An attribute-based approach to audio description applied to segmenting vocal sections in popular music songs

Shiva Sundaram and Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory (SAIL),
Dept. of Electrical Engineering-Systems, University of Southern California,
3740 McClintock Ave, EEB 400, Los Angeles, CA 90089.
Email: ssundara@usc.edu, shri@sipi.usc.edu

***Abstract*— We present a descriptive approach for analyzing audio scenes that can comprise a mixture of audio sources. We apply this method to segment popular music songs into vocal and non-vocal sections. Unlike existing methods that directly rely on within-class feature similarities of acoustic sources, the proposed data-driven system is based on a training set where the acoustic sources are grouped by their perceptual or semantic attributes. Our audio analysis approach is based on a quantitative time-varying metric to measure the interaction between acoustic sources present in a scene developed using pattern recognition methods. Using the proposed system that is trained on a general sound effects library, we achieve less than ten percent vocal-section segmentation error and less than five percent false alarm rates when evaluated on a database of popular music recordings that spans four different genres (rock, hiphop, pop, and easy listening).**

I. Introduction

The increased focus on automatic audio processing techniques for retrieval, indexing, and classification is a result of vastly improved digital media storage, delivery of content over the network, and the availability of cheaper and efficient computing. Audio processing for the above tasks involves one or more of the following related procedures: segmentation, clustering, and classification. Segmentation involves marking similar, homogeneous sections of audio. The work we present here segments popular music songs into vocal and non-vocal sections. This is primarily useful for audio information retrieval applications for annotation, browsing, summarization and creating audio thumbnails. Our implementation is based on developing a method for descriptive characterization of audio content through a data-driven approach. Typically, a data-driven approach to discerning the type (classification) and the temporal extent of an audio event (segmentation) is based on some model-free or model-based approach, in either case involving clustering and learning the characteristics of the desired audio classes [1]. The present work differs from other approaches by grouping them based on perceived description and not simply on their signal level similarities.

An obvious and widely used approach to characterizing a complex audio scene is by recognizing each of the possible constituent classes through deterministic or statistical methods. This would especially work well for closed-set problems where the number of identifiable classes are limited and a given audio event is always known to belong to one of the previously known classes. In more open problems, where the number of acoustic events and possible audio scenes are large, and perhaps unseen, it would become tremendously complex to implement such a scheme. For example more heuristic rules, such as the one implemented in [2] for classification would be required. Related work to this problem of content analysis of audio is presented in [3], [5]. Usually in these implementations, each acoustic source is labeled to be a unique audio class and the core challenge in the problem is to identify these sources. Also in methods such as in [3], [8], [10] the discrimination system and the method of analysis are based on the underlying assumption that the two classes are non-overlapping in time. Related work in description based approaches to audio typically involve application-tailored choice of descriptions trained on corresponding data. For example in [6], the author uses probabilistic descriptions specific to music. The approach proposed in [7] involves directly tying semantic level descriptions to signal level Gaussian Mixture Models built on extracted feature vectors. These approaches, are relevant to the present work, but rely on statistical similarity amongst acoustic sources and are not easily generalizable to large classes of acoustic sources.

Audio categorization implicitly or explicitly involves a change-point detection (segmentation) scheme. This is done by either recognition, classification or measuring signal statistics localized in time. In [2], [3] the authors have implemented a set of heuristic rules to classify audio into speech, music, silence and environmental sound and thus segment them. In [11] the authors have used a peak-picking scheme on the derivative of a distance signal. In [10] a classification scheme by a statistical measure of zero-crossing rate (ZCR) and root-mean-squared (RMS) of signal energy is used for segmentation. In [4] the authors have used a T^2-statistic alongwith the Bayeisan Information Criterion (BIC) to the related problem of speaker turn detection. All these techniques use only information at the signal level to perform the segmentation task. No higher level understanding of the audio is utilized.

