INTRODUCTION

The rapid advancement of multimedia tools has enabled easy capturing and editing of multimedia information including audio data. Distribution of audio data has become easier due to the growth of internet technology and availability of huge bandwidth. Furthermore, various compact coding scheme has reduced the storage and transmission cost. As a result, it is quite easy to have a large repository of multimedia data. But, efficient management of such voluminous data is very crucial to provide convenient mechanism for storing, browsing and navigation of desired data. As a result, content-based multimedia data retrieval has become an active area of research for efficient access of desired piece of data. A lot of works have been directed towards the development of content-based image and video retrieval system. Comparatively, little work has been done on the audio portion [24]. Research activities on content-based audio data management can be categorized as audio classification, audio retrieval and indexing. Among these, automatic classification is the fundamental step for any such application involving audio database.

Storing the audio data in an organized manner can act as the first step for the applications like search and retrieval, indexing. Audio data may be of various types like speech, music. Like any other collection, audio database also can be structured according to the type of its content. For example, in a digital music library, depending on the requirement data may be classified as music with voice (song) and without voice. Further the song collection may be organized based on the genre. Thus, whole collection can be put into the hierarchical form. In such context, the classification of audio data becomes an important issue and it may be carried out manually. Keeping the large volume of data in mind, an automated system for classification of audio data is desirable. Moreover, such a system will enable convenient browsing and retrieval of similar audio clips in the database against the query given by the user in the form of an audio clip.

Currently available audio search engines rely mainly on file names and embedded metadata, and do not make any use of the acoustic signal characteristics. Automatic audio signal classification system should be able to categorize the audio database into different classes like speech, instrumental, song which allows the development of useful tools for automatic organization of audio databases, segmentation of audio streams, intelligent signal analysis, intelligent audio coding, automatic bandwidth allocation, automatic equalization, automatic control of sound dynamics etc. Thus, an efficient system for automatic classification of audio signal is very much in demand.

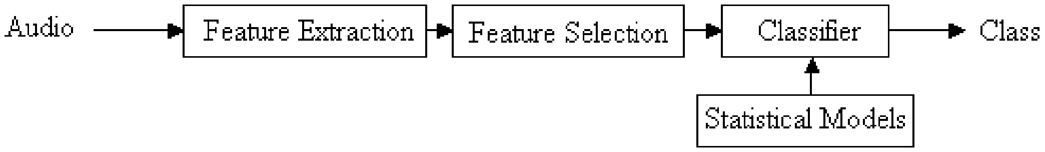

Figure 1: Block diagram of an automatic audio signal classification system

To get an understanding of content based audio classification many area like feature extraction and pattern classification have to be studied. A system extracts a set of features from the input signal to form the feature vector characterizing the content. These vectors are then fed to a classifier that employs certain rules to assign a class to the incoming vector. Figure 1. shows the block diagram of the automatic audio classification system.

AUDIO FEATURES

Before any audio signal can be classified under a given class, the features in that audio signal are to be extracted. These features will decide the class of the signal. Feature extraction involves the analysis of the input of the audio signal. The feature extraction techniques can be classified as temporal analysis and spectral analysis technique. Temporal analysis uses the waveform of the audio signal itself for analysis. Spectral analysis utilizes spectral representation of the audio signal for analysis.

A key issue in the development of an audio classification algorithm is the design of audio features that can be used to discriminate different audio classes. A good feature is expected to have a large interclass difference while maintaining a small intra-class difference. Due to the complexity of human audio perception, extraction of discriminative feature is very difficult. No feature has yet been designed which has 100% discriminative power between the classes. Though, high classification accuracy can be achieved with a combination of a set of features.

