Audio Engineering Society
Convention Paper
Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

A non-linear rhythm-based style classification for Broadcast Speech-Music Discrimination

Enric Guaus[1], Eloi Batlle[1]

[1] *Audiovisual Institute, Pompeu Fabra University, Barcelona, Spain*

Correspondence should be addressed to Enric Guaus (eguaus@iua.upf.es)

ABSTRACT

Speech-Music discriminators are usually designed under some rigid constrains. This paper deals with a more general Speech-Music Discriminator successfully used in AIDA project. The system is based on a Hidden Markov Model style classification process in which the styles are grouped into two major categories: Speech or Music. The goals of this sub-system are (1)the expandible possibilities with the addition of some new styles (like "phone female voice"), (2)the use of new rhytmical descriptors in combination with other typical ones and (3)the robustness of our speech/music discriminator in many different environments by using some Mathematical Morphology and non-linear post-processing techniques. The techniques used in our system allow a fast track in changes between styles and, thus, typical confusions in commercials can be easily cleaned. The accuracy of this system can be up to a 94.3% in broadcast radio environment.

1. INTRODUCTION

The Speech-Music discrimination problem is not new, and some of the developed systems have been quite successful. But all these systems are designed under some rigid constrains (see Sec. 2). This is the main goal of this paper: our environment has

not any kind of constrains. Broadcast audio signals (from radio stations, TV, GSM or Internet) are the input of our system and, as one can imagine, the content of this data is out of control. Mainly, we have two different kind of problems:

Channel distortions: Since the sources of audio can be very different, the system is not focused on clean audio files. As we will see in Sec.4, the used descriptors of audio are strategically chosen to be, for instance, independent of the spectrum of the input signal. Then, MP3 or GSM codifications for audio input will not produce a failure to the system

Audio content Since the content of the system is unknown, the system has to discriminate between speech and music in many different situations: radio interview, films, commercials, etc.

How do we design this system? The main idea is to create a *genre classification system* in which the musical genres are conceptually different with regard to the typical ones. For instance, a "cellular male voice" could be a genre for our system. Furthermore, the representative features of the input signals are not only based on the spectrum of the signal, but on the rhythm.

This paper is organized as follows: In Sec. 2 we will see an overview of the State of the Art. In Sec. 3 we will see an overview of the overall project in which the speech-music discriminator have to be included, and the used technology will be shown too. In Sec. 4 we will explain the system in detail and, finally, some experiments and results will be shown in Sec. 5.

2. STATE OF THE ART

The Speech-Music discrimination problem has become quite important for last years due to the automatic indexing or classification problem. Multimedia data identification and indexation has become more and more important due to the fast growth of electronic databases through Internet. Large amount of data must be automatically analyzed, and speech-music discrimination is only one step for the whole process.

2.1. Main characteristics

A lot of studies have been done in Speech-Music Discrimination. The previous work can be summarized in three different categories:

- Time domain based systems, such as zero-crossing or energy-evolution.
- Frequency domain based systems, such as cepstral coefficients.
- Mixed time-frequency domain based systems, such as 4Hz Modulation or harmonic coefficients.

This classification can be done just taking into account which kind of parameters are extracted from the input signal. But we could think in other classification scheme just taking into account the signal processing method performed to that data. Then, the Speech-Music discrimination systems can be divided in:

- Decision trees
- Neural Networks
- Gaussian Mixture Models
- Hidden Markov Models

All of the systems we will talk about in the next section should be enclosed into one of the categories of the first classification scheme as well as into one of the categories of the second classification scheme independently.

2.2. Related work

The first successful approach on the speech-music discrimination was made by John Saunders in 1996 [1]. In this study, Saunders compare different features like the *Tonality*, *bandwidth*, *excitation patterns*, *tonal duration* and *energy sequences*. The Zero-crossing rate is introduced as a significant parameter in the speech-music discrimination, and some experiments and results are presented.

Another important approach was made by Eric Sheirer and Malcom Slaney in [2]. This work presents a study of thirteen different features, derived from 8 original ones (4HzModulation, Spectral

Centroid, Cepstrum, Pulse Metric, etc.). Each one of them is supposed to be a good discriminator at once. Different sets of features have been trained and tested by using Single Gaussian Mixture Models, but the results are not spectacular. The conclusion of this work deals on more research: no significant good results are found.