Our main motivation is based on the change point detection scheme of the human auditory system. It can perform audio segmentation with ease and reasonable accuracy [12]. This

can be attributed to the fact that the auditory system as a whole depends not only on signal level characteristics but also on the semantic understanding, relevance and temporal/spatial placement of acoustic events. The segmentation task is easy because, in some way, the auditory system is able to quantitatively measure the interaction of the events presented to it in a scene. Thus when the interaction changes, a change-point in this quantitative measure can be detected.

Based on this idea, we propose to use *perceived audio descriptors*, or attributes, to measure and process the interaction of acoustic sources quantitatively and implement a change-point detection scheme. We start by grouping a general audio dataset of sound effects into high level attributes based on how humans interpret and describe audio. Then we propose a metric to quantitatively characterize a given audio clip in terms of these attributes. Finally we derive a change point detection rule based on this quantitative metric and evaluate it on segmenting vocal sections in popular music.

The next two sections describe the overall approach to the segmentation problem followed by its implementation. Experiments performed to test the accuracy are described in section IV, with the results in section V. Finally, the conclusion, additional discussion and ideas about our future work are presented in Section VI.

II. System Description

Although an audio scene may consist numerous acoustic sources, each identified by a unique linguistic name, many of them share similar perceptual qualities and thus they can be grouped under one category. For the purposes of this work, we focused on three such high level attributes, namely, *speech-like*, *harmonic* and *noise-like*. For the training data, we manually categorized the relevant audio data available in the BBC sound effects library [16] based on these three attributes. The clips were grouped according to the way they are interpreted after listening to them. We refer to these attributes as the perceived audio descriptors. For example, the sound of a vacuum cleaner and the sound of a car's engine are both considered *noise-like*. Similarly many other acoustic sources (e.g.: waves in a seashore, machine-shop tools, heavy rain, breathing sounds), have such *noise-like* characteristics. Thus, each of these sources can be grouped based on these perceived attributes regardless of the linguistic label/class or just based on their signal feature similarity/dissimilarity. Along the same lines, a wide variety of acoustic sources such as door bells, string musical instruments such as violin, guitar, (excluding percussion instruments), telephone ringtones sirens, pure tones etc. can be categorized under the group *harmonic* i.e. sources that are harmonically rich. *Speech-like* mainly covers individual speech, conversations in a crowd, laughter, and human vocalizations. Further narrow, additional descriptions are also possible. These three high level attributes were chosen because they are sufficiently distinct human interpretations of events in an auditory scene and they are sufficiently "separable" using perceptual signal-level features for the automatic classifiers.

This grouping, based on human characterization of the signal, leads to a mapping in a relatively lower dimensional representation space. We believe such a description offers scalability, and can be also effectively used as an intermediate step for conventional audio processing methods.

In practice, one can assume that the time-varying description of any audio scene typically contains acoustic events that can be considered semantically grouped under this broad description scheme. We construct this identification using a bank of classifiers (as detectors), where each one focuses on a specific attribute such as harmonicity. Then, the time series of these classifier outputs are assimilated to provide a final categorization of the audio (illustrated in Figure 1 and explained further in Section III). The descriptive labels are automatically assigned to each frame of audio by using standard pattern classifiers that are trained off-line. The technique is based on broad definitions of various audio attributes, and the training data is also chosen accordingly. For the task of tracking these attributes, we implemented a k-Nearest Neighbour (k-NN) Classifier [15]. Discrimination based on k-NN rule for classification task has been investigated previously, and found useful, in speech/music classification tasks [8]. An additional reason for choosing it here was because it belongs to a class of lazy learning algorithms, and suits our approach which makes no assumption regarding signal feature similarities of the target attributes in the feature space.

