Audio Classification Based on Content Features

Dr.Ayad A. Abdulsalam
University of Baghdad - College of Education for Women - Computer Department
ydsalam@yahoo.com

Abstract

Audio classification is the process to classify different audio types according to contents. It is implemented in a large variety of real world problems, all classification applications allowed the target subjects to be viewed as a specific type of audio and hence, there is a variety in the audio types and every type has to be treatedcarefully according to its significant properties.Feature extraction is an important process for audio classification. This workintroduces several sets of features according to the type, two types of audio (datasets) were studied. Two different features sets are proposed: (i) firstorder gradient feature vector, and (ii) Local roughness feature vector, the experimentsshowed that the results are competitive to those gotten from other popular methods inthis field, such as Zero Crossing Rate (ZCR), Amplitude Descriptor (AD), Short Time Energy (STE), and Volume (Vo). The test results indicated, that the attained averageaccuracy of classification is improved up to94.9232% for training set and 95.8666%for testing set.The classification performance of these two extracted featuresets is studied individually, and then they used together as one feature set. Theiroverall performance is investigated, the test results showed that the proposed methods give high classification rates for the audio.

Keywords: Multimedia, Audio classification, Feature extraction, Short time energy, Local Roughness features, First Order Gradient Feature.

تصنيف الصوت استنادا إلى ميزات المحتوى

د.اياد عبدالقهار عبدالسلام
جامعة بغداد - كلية التربية للبنات - قسم الحاسبات
ydsalam@yahoo.com

الخلاصة

تصنيف الصوت هو عملية عزل أنواع الصوت المختلفة بمجموعات وفقا لمحتوياتها، ويستخدم هذا التصنيف على مجموعة كبيرة ومتنوعة من مشاكل العالم الحقيقي، حيث تعمل جميع التطبيقات على ادراج ملفات الصوت المستهدفة تحت نوع معين من الصوتيتم تعريفه مسبقا، وبالتالي، هناك مجموعة كبيرةمن أنواع الصوت وكل نوع يجب أن يعامل بعناية وفقا للخصائص المميزة لهذا النوع. استخلاص الميزات هو عملية هامة لتصنيف الصوت. هذا العمل يقدم عدة مجموعات من الميزات وفقا لانواع الاصوات، تم تطبيق الدراسة على مجموعتين من الاصوات القياسية العالمية. وقمنا باقتراح مجموعتين مختلفتين من الميزات: (1) متجه سمات التدرج من الدرجة الأولى، و (2) متجهات خشونة الموضع المحلية، أظهرت التجارب أن النتائج مشجعة وهي افضل من تلك التي تم الحصول عليها من الأساليب التقليدية الأخرى والمطبقة على نفس مجموعة اصوات الاختبار مثل Zero Crossing Rate (ZCR), Amplitude Descriptor (AD), Short Time Energy (STE), and Volume (Vo). وأظهرت نتائج الاختبار أن متوسط الدقة في التصنيف تم تحسينه إلى 94.9232٪ لمجموعة التدريب و 95.8666٪ لمجموعة الاختبار. حيث تم دراسة أداء تصنيف هاتين الميزتين

المستخلصتين بشكل فردي، ثم استخدمتا معا كميزة واحدة. تم التحقق منالأداء العام، وأظهرت نتائج الاختبار أن الطرق المقترحة تعطي معدلات تصنيف عالية للصوت.

INTRODUCTION

Fast increasing of audio data files and growing the amounts of publicly available audio data, demand for practical indexing with efficient tools to enable users to retrieve the data directly,and to avoid the classical way for search files sequentially and test by hearing which spend time and efforts. Therefore, audio files classification based on features has been an interested field of research for various applications include audio segmentation, automatic speech recognition, music information retrieval,general purpose sound recognition and acoustic surveillance.

Typically, researchers developed and extracted many features for particular tasks and domains,later these features are employed for other tasks in other domains, based on these observations, we conclude that audio features may be considered independently from their original application domain.Some features are developed in this work, combined with selected a manifold set of state of the art features from literatures. The major criterion for selection is the maximization of heterogeneitybetween the features in relation to what information they carry andhow they are computed.

