Automated Discrimination of Digital Audio

Ranjan Parekh

School of Education Technology, Jadavpur University, Kolkata, India

ABSTRACT

Over the last decade there has been a huge proliferation on the use of multimedia content throughout the world. This has led to the growth of a large number digital media repositories. As such an efficient and fast mechanism for retrieval of media content from these repositories assumes fundamental importance. This paper reports the results of experimentation done on audio clips in order to extract feature information from them and utilize these to create an automatic system for discriminating speech and music. Three features, namely, silence ratio, standard deviation of ZCR and peak index in RMS histogram have been proposed. Most of the existing works involve a large number of features which not only makes the process computation intensive, but also introduces redundancies, as many of these features are dependent on each other. Furthermore, the classification tests are frequently heuristic-based and not derived from an analysis of the data. The aim of this paper is to limit on the number of these features and demonstrate that even single features by themselves can be used to attain performance accuracies to the tune of 98%.

Keywords - **Speech-music discrimination, Multimedia information systems, Content Based Storage and Retrieval, Pattern Recognition**

I. INTRODUCTION

Over the last decade there has been a huge proliferation on the use of multimedia content throughout the world. Application areas like audio-on-demand, video-on-demand, computer based training (CBT) packages, games and home entertainment, online business and corporate presentations, information kiosks and simulation packages, voice-mails etc. have led to the growth of a large number of digital media repositories all around the world. In this scenario an efficient and fast mechanism for retrieval of digital media content from these repositories assumes fundamental importance. A repository of media elements without an effective search and retrieval mechanism is comparable to a library without a catalog. Even though the information is present it is practically unavailable to somebody with a specific set of search criteria. This paper addresses the problem of audio retrieval based on pre-defined content based features. Even though a substantial amount of research work has been done in this area, most of them involve a large number of features which not only makes the process computation intensive, but also suffers from redundancies as many of these features are dependant on each other. Furthermore, the classification tests are frequently heuristic-based and not derived from an analysis of the data. The aim of this paper is to limit the number of features, which are however so chosen as to provide accuracies to the tune of 98%.

The organization of the paper is as follows: section 2 provides an overview of related work, section 3 outlines the proposed approach with discussions on overview, feature computation and classification schemes, section 4 provides details of the dataset and experimental results obtained and section 5 provides the overall conclusion and the scope for future research.

II. PREVIOUS WORK

The initial approaches for content-based audio similarity, involving comparisons between individual samples [1], had limited accuracy because of the possibility of their using different digitization parameters. Later approaches used features extracted from audio files to characterize and compare them. Loudness, indicated by the audio signal's root mean square (RMS) value [2, 3, 4, 5, 6] is defined as the following, where E is the average energy of the audio piece, N the total number of samples in the audio piece, and x_i the sample value of the *i*-th sample :

$$E = \sqrt{\frac{\sum_{i=1}^{N} (x_i)^2}{N}} \tag{1}$$

The zero crossing rate (ZCR) [2, 3, 4, 6, 7, 8] indicates the frequency at which the signal crosses the zero amplitude level. Speech being usually made up of a collection of words with gaps of silence in between, speech signals display a higher crossing frequency as compared to music. The average ZCR is defined as :

$$Z = \frac{\sum_{n=1}^{N} |\operatorname{sgn}(x_n) - \operatorname{sgn}(x_{n-1})|}{2N} \tag{2}$$

where $\operatorname{sgn}(x_n)$ is the sign of the *n*-th sample x_n and can be 1, 0 or –1, N is the total number of samples.

The silence ratio (SR) [2, 7] is a measure of the silent portions of an audio relative to its total duration. Due to the presence of background noise, "silence" usually means portions whose loudness (RMS) values lie below a certain threshold, rather than being absolute zeros. Pauses between words in speech, lead to a higher value of SR as compared to music.

Audio signals might be represented in frequency domain to highlight features related to their frequency components. The Discrete Fourier Transform (DFT) [2, 3, 7, 8, 9] provides one of the primary means for such conversion :

$$X_k = \sum_{n=0}^{N-1} x_n . e^{-jn\omega_k} \qquad (3)$$

where $\omega_k = 2\pi k / N$, x_n is a discrete signal with N samples, k is the DFT bin number.