To measure the degree of interaction between the audio attributes (descriptors) at any given time, we define the quantity *activity rate* (AR). It is defined as the number of each event (e.g., noise-like/speech-like/harmonic) detected per unit time of analysis. For example, for sections of audio with just music, the harmonic activity rate is expected to be high whereas for sections with singing/dialogues in a scene, the speech activity rate is expected to be high. The complexity of audio scenes can be described in terms of the different acoustic sources or events present in it. Thus in effect, the activity rate provides an aggregate quantitative measure of interaction of the individual events. Note that in classical speech/music discriminators, it is usually assumed that segments of speech and music are non-overlapping in time and the problem is to statistically measure the signal properties appropriately to discern one from the other (such as through a maximum likelihood scheme). In the present work, we make no such assumptions of temporal mutual exclusivity and we segment the audio based on the structure of the audio scene measured by the audio descriptors.

To evaluate the proposed scheme, we apply it to the problem of segmenting popular music songs into vocal and non-vocal sections. Note that such audio is rich characterized by vocals of one or more main singers, and possibly other background singers, along with polyphonic instrumentals. It should be pointed out that while the results indicate satisfactory segmentation of vocal sections, the framework provides a generalized approach to analyze the underlying acoustic structure in a given audio clip [9]. The chosen evaluation domain reflects one such application where obtaining sections with the vocals in a song is the goal. The next section provides details about

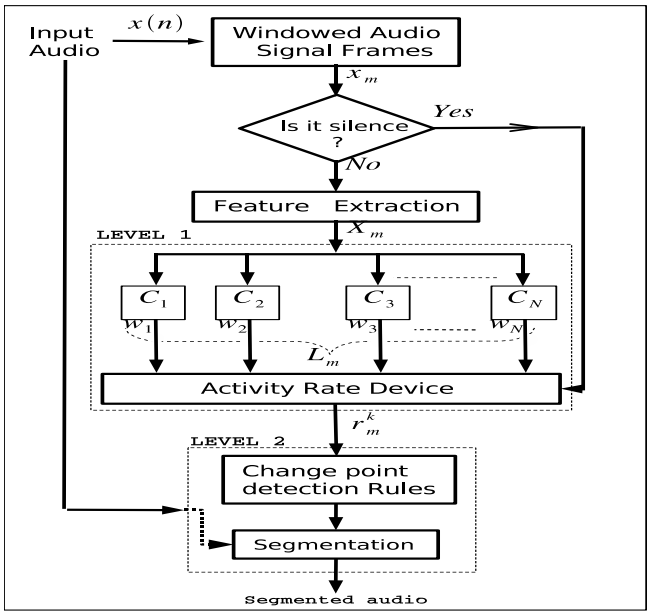

Fig. 1. Illustration of the proposed system for audio segmentation

the implementation of the proposed system.

III. Implementation

The proposed system for segmentation is depicted in Figure 1. The overall system is split into two blocks: *Level 1* and *Level 2*. The first stage of the system is a feature extraction stage. It maps a given windowed audio frame x_m of T_s duration (usually from 20 to 100 msec.) to a point X_m in a D dimensional feature space Ω (m being the time-index). The extracted feature is a popular, perceptually-motivated 37 dimensional vector comprising of 13 Mel Frequency Cepstral Coefficients (MFCC), its delta-MFCC(DMFCC) and delta-delta (DDMFCC). The features are relevant here because they model the perception of the human auditory system and they have also been successfully applied in recognition of general audio ([13] and references therein). This is used both during the training and testing. For an audio scene of T duration, this stage generates a sequence of X_m, $m \in \{1, \ldots, M\}$ (discrete-time) feature vectors. The silence segments are treated separately using the root mean-squared energy (RMS) of the signal. **Level 1** comprises N classifiers trained in a one-against all scheme. The output of the k^{th} classifier C_k is w_k, an element of the N dimensional vector L_m and $w_k \in \{0, 1\}$, $\forall\ k \in \{1, \ldots, N\}$. C_k is trained to identify (the classification process) the k^{th} label in a frame. $w_k = 1$ indicates that the classifier has classified the given frame with the k^{th} label. For example, suppose the k^{th} label is *noise-like*. Then the classifier C_i is trained on all the data that contains audio clips of acoustic sources that are *noisy* (eg: engine noise, vacuum cleaner, hair dryer, waves on a seashore etc.) and $w_k = 1$ means that the given frame is *noise-like*. In our implementation of the proposed system $N = 3$ and the labels used are $w_1 \equiv$ *Speech-like*, $w_2 \equiv$ *harmonic* and $w_3 \equiv$ *noise-like*.

