Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona
September 2011

Master in Sound and Music Computing

Genre Classification based on Predominant Melodic Pitch Contours

Bruno Miguel Machado Rocha

Supervisors:
Emilia Gómez
Justin Salamon

UNIVERSITAT POMPEU FABRA

Abstract

We present an automatic genre classification system based on melodic features. First a ground truth genre dataset composed of polyphonic music excerpts is compiled. Predominant melodic pitch contours are then estimated, from which a series of descriptors is extracted. These features are related to melody pitch, variation and expressiveness (e.g. vibrato characteristics, pitch distributions, contour shape classes). We compare different standard classification algorithms to automatically classify genre using the extracted features. Finally, the model is evaluated and refined, and a working prototype is implemented.

The results show that the set of melody descriptors developed is robust and reliable. They also reveal that complementing low level timbre features with high level melody features is a promising direction for genre classification.

Acknowledgements

I would like to thank my supervisors, Emilia Gómez and Justin Salamon, for their invaluable help and support throughout this year. Xavier Serra for giving me the opportunity to participate in this master. Perfecto Herrera and Enric Guaus for the suggestions and advices over this year. All my colleagues and members of the MTG for the fruitful discussions we had that helped to complete this thesis.

I would also like to thank my closest friends for the encouragement they gave whenever I needed. Gil for the technical assistance. Nuna for her patience and understanding.

Finally, my family, especially my parents, brother, sisters and nieces for their undeniable support in every way possible; this thesis exists because of them.

Table of Contents

List of Figures

List of Charts

List of Tables

1. Introduction

1.1 Overview

The goal of this project was to build a genre classifier using melody features extracted from predominant melodic pitch contours, estimated from polyphonic music audio signals. The classifier focuses on differentiating between distinct singing styles such as pop, jazz or opera. It uses existing technology for predominant pitch contour estimation from polyphonic material. The project also generated a ground truth dataset for system evaluation.

This chapter provides a brief introduction to the thesis, stating its goals and the motivation behind the project. In chapter 2 we review the state-of-the-art. The methodology is described in chapter 3. In chapter 4 we discuss the results and then we make some conclusions and propose future work in chapter 5.

1.2 Motivation

Listeners try to describe music with words. One of the notions most people use is *melody*, but can anyone define it? People also tend to classify the world around them into categories. Music is no exception. Musical genres are the main top-level descriptors used by music dealers and librarians to organize their music collections. Though they may represent a simplification of one artist's musical discourse, they are of a great interest as summaries of some shared characteristics in music pieces (Scaringella, Zoia & Mlynek, 2006). Possible definitions for both concepts are provided in this thesis.

"Ever since the emergence of digital signal processing, researchers have been using computers to analyze musical recordings, but it has proven more challenging than expected to recognize the kinds of aspects (...) that are usually trivial for listeners" (Poliner et al., 2007). The field of Music Information Retrieval (MIR) has significantly

contributed to the advances made in our ability to describe music through computational approaches.

The main reason behind this project was the same that led me to apply for this master: the possibility of combining my musical background with the technologies developed at the Music Technology Group (MTG) of this university.

Another motive was the possibility of classifying musical genre using mid and high level features. Figure 1 depicts all the disciplines related to MIR. Guaus (2009) states that for automatic genre classification of music we need information from digital, symbolic and semantic levels. Most attempts to classify genre have been dealing with low-level timbral and spectral features. Some have tried also tonal and rhythm features (Tzanetakis & Cook, 2002; Ellis, 2007; Gouyon et al., 2004), but, to our knowledge, melody features have not been used before. Lately, some researchers have incorporated source separation techniques to improve this task. Melody features have the advantage of making more sense to users than traditional low level features, also allowing us to easily interpret results.