AUDIO REPRESENTATION

Some general attributes should be clarify, first is the distinguishing between tones and noise. Tones are characterizedby the fact that they are “capable of exciting an auditory sensation havingpitch” [1] while noise not necessarily has a pitch. Tones may be puretones or complex tones. A pure tone is a sound wave where “the instantaneoussound pressure of which is a simple sinusoidal function in time” while a complextone contains “sinusoidal components of different frequencies” [1].The spectral composition of noise is important for its characterization. Wedistinguish between broad-band noise and narrow-band noise. Broad-band noiseusually has no pitch while narrow-band noise may stimulate pitch perception.

From a psychoacoustic point of view, all types of audio signals may be describedin terms of the following attributes: duration, loudness, pitch, and timbre.

Duration is the time between the start and the end of the audio signal of interest [2]. The temporal extent of a sound may be divided into attack, decay,sustain, and release depending on the envelope of the sound. Not all soundsnecessarily have all four phases. Note that in certain cases silence (absence ofaudio signals) may be of interest as well [2].

Loudness is an auditory sensation mainly related to sound pressure levelchanges induced by the producing signal. Loudness is commonly defined as“that attribute of auditory sensation in terms of which sounds can be orderedon a scale extending from soft to loud” with the unit *sone* [1].

Pitch: the American Standards Association defines (spectral) pitch as “that attributeof auditory sensation in terms of which sounds may be ordered on ascale extending from low to high” with the unit *mel* [3]. However, pitch hasseveral meanings in literature. It is often used synonymously with the fundamentalfrequency. An attribute related to pitch is pitch strength. Pitch strength is the “subjectivemagnitude of the auditory sensation related to pitch” [1]. Pitch perception is not only affected by the frequency content of a sound, but also by the sound pressure and the waveform [4, 5].

Timbre is the most complex attribute of sounds, according to the ANSIstandard timbre is “that attribute of auditory sensation which enables a listenerto judge that two non-identical sounds, similarly presented and havingthe same loudness and pitch, are dissimilar.” [6]. For

example, timbre reflectsthe difference between hearing sensations evoked by different musical instrumentsplaying the same musical note (e.g. piano and violin).
Timbre is a high-dimensional audio attribute and is influenced by both stationaryand non-stationary patterns. It takes the distribution of energy in thecritical bands into account (e.g. the tonal or noise-like character of sound andits harmonics structure). Furthermore, timbre perception involves any aspectof sound that changes over time (changes of the spectral envelope and temporal characteristics, such as attack, decay, sustain, and release). Preceding andfollowing sounds influence timbre as well [7].

Some features that we exploit are extracted from the shapes of waveform or spectrum of audio, an audio waveform is a time domain display, a display of amplitude vs time, while audio spectrum is a frequency domain display, a display of amplitude vs frequency[8], Figure(1) shows displaying in time and frequency domain.

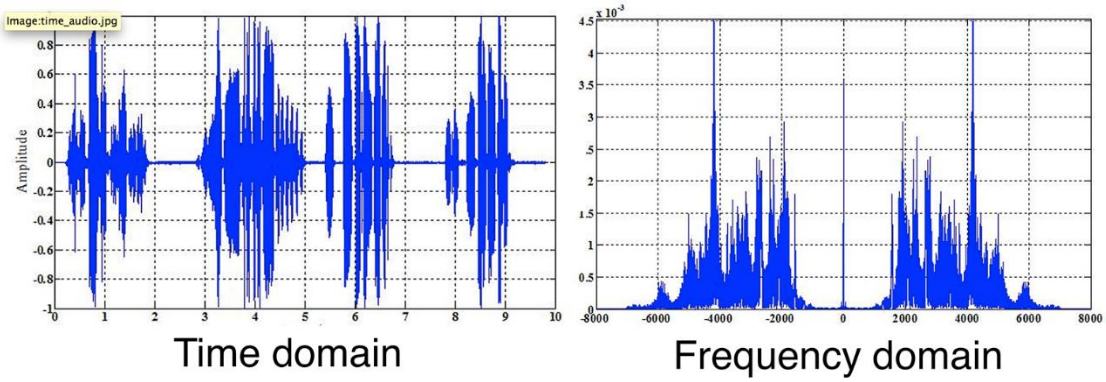

Figure (1): Time and Frequency Domain of Audio display.