The training data for the three groups totalled 7.28 hours (about 2.42 hours for each group). The clips were available as 44.1 kHz, 16-bit 2 channel uncompressed audio. For the feature extraction stage, the clips were converted to mono without changing the sampling rate. As mentioned previously, the classifiers C_1, C_2, C_3 were implemented using the non-parametric k- nearest neighbor classifier ($k = 5$) scheme [15]. Note that the realization of the $N = 3$ classifiers can be combined into one block, but they have been shown separately for the sake of clarity.

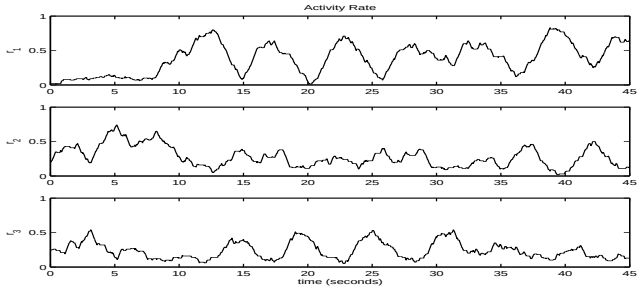

Fig. 2. The time-varying r_m vector for a $T = 45$ sec. song clip. For this particular illustration, $T_s = 20$ msec and $M_r = 100 \times T_s$

At this point, for a sequence of audio frames for an audio scene of T duration, a sequence of vectors L_m $m \in \{1, 2, \ldots, M\}$ This is the input to the Activity Rate module. If a sequence of L_m vectors is time-aligned (similar to a spectrogram where the magnitude of a sequence of Fourier coefficients are time aligned), then the number of detected events in each dimension of the L_m vector in the time-aligned representation for a given time period is a measure of the event activity rate (AR). Mathematically it can be written as:

$$r_m^k = \frac{1}{M_r} \left(\sum_{j=m-\frac{M_r}{2}}^{j=m+\frac{M_r}{2}} I\{w_{k,j} = 1\} \right), \tag{1}$$

$$r_m = (r_m^1, r_m^2, \ldots, r_m^N) \tag{2}$$

where $I\{\cdot\}$ is an indicator function and $w_{k,j}$ is the k^{th} dimension of L_m. M_r is the duration for measuring the AR and typically $T_s < M_r << T$ and $M_r \approx 50$ to 100 times T_s.

The output of the AR device is again a $N = 3$ dimensional signal r_m that takes continuous values between 0.0 and 1.0. A value close to 0.0 indicates no activity and 1.0 indicates high activity in terms of the corresponding decriptor. As an example, Figure 2 shows the r_m vector for a 45 second clip of the song *Don't Know Why* by Norah Jones. In this clip, the song starts with soft music (comprising a piano, a stringed instrument, and snare drums) and the singer starts singing at the 10^{th} second. Studying the activity rate plot, it can be seen that the speech-like activity rate (r_1) increases around the 10^{th} second and the harmonic activity rate (r_2) is high initially and lower when the singer starts to sing. As the singer sings each verse of the song, the rate (r_1) alternates between high and low values. Note a different trend in the plot between the 26^{th} and 36^{th} second, as compared to the segment between the 15^{th} and 25^{th} second. This is observed because the singer actually sings the second part of the first verse differently from the first part. Similarly in the r_2 plot, the piano notes briefly come to the foreground in the clip between the 36^{th} and 38^{th} second indicating fluctuations in

the detected harmonic activity rate. The trends in the noise-like activity rate (r_3) also brings out such variations as the song proceeds. If a windowed audio segment is determined to be silence (by appropriate thresholding of the RMS energy signal) then the frame's corresponding L_m vector is not used to calculate the activity rate.