At this point, the basic features and procedures for discrimination are presented, and future works will only introduce new features or little deviations of these main ideas. Wu Chow and Liang Gu present us a set of features derived from Harmonic Coefficient and its 4Hz Modulation in [3]. This approach is based on a two-level processing structure, one for singing/non singing musical signals detection and the other for the typical speech-music discrimination. After a rule-based post-filtering smoothing algorithm, significant enhancements are obtained for complex audio streams. Karnebck presents an exhaustive study on Low Frequency Modulations in [4], and Berenzweig and Ellis present some new statistical features (defined in [5]) embedded in a simple HMM for distinguishing between singing and instrumental music in [6]

Some comparisons on the methods mentioned above have been made. M. J. Carey [7] has tested most of those different features. The cepstral coefficients and delta cepstral coefficients seem to be the most successful parameters, while the zero-crossing and the energy (mean and variance values across the time) are not so important. Cepstra and delta cepstra can give us an equal error rate of about 1.2%, slightly far of the 6% of equal error rate by using the zero-crossing coefficients.

Finally, a study of the State of the Art has been made by the *Audio Research Group* in *Tampere University of Technology*.

3. ENVIRONMENT

The Speech-Music Discrimination system is one of the requirements for the AIDA (Automatic IDentification of Audio) project[1]. In the next sections, we will see a brief overview of the AIDA project and where the Speech-Music Discriminator is located, and a brief description of the used technique (AMADEUS)..

[1]The AIDA project is founded by SDAE (Sociedad Digital de Autores y Editores)

3.1. AIDA system overview

The major goal for the AIDA project is the automatic recognition of broadcast audio. This process could appear quite simple, but it becomes more and more complicated when huge audio databases must be managed. Some techniques are applied for reducing that large amount of data, but the complexity of the system is increased as well. Furthermore, the system must be robust to many different real situations: noise and other non-linear distortions are always present. Then, the problem becomes quite difficult to solve and a lot of considerations must be taken into account.

Hidden Markov Model techniques are used for this purpose. By using HMM, the system is not a TRUE/FALSE identification process, but a non-linear similarity measure is obtained. In this context, we find the *identified song* as the *most similar* song. When this *most similar song* is found under some other constrains, it is considered as the *identified song*. This technique is quite useful for other similarity applications such as rhythmical similarity. The system must be robust to multiple distortions from the input signal as well. Is widely known that almost radio-stations apply different distortions to the audio signal in order to increase the listener's attention. The most common radio distortions are Compressor/Limiter, Stereo Base-width, Exciter/Enhancer and Pitching. The system must be source-independent. Different sampling frequencies, bit depth or codification must not affect the robustness of the system. Signals from cellular phones must be identified as well as MP3 files or direct real-time streaming.

Our Speech-Music Discrimination system is the first step in this automatic recognition process. If data like news, interviews or films goes through AIDA, a lot of data will be mark as *unknown*. Then, the system can easily be saturated. The aim of our Speech-Music Discriminator is let only the music pass through AIDA.

3.2. AMADEUS system overview

Finally, the system is implemented by using the AMADEUS technology. The AMADEUS technology has been developed by the Music Technology Group (MTG), at the Pompeu Fabra University (UPF). It is just a set of classes implemented in C++

style	description	main group
mal	Male voice	speech
fem	Female voice	speech
cma	Cellular male voice	speech
cfe	Cellular female voice	speech
cla	Classical music	music
cop	Copla & Author music	music
ele	Electronic soft music	music
jaz	Jazz music	music
pop	Pop music	music
roc	Hard rock music	music
tec	Tecno & Dance music	music
sil	Silence	speech

Table 1: Definition of different styles

with all the needed features for the HMMs training and calculations in real-time [8].

4. DESCRIPTION

In few words, the systems is designed as a rough musical genre classification. But the genres have not any special musical meaning. A genre is, from our point of view, a group of audio signals with some common (spectral, timbrical or rhythmical) features. The selected genres are grouped into two main groups, *Speech* or *Music*, as shown in Table 1. The developing process is clearly divided in three parts:

- Data acquisition
- Parametrization of both training and test audio data.
- Training Process
- Real-time Recognition and graphical interface.

The training process is made by using the HTK software and the whole process is oriented to the HTK philosophy. The Wavesurfer software is also used for creating labels.

4.1. Data Acquisition

A lot of audio data must be recorded and manually labeled for a successful training process. As the system will work in a real-time broadcast audio environment (from many different radio stations), many excerpts of radio broadcast audio have been recorded.

