SPEECH DETECTION ON BROADCAST AUDIO

Ünal Zubari[1]*, Ezgi Can Ozan*[1,2]*, Banu Oskay Acar*[1]*, Tolga Ciloglu*[2]*, Ersin Esen*[1]*, Tuğrul K. Ateş*[1,2]
and Duygu Oskay Önür[1]

[1]Video and Audio Processing Group, TUBITAK UZAY, METU Campus 06531 Ankara Turkey,
phone: + 90 (312) 210 1310, fax: + 90 (312) 210 1315
email:{unal.zubari, ezgican.ozan, banu.oskay, ersin.esen, tugrul.ates, duygu.oskay}@uzay.tubitak.gov.tr
[2]Department of Electrical and Electronics Engineering , METU, 06531 Ankara Turkey,
phone: +90 (312) 210 2302 fax: +90 (312) 210 2304 email: ciltolga@eee.metu.edu.tr

ABSTRACT

Speech boundary detection contributes to performance of speech based applications such as speech recognition and speaker recognition. Speech boundary detector implemented in this study works on broadcast audio as a pre-processor module of a keyword spotter. Speech boundary detection is handled in 3 steps. At first step, audio data is segmented into homogeneous regions in an unsupervised manner. After an ACTIVITY/NON-ACTIVITY decision is made for each region, ACTIVITY regions are classified as Speech/Non-speech via Gaussian Mixture Model (GMM) based classification. GMM's are trained using a novel feature, Spectral Flow Direction (SFD), and an improved multi-band harmonicity feature in addition to widely used Mel Frequency Cepstral Coefficients (MFCC's).

1. INTRODUCTION

Broadcast audio contains various types of audio classes. Segmentation and classification of broadcast audio is important for multimedia indexing and automatic speech recognition (ASR) applications. Detecting speech boundaries in broadcast audio contributes to the performance of ASR applications by means of enabling feature normalization and adaptation in a specific region of the audio.

Early works on speech detection were based on speech/music discrimination. Scheirer and Slaney [1] presented and examined large amount of low level features to differentiate between speech and music. Saunders [2] worked on broadcast audio and suggested the features zero-crossing rate (ZCR), energy contour for speech/music classification. Although such methods performed well for speech/music discrimination, detection of speech regions on general audio requires more classes for covering the data other than speech and music, and requires precise boundary detection of the underlying classes. Considering these aspects, the problem of detecting speech regions is essentially a content based retrieval problem and is processed in that context in the works of Lu et al. [3, 4], Li et al. [5], Minami et al. [6] and Zhang and Kuo [7].

One of the problems in classifying general audio is the accuracy losses due to misplacing of class boundaries. Li et al. [5] defined these losses as *"border effect"* and suggested a segmentation pooling algorithm with pause detection to solve this problem. Our suggestion to the problem is the initial unsupervised segmentation which locates boundaries as the points where power changes dramatically.

Martin and Breebaart [8] suggested the use of temporal features in combination with frame level features. Saunders [2] proposes using zero crossing rate and power contour as features. We have used both dynamic (SFD) and frame level features (MFCC and band harmonicity).

This work is implemented as a part of a keyword spotting algorithm and aims to maximize the speech region detection while minimizing the non-speech segments detected as speech.

Our main contributions can be summarized as;

i- Unsupervised, power based segmentation for detecting class boundaries.

ii- Activity/non-activity classification for detecting pauses

iii- Two new features; SFD and band harmonicity for speech/music/other classification

2. SYSTEM OVERVIEW

Speech boundary detection is processed in 3 steps. First, broadcast audio is segmented into homogeneous regions using a power based unsupervised algorithm. An homogenous region is an audio segment, which contains only one audio class.

Secondly, each audio segment is examined for an audio activity. The segments that do not involve any activity are considered to be pauses. The regions labelled as *ACTIVITY* are then classified into speech and non-speech classes.