The frequency range of an audio signal is computed by taking the difference between the highest and lowest non-zero frequency components. Since sound generated from the human voice box (larynx) usually does not exceed 7 kHz in frequency, in contrast to musical sounds from instruments which can cover the entire audible range from 20 Hz to 20 kHz, average frequency range of speech signals are lower than those of music signals. Frequency range is generally indicated by the amplitude of the centroid of the frequency domain waveform.

Another frequency domain feature of an audio signal is harmonicity [2, 10]. Harmonicity refers to the proportion of the audio signal which is harmonic i.e. signal components having frequency values which are multiples of a lower or base frequency. Music has been observed to be much more harmonic than speech, in fact harmonicity itself is believed to produce the sensation of "musical sound" in human ears and brain. Often the base frequency is the lowest or fundamental frequency of the audio signal

It is difficult to compute fundamental frequency of an aperiodic signal, where the concept of "periodicity" is not very well defined. One popular way of doing so is by using the "cepstrum" [11, 12]. It relies on the fact that if the original audio signal contains a number of harmonic components, then its frequency spectrum would show peaks at frequency values corresponding to the harmonics. A second transform of the spectrum waveform to the frequency domain (through Fourier Transform) would indicate a peak corresponding to the periodicity in the spectrum, which in turn indicates the fundamental frequency, being a measure of the gap between the peaks in the spectrum. This double frequency transform of the audio signal is referred to as the "cepstrum"

Most of the authors use a combination of multiple features to characterize audio content. In [6] four features based on the ZCR – variance of the derivative, third central moment, a threshold value and a skewness measure, have been used. In [4] features like energy function, average ZCR, fundamental frequency and spectral peaks have been used to achieve a performance of 95%. In [5] the authors use 13 features, related to the power spectrum, ZCR, cepstrum and their variances. [9] uses energy spectral based features like spectral centroid, flux and moments, as well as pitch, harmonicity and cepstral coefficients. [10] uses loudness, pitch, brightness, bandwidth and harmonicity as features.

III. Proposed Approach

Silence Ratio

As already mentioned, the silence ratio (SR) indicates the proportion of the sound piece that is silent. Silence is defined as a period within which the absolute amplitude values of a certain number of samples are below a certain threshold. The silence ratio is calculated as the ratio between the sum of silent periods and the total length of the audio piece. There are two critical issues involved. The first is how to decide if a sample is silent. In this work silence is defined in terms of an experimentally determined RMS threshold value. Let R_T represent this threshold value of RMS. The second issue is to decide how many consecutive silent samples would qualify for a silent zone in the audio. In this work audio clips are considered to be made of a collection of audio frames each of 20 ms duration. The instantaneous accuracy is fixed at 20 ms because the human perceptual system is generally not more precise, and moreover because speech signals remain stationary for 5–20 ms [13]. A silent zone is said to occur when all audio samples within an audio frame are silent. Thus if n_f is the total number of samples in an audio frame of 20 ms duration, and x_i is the sample value of the *i*-th sample, then a frame is designated as silent when the following condition is true.

$$\sqrt{\frac{\sum_{i=1}^{N}(x_i)^2}{n_f}} < R_T \qquad (4)$$

Silence ratio S_R is calculated as the ratio of the number of silent frames N_{sf} to the total number of frames N_f in an audio clip.

$$S_R = \frac{N_{sf}}{N_f} \qquad (5)$$

Since human speech contains gaps between words and sentences, speech files in general contains larger number of silent zones than music files..

Standard Deviation of ZCR

A statistical measure derived from ZCR, given by equation (2), has been used as a feature for discrimination in this work - standard deviation value (σZ). An audio file is partitioned into a number of logical frames, each frame consisting of a fixed number of audio samples. For each frame the ZCR is computed over all the samples of the frame. Each frame is thereafter represented by a single ZCR value. The standard deviation value of ZCR over all audio frames in a file is calculated to generate the σZ value. The values were found to be lower for music than for speech

Peak Index in RMS Histogram

The third feature used is based on the RMS value, given by equation (1). RMS has been shown to be independent of ZCR values [3] as the former depends on the signal amplitude while the latter on the signal frequency. An audio clip is partitioned into a number of audio frames, each of 20 ms duration. For each frame RMS value is computed over the samples of the frame. A 256-bin histogram is computed by plotting the number of frames in an audio file against the RMS bin value. It is observed that the peak value in the RMS histogram occurs at lower RMS index value for speech than for music.