Descriptor	L	Value	∇	∇^2
MFCC	12	*	*	*
Energy	1	*	*	*
4Hz Modulation	1	*	*	*
Zero Crossing Rate	1	*	*	*
Spectral Centroid	1	*	*	*
Spectral Flatness	1	*	*	*
Voice2White	1	*	*	*

Table 2: List of available descriptors

Finally, this data have been edited and many different audio files have been produced. All these audio files are 1 minute long with $f_s = 22050Hz$, 16 bits and mono. But the main characteristic of these audio files is that they belong into one specific musical genre, that is, each file belongs exclusively to an specific genre from the beginning to the end. Finally, Different HMM models will be defined for different genres, so the models will hold more accurate descriptions of the music.

4.2. Parametrization

From now on, we have a lot of audio files, and each one of them can be associated with a specific musical genre.The parametrization process should transform all these audio files into a set of description files. Table 2 shows us the available descriptors we can use, and we describe the right selection in Sec. 5. The parametrization is made by an *AMADEUS* application which generates an HTK-format file `*.htk` for each audio file.

4.2.1. Voice2White

A part of the other well known descriptors, we define a *new* one, the Voice2White ratio. The Voice2White ratio is a measure of the energy inside the typical speech band ($300Hz..4KHz$) respect the energy of the whole audible margin (in case of $f_s = 44100Hz$) or global band (in case of $f_s < 44100Hz$). From a mathematical point of view:

$$v2w = 10log_{10}\frac{\sum_{f_i=300}^{4500} B_{f_i}}{\sum_i B_i} \quad (1)$$

By using this descriptors, the accuracy of the system can be incremented in a 4%.

4.2.2. Rhythm Transform

In order to make the system robust to multiple spec-

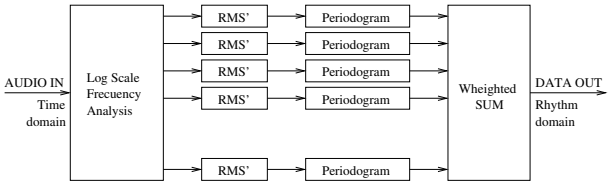

Fig. 1: Block diagram for *Rhythm Transformation*

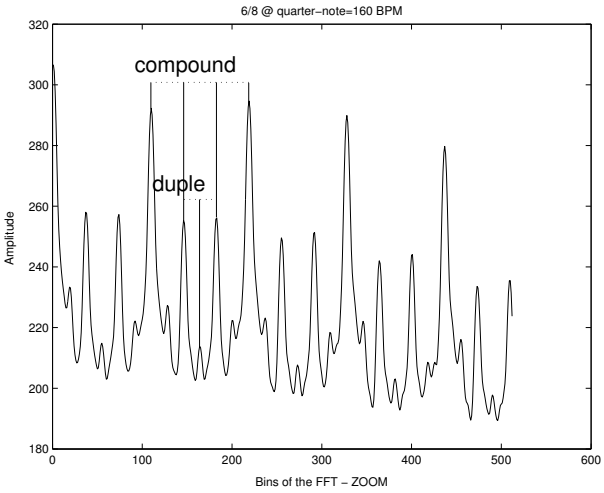

Fig. 2: Example of data in *rhythm domain*

tral codifications such as GSM or MP3, a new rhythmical descriptor is used [9]. The method is based on the periodogram computation of the multi-band pre-processed input data (See Fig. 1). The whole process is what we call *Rhythm Transformation* and it transforms audio data from time domain to a so called *rhythm domain*. The goal of this method is that data in *rhythm domain* can be interpreted as frequency domain information (for BPM detection) as well as time domain information (for meter detection). In Fig. 2, an example of data in *rhythm domain* is shown. Some musical information, such as the meter (simple or compound, duple or triple, swinged or non-swinged), can be extracted from input audio too. In our case, the LPC coefficients of data in rhythm domain are used. By using this descriptor, the accuracy of the system is increased specially when the input signal is Classical Music (see Sec. 5).