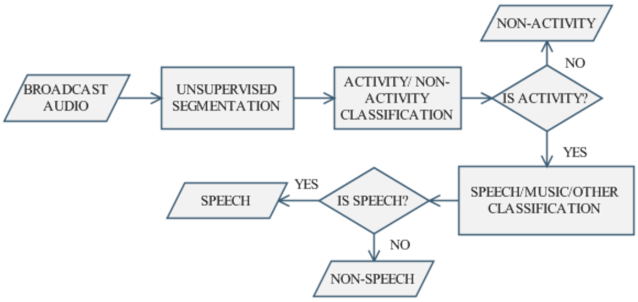
NON-ACTIVITY
NO
BROADCAST AUDIO
UNSUPERVISED SEGMENTATION
ACTIVITY/ NON-ACTIVITY CLASSIFICATION
IS ACTIVITY?
YES
SPEECH
YES
IS SPEECH?
SPEECH/MUSIC/OTHER CLASSIFICATION
NO
NON-SPEECH

Figure 1 – System Flow Diagram

3. HOMOGENEOUS REGION DETECTION USING UNSUPERVISED SEGMENTATION

Homogeneous region detection is the problem of dividing audio into small segments so that the type of the audio event does not change within the segment. Segmentation is based on the points of power change. Experiments on the proposed method show that 99% of the obtained segments consist of single audio class. The average length of the segments is found to be 400 ms.

3.1 Segment Boundary Detection

To determine segment boundaries, first, powers of non-overlapping 10 ms length frames are computed. Two adjacent windows are moved along the whole data with one frame slide and powers for each window are computed. These values are used to compute the power ratio of two windows by dividing the greater one by the lesser. Applying this procedure with one frame slide, a sequence of power ratios is obtained. The peak locations of this sequence having a power ratio greater than a threshold are set to be segment boundaries. Any segment shorter than 200 ms is merged to the adjacent segment with closest power value.

Segment Boundary Detection

```
Require: F: Frame energy sequence for audio data $FP = \{f_1 \dots\dots..f_N\}$
Ensure: SB: the set of segment boundaries detected in F
SB ←Ø , DP ←Ø
Begin
    1)    //Divide
    do for each $f_i$ in FP {
        Select two windows of length wndsize i.e;
    $W_i = \{f_{i-wndsize}, \dots\dots, f_i\}$ , $W_2 = \{f_i, \dots\dots, f_{i+wndsize}\}$
        Compute power ratio PRi between W1 and w2;
        $PR = (max\ (PW1, PW2))\ /\ (min\ (PW1, PW2))$
        if (PR > Pth)   RS ←PR
        else            RS ←0
    }
    DP = localmax(DP)
    2)    //Combine
    do for each $DP_i$ in DP {
        Let S be the segment between $DP_i$ and $DP_{i+1}$, X be the segment on the left of S and Y be the segment on the right of S;
        Let $E_s$, $E_x$, and $E_y$ be the segment energies of S,X and Y respectively ,

        if( ($DP_{i+1}$-$DP_i$) < 200msec) ) {
          if (abs(Ex-Es) < abs(Ey-Es))
                Merge X and S; SB←DPi+1
          else
                Merge Y and S; SB ←DPi; i++;
        }else
           SB←DPi U DPi+1
    }
End
```

3.2 Parameter Selection

The window size and division threshold are selected experimentally. To determine the window size, the percentage of segments consisting of single audio class, (Figure 2) and the average length of segments, (Figure 3) both as determined by the proposed method, are considered. A high percentage of unique audio class in a segment is desired. This is achieved by reducing the window size. On the other hand, longer segments are preferred, since audio classification is based on the features extracted from a segment. A longer segment yields more information about the audio class of that segment. This is achieved by increasing the window size. In this paper, a window size of 20 frames (200ms) is determined.