Audio features are commonly extracted from short-term frame level and the long-term clip level [19]. The concept of audio frame comes from traditional speech signal processing, where a frame usually covers a length of around 10 to 40 ms within which the signal is assumed to be stationary. However, to reveal the semantic meanings of an audio signal, an analysis over a longer interval is more appropriate. A signal with such a long interval is called an audio clip, usually with a length ranging from one second to several tens of seconds. Clip-level features usually describe how frame-level features change over a time window. Depending on different applications, the length of an audio clip could be fixed or not. Audio frames and clips may overlap in time with their predecessors. Some common frame-level and clip-level features are introduced below. If x(t) is the signal, then it is divided into N frames, each of length r, and can be denoted as $\{x_i(t) : 1 \leq t \leq r, \quad 1 \leq i \leq N\}$. However, to understand the semantic meanings of an audio signal, an analysis over a longer interval is needed. A signal with such a long interval is called an audio clip; usually the length ranges from one second to several tens of seconds. Clip-level features reflect how frame-level features change over a time window. Very often frame level features are analyzed to derive the clip level features. Audio frames and clips may overlap in time with their predecessor. Dividing a signal into frames and clips may be taken as the pre-processing step as shown in Figure 2.

As feature extraction plays an important role in classification of an audio signal. Hence it becomes more important to select the features that help the classification process more efficient.

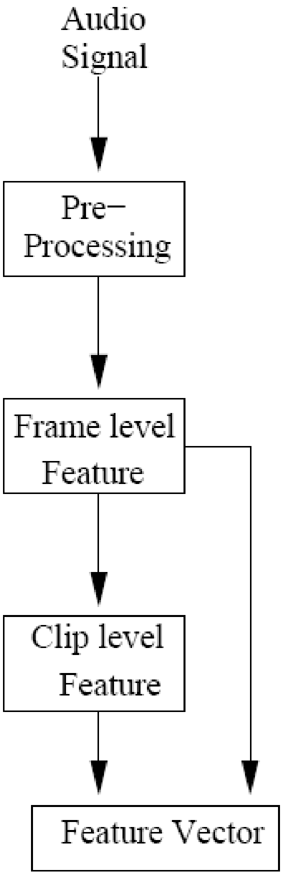

Figure 2: Frame level and Clip level Feature

Frame Level Features

Varieties of features have been proposed by the researchers which may be categorized as low level features, perceptual/ psychoacoustic features etc. Low level feature are mostly time domain and frequency domain features. Some of the features are as follows.

i) **Distance to voicing:** It estimates [3] the voicing level of the signal and if the level is above a certain threshold then it is marked as voice. The distance to voicing is defined as the distance of the current frame from closest voiced frame. If it is zero then the frame is voiced frame and in case of large distance (greater than a threshold) is consider as unvoiced frame.

ii) **Short-Time Energy:** Short-Time Energy (STE) is a simple time-domain feature, which is widely used by researchers. The energy of the n^{th} frame is defined as

$$E_n = \frac{1}{r} \sum_{m=1}^{r} [x_n(m)]^2 \qquad (1.1)$$

where, E_n denotes the nth frame energy, r is the frame length, and $x_n[m]$ represents the mth sample in the nth frame. Li [15] and Zhang [2] have worked with this feature. STE is suitable for discrimination between speech and music [5, 23, 8]. Speech consists of words interleaved by silence which gives variation of STE that is quite different from the pattern obtained for music. Instead of short-time energy, some studies [13, 22] have used root-mean-square (RMS) value, where

$$\mathrm{RMS}_n = \sqrt{}\mathrm{E}_n \qquad (1.2)$$

The RMS approximates the loudness of the signals and it can also be used to discriminate speech and music.

iii) **Zero-Crossing Rate:** The Zero-Crossing Rate (ZCR) is an extremely popular short time feature which depicts the concentration of energy in the spectrum. It is defined as the number of times the signal crosses zero (i.e., changes sign) within the frame:

$$ZCR_n = \sum_{m=2}^{r} sign[x_n(m-1) * x_n(m)] \qquad (1.3)$$

where, r is the number of samples in the nth frame and

$$\mathit{sign[v]} = \begin{cases} 1, \text{ if v} < 0 \\ 0, \text{ otherwise} \end{cases} \qquad (1.4)$$