4.3. Training

From now on, we have a set of `*.wav`, `*.htk` and

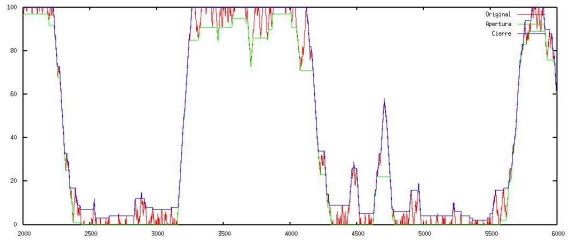

Fig. 3: *opening* and *closing* operations

`*.lab` for each file (audio, parametrization and labels). After some experimental tests, the best results are obtained by using models as follows:

Number of means and vars per state: This value is fixed by the descriptors' selection (see Sec. 5 for details).

Number of States: Our system will have 3 states, that is, only one state plus the input state plus the output state.

Number of Gaussian Mixtures: The model will be created, initially, with only 1 Gaussian mixture. After the initialization, we will increase the number of Gaussian Mixtures up to 16.

Left to right model: The Transition matrix will not allow backward paths.

Finally, the training process is started and the trained models are saved for a future use in our real-time application.

4.4. Mathematical Morphology

As we have discussed before, the use of HMM does not give us a digital output, that is, a *Speech* or *Music* label. Then, some post-processing techniques are needed. After some tests and discussions, some Mathematical Morphology techniques are used. Mathematical Morphology[10] is a set of non-linear techniques based on *maximum* and *minimum* operators. The basic operations of mathematical morphology are: *dilation*, *erosion*, *opening* and *closing*.

In Fig. 3 there is a comparison between the *opening* and *closing* operations applied to an input signal. Low values are assigned to *Speech* and high values are assigned to *Music*. In the context of the AIDA

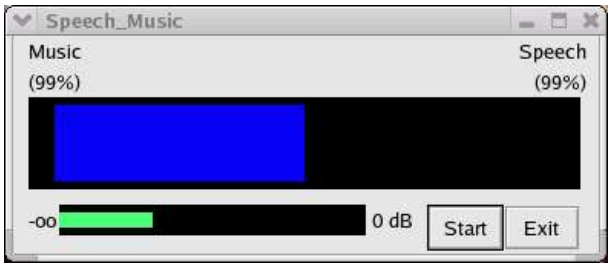

Fig. 4: Graphical User Interface

project, we have to assume that not false positive values are allowed. Let's define a "false positive hit" when the label *Speech* is the output of the system for a musical signal at the input. On the other hand, as we can see in Fig. 3, the opening operation gives results below the original points, while the closing operation gives results above the original points. Then, the input signal can be cataloged as *Speech* under two conditions:

1. The result of the opening operation is exactly 0%.
2. The result of the closing operation is under a threshold, manually selected (5% in our case).

4.5. Graphical User Interface

As the system is implemented by using the AMADEUS technology, some graphical results can be shown (see Fig. 4). This monitoring tool is a real-time implementation in a Pentium IV, 2.5MHz, 512Mb RAM and Red-Hat 9 operating system.

5. EXPERIMENTS AND RESULTS

5.1. First approach

The first tests we made were based on all the descriptors shown in Table 2. This is a really large amount of data and, of course, the process could not be executed in real-time. Then, some tests were made in order to select only the representative descriptors and get an optimized version. The parameters used in each test are shown in Table 3, and the results of the tests for all the configurations and all the audio files are shown in Table 4. The tests are made against a set of 11 real audio recordings. Each audio file is 10 minutes long and manually labeled for this purpose. All these preliminary tests

#Id	Parameters used
A	MFCC
B	Energy
C	Δ Energy
D	Δ Energy
E	4Hz Modulation
F	Δ4Hz Modulation
G	Δ4Hz Modulation
H	Spectral Centroid
I	ΔSpectral Centroid
J	ΔSpectral Centroid
K	Spectral Flatness
L	ΔSpectral Flatness
M	ΔSpectral Flatness
N	Zero Crossing
O	ΔZero Crossing
P	ΔZero Crossing
Q	Voice to White
R	ΔVoice to White
S	ΔVoice to White
T	$A + B + C + \ldots + S$
U	$A, E, H, K, N, Q, \Delta MFCC, \Delta MFCC$
V	$N + O$
W	$U + O$
X	$H + I$
Y	$K + L$
Z	$Q + R$
AA	$E + F$
AB	$U + C + F + I + L + O + R$
AC	AB with $frame = 1000[ms]$

Table 3: Descriptors used for initial tests

(except for the last one) have been made with the next properties: $f_s = 22050[Hz]$, $Frame = 200[ms]$, $hopsize = 50[ms]$, 3 state model (1 state + input state + output state) and 16 mixture models. The best results are obtained with the AB descriptors combination, getting an accuracy of 83.0% (The AC set of descriptors combination has been discarded for computational problems).