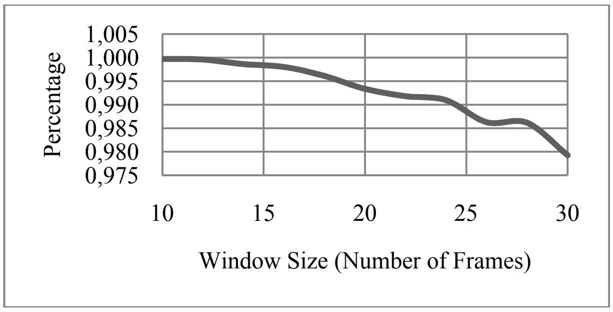

Figure 2 – Single Class Segment Percentage vs. Window Size

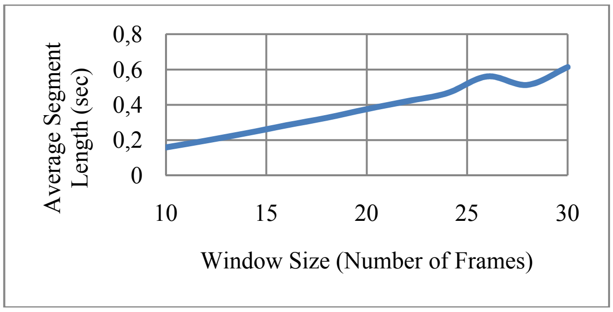

Figure 3 – Avg. Segment Length vs. Window Size

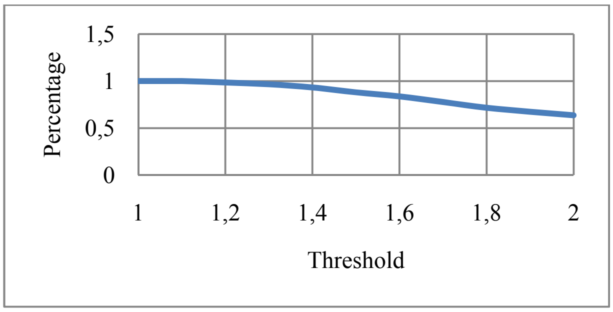

Figure 4 - Single Class Segment Percentage vs. Threshold

To determine the division threshold, the number of segment boundaries in the ground truth, having a power ratio greater than a selected threshold, (Figure 4) is considered. The desired case is to select a sufficiently high threshold, since higher thresholds yield longer segments. In this paper, a threshold of 1.2 is determined.

4. ACTIVITY/NON-ACTIVITY CLASSIFICATION

A non-activity region is defined as the region that does not carry any event information, i.e. the regions of the background noise in an outdoor scene, the background music between sentences of a dialog in a movie, silence regions, etc... Elimination of these regions provides faster execution of event classification algorithms on audio and better precision rates.

Activity/Non-activity decision of a segment is based on the comparison of segment power to both short and long term power levels. For short term representation of segment power, segments are represented by a 5 dimensional feature vector consisting of the powers of the segment and of its two predecessors and the two successors. The feature vector is normalised by the long term power (LTP).

LTP is the average of the powers of the segments between two LTP boundaries. The LTP boundaries are set where “significant” changes in the power levels of different audio environments occur. For example, a silent movie scene coming after a loud commercial is where an LTP boundary exists. The first scene consists of long silent regions, while the second one has loud background music between the speech regions.

4.1 Training Stage:

4.1.1 LTP boundary detection

To determine the LTP boundaries, two adjacent windows of “non-activity” labelled segments are moved along the segmented data. The peaks of the ratios of the averages inside those windows are filtered by a threshold, and the resulting points are set as the boundaries.

4.1.2 Training

LTP value is computed using the powers of activity regions inside the LTP boundary. Region features are normalised by the LTP of the corresponding region.

As the region features are obtained, 32-mixture GMM’s are trained for each class.

4.2 Classification Stage

4.2.1 LTP boundary detection

While non-activity regions are defined in the training stage, in the test stage a pre-classification step is implemented for a course estimation of non-activity regions.

The power of each segment in the feature vector is normalised by the power of whole data. GMM based classification is performed using computed features which results in coarse estimation of activity/non-activity regions. Using coarsely estimated non-activity region information, LTP boundaries are estimated as defined in the training stage.

4.2.2 Test

LTP values are computed using the powers of activity labelled regions inside the LTP boundaries. The region features are normalised by the LTP value of the corresponding region. Finally, region classification is performed using GMM’s obtained during the training stage.

Proposed method has been tested on a dataset of 1 hour broadcast audio. 94.6% of Activity detection accuracy and 94.2% of Non-activity detection accuracy have been obtained.

4.3 Parameter Selection

To determine the window size and division threshold parameters for LTP boundary detection, a cost function (*avgF1*) is given in equation (1).

$$avgF1 = \frac{2\times F1_A \times F1_N}{F1_A + F1_N} \quad (1)$$

Where $$F1_A = \frac{2\times R_A \times P_A}{R_A + P_A} \quad F1_N = \frac{2\times R_N \times P_N}{R_N + P_N} \quad (2,3)$$

R_A: Recall of Activity.
P_A: Precision of Activity.
R_N: Recall of Non-Activity.
P_N: Precision of Non-Activity.

