Audio Classification in Speech and Music by using Neural Network: Multilayer Perceptron

Ei Sandar Myint, Nwe Ni

University of Computer Studies, Yangon
becute14@gmail.com

Abstract

Audio classification can be used in many different applications. Rapid increase in the amount of audio data demands for an efficient method to automatically segment or classify audio stream based on its content. This paper focuses the attention on audio classification in music and speech. This audio classification system consists of three processing stages: feature extraction, training and classification. Spectral flux, short time energy and cepstrum coefficient are used to classify input audio into two types: speech and music. In this paper, single type of audio classification system is based on Multilayer Perceptron (MLP) neural network. Back propagation algorithm is used to perform training process. Simulation results are also included in this paper. It can classify audio files combining speech and music.

1. Introduction

Development in Internet and broadcast enables users to enjoy large amounts of multimedia content. With this rapidly increasing amount of data, user requires an efficient method to automatically aunt or classify audio stream based on its content.

Digital signal processing is the processing of signals by digital means. A signal in this context can mean a number of different things. Automatic discrimination between speech and music has become an interest topic in the last few years. A variety of system for audio segmentation and/ or classification have been proposed and implemented in the past for the needs of various applications. Each of these used different features and pattern classification techniques and describes results on different material [1].

Classification of audio signals according to their content has been a major concern in recent years. There have been many studies on audio content analysis, using different features and different methods.

Neural network is a useful for various applications that requires extensive categorization. Several applications of neural networks have been proven successful or partially successful in the areas of speech classification, character recognition, image compression, medical diagnosis, and financial and economic prediction. The power of parallel processing in neural network and their ability to classify the data based on selected features provide promising for speech classification [2].

In this paper we present an audio classification system which can classify between music and speech.

The structure of this paper is as follows. Section 2 describes related works. Section 3 introduces the necessary background theory used in this system. In Section 4, the implementation of the system is presented. Finally we conclude the paper in Section 5.

2. Related Works

In this section, the work in the literature related to audio classification is described.

Saunders [3] addresses the issue of single type audio classification for FM radio. Zero crossing rate (ZCR) and short time energy are used to classify input audio into two types: speech and music.

Scheirer and Slaney [4] use thirteen features in time, frequency, spectrum and cepstrum domains and achieve better classification. Based on Scheirer's conclusion, Carey et al. [5] compares audio features for speech and music discrimination. They find that simple audio features, such as pitch and amplitude, have significant differences between music and speech. Since then, many approaches have been proposed to classify single type audio using different audio features and classifier [6, 7, 8, 9, 10].

3. Background

This is briefly introduced the necessary background theory used in this system.

3.1 Signal Processing

Today all signal processing is done in the digital domain which leads to the fact that the signal has to be sampled before being processed, i.e. transformed from a continuous to a discrete signal by sampling the signal in equally spaced points in time.

In the following, the characteristics of different types for sounds are presented. They are the bases for audio classification [11].

Speech Characteristics

- Frequency Range 100 to 7000 Hz
- Higher silence ratios (SRs)
- Have a special pitch

Music Characteristics

- Frequency range from 16 Hz to 16000 Hz
- "Loud" music has energy 10 kHz and above.
- Long harmonic tracks
- Rare periods of silence.

3.2. Feature Extraction

Feature extraction is the process of converting and audio signal into a sequence of feature vectors carrying characteristics information about the signal. These vector are used by classification algorithms.

There have two categories of audio features. They are Time-Domain Features and Frequency-Domain Features. Time-Domain Features are short time energy (STE), zero crossing rate (ZCR), high zero crossing rate ratio (HZCRR) and etc. And then, Frequency-Domain Features are spectral flux (SF), Cepstrum coefficient, mel frequency cepstrum coefficients (MFCC) and etc.

Short time energy

Short time energy is the discrimination between silence and non silence. It indicates how rapidly changes the frequency spectrum, with particular attention to the low frequencies (up to 2.5 kHz), and it generally assumes higher values for speech. The equations for short time energy [12] is,

$$E_n = \frac{1}{N}\sum_m [x(m)]^2 \quad (1)$$

where, x(m) = discrete time audio signal
n = time index of the short time energy

Spectral Flux

This feature measures frame-to-frame spectral difference. Thus, it characterizes the changes in the shape of the spectrum. Speech goes through more drastic frame-to-frame changes than music. Speech alternates periods of transition and periods of relative stasis, whereas music typically has a more constant rate of change. As the result the spectral flux value is higher for speech, particularly unvoiced speech, than it is for music [4].