These are not really good results. Some short-time false positive hits makes the accuracy go down. This problem will be arranged with the post-processing techniques.

5.2. Rhythm Transform

The inclusion of the *rhythm transform* descriptor

#Id	1	2	3	4	5	6	7	8	9	10	11	$\bar{x}$	τ
A	72.3	68.5	69.1	81.6	65.7	66.5	60.3	71.9	70.7	74.6	76.4	70.69	5.46
B	46.8	54.2	55.2	51.0	60.6	53.6	63.8	53.0	54.3	65.0	54.8	55.66	5.17
C	40.7	49.0	52.2	45.7	51.9	44.7	58.3	50.5	50.3	60.7	48.7	50.06	5.17
D	40.7	49.0	52.2	45.7	51.9	44.7	58.3	50.5	50.3	60.7	48.7	50.20	5.38
E	40.7	48.9	50.7	45.7	51.9	44.7	58.1	50.5	50.3	60.7	48.6	50.08	5.39
F	40.8	49.1	52.2	45.8	52.0	44.8	58.2	50.6	50.1	60.6	48.6	50.20	5.39
G	42.3	48.7	49.5	47.5	49.5	45.7	54.8	47.9	49.4	56.2	48.4	49.08	3.64
H	50.3	55.2	74.7	58.2	72.2	57.8	65.0	65.1	51.7	75.5	66.3	62.90	8.51
I	40.7	49.0	49.1	45.7	51.7	44.7	58.3	50.5	50.3	60.7	48.7	49.94	5.42
J	40.7	49.0	44.8	45.7	45.7	44.7	58.3	50.5	49.1	60.7	48.7	48.90	5.66
K	63.7	50.8	66.3	83.3	67.4	67.7	70.4	77.5	55.9	80.8	76.9	69.10	9.67
L	53.2	49.2	52.1	52.9	51.7	52.0	48.5	49.3	51.5	52.4	52.3	51.37	1.53
M	44.7	50.5	52.0	49.1	51.8	50.8	57.4	54.2	54.0	56.8	52.7	52.18	3.39
N	60.4	56.0	63.4	52.6	66.6	43.3	68.9	66.9	51.9	76.9	61.6	60.77	8.93
O	56.4	60.4	54.0	61.9	56.2	57.7	63.8	54.5	62.3	59.9	57.8	58.62	3.10
P	41.3	49.0	52.1	46.2	51.9	44.8	58.3	50.9	50.3	60.7	48.7	50.38	5.31
Q	67.4	42.7	56.5	52.6	56.0	56.0	53.6	67.9	43.8	63.1	61.4	56.45	7.93
R	54.6	42.9	56.2	50.7	47.9	42.5	50.3	48.0	53.1	55.0	47.0	49.80	4.45
S	43.0	50.9	52.9	46.3	52.6	45.3	59.6	49.3	50.1	59.4	48.2	50.68	5.04
T	89.0	82.5	83.9	94.6	75.5	74.3	79.7	88.3	88.2	93.2	91.4	84.63	6.80
U	87.7	80.4	82.5	94.3	72.4	71.4	71.1	83.5	78.6	89.3	92.6	82.16	7.91
V	73.3	71.0	76.0	66.6	81.6	45.6	79.2	74.6	66.9	81.7	69.9	71.49	9.61
W	89.4	81.6	81.4	94.8	73.0	71.3	72.9	84.9	78.0	89.9	92.7	82.70	7.93
X	57.7	61.0	77.4	61.0	74.9	61.2	69.0	67.2	50.5	81.7	68.5	66.37	8.80
Y	63.8	50.8	66.3	83.6	67.4	70.4	77.6	55.9	80.8	76.9	69.1	69.10	9.60
Z	67.6	42.5	55.9	52.6	53.7	56.2	44.0	68.3	48.9	62.8	61.6	55.82	8.29
AA	40.7	48.8	50.7	45.7	51.9	44.7	58.1	50.5	50.3	60.7	48.7	50.07	5.39
AB	87.1	81.6	82.0	93.7	75.2	70.5	76.9	84.3	78.0	91.8	92.0	83.00	7.22
AC	86.9	86.3	83.2	93.2	84.3	74.3	81.4	91.0	78.1	95.1	88.9	85.69	6.03

Table 4: Evaluation results for all the combinations

in our experiment is to make both the system robust against frequency manipulations and increase the accuracy for Classical Music. Previous audio test files have no excerpts of Classical Music. In fact, the system labels as *Speech* the classical music files. The test is configured with the descriptors shown in Table. 5. The used files are the same than those defined for precious experiments, but with some excerpts of classical music included. Then, the length of the files is now about 15 minutes. Results are shown in Table 6. Although results are less impressive than the previous ones, classical music can be included in our system. On the other hand, the error is basically introduced for the short-time false positive hits.