Table 1 shows the cost function values with different window sizes and thresholds. LTP boundary detection method increases especially the recall rate of Non-Action.

Table 1 – LTP method with different parameters

Window Size	*Threshold*	*avgF1*	*Recall of Non-Action*
LTP OFF	*na*	0,9090	0,9110
10	**2**	**0,9199**	**0,9436**
10	1,5	0,9186	0,9366
10	1	0,9133	0,9371
20	2	0,9072	0,9355
20	1,5	0,9178	0,9327
20	1	0,9147	0,9281

LTP Boundary Detection

```
Require: SB: the set of segment boundaries
Ensure: LTPB: set of LTP boundaries detected in SB
SB ←Ø , LTPB ←Ø
Begin
    NAS: Non-Activity Segment
    do for each $NAS_i$ in SB {
        Select two windows of length ltpwndsize i.e;
    $W_1 = \{NAS_{i-wndsize}, \dots \dots, NAS_i\}$,
    $W_2 = \{NAS_i, \dots \dots, NAS_{i+wndsize}\}$
        Compute power ratio PRi between W1 and w2;
        $PR = (max\ (PW1, PW2)) / (min\ (PW1, PW2))$
        if (PR > Pth)   RS ←PR
        else            RS ←0
    }
    DP = localmax(DP)
    LTPB←DPi U DPi+1
End
```

5. SPEECH/NON-SPEECH CLASSIFICATION

5.1 Features

6 MFCC and their Δ values are used in combination with spectral flow direction (SFD) and band harmonicity features for speech/non-speech classification. MFCC features are extracted from 20Hz-4000Hz frequency range using 12 channels at 8 kHz sampling frequency. Features are extracted from 25 ms frames with 15 ms overlap. The extraction of SFD and harmonicity features is defined at sections 5.1.1 and 5.1.2, respectively.

5.1.1 Spectral Flow Direction

The temporal behavior of the spectral peaks is a characteristic for each audio class. For example, variation of spectral peak locations in time is greater for speech regions as compared to music. An example of a typical flow of the spectral peaks in speech and music is shown in Figure 5.

Some recent work [8, 9] uses this information to classify audio into speech and non-speech classes. Although extracting the spectral peaks is straightforward, tracking the spectral peaks is computationally expensive. To represent the temporal behavior of the spectral peaks, the spectral flow direction feature is defined.

SFD is defined as the frequency lag providing the maximum correlation between the spectra of two sequential frames, as given in equation (4). 512-point spectral representation is computed for 8 kHz audio. For this resolution of spectrum, the maximum allowed lag value *w* between two adjacent frames is decided to be 10 points. The cross-correlation estimation is performed on [*w, N/2-w*] frequency bins of the spectrum of frame *n*.

$$SFD(n) = argmax_l\left(\sum_{j=w}^{\frac{N}{2}-w} s(n,j) \times s(n+1,j+1)\right) \quad l = -w, \ldots, w \quad (4)$$

Where, *N* is the FFT length, *l* is the lag amount, *s(n,m)* is the energy at frequency bin *m* of frame *n*.

As seen in Figure 5, the flow direction tends to be zero in musical sounds. It has small variations in speech and high variations in noisy data. Using this information, we have calculated two values from the SFD values. The first one is the number of zero SFD's in a 20-frame window, and the second one is the sum of the absolute differences in a 20-frame window.

5.1.2 Band Harmonicity

Harmonicity [10] is a well known audio feature commonly used in content-based audio classification [11]. It is generally calculated using comb filtering method, which requires calculation of fundamental frequency [10, 12]. A new harmonicity calculation method based on the short-time Fourier transform (STFT) of the magnitude spectrum (2nd spectrum), which does not require fundamental frequency calculation, is proposed. For a perfectly harmonic signal, the harmonic peaks are located at the integer multiples of fundamental frequency *F*. FFT of the magnitude spectrum for this harmonic signal, has a peak located at *F*.

The energy of the 2nd spectrum is concentrated at fundamental frequency for harmonic signals while the energy is spread over the spectrum as the harmonicity decreases. Harmonicity is defined as the ratio of the energy value at the peak location of 2nd spectrum over the total energy.