$$X_f = \sum_{f=1}^{m} (S(f) - S(f-1))^2 \quad (2)$$

where, S = input signal
f = frame number
m = no of sample

Cepstrum Coefficient

The cepstrum coefficients, evaluated using

$$c(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi} \log |X(e^{jw})| e^{jwn} dw \quad (3)$$

Cepstrum coefficients $c_j(n)$, suggested [14] as good speech detectors, have been computed for each frame, then the mean value c_μ (n) and the standard deviation $c_\sigma(n)$ have been calculated

$$C_\mu = c_\mu(9).c_\mu(11).c_\mu(13) \quad (4)$$

$$C_\sigma = c_\sigma(2).c_\sigma(5).c_\sigma(9).c_\sigma(12) \quad (5)$$

3.3. Neural Network

A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects. Knowledge is acquired by the network from its environment through a learning process. Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge. The network consists of a set of source nodes that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes [15].

3.3.1 Multilayer Perceptrons (MLP)

The input signal propagates through the network in a forward direction, on a layer-by-layer basis. These neural networks are commonly referred to as multilayer perceptrons, which represent a generalization of the single layer perceptron.

MLP has been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the back-propagation algorithm. This algorithm is based on the error-correction learning rule [15].

3.3.2 Back Propagation Algorithm

The back-propagation algorithm is a supervised learning procedure which involves the representation of a set of pairs of input and output patterns. Back propagation algorithm is the most widely used algorithm to perform training, particularly on large problems. It is a method to find weights for a

multilayered feed forward network. It consists of two passes through the different layers of the network: a forward pass and backward pass. In the forward pass, an activity pattern (input vector) is applied to the sensory nodes of the network, and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the backward pass, the synaptic weights are all adjusted in accordance with an error-correction rule [15].

4. System Implementation

This section presents the implementation of the system and simulation results are also included. This system is implemented by MATLAB 7.0.4.

4.1. System Architecture

In this section we present a single type of audio classification system based on Multi-Layer-Perceptron (MLP) neural network. We focus the attention on audio classification in music and speech.

System architecture is as shown in Figure 1.

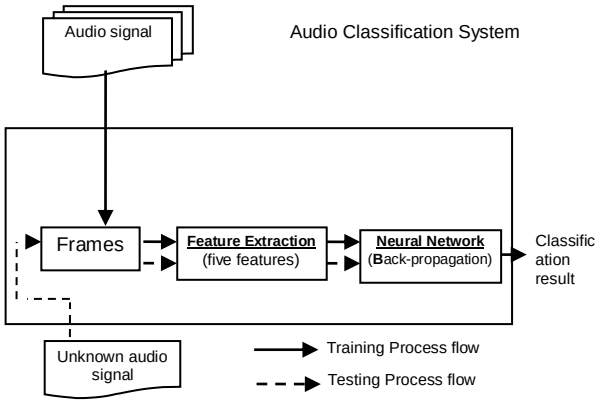

Figure 1. System Architecture

Figure 1 shows the feature extraction and the training or testing processes of this system. First, audio signal is divided into frames which are applied for feature extraction. The extracted features are trained in Neural Network by using back-propagation algorithm. In testing process, unknown audio signal is divided into frames as inputs which are applied for feature extractions. These features are used to classify with Neural Network. The result may be Speech or Music.

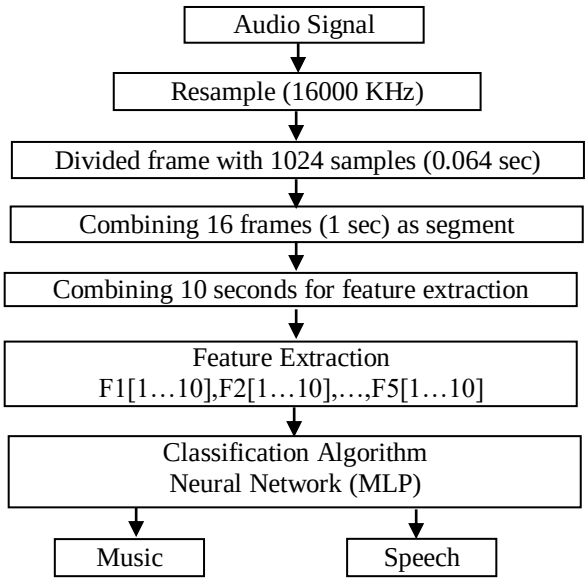

Figure 2. Overview of the system

In figure 2, audio signals with 3 minutes long files are accepted as input files. There have been computed considering 16 frames by 1024 points each (sampling frequency f_s = 16000 Hz), with a total observing time of about 1second. And then 16 frames (16*0.064≈1sec) are extracted to get the features for these frames. These 16 frames are considered as one segment (1sec). And then 10 segments (10 sec) are combined for feature extraction. Because features (F1, F2,..., F5) are extracted from each segment, feature vectors: F1[1…10], F2[1…10],…, F5[1…10] are obtained for 10 seconds.