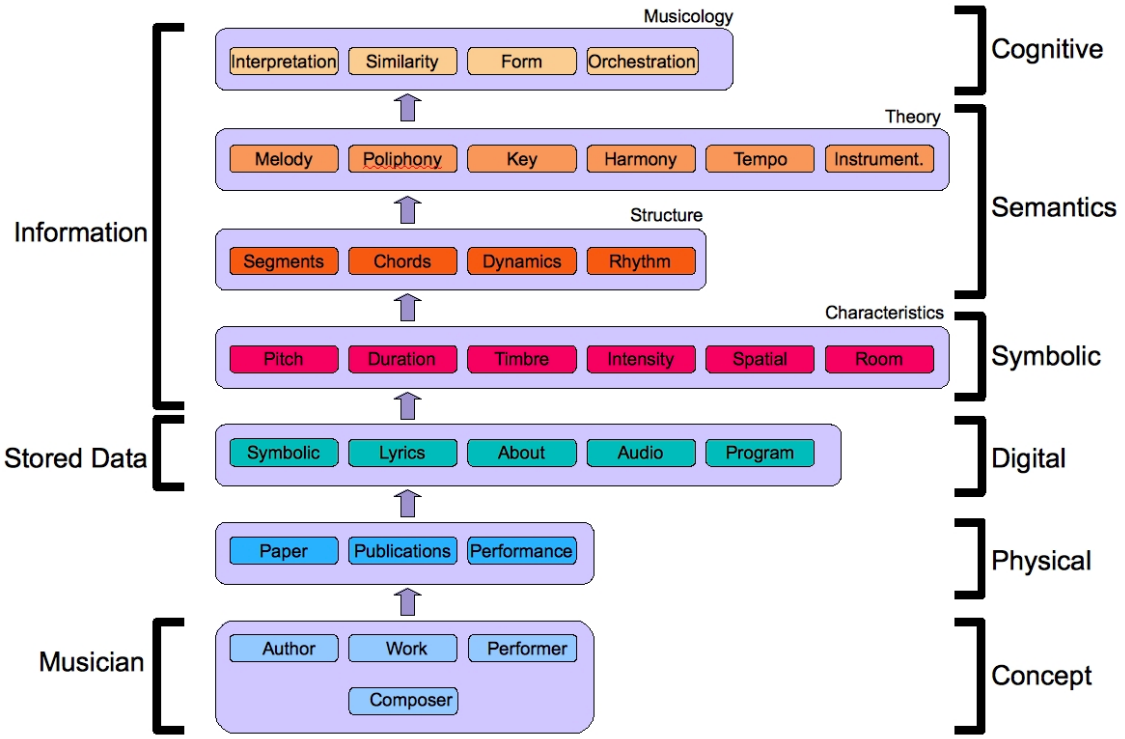

Figure 1: Simplified version of the MIR map proposed by Fingerhut and Donin (as depicted in Guaus, 2009)

1.3 Goals

The main goals of this thesis are:

- Provide a state-of-the-art review in the fields of the singing voice, melody description and genre classification;
- Build a genre classifier using melody features extracted from predominant melodic pitch contours, estimated from polyphonic music audio signals;
- Achieve a set of reliable melody descriptors;
- Generate a ground truth dataset for system evaluation;
- Evaluate our method employing different datasets and compare it to other approaches;
- Fuse low level and mid/high level descriptors to see if the accuracy of the system improves;
- Discuss the results, finish off the work carried out and propose future work.

2. State-of-the-art Review

This state-of-the-art review aims to analyse current research on the most significant areas related to this thesis. It covers subjects such as the acoustics of the singing voice (section 2.2), different singing styles (section 2.3), automatic melody description (section 2.4) and genre classification (section 2.5).

2.1 Definitions

Before going further, we need to clarify some relevant terms used in this work that have no consensual definition.

a) Melody

Melody is a musicological concept based on the judgment of human listeners (Poliner et al., 2007), and its definition can change according to culture or context. The same authors propose the definition adopted in this work:

> "(…) the melody is the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that the listener would recognise as being the 'essence' of that music when heard in comparison".

b) Musical Genre

Musical genre is a concept that is discussed by every music lover and, generally, never agreed on. For simplicity, we adopt Guaus (2009) definition in this paper:

> "the term used to describe music that has similar properties, in those aspects of music that differ from the others".

2.2 Acoustics of the Singing Voice

a) The Voice Source

The voice organ includes the lungs, larynx, pharynx, mouth and nose. Analysing the voice organ as a sound generator, three main parts can be distinguished: a generator (the respiratory system), an oscillator (the vocal folds) and a resonator (the vocal tract). An air stream is generated when the lungs are compressed, if the airways are open. This air stream sets the vocal folds vibrating, creating a pulsating air flow, the *voice source*, which is controlled by the air pressure in the lungs and the vocal folds. The voice source is a chord of simultaneously sounding tones of different frequencies and amplitudes. This sound is filtered by the vocal tract, which has the function of acoustically forming the output sound. We call the vocal tract resonances *formants*. Each formant produces a peak in the frequency curve of the vocal tract. The properties of the voice source plus the frequencies of the formants determine the vowel quality and the personal timbre we perceive in a voice (Sundberg, 2000).