The ZCR can be viewed as a measure of the dominant frequency [21, 20, 23]. Since unvoiced speech typically has much higher ZCR values than voiced speech, ZCR can be used to distinguish between voiced and unvoiced speech [9]. Scheirer [5] have used ZCR to classify audio between speech/music. Tzanetakis [14] has used it to classify audio into different genres of music.

iv) **Pitch/Fundamental Frequency:** The sound that comes through vocal tract starts from the larynx where vocal cords are situated and ends at mouth. The vibration of the vocal cords and the shape of the vocal tract are controlled by nerves from brain. The sound, which we produce, could be categorized into voiced and unvoiced sounds. During the production of unvoiced sounds the vocal cords do not vibrate and stay open whereas during voiced sounds they vibrate and produce what is known as glottal pulse. A pulse is a summation of a sinusoidal wave of fundamental frequency and its harmonics (Amplitude decreases as frequency increases). The fundamental frequency of glottal pulse is known as the pitch.

The fundamental frequency, often referred to simply as the fundamental and abbreviated f_0 or F_0, is defined as the lowest frequency of a periodic waveform. In terms of a superposition of sinusoids (e.g. Fourier series), the fundamental frequency is the lowest frequency sinusoidal in the sum. All sinusoidal and many non-sinusoidal waveforms are periodic, which is to say they repeat exactly over time.

A harmonic sound consists of a series of major frequency components including the fundamental frequency and its integer multiples. Pitch, a perceptual term, is also used to represent the fundamental frequency. Typically, pitch frequency for a human being is between 50-450 Hz, whereas for music the value can be much larger [19]. Pitch can be determined by locating peaks from the autocorrelation function [9]. Some other efforts have been made to get estimation of fundamental frequency. Short-time fundamental frequency (SFuF) has been computed in [17, 18], when signal is harmonic, SFuF is equal to the fundamental frequency else SFuF is zero. Subband-based pitch detection has been tried in [2, 16].

v) **Spectrum Centroid:** It is used in the work of Li [15] to classify speech and noise. The same feature is adopted in [14] to classify music into different genres. The spectral centroid (SC), also known as brightness, can be defined as

$$SC_n = \frac{\sum_{k=1}^{K-1} k|A(n,k)|^2}{\sum_{k=1}^{K-1} |A(n,k)|^2} \qquad (1.5)$$

where SC_n is the spectral centroid corresponding to the nth frame, A(n, k) is the DFT of the nth frame and K is the order of the DFT.

vi) **Spectrum Spread:** Spectrum spread (SS) is closely related with spectrum centroid. It is defined as magnitude-weighted average of the differences between the spectral components and the centroid [13], i.e.,

$$SS_n = \sqrt{\frac{\sum_{k=1}^{K-1}[(k - SC_n)^2.|A(n,k)|^2]}{\sum_{k=0}^{K-1}|A(n,k)|^2}} \qquad (1.6)$$

vii) **Spectral Rolloff:** It is defined as the 95th percentile of the spectral energy distribution [5], i.e., a value R_n such that

$$\sum_{k=1}^{R_n}|A(n,k)| = \alpha\sum_{k=1}^{K}|A(n,k)| \qquad (1.7)$$

where $\alpha = 0.95$. R_n measures the skewness of the spectral shape. In case of voiced speech or music the energy mainly contained in the low frequency band whereas for the unvoiced speech major portion of the energy is located in high frequency range of the spectrum.

viii) **Spectral Flux:** It is the measure of spectral change between two adjacent frames, which is defined as

$$SF_n = \sqrt{\sum_{k=1}^{K}(|A(n+1,k)| - |A(n,k)|)^2} \qquad (1.8)$$

Flux is an important feature for the separation of music from speech [1]. Speech alternates periods of transition (consonant-vowel boundaries) and periods of relative stasis (vowels), while music typically has a more constant rate of change.

ix) **Mel-Frequency Cepstral Coefficients:** Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFCC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs are also found useful in audio classification applications [12]. Two properties of MFCCs, one being that the first coefficient is proportional to the audio energy and the other being that there is no correlation among different coefficients, make MFCCs attractive in audio classification [11].