A short description of the five selected features (such as F1, F2, F3, F4, F5) are as follows. Feature F1 is computed by the spectral flux. Equation (1) is used in this feature. As an example, the value of F1 feature vector is shown in Table 1.

Table 1: Feature for spectral flux(F1[1…10])

Music	Speech
119.18	407.2
126.44	571.34
166.38	297.82
133.43	182.04
107.51	271.08
110.86	181.19
99.227	374.56
149.82	34.683
127.56	650.82
153.01	119.36

Feature F2 and F3 are related to the short-time energy. Feature F2 is computed as the standard deviation of the absolute value of the signal, and it is

generally higher in speech. Equation (1) is used for this feature. Feature F3 is the minimum of the short-time energy. Equation (1) is this used in this feature. Features F4 and F5 are computed by using cepstrum coefficients equations (4) and (5).

These features are used as the inputs of neural network for training process. We make the samples 3 minutes long audio files for each training file. Feature vectors are extracted from music (such as country, classical, rap, jazz, etc.) and speech (from BBC news). There are two output neurons corresponding to speech and music. Output value for music is set as 10 and output value for speech is set is 01.

Simplified neural network architecture for this classification system is shown in Figure 3.

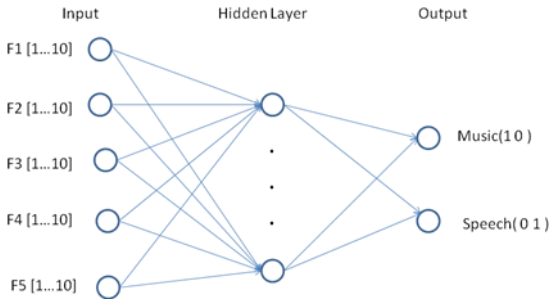

Figure 3. Simplified Neural Network Architecture for speech and music classification

After training this neural network can classify as music or speech for unknown audio signal. The selection of the test files is carried out "manually," that is, each file is composed of many pieces of different types of audio (different speakers over different environmental noise, different kinds of music such as country, classical, pop, rap, jazz, etc.) concatenated in order to have a three minutes segment of speech followed by a five minutes segment of music, and so on. All the contents of speech files have been recorded from BBC news.

4.2. Simulation Results

In this work, is developed an audio classification system that discriminates between speech and music segments in broadcast news audio. In simulation, to train audio files, 37 files for music (different types of pure music such as classic, country, pop and jazz) and 30 files for speech (news from BBC) are used. All files are 3 minutes long audio files. In this paper, the speech with 30 files (540 training data) is used to train in neural network. The music with 37 files (711 training data) is used to train in neural network.

And unknown audio files are extracted into ten second long audio signals. Feature vectors from each extracted file are fed into the neural network to classify speech and music.

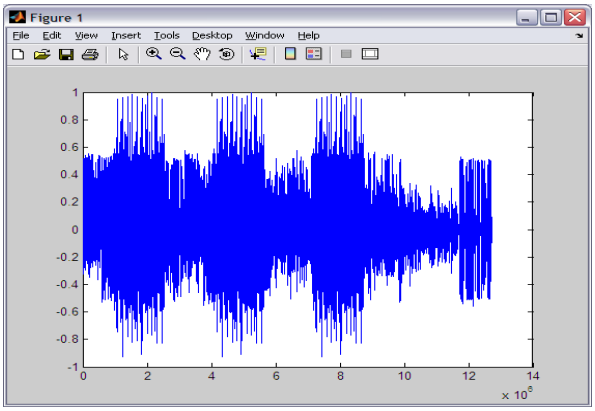

Figure 4. input audio signal which contains music and speech (5 min long)

Audio file (' .wav') including music and speech signal is shown in Figure 4. Music signals have variations of amplitude (high and low) and in speech signals, it goes average amplitude. Figure 4 shows the input audio signal which contains music and speech for 5 minutes long. Unknown audio may be speech/music alone or may be the combination of both. It may be varying in length.