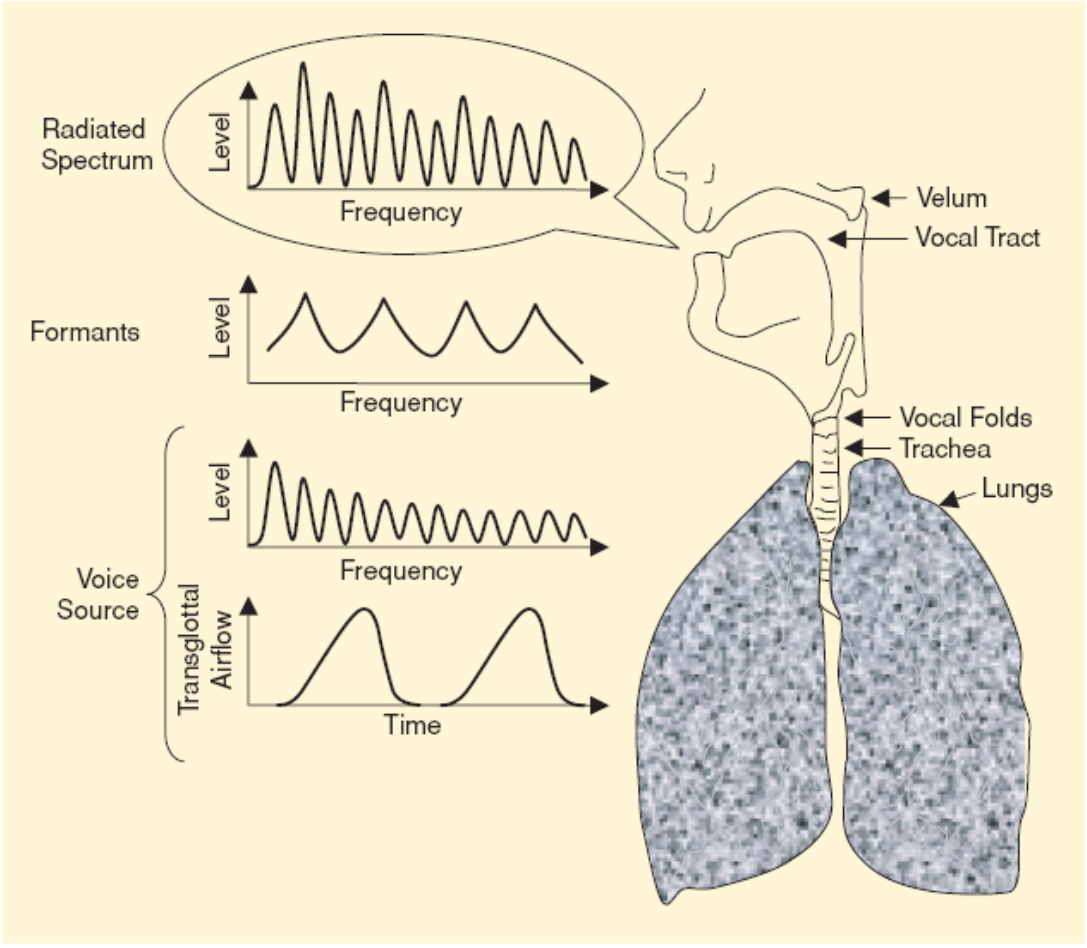

Figure 2: The voice organ (Bonada & Serra, 2007)

b) Pitch

Pitch is a perceptual concept and is distinct from fundamental frequency. For the sake of simplicity, in this thesis we use both concepts interchangeably.

The pitch of the human voices is determined by the frequency with which vocal folds vibrate. The artistically acceptable range for most singers is two to two-and-a-half octaves although they can produce notes of higher and lower pitch (Bunch, 1997). According to Sundberg (2000), the approximate ranges covered by the main voice types are:

Voice Type	Frequency Range
Bass	80 – 330 Hz
Tenor	123 – 520 Hz
Alto	175 – 700 Hz
Soprano	260 – 1300 Hz

Table 1: Approximate frequency ranges covered by the main voice types (Sundberg, 2000)

The vocal folds vibrate in different modes according to the pitch that is being produced. These modes are called *vocal registers*. There are at least three registers: vocal fry, chest (or modal) and falsetto.

c) Vibrato

One of the most controversial aspects of singing is the role played by vocal tremulousness: should the voice be steady or should it exhibit some form of tremulousness? (Stark, 1999). Stark concludes that vocal tremulousness has been an important component in good singing since at least the early Baroque period.

The most familiar form of tremulousness is today known as *vibrato*. Vibrato is a voice source characteristic of the trained singing voice. It corresponds to an almost sinusoidal modulation of the fundamental frequency (Sundberg, 1987). According to Seashore (1938/1967), its rate and extent are fairly constant and regular. Bunch states that a good

singer will have an average vibrato of five to eight regular pulsations per second, referring studies by Seashore and other researchers. Seashore affirms the average extent of the pitch pulsation for good singers is a semitone, although there can be a variation of individual vibrato cycles of 0.1 to 1.5 of a tone from the singer's characteristic average. Benade's study (as cited in Stark, 1999) claims that vibrato may aid in the inteligibility of the vowels by sweeping the vowel formants, and may also give the voice greater audibility, due to the independence of vibrato from rhythmic patterns of the music.

An important vocal ornament that may be related to the vibrato is the *trill*. Stark (1999) defines it as "a rapid alternation between two adjacent notes, in which the voice oscillates between separate pitches, perhaps using the same impulse that drives the vibrato".

2.3 Singing Styles

Different styles of singing apparently involve different manners of using the voice (Thalén & Sundberg, 2001). In this work we concentrate on four musical genres that are associated with different approaches to singing: pop, vocal jazz, flamenco and opera. Reasons for choosing these styles will be presented later in this thesis (section 3.1).

a) Pop

Originated in Britain in the mid-1950s, the term "pop music" described the new youth music influenced by rock-and-roll (Middleton et al., 2011).
Middleton (2000) summarized the generally agreed core tendencies of pop singing: short phrases often much repeated, falling or circling in shape, usually pentatonic or modal; call-and-response relationships between performers; off-beat accent, syncopation and rhythmically flexible phrasing; a huge variety of register and of timbre.

b) Vocal Jazz

Jazz is a style characterised by syncopation, melodic and harmonic elements derived from the blues, cyclical formal structures and a supple rhythmic approach to phrasing

known as swing (Tucker & Jackson, 2011). Most of 20th century's great vocalists performed in the jazz idiom, establishing the style known today as vocal jazz.

Louis Armstrong's contribution to the evolution of jazz singing was essential. He was able to fashion a singing that was very close to his speech, using a similar technique to some of the early blues singers, but his singing removed any residual 'classical' tendencies from popular singing, making it ultimately susceptible to swing in the same way as instrumental music. The sustained and cultured tone of a conventional singer is less likely to facilitate swing than a speech-like shaping of syllables, words and phrases. He showed that being a horn player or a singer were not so very different from each other, and that the basic requirements of singing were to do with feel and personality (Potter, 2000).

c) Flamenco

Flamenco is the generic term applied to a particular body of *cante* (song), *baile* (dance) or *toque* (solo guitar music), mostly emanating from Andalusia in southern Spain (Katz, 2011).

Merchán (2008) makes some conclusions about the behaviour of flamenco melodies: short intervals (2nd, 3rd) are very common and most of the movements are adjacent degrees; short pitch range (up to a 6th); high degree of ornamentation (melisma) for improvisation.

Vibrato in flamenco differs from other styles, as it is hardly distinguishable from a melismatic ornament and it is unstable.