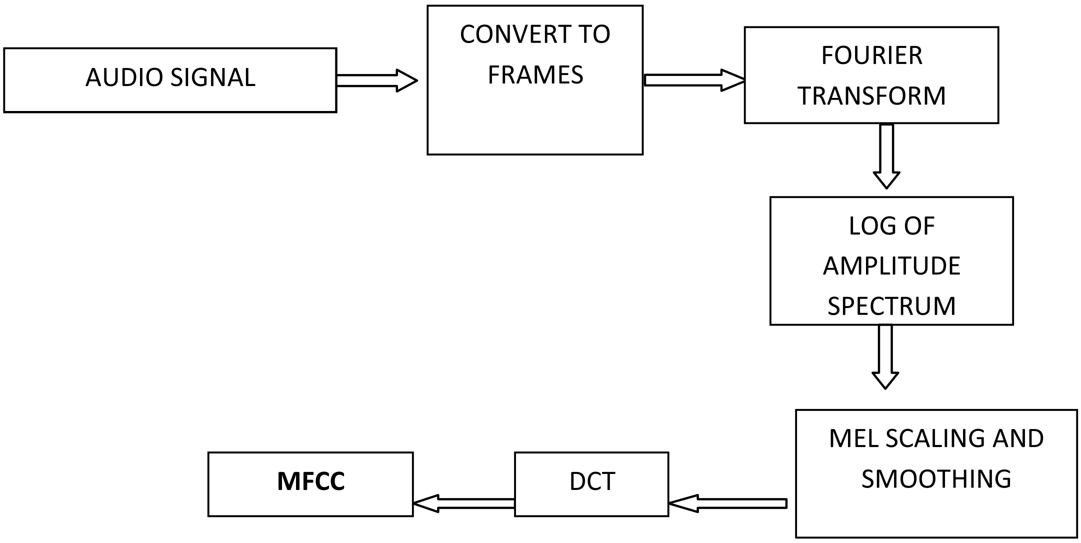

Figure 3: Steps for computing MFCC

STEPS FOR COMPUTING MFCC

1. Audio signal is divided into number of frames of fixed duration. Frames may be overlapping.
2. Amplitude Spectrum of each frame is obtained by applying DFT and logarithm of amplitudes is taken.
3. The spectrum is smoothened to make it perceptually meaningful using Mel frequency which is calculated as

$$f_m = 2595 * \log_{10}\left(1 + \frac{f}{100}\right) \qquad (1.9)$$

4. The elements in the smoothened Mel-spectra vector are highly correlated. To decorrelate and to reduce the number of parameters DCT is performed to obtain MFCCs and first 13 co-efficients are taken as features for the frame.
5. After computing the MFCCs of all frames, the vector comprising of the average value corresponding to each co-efficient form the feature descriptor.

x) **LPC Related Features:** The basic idea behind the linear predictive coding (LPC) analysis [10, 9] is that a speech sample can be approximated as linear combination of past speech samples. By minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients is determined. Speech is modeled as the output of linear, time-varying system excited by either quasi-periodic pulses (during voiced speech), or random noise (during unvoiced speech). The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time-varying system representing vocal tract. LPC is an alternative to the short-time Fourier transform for speech analysis. A frame-by-frame audio classification method is proposed in [8], the proposed method has employed 10th order line spectrum frequency (LSF) coefficients and the differential FSFs. In [8], LSF analysis is employed to refine the classification results of a pre-classifier.