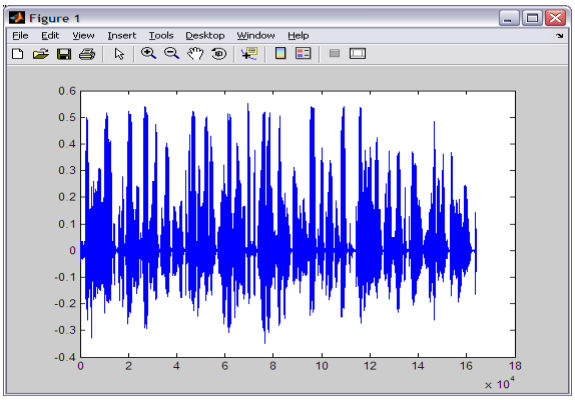

Figure 5. Extracted signal (10 sec long)

Figure 5 shows the signals extracted from the input signals shown in Figure 4. Extracted signals are round about 10 seconds long. Features from above signals are extracted and then tested in the neural network. Classified output result of input signal (Figure 4) is shown in Figure 6.

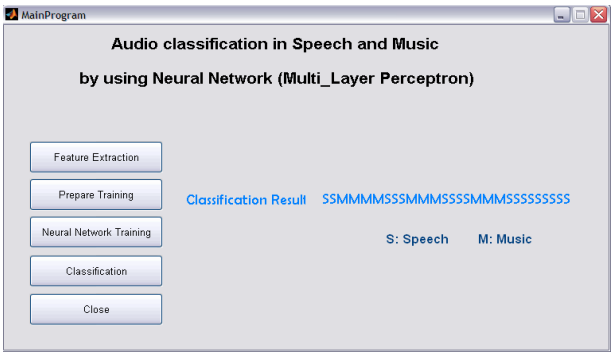

Figure 6. Classified output result

The accuracy of the classification process is calculated based on the percentage of correctly identified outputs. Evaluation of classification system is based on a number of criteria. Classification accuracy is the most important measure of system performance. The audio classification accuracy is computed using the following formula:

$$Accuracy = 100 * \frac{x}{y} \quad (5)$$

where, x = no. of audio signals classified accurately

y = total number of audio signals fed

Accuracy of the simulation results for 2 training data sets is shown in table 2. In this table, pop music which does not including in training is analyzed for testing phase.

Table 2. Simulation results of testing accuracy

Data Set	No. of hidden neurons	Testing (Music)	Testing (Speech)
Set 1(827 training data)	4	99.725%	100%
	8	97.624%	100%
Set 2 (1184 training data)	4	100%	100%
	8	97.659%	100%

5. Conclusion

In this work, the audio classification system is implemented by applying Multilayer Perceptron neural network approach. It can discriminate between speech and music segments. According to table 2, accuracy for speech is higher than accuracy of the music. In training audio files are the pure speech and pure music files. Therefore classification will be classified for only those types of input signals. Signals are segmented into 10 seconds long signals. So, the short signal less than 10 seconds will not be classified correctly. This system is not suitable to classify background music. Speech superimposed on music cannot be classified in this system.

Reference

[1] Bugatti, A. Flammini and P. Migliorati, "*Audio Classification in Speech and Music: A Comparison Between a Statistical and a Neural Approach*", EURASIP Journal on Applied Signal Processing 2002:4, 372–378.

[2] P. Vanroose, *"Blind Source Separation Of Speech And Background Music For Improved Speech Recognition"*, the E.U. 5th Framework project "MUSA", 2008.

[3] J. Saunders, "Real-time discrimination of broadcast speech/music," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, pp. 993–996, Atlanta, Ga, USA, May 1996.

[4] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 1331–1334, Munich, Germany, April 1997.

[5] M.J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *ICASSP*, April 1999.

[6] "Contentbased classification, search, and retrieval of audio," *IEEE Multimedia*, 1996.

[7] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal "Speech/music discrimination for multimedia applications," in *ICASSP*, June 2000.

[8] L. Lu, H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," in *Proc. 9th ACM Int. Conf.on Multimedia*, 2001.

[9] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," IEEE Trans. on Speech and Audio Processing, 2000.

[10] S.-Z. Li and G. Guo, "Content-based audio classification and retrieval by support vector machines," in *PRCM (invited talk)*, 2000.

[11] G. Lu and T. Hankinson, " *A Technique towards Automatic Audio Classification and Retrieval"* Gippsland School of Computing and Information Technology , Monash University, Churchill, Vic 3842, Australia.

[12] G. Theodoros, K. Dimitrios, A. Andreas, and T. Sergios, *" Violence Content Classification Using Audio Features"* Department of Informatics and Telecommunications, National and Kapodistrian, University of Athens, Panepistimiopolis, Ilissia Athens 15784.

[13] L. Rabiner and B. Juang, *Fundamentals of peech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[14] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[15] S. Haykin "Neural Networks , a comprehensive foundation, second edition", Upper Saddle River, NJ 07458, 1999.