The dynamics in flamenco are very irregular. A phrase may begin with very soft utterances and end with an intense flow of voice, changing suddenly in between.

d) Opera

The word *opera* can be generically defined as a staged drama in which accompanied singing has an essential function, portraying the actions and emotions of the characters (Arnold et al., 2011).

Opera singing exhibits great variability in several aspects, such as pitch range, dynamics, or melodic contour length. Vibrato is regular and tuning is good, as it is a genre commonly performed by trained professional singers.

2.4 Automatic Melody Description

Most listeners are able to recognise and sing or hum the melody in a polyphonic piece of music. However, performing this task automatically using computers is still regarded by researchers as an unsolved problem. While the pitch of a single note is consistently represented as a waveform with a more or less stable periodicity, polyphonic music will often have overlapping notes with different fundamental frequencies and their respective series of harmonics, that can actually coincide, which appears to be at the core of musical harmony (Poliner et al., 2007). Figure 3 shows the basic processing structure of typical melody transcription systems.

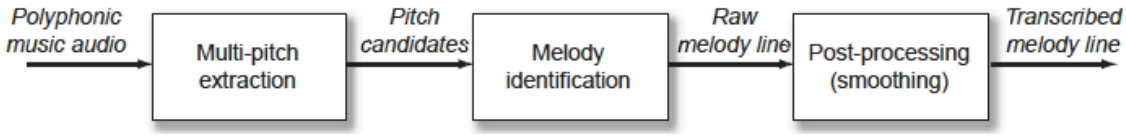

Figure 3: Basic processing structure underlying all melody transcription systems (Poliner et al., 2007)

All approaches to melody transcription face two problems: identifying a set of candidate pitches that appear to be present at a given time, then deciding which (if any) of the pitches belongs to the melody. In 2007, Poliner et al. reviewed some melody transcription systems that participated in previous MIREX contests, concluding there was a common processing sequence to most of these systems (Figure 3), resumed in Salamon (2008):

- ⇒ Multi-pitch extraction: from an audio input, a set of fundamental frequency candidates for each time frame is obtained.
- ⇒ Melody identification: selecting the trajectory of F0 candidates over time which forms the melody.
- ⇒ Post processing: remove spurious notes or otherwise increase the smoothness of the extracted melody contour.

A current trend in melody extraction systems (and in other music information retrieval disciplines) is the adoption of source separation methods to help in this process. Harmonic/Percussive Sound Separation (HPSS) was used by Tachibana et al. (2010)

and Hsu and Jang (2010) in the MIREX 2010 evaluation campaign, while Durrieu, Richard and David (2008) proposed Non-Negative Matrix Factorization techniques.

Despite the source separation trend, salience based methods are still amongst the best performing systems. In this work, we used the salience based method designed by Salamon and Gómez (2010), which achieves results equal to current state-of-the-art systems and was one of the participants in the MIREX contest of 2010.

2.5 Genre Classification

Music genre classification is a categorization problem that is object of study by different disciplines, such as musicology, music industry, psychology or music information retrieval (Guaus, 2009). Automatic music genre classification is the task of assigning a piece of music its corresponding genre.

One of the most relevant studies on automatic musical genre classification was put together by Tzanetakis and Cook (2002). In this paper, the researchers propose the extraction of timbral texture features, rhythmic content features and pitch related features. For classification and evaluation, the authors use Gaussian mixture model (GMM) classifiers and K-nearest neighbour (K-NN) classifiers. The dataset is composed of 20 musical genres and three speech genres with 100 samples of 30 seconds per genre. The 20 musical genres are then divided into three smaller datasets (genres, classical and jazz). Many experiments and results are discussed and the accuracy of the system reaches 61% of correct classifications in the bigger genres dataset, 88% in the classical one and 68% in the jazz one.

Since then, different sets of descriptors and classification techniques have been used to improve the accuracy of the classification. As in the recent years' submissions to MIREX melody extraction contest, audio source separation techniques have also been used in music genre classification systems (Rump et al., 2010), although timbral features and their derivatives continue to be the most used (Ellis, 2007; Langlois & Marques, 2009; Genussov & Cohen, 2010; Bergstra, Mandel & Eck, 2010).

Panagakis, Kotropoulos, and Arce (2009) proposed a robust music genre classification framework combining the properties of auditory cortical representations of music recordings and the power of sparse representation-based classifiers. This method achieved the best accuracies ever on two of the most important genre datasets: GTZAN (92%) and ISMIR2004 (94%).

Music genre classification can be considered as one of the traditional challenges in the music information retrieval field, and MIREX has become a reference point for the authors, providing a benchmark to compare algorithms and descriptors with exactly the same testing conditions (Guaus, 2009).

Concerning melodic features, they have been used for automatic genre classification of MIDI recordings. McKay (2004) collected statistics about melodic motion and intervals and used them to implement features. However, to our knowledge, there have never been any attempt to classify audio recordings using features based on melodic contours.

3. Methodology

Our framework followed the basic process of building a Music Information Retrieval classifier: dataset collection, feature extraction, machine learning algorithm, evaluation of the trained system (Guaus, 2009). Figure 4 shows a block diagram for a basic system.