Clip Level Features

Some clip-level features are designed to characterize the temporal variation of frame-level features (i.e., they are calculated using frame-level features), whereas others are designed without the use of frame-level features. The most widely used clip-level features are the statistical mean and variance values of frame-level features such as the energy and ZCR. In addition, there are many other clip-level features as introduced below.

i) **Low Short-Time Energy Ratio:** It is a variation of STE feature proposed in [5, 2, 7], which is defined as

$$LSTER = \frac{1}{2N}\sum_{n=1}^{N}(sign[0.5\overline{E} - E_n] + 1) \tag{1.10}$$

where N is the total number of frames, n is the frame index, E_n is the STE of the nth (see. eqn. (1.1)), sign[·] is the sign function defined in eqn. (1.4), and $\overline{E}$ denotes the average STE in a 1 second window. Therefore, LSTER is defined as the ratio of number of frames whose STE is below 0.5-fold average STE in a 1 second window. In general there are more silence frames in speech than in music. So, the LSTER value of speech is much higher than music.

ii) **Low Feature-Value Ratio:** The technique of LSTER is generalized to any feature and it is called as low feature-value ratio (LFVR) [2]. LFVR is defined as

$$LFVR = \frac{1}{2N}\sum_{n=1}^{N}(sign[0.5\overline{FV} - FV_n] + 1) \tag{1.11}$$

where FV_n is the feature value of the n^{th} frame, $\overline{FV}$ is the average FV in a processing window.

iii) **High Zero-Crossing-Rate Ratio:** It is proposed in [5, 2, 7], which is a variation of ZCR. This feature mainly used to discriminate speech and music and mathematically it is defined as

$$HZCRR = \frac{1}{2N}\sum_{n=1}^{N}(sign[ZCR_n - 1.5\overline{ZCR}] + 1) \tag{1.12}$$

where, $\overline{ZCR}$ is the average of the ZCR in a 1 second window. Thus, HZCRR is defined as the ratio of the number of frames whose ZCR is above 1.5-fold average zcr-crossing rate and number of frames in a 1 second window.

Since there are always some silence intervals in a speech, the occurrence of null zero crossings (i.e., ZCR = 0) can be used to identify speech. Based on this, the ratio of null zero-crossings can be used to identify speech segments [6]. However, the accuracy of this feature may be affected by the presence of background noise or a change of speaking rate.

iv) **High Feature-Value Ratio:** It is generalization of HZCRR like LFV R [2], which is defined as

$$HFVR = \frac{1}{2N}\sum_{n=1}^{N}(sign[FV_n - 1.5\overline{FV}] + 1) \tag{1.13}$$

HFV R defines the ratio of number of frames whose feature value is above 1.5-fold average feature value and number of frames in a processing window.

v) **Rhythm and Tempo:** In [5], a feature is proposed to compute the rhythmicness of a signal. In this method, a signal is divided into six bands and finds the peaks in each envelope of each band. Here, the peaks roughly represent the perceptual onsets. Parallel comb filters are used in [4] to analyze the tempo of a music signal and also extract the beat from the signals. Beat spectrum is computed to characterize the rhythm and tempo of a signal [3].

vi) **Other Features:** There are many other clip-level features for describing the audio signal. One of them is noise frame ratio [2], which is defined as the ratio of the number of noise frames to the number of frames in a given audio clip. Entropy modulation [75], isi used to measure the “disorderness” of a signal. Entropy modulation of speech is higher than music.

Feature Selection

From a large set of features it is important to select particular set of features that would determine the nature and hence the class of the audio signal. These features determine the dimensionality in the feature space. It is important therefore to select an optimum number of features that not only keeps accordance with the accuracy and the level of performance but also reduces the computation costs. Thus there is no point in increasing the number of features as it would not have a drastic impact on the accuracy but would pave for more complexities in computation. Therefore a selected feature must have the following properties,

1) **Invariance to irrelevancies:** Any good feature should exhibit invariance to irrelevancies such as noise, bandwidth or the amplitude scaling of the signal. It is also upon the classification system to consider such variations as irrelevant to achieve better classification across a wide range of audio formats.

2) **Discriminative Power:** The purpose of feature selection is to achieve discrimination among different classes of audio patterns. Therefore a feature must take round about similar values within the same class but different values across different classes.