Thus, it can be seen that the interaction between the acoustic sources in an audio scene can be quantitatively measured using the activity rate. This is the *descriptive* approach presented in this paper. The time variations in the activity rate (AR) signals quantitatively describe the audio scene locally in time. The description is generalized in the sense that the N categories cover the type of acoustic source irrespective of the actual signal statistics. While direct correlation between the trends in the Activity Rate to changes in the audio clip can be drawn by audio-visual inspection, we provide one such secondary analysis for segmenting popular music songs.

Level 2: This level focuses on using the proposed attribute-based audio descriptors for specific categorical classification. Specifically, we consider the application of segmenting audio into vocal and non-vocal sections (generically, referred from here on as 'speech-like' and 'non-speech like', although it includes both sung and spoken forms). In frame based analysis, we observe an audio segment W of T_w duration as a sequence of independent frames of short duration. Let this set of observed frames be given by the set $\{X_1, X_2, \ldots, X_R\}$. In practice each observation $X_i, i \in \{1, \ldots, R\}$ is a point in the feature space and represents about 20 to 50 msecs of windowed audio samples. A typical probabilistic detection scheme involves estimating $P(W$=Speech-like$)$ and $P(W$=non-Speech like$)$ which gives us the rule,

$$\text{if, P}(W\text{=Speech-like}) > \text{P}(W\text{=non-Speech like}) \quad \text{then } W \text{ is a vocal section}$$

Let $I\{X_i = \textit{speech-like}\}$ represent classification of a frame X_i as *speech-like*. In an observation of only R vectors, $P(W$=$Speech)$ can be estimated by the classical definition [14],

$$\text{P(W=Speech like)} = \frac{\sum_{\forall i}(I\{X_i=\textit{speech-like}\})}{R}, \text{and}$$

$$\text{P(W=non-Speech like)} = \frac{\sum_{\forall i}(I\{X_i=\textit{harmonic-like}\})}{R} + \frac{\sum_{\forall i}(I\{X_i=\textit{noise-like}\})}{R}$$

Thus from the inequality 3 we get, for a given R for a segment W we conclude that it is a vocal segment if

$$\sum_{\forall i} I\{X_i=\text{speech-like}\} > \left(\begin{array}{c} \sum_{\forall i} I\{X_i=\text{harmonic-like}\} \\ +\sum_{\forall i} I\{X_i=\text{noise-like}\} \end{array} \right)$$

Therefore we obtain a time-localized voting scheme for detection of a vocalized segment in a test clip. By the definition of the event activity rate (AR) in equation 1 we get the rule for change point detection : *if* $s > 0$ *then segment is a vocal section*, where,

$$s = r_1 - r_2 - r_3, \quad \text{i.e,}$$

$$s = (\textit{speech-like AR}) - (\textit{harmonic AR}) - (\textit{noise-like AR})$$

TABLE I

K-NN CLASS. ACCURACY (%) *speech-like(S-l), harmonic(H), noise-like(N-l)*

	%split Train/Test Size = 80/20			%split Train/Test Size = 90/10		
classified as →	*S-l*	*H*	*N-l*	*S-l*	*H*	*N-l*
S-l	94.44	2.96	2.60	94.72	2.66	2.62
H	1.08	97.30	1.62	0.94	97.56	1.50
N-l	0.42	0.67	98.91	0.47	0.55	98.98

In the implementation, we use $R = 100$ (from the value set for M_r). Note that this combination highlights the segments that contain the vocals of the song. It basically draws inference based on a set of individual classification results. The classification accuracy of the k-NN classifier for two set of instances is listed in Table I. We also include a "fill-in" procedure where two segments that are not more than 1 sec. apart are combined [13]. Other combination schemes are also possible depending on the application. For example, for larger N, the r_m can be directly used as a low-dimensional feature vector for classifying auditory scenes. The next section discusses the performance of this rule to mark sections of the songs with the vocals.