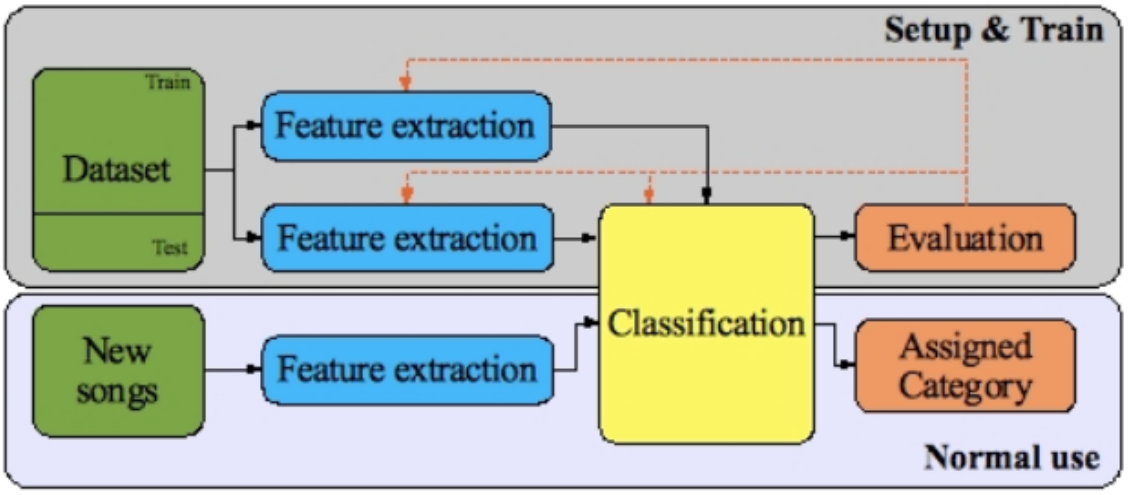

Figure 4: Block diagram for a basic automatic classification system (Guaus, 2009)

3.1 Dataset Collection

For the purpose of this thesis, we wanted to focus on genres that have a clear melodic line and distinct characteristics. Thus, we decided to compile a new dataset. First, we had to decide on the musical genres to include. We tried to choose very different genres that cover a broad scope, in which the vocals carry the melody. After some discussion, the chosen genres were pop, vocal jazz, flamenco and opera. We decided to have an instrumental music genre in order to see if there is much difference between the extracted melodies from a physical instrument and the singing voice. Instrumental jazz can be a good source of comparison due to the fact that some of its performers were also singers, such as Chet Baker or Louis Armstrong. Fifty excerpts with the duration of approximately 30 seconds were gathered for each genre. In order to minimize possible errors of the predominant melody extractor, voice is predominant in the chosen excerpts. We tried to keep the number of snippets per artist below 2 when possible. We now describe the excerpts selected for each genre. A summary of the dataset is presented in the end of this section.

a) Pop

As the boundaries between pop and other genres such as rock are very thin, we tried to focus on songs and artists that most people would consider to be "pop". For this database, excerpts ranging in time from the 1980s to the current year were obtained. From these, 26 are sung by females and 24 by males.

b) Flamenco

Flamenco was chosen not only because it is widely studied in Spain, but also for its vocal technique, which is very particular, and the type of songs that vary a lot, based on the different *palos* (Fernandez, 2004). In this genre, the equilibrium between female and male excerpts was more difficult to obtain. In the end, 34 male and 16 female sung snippets were chosen.

c) Opera

"Classical" or "erudite" music is represented in this dataset by its most notable form involving singing: the opera. All periods of opera are represented here, from baroque arias to modern ones. 28 excerpts are sung by females, while 22 are sung by males.

d) Vocal Jazz

For this genre, we gathered excerpts ranging in time from the 1950s to the 21st century. Its singing style has a lot in common with pop, which makes it a hard task to distinguish between both singing styles, even for humans (Thalén & Sundberg, 2001). For that reason, artists such as Frank Sinatra or Nat King Cole are not present in the database. From the fifty excerpts gathered, 36 are sung by females and 14 by males.

e) Instrumental Jazz

As this is a huge genre and this thesis is mainly concerned with vocal melodies, we focused on getting excerpts which have clear mid-tempo melodies, some of them very similar to the ones in the vocal jazz excerpts. Saxophonists or trumpeters, with the exception of one trombonist and one flutist, play most of the melodies in the excerpts.

Genre	No. Excerpts	Duration	Male	Female
Pop	50	30s	24	26
Flamenco	50	30s	34	16
Opera	50	30s	22	28
Vocal Jazz	50	30s	14	36
Instrumental Jazz	50	30s	50	0

Table 2: Summary of the initial dataset

3.2 Melody Estimation

After building the database, we used Salamon and Gómez's method (2011) to extract the melodies from the polyphonic excerpts. In the first block of the system the audio signal is analysed and spectral peaks (sinusoids) are extracted, which will be used to construct the salience function in the next block. This process is comprised of three main steps: pre-filtering, transform and frequency/amplitude correction. A time-domain equal loudness filter is applied in the pre-filtering stage to attenuate spectral components belonging primarily to non-melody sources. Next, a spectral transform is applied and the peaks of the magnitude spectrum are selected for further processing. In the third step the frequency and amplitude of the selected peaks are re-estimated by calculating the peaks' instantaneous frequency using the phase vocoder method.

The spectral peaks are then used to compute a representation of pitch salience over time, a *salience function*. This salience function is based on harmonic summation with magnitude weighting. In the next block, the peaks of the salience function are grouped over time using heuristics based on auditory streaming cues. This results in a set of pitch contours, out of which the contours belonging to the melody need to be selected. The contours are automatically analysed and a set of contour characteristics is computed.

In the final block of the system, these characteristics are used to filter out non-melody contours. Contours whose features suggest that there is no melody present (voicing detection) are removed first. The remaining contours are used to iteratively calculate an

overall melody pitch trajectory, which is used to minimise octave errors and remove pitch outliers. Finally, contour salience features are used to select the melody F0 at each frame from the remaining contours.

In Figure 5, we can perform a visual inspection on how melodic contours from different genres can look very diverse.