IV. EXPERIMENTATIONS

Training Phase

For experimentations audio samples from Dan Ellis' audio database has been used, available at http://www.ee.columbia.edu/~dpwe/sounds/musp/music-speech-20060404.tgz. The training set contains 60 speech samples, labeled here as S01 to S60, and 60 music samples, labeled as M01 to M60. The digitization parameters are 22050 Hz sample rate, 16-bit mono. Each audio file is of 15 second duration. The music files are of both vocal and instrumental types. The vocal type involves songs sung by single or multiple artists along with accompanying music, while the non-vocal type contains only instrumental music involving piano, violin, drums etc., The following files contain only instrumental music: M01, M02, M15, M16, M17, M21, M23, M24, M32, M33, M34, M38, M39, M40, M41, M42, M47, M54, M55, M56, M57, M59.

For characterizing the audio each 15 second clip is divided into a collection of 750 non-overlapping audio frames, each of 20 ms duration. The instantaneous accuracy is fixed at 20 ms because the human perceptual system is generally not more precise, and moreover because speech signals remain stationary for 5–20 ms [13]. The maximal interval for measuring speech characteristics should therefore be limited to intervals of 20 ms. At a sample rate of 22050 Hz, each audio frame consists of 441 samples.

Silent zones have been calculated by computing the RMS value of audio samples in each frame and by considering a frame to be silent if the RMS value is less than an experimentally determined threshold value of 0.01. The number of silent frames is then summed over the entire file. The percentage of silent audio frames is considered as a feature for discrimination. Since human speech contains gaps between words and sentences, speech files in general contains larger number of silent zones than music files. Fig. 1 shows the plot of the number of silence zones for each of the 60 speech files and 60 music files. For speech the maximum value is found to be 260, the minimum value 0 while most of the music files has silence zones between 15 and 0, the notable exceptions being M12 (48) where the song fades off gradually towards the end, and M54 (141) which contains a piano sequence which becomes almost inaudible towards the middle.

A statistical measure derived from ZCR have been used as a feature for discrimination in this work - standard deviation value (σZ). An audio file is partitioned into a number of logical frames, each frame consisting of a fixed number of audio samples. For each frame the ZCR is computed over all the samples of the frame. Each frame is thereafter represented by a single ZCR value. The standard deviation value of ZCR over all audio frames in a file is calculated to generate the σZ value. As shown in Fig. 2, the standard deviation of ZCR was much higher for speech having a maximum value of 0.1753, a minimum of 0.0396, than music with a maximum of 0.1270, minimum of 0.0110. So there appears to be a demarcation line at 0.1 below which most samples were music and above which most samples were speech.

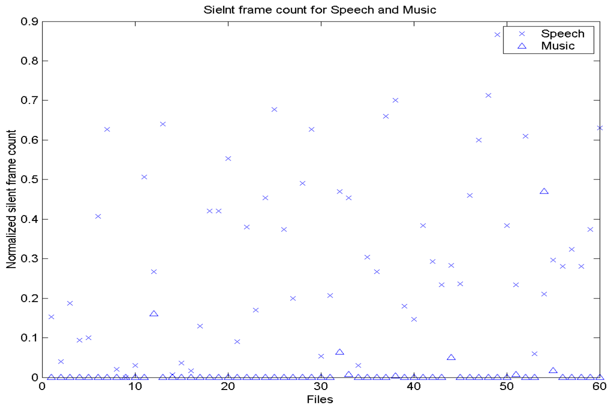

Figure 1. Silent frame count for Speech and Music

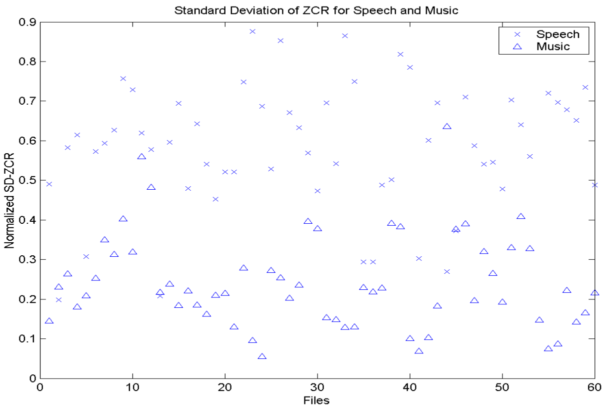

Figure 2. SD-ZCR values for Speech and Music

The third feature used is based on the RMS value, which is a measure of the average energy content of the audio signal. In general it was observed that peaks in RMS histogram occurs at lower values for speech than for music. The RMS index values are plotted for speech and music, as shown in Fig. 3. It shows that speech was clustered at the bottom of the graph with index values less than 12, while music was spread evenly across the middle and upper half of the plot.