PROPOSED SYSTEM LAYOUT

The proposed system consists of three main modules, as shown in Figure (2). Modules consist of several steps. The workflow of the proposed system is given in the following strategy:

Stage 1) Audio preprocessing part: this stage is used to generate a standard audio format, to be ready for extracting features, then extract features from audio objects stored in an audio database. Feature extraction aims to reducing the amount of data and extracting meaningful information from the signal for a particular retrieval task. It passesthrough the following two main sub-steps:
a. Noise removing by using standard audio filters.
b. Audio Normalization: to transform the audio waves to a standard form, with uniform width and height (dimension compensation).
The features are extracted once from all objects in the database and stored in a features database.
Stage 2) Query part is the main interface between user and system, it's used for formulating queries, there are different types of queries. Usually, the user provides the system with a query that contains one or more audio objects of interest. After formulation of a query, features are extracted from the query object(s) by the same procedure as in the first stage. The resulting features haveto be compared to the features stored in the features database in order to find objects with similar properties.
Stage 3) Retrieval part, Recognition and Verification or matching stage, to make a decision about the claimed individual identity, depending on the strength of the extracted features, the proposed system tested the matching accuracy before and after encoding to estimate the difference between the two cases. The crucial step in the retrieval module is similarity comparison which estimates the similarity of different feature-based media descriptions.

Similarity judgments usually base on distance measurements. The vector space model is using in this work. The basic assumption of this model is that the numeric values of a feature may be regarded as a vector in a high-dimensional space. Consequently, each feature vector denotes one position in this vector space. Distances between feature vectors may be measured by Euclidean distance metric [9]. Similarity measurement is performed by mapping distances in the vector space to similarities. We expect that similar content is represented by feature vectors that are spatially close in the vector space.