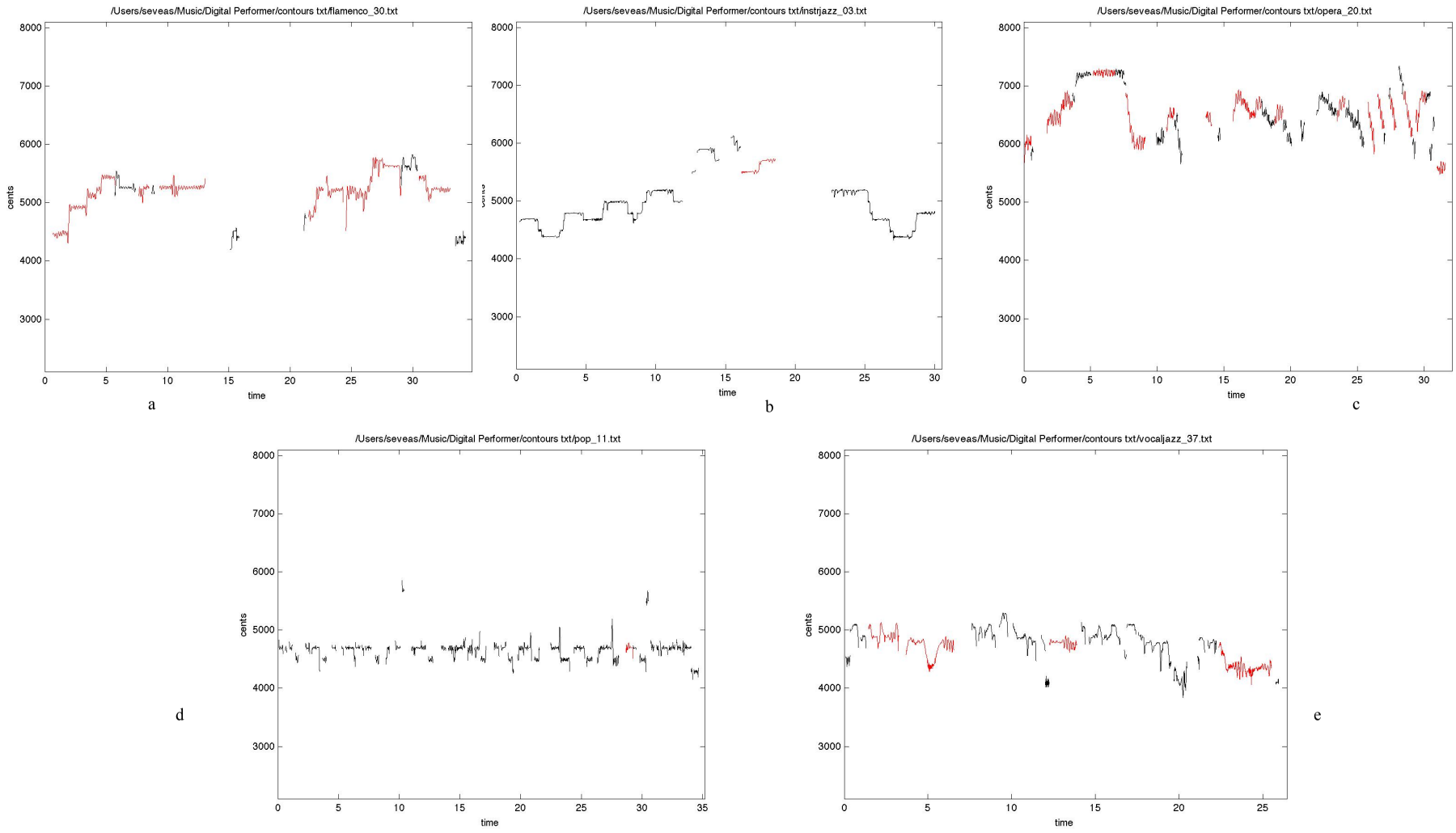

Figure 5: Some examples of melodic contours: a-flamenco; b-instrumental jazz; c-opera; d-pop; e-vocal jazz. Red indicates the contours where vibrato was detected.

3.3 Feature Extraction

For extracting relevant features from the estimated melodies, we used a series of descriptors. These were derived from the results of Salamon's algorithm, which outputs features for each contour of each audio file. The most relevant features extracted are:

- Length;
- Pitch height for each frame;
- Mean pitch height and standard deviation;
- Vibrato presence, extent, rate and coverage (proportion of pitch contour where vibrato is present).

Global descriptors concerning each file, which we will cover in more detail, were computed from these values. We implemented these descriptors in Matlab.
A list of all 89 descriptors is provided in the end of this section.

3.3.1 Pitch Descriptors

a) Pitch Range

For each contour, we retained the pitch values of the first and last frames, as well as the highest and lowest pitch values. From the absolute difference between the former two values, we computed the pitch range for each contour. Then, for each file, we calculated its mean, standard deviation, skewness and kurtosis values. A global pitch range was also estimated from the highest and lowest pitch values in each file.

b) Highest and Lowest Pitch Values

The highest and lowest pitch values for each contour were used as descriptors, as well as their mean, standard deviation, skewness and kurtosis values. The highest and lowest values in pitch for each file were also used as descriptors.

c) Pitch Height and Interval

From the mean pitch height and standard deviation of each contour, we computed the mean, standard deviation, skewness and kurtosis of these values for each file. A

different descriptor derived from these values is what we call “interval”, which we considered to be the absolute difference in cents between the mean pitch height of one contour and the previous one. Its mean, standard deviation, skewness and kurtosis were also computed.