Since the fundamental frequency is not known, the location of this peak has to be estimated. To eliminate the effect of low frequency components, a maximum expected fundamental frequency value is defined, and the peaks corresponding to a frequency above that value are eliminated. Figure 6, shows the spectrum of a 50 ms length audio signal and the estimated harmonicity value.

While the harmonic properties of speech are prominent at low frequencies, the harmonic peaks could be located in a broader frequency range for musical sounds. To differentiate between those two classes, the harmonicity is represented in 3 dimensions, where each dimension corresponds to the harmonicity value for the bands; [0, 1000], [1000, fs/2] and [0, fs/2] where fs is the sampling frequency.

5.2 Classification Method

We have defined three classes; 'speech', 'music' and 'other' to represent different events found in general broadcast audio data. The music class has been defined to contain music and music like sounds (phone ring, engine sounds, etc). The 'other' class has been defined to contain events such as crowd, laughter, animal sounds, water sounds, explosions, gunshots, various noise, etc. An 8-mixture GMM is trained for each class. The training data consist of 1000 seconds of speech, 690 seconds of music and 612 seconds of 'other'.

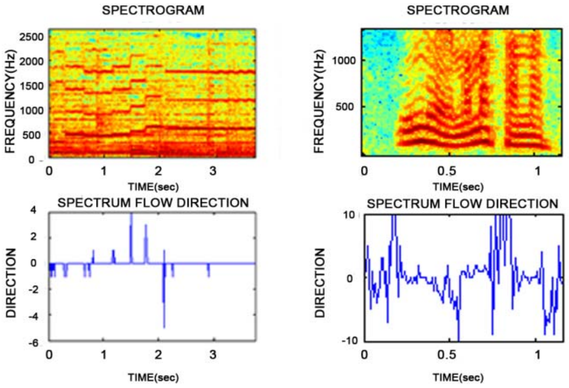

Figure 5 – Spectrogram and SFD values of a musical sound (left), and speech (Right).

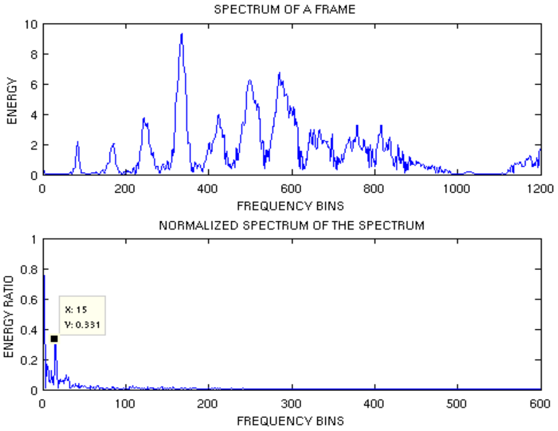

Figure 6 – A typical magnitude spectrum of a harmonic audio frame (Top), Normalized magnitude spectrum of the magnitude spectrum in the top panel (Bottom).

The classification is performed for each segment. Frame probabilities are computed for each class and the class that yields the maximum sum of frame probabilities over the segment is selected as the segment class.

After the segments are classified into speech, music and 'other', a second pass is applied to classification results to combine the adjacent segments of the same type and to eliminate insignificant silence and non-speech segments (<0.3 seconds) between speech segments. Existence of short silence regions is common in general conversations so these regions can be added to the confining speech regions.

6. EXPERIMENTAL RESULTS

Tests have been performed on a dataset consisting of different broadcast audio sources which contain various types of audio; such as movies, commercials and news, recorded from six different TV channels continuously. The test data has been labelled by hand on segment basis. Since broadcast audio may contain more than one class at a time, 6 classes including speech(s), music (m), singing (h), other (o), silence (sil) and background noise (n) are used in combination to label each segment. Size of test data for each class is represented in Table 3.

Table 2 – Test dataset.

Audio Class	*Length (seconds)*
singing + speech +music (hsm)	55.80
singing + music (hm)	98.20
pure speech (s)	2427.06
speech + music (sm)	2411.18
speech + music + other (smo)	163.10
speech + other (so)	568.92
pure music (m)	1252.53
music + other (mo)	187.30
background noise (n)	261.90
other (o)	1055.07
silence (sil)	444.11
TOTAL	8926.17

To measure the performance of our speech/non-speech classifier and to observe the effect of the SFD and band harmonicity features, comparative tests have been applied with different feature sets. The results in Table 4 show the classification performance of MFCC features on "activity" labelled segments containing single class. The results obtained by adding the SFD and band harmonicity features are presented in Table 5.