5.3. Mathematical Morphology

As mentioned before, non-linear mathematical morphology techniques are applied in the post-processing part of the system. With the inclusion of these techniques, we can avoid the system fails for short-time false positive hits. The "short-time" period is selected according to the length of the structural element for the *opening* and *closing* operators. Furthermore, with mathematical morphology techniques applied, we can exactly define the point in which the system labels the input audio as *Speech*. We won't consider as an error all those audio parts with speech, music or both speech and music (news,

#Id	1	2	3	4	5	6	7	8	9	10	11	$\bar{x}$	τ
A	82.7	75.9	81.3	86.4	80.1	71.2	81.6	82.9	78.1	87.9	81.9	80.9	4.43

Table 6: Evaluation results for rhythm tests

Descriptor	L	Value	∇	∇^2
4Hz Modulation	1	*	*	
Spectral Centroid	1	*	*	
Spectral Flatness	1	*	*	
Zero Crossing	1	*	*	
Voice to White	1	*	*	
MFCC	12	*	*	*
Rhythm Transform	12	*		

Table 5: Descriptors used for rhythm tests

commercials, films, etc.) labeled as *Music*.

Taking into account all the considerations, we can assure that the accuracy of the system, with post-processing techniques applied, can be up to 94.3%.

Finally, some GSM audio files have been tested in our system. It is really difficult to give an exact number for the accuracy in this case. As the input audio files are obtained just recording audio with a cellular phone near a loudspeaker, the quality of the GSM codification is unknown. We have seen that better results are obtained when we use the Rhythm Transform descriptors: the accuracy is near 85%. Results are right, but more efforts have to be made in that sense.

6. CONCLUSIONS

Some new techniques for the speech-music discrimination problem have been presented in this paper. The inclusion of these techniques allows the system to be independent of the audio source and codification. Then, no restrictions are assumed. Our research has been focused on the rhythmic aspects of audio, due to the typical compression algorithms are frequency-based: the frequency and the timbre of audio are usually modified but, from now on, we have never heard a compression technique that modifies the rhythm of the audio. Non-linear techniques such as Mathematical Morphology are presented, and interesting results are shown too. Although these results are quite good, more research is needed in this field.

7. REFERENCES

[1] J. Saunders: *Real-Time Discrimination of Broadcast Speech/Music*, Proc. ICASSP, 1996, pgs. 993-996.

[2] E. Scheirer and M. Slaney: *Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator*, Proc. ICASSP, 1997, pgs. 1331-1334.

[3] W. Chow, and L. Gu: *Robust Singing Detection in Speech/Music Discriminator Design*, Proc. ICASSP, 2001, pgs. 865-868.

[4] Stefan Karnebck: *Discrimination between Speech and Music based on a Low Frequency Modulation feature*, Proceedings of Eurospeech, 2001.

[5] Gethin Williams and Daniel P.W. Ellis: *Speech/Music Discrimination based on Posterior Probability Features*, Proceedings of Eurospeech, 1999, pgs. 687-690.

[6] Adam L. Berenzweig and Daniel P.W. Ellis: *Locating Singing Voice Segments within Music Signals*, Proc. ICASSP, 2001.

[7] M. J. Carey, E. S. Paris and H. Lloyd-Thomas: *A comparison of features for speech,music discrimination*,Proc. ICASSP, 1999, pgs. 149-152.

[8] E. Batlle, J. Masip and E. Guaus: *AMADEUS: A scalable HMM-based audio information retrieval system*, ISCCSP, 2004.

[9] Enric Guaus, Eloi Batlle: *Visualization of metre and other rhythm features*, Proc. ISSPIT, 2003.

[10] R. Haralick, S. Sternberg, X. Zhuang: *Image Analysis using Mathematical Morphology*, IEEE PAMI 9, 532, 1987.