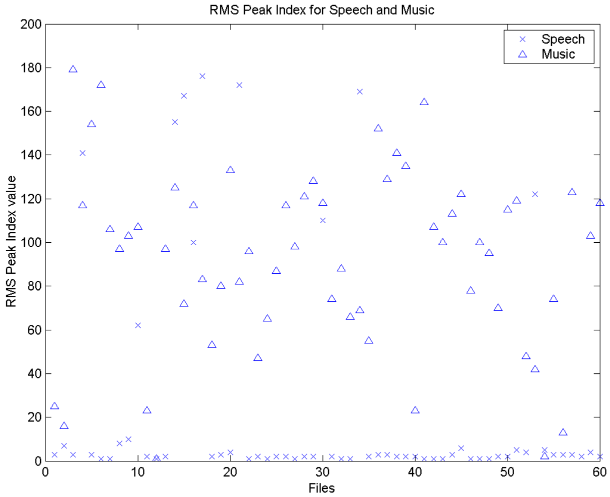

Figure 3. RMS peak index for Speech and Music

Testing Phase

The test data set consists of 20 speech samples from the same database, labeled here as T01 to T20, 20 vocal music samples, labeled as T21 to T40 and 20 non-vocal (instrumental) music samples labeled as T41 to T60. The digitization parameters are 22050 Hz sample rate, 16-bit mono. Each audio file is of 15 seconds duration. The same 20 ms audio frame size has been used.

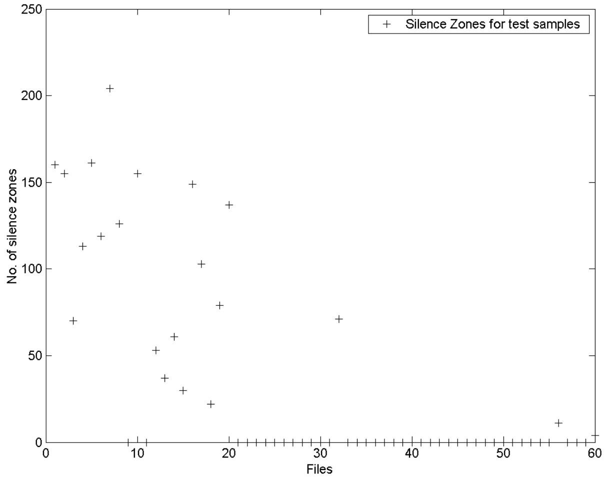

Figure 4. Plot of silence zones for test clips

The plot of silence zones for the 60 test samples is shown in Fig. 4. Based on the training set observations, a threshold value of 15 was used to discriminate between speech and music, i.e. if number of silent zones < 15, the file is classified as music, otherwise as speech. Table 1 shows that based on silent zones alone, 18 of 20 speech samples could be identified correctly – accuracy of 90%.

Table 1. Confidence grid based on silent zones

S/N	SZ	P	S/N	SZ	P	S/N	SZ	P
T01	160	S	T02	155	S	T03	70	S
T04	113	S	T05	161	S	T06	119	S
T07	204	S	T08	126	S	T09	0	M
T10	155	S	T11	0	M	T12	53	S
T13	37	S	T14	61	S	T15	30	S
T16	149	S	T17	103	S	T18	22	S
T19	79	S	T20	137	S	T21	0	M
T22	0	M	T23	0	M	T24	0	M
T25	0	M	T26	0	M	T27	0	M
T28	0	M	T29	0	M	T30	0	M
T31	0	M	T32	71	S	T33	0	M
T34	0	M	T35	0	M	T36	0	M
T37	0	M	T38	0	M	T39	0	M
T40	0	M	T41	0	M	T42	0	M
T43	0	M	T44	0	M	T45	0	M
T46	0	M	T47	0	M	T48	0	M
T49	0	M	T50	0	M	T51	0	M
T52	0	M	T53	0	M	T54	0	M
T55	0	M	T56	11	M	T57	0	M
T58	0	M	T59	0	M	T60	4	M