3.3.2 Vibrato Descriptors

a) Ratio of Vibrato to Non-Vibrato Contours

This descriptor is computed by counting the number of contours in which vibrato is detected and dividing it by the total number of contours for each file.

b) Vibrato Rate, Extent and Coverage

Vibrato is detected in a contour when there is a low-frequency variation in pitch between five and eight cycles per second. This value is the vibrato rate output by the algorithm for each contour. The extent in cents (100 cents is a semitone) is also computed, as well as the coverage, which is the percentage of each contour in which vibrato is detected. For all these features, we calculated the mean, standard deviation, skewness and kurtosis as descriptors.

3.3.3 Shape Class Descriptors

Charles Adams (1976) proposed a new approach to study melodic contours, defining them as "the product of distinctive relationships among the minimal boundaries of a melodic segment". "Minimal boundaries are those pitches which are considered necessary and sufficient to delineate a melodic segment, with respect to its temporal aspect (beginning-end) and its tonal aspect (tonal range)".

Following his approach, we computed the initial pitch (I), the final pitch (F), the highest pitch (H) and the lowest pitch (L) for each contour. Adams also referred three primary features as essential to define the possible melodic contour types: the *slope* of the contour (S), which accounts for the relationship between I and F; the *deviation* (change

of direction) of the slope of the contour (D), indicated by any H or L which is different than I or F; the *reciprocal* of deviation in the slope of the contour (R), which expresses the relationship between the first deviation and I, whenever there is more than one deviation.

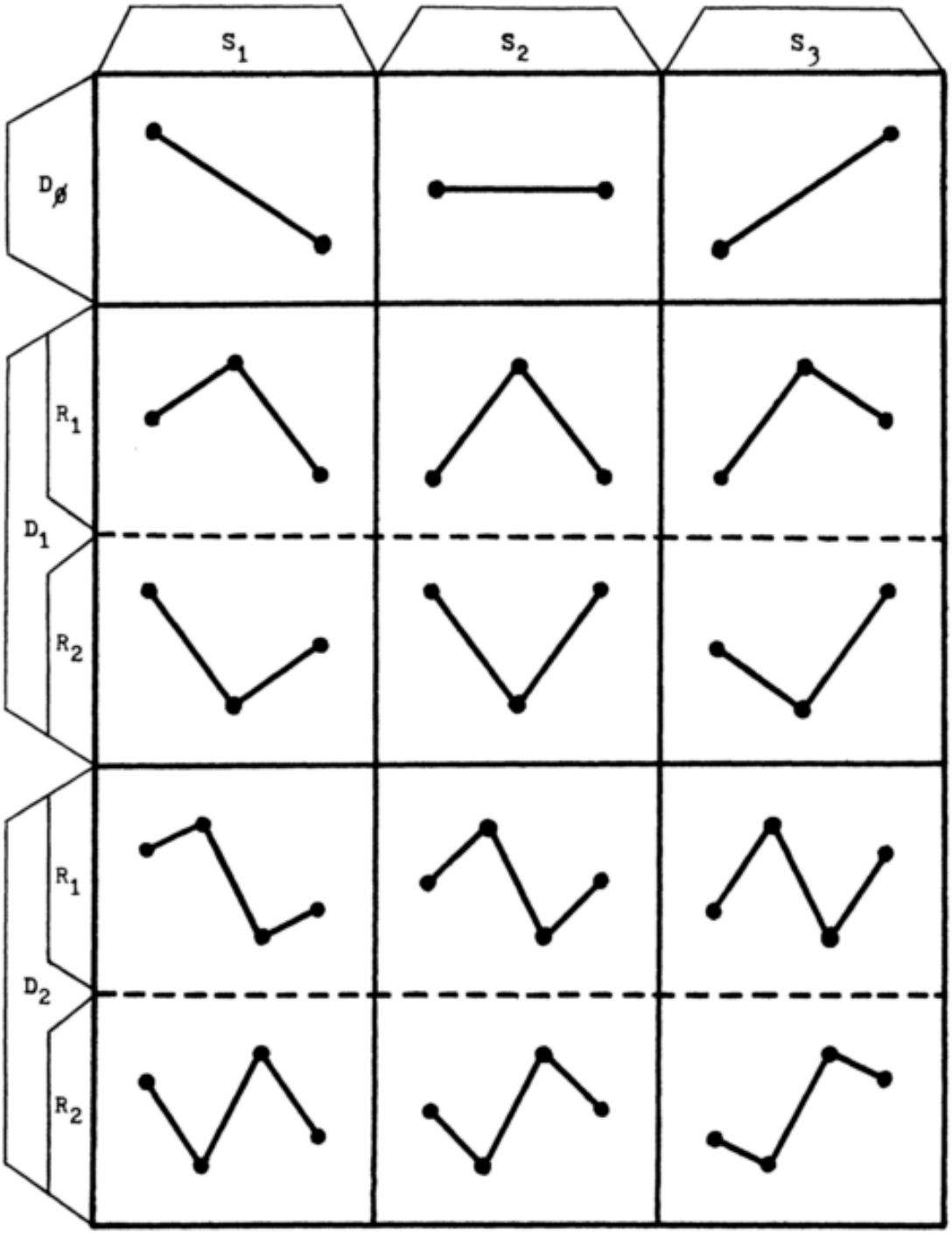

Figure 6: Graphic representation of the fifteen melodic contour types (Adams, 1976)

"Thus, the product of distinctive relationships (features) among the minimal boundaries of a melodic segment defines fifteen melodic contour types" (see Figure 6). Each contour was assigned one of these types.