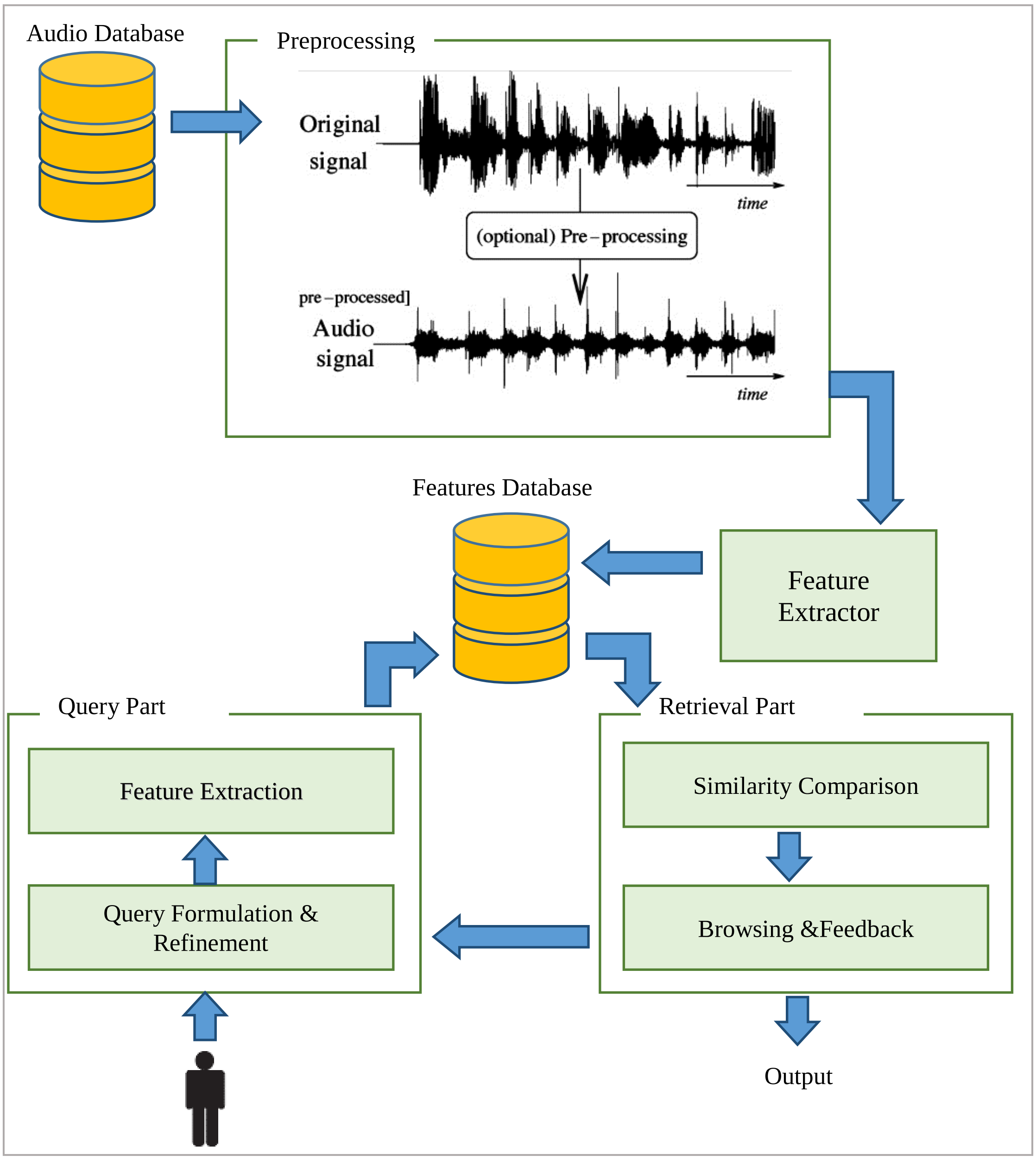

Figure (2): Proposed System Structure

FEATURE EXTRACTION

Feature extraction is the task that every machine learning and pattern recognition systems contain. In pattern recognition, the concept feature means a function of single or setof measurements, that quantifies a property orcharacteristic of an object. Featureextraction is a particular form of data represented in meaningful way to reduce thesize of these data, and describe them accurately [10], it represents a critical stage, because it deals with how toextract optimal feature that describes audio content essentially, hence, it is one ofimportant challenge of the computer multimedia issues. In the proposed system, many feature types in audio field are dealt with, like temporal features that extracted from the temporal domain which is the native domain for audio signals, all temporal features have in common that they are extracted directly from the raw audio signal, without any preceding transformation, consequently, the computational complexity of temporal features tends to be low. And physical frequency features, the group of frequency domain features is the largest group of audio features, all features in this group have in common that they live in frequency or autocorrelation domain.The following features are used in proposed system:

Zero Crossing Rate (ZCR): One of the cheapest and simplest features is the zero crossing rate, which is defined as the number of zero crossings in the temporal domain within one second [11]. This feature has been used heavily in both speech recognition and music information retrieval, ZCR is defined formally as in equation (1):

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1}(s_t\, s_{t-1}) \qquad (1)$$

wheresis a signal of length *T*, in some cases only the "positive-going" or "negative-going" crossings are counted, rather than all the crossings since, logically, between a pair of adjacent positive zero-crossings there must be one and only one negative zero-crossing.

Amplitude Descriptor (AD): The amplitude is the height from the center line to the peak (or to the trough), or we can measure the height from highest to lowest points and divide that by 2 [11]. The amplitude descriptor separates the signalinto segments with low and high amplitude by an adaptive threshold (a levelcrossingoperation). The duration, variation of duration, and energy of thesesegments make up the descriptor. AD characterizes the waveform envelope interms of quiet and loud segments. It allows to distinguish sounds with characteristicwaveform envelopes.

Short Time Energy (STE):The energy associated with speech is time varying in nature. Hence the interest for any automatic processing of speech is to know how the energy is varying with time and to be more specific, energy associated with short term region of speech [11]. By the nature of production, the speech signal consists of voiced, unvoiced and silence regions. Further the energy associated with voiced region is large compared to unvoiced region and silence region will not have least or negligible energy. Thus short term energy can be used for voiced, unvoiced and silence classification of speech as shown in figure (3).
Let the samples in a frame of speech are given by ***"n=0 to n=N-1"***, where ***" N "*** is the length of frame (samples), then for energy computation the speech will be zero outside the frame length. Then for energy computation amplitude of the speech samples will be zero outside the frame [12]. Accordingly we can write above mentioned relation as in equation (2).

$$E_T = \sum_{n=0}^{N-1} s^2(n) \qquad (2)$$

Volume (Vo): Volume is a popular feature in audio retrieval, for example in silence detection and speech/music segmentation [11]. Volume is sometimes called loudness, as in [12]. We use the term loudness for features that model human sensation of loudness. Volume is usually approximated by the root-mean-square (RMS) of the signal magnitude within a frame. Consequently, volume is the square root of STE. Both, volume and STE reveal the magnitude variation over time [11].