IV. EXPERIMENTS

A collection of 67 full-length assorted songs were used to asses the segmentation performance of the proposed system. The tracks belonged to 4 genres: Rock, Pop, Hip-Hop and Easy Listening, covering a variety of artists like *Red Hot Chilli Peppers, Cake, Bangles, Chris Isaak, Wyclef Jean, Mark Knopfler, Tom Waits, Enya* etc. The uncompressed audio tracks were directly obtained as mono from the original commercially available CDs at 44.1 kHz sampling rate.

For each track, a binary signal of the same duration was obtained at the output. The binary signal has a value of 1 for sections of audio determined to have vocals, and 0 otherwise. This output was then converted into time markers that mark the corresponding sections on a waveform. A Graphical User Interface (GUI) was used to display the waveform and the markers aligned in time. The *error* and *false alarm* percentage were calculated manually by listening to the songs and checking the actual vocal sections of the songs against sections marked by the system. The error and false alarm rate were calculated by using the following formulae:

$$\%\text{Error} = \frac{\text{no. of vocal sections in a song not marked by the system}}{\text{total no. of segments in the song}}$$

$$\%\text{ False Alarm} = \frac{\text{no. of non-vocal sections marked by the system}}{\text{total no. of segments in the song}}$$

The tolerance for segmentation was 1 sec. i.e., if the markers were off by ± 1 sec., then it was not considered as an error. This tolerance value can be determined from the size of M_r to calculate the activity rate. In the proposed implementation $T_s = 20$ msec. and $M_r = 100 \times T_s$ with 50% overlap between frames for estimating r_m^k. It is of the same order of acceptance measure suggested by the authors in [12].

V. RESULTS AND DISCUSSION

The error and false alarm rates for segmentation of the songs using the proposed system, grouped by genre, is tabulated in Table II. It can be seen that the estimates are consistent through the various genres. The overall false alarm and error rate

TABLE II

FALSE ALARM & ERROR RATES OF POPULAR MUSIC SEGMENTATION

Genre	No. of Songs	% False Alarm	% Error
Rock	14	4.28	5.97
Pop	20	5.09	6.63
Hip-Hop	13	3.39	8.45
Easy Listening	20	3.37	6.66

was found to be about 4.0% and 7.0% respectively. Although direct comparision with other speech/music discriminators is not possible due to differences in training data, implementation and domain of application our results are comparable to other systems such as [8], [10]. While the results show low error values, some specific problems arose in certain cases causing relatively high error. They are discussed below:

Observed sources of error: In some Rock songs, there were sections with high intensity music (lead guitars+drums) along with screaming voices that were not segmented correctly. This is because relatively high values were obtained for all the three Activity Rate signals r_1, r_2, r_3 and these resulted in low (approximately 0) values of s (refer to eqn. 1). Also, certain instances of extreme pitch levels of voices were not correctly marked. Similar problems with segmentations arose in songs of other genres that had extended duration of singing notes, accompanied by a musical instrument such as trumpet or violin. In certain cases, due to fade in and fade out of the singer's voices the exact time instant of the segmentation boundaries was off by a few seconds (termed as *border effect* [13]).

VI. Conclusions and Future Work

In this work, we proposed an attribute-based approach to quantitatively measure the changes in an audio scene and applied it to segment popular music tracks into sections with and without vocals. The idea is motivated by the fact that the human auditory system can instantly identify changes in the scene by tracking changes in the interaction of the different acoustic sources. The work in this paper presents *activity rate* (AR) as a metric to quantitatively measure the interaction of different sources. This measure does not identify individual sources, but measures the activity of different semantically-related attributes of sources over time. These are based on how the sources are perceived, and not necessarily on the similarities in signal properties. We then use these attributes for categorical classification; specifically we consider segmenting music into vocal and non-vocal sections. We perform the segmentation without assuming that the vocal and non-vocal sections of audio are non-overlapping in time.