Table 3 – Classification results using features mfcc.

GT \RESULT	*Speech*	*Music*	*Other*	*Total*
Speech	**0.9358**	0.0476	0.0165	1.000
Music	0.0693	**0.5259**	0.4048	1.000
Other	0.0652	0.1794	**0.7555**	1.000

Table 4 – Classification results using features mfcc + SFD + harmonicity.

GT \RESULT	*Speech*	*Music*	*Other*	*Total*
Speech	**0.9630**	0.0205	0.0164	1.000
Music	0.0792	**0.7908**	0.1299	1.000
Other	0.1418	0.0847	**0.7735**	1.000

Table 5 – Classification results using features mfcc + SFD + harmonicity.

GT \RESULT	*s*	*m*	*o*	*NA*	*Total*
hsm	0.941	0.045	0.010	0.004	1.000
hm	0.519	0.401	0.064	0.016	1.000
s	**0.936**	0.010	0.010	0.044	1.000
sm	0.864	0.090	0.026	0.020	1.000
smo	0.754	0.126	0.103	0.017	1.000
so	0.733	0.096	0.152	0.019	1.000
m	0.079	**0.698**	0.116	0.107	1.000
mo	0.113	0.257	0.585	0.045	1.000
n	0.084	0.026	0.145	0.745	1.000
o	0.149	0.072	**0.671**	0.108	1.000
sil	0.124	0.008	0.023	**0.845**	1.000

Overall performance results are presented in Table 6; this is the classification result for all classes after the second pass has been applied. Since the second pass refinement adds small silence regions to speech segments, overall non-activity detection accuracy is lower. As it can be seen, the classifier selects speech class for hybrid regions which include speech. Therefore, the proposed system is shown to be appropriate for speech detection problem. When all the classes containing speech and singing are considered as speech the precision rate is found to be 0.9339, when classes containing speech are considered as speech the precision rate is found to be 0.9144.

7. CONCLUSION

In this paper, a new approach for detection of speech region boundaries has been proposed. Activity / Non-Activity regions are defined and a method for classification of these regions has been described. Two new features have been proposed and their usefulness has been verified with tests. As a result of the proposed approaches, significant improvement has been observed. New class models for combined classes such as speech with background music and speech with noise, will be studied as a future work.

8. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1997, pp. 1331–1334.

[2] J. Saunders, "Real time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1996,pp. 993–996.

[3] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10,no. 4, pp. 504–516, Oct. 2002.

[4] L. Lu, H.-J. Zhang, and S. Li, "Content-based audio classification and segmentation by using support vector machines," *ACM Multimedia Syst. J.*, vol. 8, no. 6, pp. 482–492, Mar. 2003.

[5] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, Vol. 22, No. 5, pp. 533–544, 2001.

[6] K. Minami, A. Akutsu, H. Hamada, and Y. Tomomura, "Video handling with music and speech detection," *IEEE Multimedia,* Vol. 5, No. 3, pp. 17–25, 1998.

[7] T. Zhang and C.-C. J. Kuo, "Video content parsing based on combined audio and visual information". *In Proc. SPIE 1999*, vol. 4,1999, pp. 78–89.

[8] K. Seyerlehner, G. Widmer, Tim Pohle and Markus Schedl, "Automatic music detection in television productions". *In Proceedings of the Int. Conf. on Digital Audio Effects* (DAFx-07), 2007.

[9] T. Taniguchi, M. Tohyama, and K. Shirai, "Detection of speech and music based on spectral tracking," *Speech Communication*, vol. 50, no. 7, pp. 547–563, 2008.

[10] H.Kim, N.Moreau, and T.Sikora, *MPEG-7 audio and beyond*. Wiley, West Sussex, England, 2005.

[11] D. Mitrovic, M. Zeppelzauer and C. Breiteneder, "Features for content-based audio retrieval". *Advances in Computers,* vol. 78, pp 71-150, 2010.

[12] S. H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics-based features for audio". *In IEEE Proc. of ICASSP,* May 2004, vol. 4, pp. 321–324.