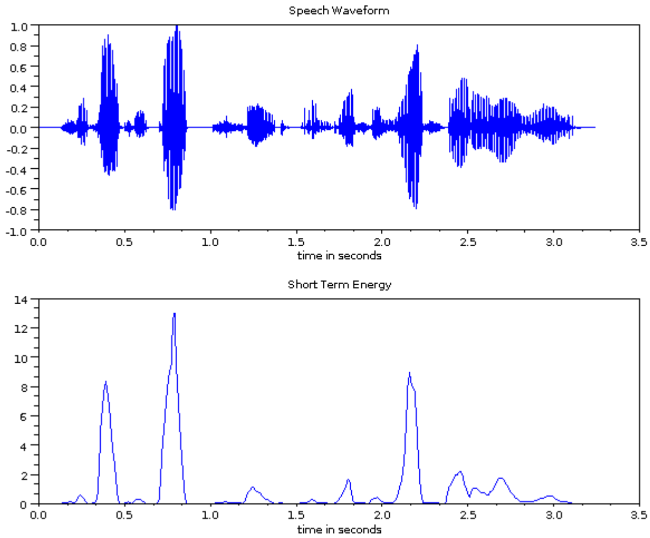

Figure (3): Short term energy for the speech signal

First Order Gradient Feature Vector F1: It is a suggested features of a set of first order derivatives in audio signals are implemented using the magnitude of the gradient. For a function S(x,y), the gradient of S at coordinates (x,y) is defined as the two- dimensional vector as in equation (3):

$$\text{Grad (S)= } [G_xG_y] \qquad (3)$$

Where G_x and G_y are the horizontal and vertical derivatives, respectively. This vector holds geometrical information that points to the direction of the greatest rate of change in S at (x,y). At each point in the audio signal, the resulting gradient approximations can be combined to give the gradient magnitude, using formula in equation (4):

$$G = \sqrt{G_x^2 + G_y^2} \qquad (4)$$

Local Roughness Features F2: It's other suggested feature depends on a roughness which is a component of wave surface texture. It is measured by the deviations in the wave element of the normal vector of real surface from its ideal form. The surface is considers as rough surface when the deviations are large, otherwise it is smooth. The not noisy waves have high heterogeneity nature, so the roughness is different from part to part. Local roughness features are suggested here to cover this issue. The set of features depends on the local variations in the wave values relative to time element. The original wave is divided into slices of time, the differences between the wave samples center and the samples surrounding it considered as local roughness measure. After the local sample differences are computed for each central element, two vectors are constructed to hold the minimum and maximum of these difference respectively.

TEST RESULTS

In order to demonstrate the efficiency of the proposed classification system for the audio files, a number of experiments have been performed, proposed method examined on 2280selectedsamples that collected from (Audio database of UMass Amherst Libraries), and(Gilmore Music database fromYale University Library), that distributed over 12 selected classes, each class consists of 190 samples. This dataset is specified for researchers to study the details of textural features of audio.

In the first experiment, 25 training samples are selected randomly for each class, then all 190 samples are tested. Three stages are performed, first stage traditional features vectors (ZCR, AD, STE and VO) are using, second stage F1 proposed feature vector is added to the first collection, third stage F2 proposed feature vector is added to the second stage, Table (1) shows the percentage of correctly classified under the different features collections.

Table (1): The percentage of correctly classified samples under the different features collections, using 25 training samples for each class.

Class	*ZCR, AD, STE and VO*		*ZCR, AD, STE , VO and F1*		*ZCR, AD, STE, VO, F1, and F2*	
	Training%	*Testing%*	*Training%*	*Testing%*	*Training%*	*Testing%*
1	88	90.30303	92	93.333333	96	96.969697
2	84	89.090909	92	92.727273	92	95.757576
3	88	88.484848	88	91.515152	92	95.151515
4	88	89.69697	96	93.939394	100	97.575758
5	80	87.272727	84	90.909091	92	95.151515
6	84	86.666667	92	93.939394	92	96.969697
7	92	87.272727	92	89.090909	92	93.939394
8	92	86.060606	96	88.484848	96	94.545455
9	84	86.666667	96	87.878788	96	92.727273
10	80	84.84848	88	86.060606	92	90.909091
11	92	88.484848	96	93.939394	100	97.575758
12	88	85.454545	96	90.909091	96	96.363636
Average	86.66667	87.52525	92.33333	91.06061	94.66667	95.30303

Training samples increased in second experiment to be 40 samples, also they are selected randomly for each class, and same stages with the features collections are treated. Table (2) shows the average of percentages of correctly classified data set.

