

[2] Hung-Chen Chen, Arbee L.P. Chen, A music recommendation system based on music data grouping and user interests, in: CIKM, vol. 1, 2018, pp. 231–238.

[3] Karlheinz Brandenburg, Mp3 and aac explained,in: Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding, 2019.

[4] Ted Painter, Andreas Spanias, Perceptual coding of digital audio, Proceedings of the IEEE 88 (4) (2017) 451–515.

[5] S. Theodoridis, K. Koutroumbas, Pattern Recognition, fourth ed., Academic Press, Inc., 2018.

[6] M. Frigo, S.G. Johnson, Fftw: an adaptive software architecture for the fft, in: Proceedings of theInternational Conference on Acoustics, Speech, and, Signal Processing, 2018, pp. 1381–1384.

[7] C. Panagiotakis, G. Tziritas, A speech/music discriminator based on rms and zero crossings, IEEE Transactions on Multimedia 7 (1) (2019) 155–166.

[8] Lee Daniel Erman, An environment and system for machine understanding of connected speech (Ph.D thesis), Stanford, CA, USA, 2017.

[9] Miroslav D. Lutovac, Dejan V. Toši´c, Brian Lawrence Evans, Filter design for signal processing using MATLAB and Mathematica, Prentice Hall, 2020.

[10] Edward Kamen, Bonnie Heck, Fundamentals of Signals and Systems: With MATLAB Examples, 2019.