Sample T09 is identified wrongly because there are no gaps in the spoken sample while in T11 presence of background noise has blanked out the gaps. Out of 40 music samples 39 has been correctly identified – an accuracy of 97.5%. T32 has been wrongly classified because even though it is a song, there are distinct gaps in the sequence. (S=speech, M=music)

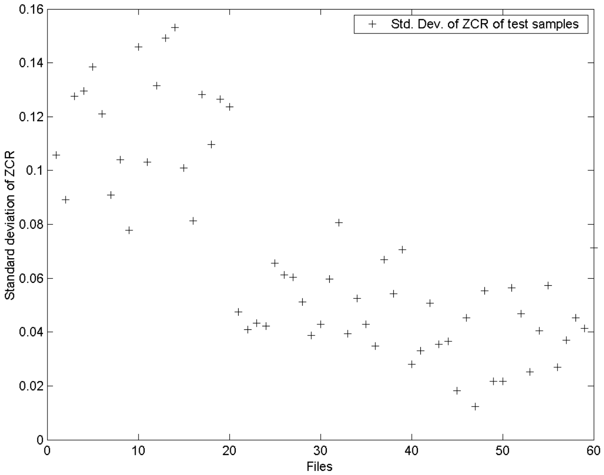

Figure 5. Plot of standard deviation of ZCR for test clips

For standard deviation of ZCR, we take the threshold value as 0.1 below which samples would be considered as music and above which samples would be considered as speech. Fig. 5 shows the plot of the standard deviation of ZCR for test samples. The corresponding confidence grid is shown in Table 2. Out of 20 speech samples, 16 are correctly identified giving an accuracy of 80%. Out of 40 music samples, all have been correctly identified giving an accuracy of 100%.

Table 2. Confidence grid based on SD of ZCR

S/N	σ ZCR	P	S/N	σ ZCR	P
T01	0.1058	S	T02	0.0892	M
T03	0.1275	S	T04	0.1295	S
T05	0.1384	S	T06	0.1210	S
T07	0.0909	M	T08	0.1041	S
T09	0.0779	M	T10	0.1459	S
T11	0.1032	S	T12	0.1316	S
T13	0.1493	S	T14	0.1531	S
T15	0.1010	S	T16	0.0813	M
T17	0.1283	S	T18	0.1097	S
T19	0.1266	S	T20	0.1237	S
T21	0.0474	M	T22	0.0410	M
T23	0.0433	M	T24	0.0422	M
T25	0.0655	M	T26	0.0612	M
T27	0.0603	M	T28	0.0512	M
T29	0.0387	M	T30	0.0430	M
T31	0.0597	M	T32	0.0807	M
T33	0.0393	M	T34	0.0526	M
T35	0.0428	M	T36	0.0348	M
T37	0.0668	M	T38	0.0542	M
T39	0.0706	M	T40	0.0281	M
T41	0.0331	M	T42	0.0508	M
T43	0.0354	M	T44	0.0366	M
T45	0.0183	M	T46	0.0454	M
T47	0.0124	M	T48	0.0553	M
T49	0.0217	M	T50	0.0218	M
T51	0.0565	M	T52	0.0468	M
T53	0.0253	M	T54	0.0406	M
T55	0.0573	M	T56	0.0270	M
T57	0.0371	M	T58	0.0454	M
T59	0.0413	M	T60	0.0712	M

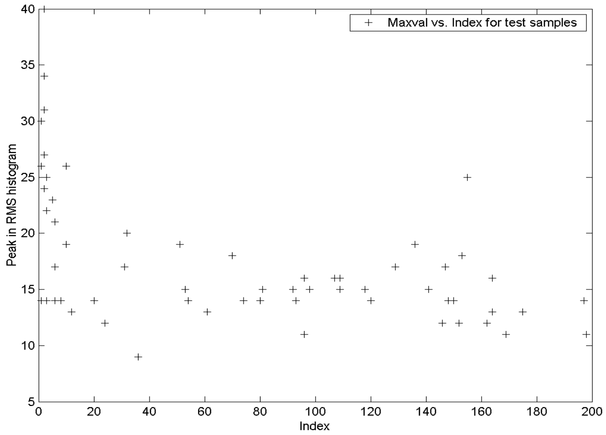

Figure 6. Plot of peak vs. index of RMS histogram for test clips

Fig. 6 shows the peak vs. index in RMS histogram of test samples. Based on the training samples we fix up a threshold that if the index is less than or equal to 12, it is classified as speech, otherwise music. Table 3 shows the corresponding confidence grid.