Table (2): The average of percentages of correctly classified samples under the different features collections, using 40 training samples for each class.

Class	*ZCR, AD, STE and VO*		*ZCR, AD, STE , VO and F1*		*ZCR, AD, STE, VO, F1, and F2*	
	Training%	***Testing%***	***Training%***	***Testing%***	***Training%***	***Testing%***
Average	89.31231	89.7621	91.27372	91.8282	94.9232	95.8666

DISCUSSION AND CONCLUSIONS

Table (1) shows the percentage of correctly classified samples of all the tested samples under the use of four traditional features, then two suggested features (First Order Gradient Feature Vector F1, and Local Roughness Features F2) respectively. Results are:

With four traditional features model the highest percentage of correctly classified sample achieved with 25 random selected training samples, the values are 86.66667%for training and 87.52525% for testing.

With the adding F1 features vector model the highest percentage of correctly classified sample achieved with 25 random selected training samples, and the values are 92.33333% for training and 91.06061% for testing. At the same time there is a higher result when adding F2 features vector, the accuracy values will be 94.66667% for training and 95.30303% for testing.
Table (2) shows the average of correctly classified samples by using same procedure in first experiment with 40 random selected training samples per class. The results are more stable.

We conclude that the suggested features increase the classification rate, and the accuracy was as maximum as possible when all six features are implemented.

REFERENCES

[1] ANSI. Bioacoustical Terminology, ANSI S3.20-1995 (R2003). AmericanNational Standards Institute, New York, 1995.
[2] H. Jiang, J. Bai, S. Zhang, and B. Xu. "Svm-based audio scene classification".In Proceedings of the IEEE International Conference on NaturalLanguage Processing and Knowledge Engineering, Wuhan,China, Oct. IEEE, 2005.
[3] C. Papaodysseus,G.Roussopoulos,D. Fragoulis, T. Panagopoulos, and C. Alexiou, "A New Approach to the Automatic Recognition of Musical Recordings", J. Audio Eng. Soc., vol. 49, 2001.
[4] Pedro Cano and Eloi Batlle, "A Review of Audio Fingerprinting", Journal of VLSI Signal Processing, Springer Science + Business Media, Inc. 2005.
[5] G. Agostini, M. Longari, and E. Pollastri. "Musical instrument timbresclassification with spectral features",In Proceedings of the IEEE Workshop on Multimedia Signal Processing, Cannes, France, IEEE,Oct.2001.
[6] J.J. Aucouturier, F. Pachet, and M Sandler. "The way it sounds: timbremodels for analysis and retrieval of music signals", IEEE Transactions on Multimedia, Dec. 2005.
[7] R. Cai, L. Lu, A. Hanjalic, H.J. Zhang, and L.H. Cai. "A flexible frameworkfor key audio effects detection and auditory context inference", IEEETransactions on Speech and Audio Processing, May 2006.
[8] S. Esmaili, S. Krishnan, and K. Raahemifar. "Content based audio classificationand retrieval using joint time-frequency analysis", In Proceedings ofthe IEEE International

Conference on Acoustics, Speech, and Signal Processing, volume 5, Montreal, Canada, IEEE, May 2014.
[9] Y. Zhu and M.S. Kankanhalli. “Precise pitch profile feature extractionfrom musical audio for key detection”,IEEE Transactions on Multimedia, Jun. 2006.
[10]C. Panagiotakis and G. Tziritas.” A speech/music discriminator based on rms and zero-crossings”, IEEE Transactions on Multimedia, Feb. 2015.
[11] T. Wold, D. Blum, and J. Wheaton. “Content-based classification, search, and retrieval of audio”. IEEE Multimedia, 2006.
[12] DaliborMitrovic, Christian Breiteneder, “ Features for Content-Based Audio Retrieval”, Vienna University of Technology, Advances in Computers Vol. 78, 2010.