3) **Uncorrelated to other features:** It is very important that there are no redundancies in the feature space. Each new feature that is selected must give altogether different information about the signal as possible. This helps in better computation efficiency, improved performance and optimization of cost [1].

REFERENCES

[1] J. J. Burred and A. Lerch, “Hierarchical Automatic Audio Signal Classification”, *J. Audio Eng. Soc*, Vol. 52, pp. 724-739, July/August 2004.

[2] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. “Content analysis for audio classification and segmentation”. *Speech and Audio Processing, IEEE Transactions* on, 10(7):504–516, 2002.

[3] Jonathan Foote and Shingo Uchihashi. “The beat spectrum: A new approach to rhythm analysis”. In *ICME*, 2001.

[4] Eric Scheirer. “Tempo and beat analysis of acoustic musical signals”. *The Journal of the Acoustical Society of America*, 103:588, 1998.

[5] Eric Scheirer and Malcolm Slaney. “Construction and evaluation of a robust multifeature speech/music discriminator”. In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 *IEEE International Conference* on, volume 2, pages 1331–1334. IEEE, 1997.

[6] C. Panagiotakis and G. Tziritas, “A speech/music discriminator based on RMS and zero-crossings," *IEEE Trans. Multimedia*, vol. 7, pp. 155-166, Feb. 2005.

[7] Hao Jiang, Tong Lin, and Hong-Jiang Zhang. “Video segmentation with the assistance of audio content Analysis”. *In Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1507–1510. IEEE, 2000.

[8] Khaled El-Maleh, Mark Klein, Grace Petrucci, and Peter Kabal. “Speech/music discrimination for multimedia applications”. *In Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 2445–2448. IEEE, 2000.

[9] Douglas O’shaughnessy. “Speech communication: human and machine”. *Universities press*, 1987.

[10] Lawrence R Rabiner and RonaldWSchafer. “Digital processing of speech signals”, volume 19. *IET*, 1979.

[11] Z. Liu and Q. Huang, \Content-based indexing and retrieval-by-example in audio," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, Jul.-Aug. 2000, pp. 877-880.

[12] S. Z. Li, “Content-based audio classification and retrieval using the nearest feature line method," *IEEE Trans. Speech Audio Process*., vol. 8, no. 5, pp. 619-625, Sep. 2000.

[13] Erling Wold, Thom Blum, Douglas Keislar, and JamesWheaten. “Content-based classification, search, and retrieval of audio”. *MultiMedia, IEEE*, 3(3):27–36, 1996.

[14] George Tzanetakis and Perry Cook. “Musical genre classification of audio signals”. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[15] Dongge Li, Ishwar K Sethi, Nevenka Dimitrova, and Tom McGee. “Classification of general audio data for content-based retrieval”. *Pattern recognition letters*, 22(5):533–544, 2001.

[16] C-C Lin, S-H Chen, T-K Truong, and Yukon Chang. “Audio classification and categorization based on wavelets and support vector machine”. *Speech and Audio Processing, IEEE Transactions on*, 13(5):644–651, 2005.

[17] Tong Zhang and C-C Jay Kuo. “Audio content analysis for online audiovisual data segmentation and Classification”. *Speech and Audio Processing, IEEE Transactions* on, 9(4):441–457, 2001.

[18] Tong Zhang and C-CJ Kuo. “Hierarchical classification of audio data for archiving and retrieving”. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 3001–3004. IEEE, 1999.

[19] Yao Wang, Zhu Liu, and Jin-Cheng Huang. “Multimedia content analysis-using both audio and visual clues”. *Signal Processing Magazine, IEEE*, 17(6):12–36, 2000.

[20] Kristopher West and Stephen Cox. “Features and classifiers for the automatic classification of musical audio signals”. In *ISMIR*, 2004.

[21] Benjamin Kedem. “Spectral analysis and discrimination by zero-crossings”. *Proceedings of the IEEE*, 74(11):1477–1493, 1986.

[22] Costas Panagiotakis and Georgios Tziritas. “A speech/music discriminator based on rms and zero-crossings”. *Multimedia, IEEE Transactions on*, 7(1):155–166, 2005.