As mentioned previously the framework presented here is not limited to building a binary speech/music discriminator type system. It provides a way to analyze the underlying acoustic structure in a given audio clip and also a way to annotate and highlight relevant sections. This is very useful for audio summarization and thumbnailing applications. The application of this system is not limited to segmentation. As an example, the activity rate (AR) signals can be used to organize a large audio database. A user can make a query to such a database using the small-dimensional AR signals and the relevant audio clips can be returned to the user. Of course, one would require a larger dimension ($N > 3$) description for this application and the categories need to be appropriately chosen. Splitting the large category *noise-like* into *machine-noise* (e.g: engines noise, vacuum cleaner, hair dryer etc.) and *non-machine-noise* (e.g: seashore, breathing sound, heavy rain, clothes rustling etc.), or having additional categories such as *impulsive* sounds (e.g: explosions, gunfire, knocking, clock ticks etc.) are some ways to increase the dimensionality of the descriptions. Our future work would involve further investigating this categorization of acoustic sources through perceptual/language descriptions similar to the ideas presented here.

The main principle which sets the stage for this work is that humans can describe auditory scenes using language which is a representation of the semantic information captured from an audio clip. Analysis of more complex and rich scenes with large number of acoustic sources can be potentially implemented by increasing the number of audio descriptors and seeking quantitative measures such as the activity rate to adequately characterize them. This is a topic of our on going work.

References

[1] M. Davy and S.J. Godsill, "Audio Information Retrieval: A bibliographical Study", Tech. Report CUED/F-INFENG/TR.429, Signal Proc. Group, Cambridge Univ. Eng. Dept, February 2002.

[2] L. Liu, H.J. Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation," IEEE Trans. on Speech and Audio Processing, Vol.10,No.7, October, 2002.

[3] T. Zhang, J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and classification," in IEEE Trans. on Speech and Audio Processing Vol.9, No.4, May, 2001.

[4] B. Zhou, J. H.L. Hansen "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion," In Proc. ICSLP (2000), Beijing, China. October 16-20, 2000.

[5] G. Guo and S. Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector Machines,"in IEEE Trans. on Neural Networks, Vol.14, No.1, January, 2003.

[6] B. Whitman, "Semantic Rank Reduction of Music," IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics, New Paltz, NY, USA. October 19-22, 2003

[7] M. Slaney, "Semantic-Audio Retrieval," Presented at the ICASSP, Orlando, Florida, USA. May 13-17,2002.

[8] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," In Proc. of the ICASSP, Munich, Germany. April,1997

[9] K.D. Martin,E.D. Scheirer, and B.L. Vercoe, "Music content analysis through models of audition," In Proc., ACM Multimedia '98 Workshop on Content Process of Music for Multimedia App., Bristol UK, (Sept.1998).

[10] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," in IEEE Trans. on Multimedia, Vol.7, No.1, February, 2005.

[11] G. Tzanetakis and P. Cook, "A Framework for Audio Analysis based on Classification and Temporal Segmentation", In Proc. Euromicro, Workshop on Music Technology and Audio processing, Milan, Italy, 1999

[12] G. Tzanetakis and P. Cook, "Multifeature Audio Segmentation for Browsing and Annotation," In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 1999.

[13] D. Li, I.k. Sethi,N. Dimitrova,T. McGee, "Classification of general audio data for content-based retrieval," in Pattern Recog. Letters, Vol.22, 533-544, 2001.

[14] A. Papoulis, S. U. Pillai "Probability, Random Variables and Stochastic Processes," 4^{th} Edition, McGraw Hill.

[15] R. O. Duda, P. E. Hart, D.G. Stork , "Pattern Classification," Wiley-Interscience; 2nd edition,October, 2000

[16] "The BBC Sound Effects Library- Original Series." http://www.sound-ideas.com