Table 3. Confidence grid based on peak index in RMS histogram

S/N	I	P	S/N	I	P	S/N	I	P
T01	6	S	T02	2	S	T03	3	S
T04	10	S	T05	3	S	T06	10	S
T07	5	S	T08	2	S	T09	164	M
T10	3	S	T11	12	S	T12	1	S
T13	8	S	T14	6	S	T15	6	S
T16	2	S	T17	1	S	T18	152	M
T19	2	S	T20	2	S	T21	153	M
T22	141	M	T23	155	M	T24	96	M
T25	61	M	T26	96	M	T27	148	M
T28	198	M	T29	74	M	T30	147	M
T31	70	M	T32	1	S	T33	150	M
T34	146	M	T35	197	M	T36	81	M
T37	98	M	T38	164	M	T39	109	M
T40	169	M	T41	80	M	T42	109	M
T43	129	M	T44	96	M	T45	107	M
T46	32	M	T47	92	M	T48	31	M
T49	51	M	T50	54	M	T51	24	M
T52	175	M	T53	93	M	T54	118	M
T55	53	M	T56	36	M	T57	136	M
T58	162	M	T59	120	M	T60	20	M

Out of the 20 speech files 18 files have been identified correctly giving an accuracy of 90%. Out of 40 music files, 39 have been identified correctly, giving an accuracy of 97.5%.

V. CONCLUSION

This work demonstrates that audio discrimination need not necessarily deal with a large number of features. Single features taken at a time are able to provide reliable accuracy of more than 97%. Since a small number of features are involved, such discriminators might be included with existing application packages where searching and retrieving of digital audio is required e.g. commercially available multimedia authoring packages to help content developers to search for audio content quickly. Audio on Demand applications also can make use of such discriminators as an initial step for categorizing audio content. Future work will involve categorizing other databases as well as trying to discriminate between different types of music e.g. vocal / non-vocal, and recognize presence of specific instruments e.g. drums and flute.

REFERENCES

[1] S. Subramanya et al, “Transform based indexing of audio data for multimedia databases”, *Proc. of IEEE Intl. Conf. on Multimedia Computing and Systems*, 1997, 211-218.

[2] G. Lu, “Indexing and Retrieval of Audio: A Survey”, *Multimedia Tools and Applications*, 15, 2001, 269-290.

[3] C. Panagiotakis, G. Tziritas, “A speech/music discriminator based on RMS and zero crossings”, *IEEE Transaction on Multimedia*, 7(1), 2005, 155-166.

[4] T. Zhang, J. Kuo, “Audio content analysis for on-line audiovisual data segmentation and classification”, *IEEE Transactions on Speech Audio Processing*, 9(3), 2001, 441-457.

[5] E. Scheirer, M. Slaney, “Construction and evaluation of a robust multi-feature speech/music discriminator”, *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, 1331-1334.

[6] J. Saunders, “Real-time discrimination of broadcast speech/music”, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal processing (ICASSP)*, 1996, 993–996.

[7] L. Lu, H-J Zhang, “Content Analysis for Audio Classification and Segmentation”, *IEEE Transactions on Speech and Audio Processing*, 2002, 504-516.

[8] Khaled El-Maleh, M Klein, G Petrucci, P Kabal, “Speech/music discrimination for multimedia applications”, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal processing (ICASSP)*, 2000, 2445-2448.

[9] G. Tzanetakis, P. Cook, “A framework for audio analysis based on classification and temporal segmentation”, *Proc. of 25th Euromicro Conf. on Music Technology and Audio Processing*, 1999, 61-67.

[10] E. Wold, Thom Blum, Douglas Keislar, James Wheaton, “Content-based classification, search, and retrieval of audio”, *IEEE Multimedia*, 3(3), 1996, 27–36.

[11] P. Moreno, R. Rifkin, “Using the fisher kernel method for web audio classification”, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal processing (ICASSP)*, 2000, 2417-2420.

[12] M. Seck, F. Bimbot, D. Zugah, and B. Delyon, “Two-class signal segmentation for speech/music detection in audio tracks”, *Proc. of 6th European Conf. on Speech Communication and Technology (Eurospeech)*, 1999, 2801–2804.

[13] A. S. Spanias, “Speech coding: A tutorial review”, *Proceedings of the IEEE*, 82(10), 1994, 1541–1582.