[23] J. Saunders. “Real-time discrimination of broadcast speech/music”. In *IEEE Conf. on Acoustics, Speech, Signal Processing*, pages 993–996, 1996.

[24] T. Zhang and C. C. Jay Kuo. “Content-based classification and retrieval of audio”. *In SPIE Conf. on Advanced Signal Processing Algorithms, Architectures and Implementations VIII*, 1998.

[25] E. Zwicker and H. Fastl. “Psychoacoustics: Facts and models”. *Springer Series on Infromation Science*, 1999.

[26] M. Zuliani, C. S. Kenney, and B. S. Manjunath. “The multiransac algorithm and its application to detect planar homographies”. *In IEEE Conf. on Image Processing*, 2005.

[27] Chao Zhen and Jieping Xu. “Solely tag-based music genre classification”. *In Web Information Systems and Mining (WISM), 2010 International Conference on*, volume 1, pages 20–24. IEEE, 2010.

[28] Ming Zhao, Jiajun Bu, and Chun Chen. “Audio and video combined for home video abstraction”. *In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, volume 5, pages V–620. IEEE, 2003.

[29] Yibin Zhang and Jie Zhou. “Audio segmentation based on multi-scale audio classification”. *In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on*, volume 4, pages iv–349. IEEE, 2004.

[30] T. Zhang. “Semi-automatic approach for music classification”. *In SPIE Conf. on Internet Multimedia Management Systems*, pages 81–91, 2003.

[31] G. Yu and J.-J. Slotine. “Audio classification from time-frequency texture”. *In IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pages 1677 – 1680, 2009.

[32] Changsheng Xu, Namunu Chinthaka Maddage, and Xi Shao. “Automatic music classification and summarization”. *Speech and Audio Processing, IEEE Transactions on*, 13(3):441–450, 2005.

[33] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S Huang. “Comparing mfcc and mpeg-7 audio features for feature extraction, maximum likelihood hmm and entropic prior hmm for sports audio classification”. *In Acoustics, Speech, and Signal Processing, 2003. IEEE International Conference on*, volume 5, pages V–628. IEEE, 2003.

[34] Stuart N Wrigley, Guy J Brown, Vincent Wan, and Steve Renals. “Feature selection for the classification of crosstalk in multi-channel audio”. *In Proc. EuroSpeech. International Speech Communication Association*, 2003.

[35] C. West and S. Cox. “Finding an optimal segmentation for audio genre classification”. *In Int. Sym. on Music Information Retrieval*, 2005.

[36] C. West and S. Cox. “Finding an optimal segmentation for audio genre classification”. *In Int. Sym. on Music Information Retrieval*, 2005.

[37] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. “Large margin methods for structured and interdependent output variables”. *In Journal of Machine Learning Research*, pages 1453–1484, 2005.

[38] Wei-Ho Tsai and Duo-Fu Bao. “Clustering music recordings based on genres”. *In Information Science and Applications (ICISA), 2010 International Conference on*, pages 1–5. IEEE, 2010.

[39] P. H. S. Torr and A. Zisserman. Mlesac: “A new robust estimation with application to estimating image geometry”. *Journal of Computer Vision and Image Understanding*, 78(1), 2000.

[40] Andrey Temko, Climent Nadeu, Duˇsan Macho, Robert Malkin, Christian Zieger, and Maurizio Omologo. “Acoustic event detection and classification”. *In Computers in the Human Interaction Loop*, pages 61–73. Springer, 2009.

[41] Mathias Stager, Paul Lukowicz, and Gerhard Troster. “Implementation and evaluation of a low-power sound-based user activity recognition system”. *In Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, volume 1, pages 138–141. IEEE, 2004.

[42] Mathias Stäger, Paul Lukowicz, Niroshan Perera, Thomas von Büren, Gerhard Tröster, and Thad Starner. Soundbutton: “Design of a low power wearable audio classification system”. *In ISWC*, pages 12–17, 2003.