The contour pitch is described with a resolution of 10 cents. This resolution is too high to compute the shape class directly as an almost straight contour which should belong to class S2D0 could be wrongly classifier as S2D1R1 due to very subtle pitch variation. Similarly, if we quantise the contours to a resolution that is too low we risk losing the shape class altogether. In the end a resolution of one quarter-tone (50 cents) was found to be adequate. The distributions of shape classes were then computed and used as descriptors.

3.3.4 Length-based Descriptors

a) Length

After estimating the duration of each contour, we computed the mean, standard deviation, skewness, kurtosis and the maximum for each file.

b) Length-based Descriptors

Although length itself proved not to be a very useful descriptor, it was helpful to build a series of other descriptors. These are features computed taking into consideration only the longest contours in each file. Apparently, pitch and vibrato related features vary depending on the length of the contours. This may also help to eliminate some noise in the melody estimation.

Pitch (27) descriptors	Contour Pitch Range (mean, standard deviation, skewness and kurtosis)*
	Global Pitch Range
	Contour Highest Pitch (mean, standard deviation, skewness and kurtosis)*
	Contour Lowest Pitch (mean, standard deviation, skewness and kurtosis)*
	Global Highest Pitch*
	Global Lowest Pitch
	Contour Pitch Height Mean (mean, standard deviation, skewness and kurtosis)
	Contour Pitch Height Standard Deviation (mean, standard deviation, skewness and kurtosis)
	Contour Interval (mean, standard deviation, skewness and kurtosis)
Vibrato (13) descriptors	Global Ratio of Vibrato to Non-Vibrato Contours
	Contour Vibrato Rate (mean, standard deviation, skewness and kurtosis)*
	Contour Vibrato Extent (mean, standard deviation, skewness and kurtosis)*
	Contour Vibrato Coverage (mean, standard deviation, skewness and kurtosis)*
Shape Class (15) descriptors	Distributions of Shape Classes (15 types)
Length-based (34) descriptors	Contour Length (mean, standard deviation, skewness and kurtosis)*
	Length of the longest contour per file
	All previous descriptors marked with *

Table 3: List of 89 melody descriptors

3.4 Genre Classification

To perform the classification we used the data mining software Weka (Hall et al., 2009). This software allows the user to choose several filters, very useful for us to understand which are the most important features for a successful classification. It also permits the user to apply different types of classifiers and compare the results between them.

3.4.1 Attribute Selection Methods

The problem of feature selection is well known in machine learning. Given a particular classification technique, it is conceivable to select the best subset of features satisfying a given "model selection" criterion by exhaustive enumeration of all subsets of features (Guyon et al., 2002). After feeding Weka all the computed descriptors, we applied two supervised attribute selection methods. Both are composed of two algorithms: an evaluator and a searcher.

The first method is *CfsSubsetEval + BestFirst. CfsSubsetEval* (Hall, 1999) evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. *BestFirst* searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done.

The second method is *SVMAttributeEval + Ranker*. *SVMAttributeEval* (Guyon et al., 2002) evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is handled by ranking attributes for each class seperately using a one-vs-all method and then "dealing" from the top of each pile to give a final ranking. *Ranker* ranks attributes by their individual evaluations. It is computationally more expensive but allows the user to choose the number of descriptors he wants to keep.

3.5 Evaluation

3.5.1 Evaluation Methodology

To evaluate this work, we decided to implement a baseline approach with which we could compare. A typical approach is to extract Mel Frequency Cepstral Coefficients (MFCC) features and perform the genre classification (Scaringella, Zoia, & Mlynek, 2006). We extracted the first 20 coefficients using the Rastamat Matlab toolbox (Ellis, 2005). The samples were chopped into frames of about 23 ms with 50% overlap and we used 40 Mel frequency bands, up to 16 kHz (following Pampalk, Flexer, & Widmer, 2005). Then, we computed the means and variances for each coefficient, ending with a total amount of 40 descriptors.

We also tried to bind both melodic and MFCC features into the same vector and examine the results to see if it is advantageous to apply this early fusion technique.

3.5.2 Datasets

After initial results were obtained using the 250 excerpt dataset, we expanded it further to include 500 excerpts, 100 per genre. The extended dataset has 100 excerpts of each of the five musical genres, increasing the total number of snippets to 500. The same rules were applied to the collection of the new samples.

We also decided to test the system on an unprepared dataset. The chosen one was the GTZAN Genre Collection (Tzanetakis & Cook, 2002), which consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The genres are: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock.

4. Results

In this chapter we present a comparison of quantitative results between our system and the baseline approach. Results for the initial and the extended datasets are shown, using three approaches, two attribute selection methods and five classifiers.

4.1 Initial Dataset

4.1.1 Attribute Selection: *CfsSubsetEval* + *BestFirst*

Chart 1 shows the results for the initial dataset of 250 excerpts, using three approaches: melodic features (our approach), MFCC features (baseline approach) and a fusion of both. For all of them, the same attribute selection filter was applied. This filter uses the *BestFirst* search method and *CfsSubsetEval* evaluator algorithm from Weka. A 10-fold cross-validation scheme was used for evaluating the performance of all classifiers. One can find between brackets the number of features selected by the filter.

In the case of melodic features, the following 14 are the ones that were selected by the filter:

Mean Pitch Range
Std Dev Length
Std Dev Vibrato Coverage
Shape Class 1
Shape Class 3
Shape Class 4
Highest Maximum Big Length
Std Dev Interval
Mean Highest
Mean Vibrato Rate Big Length
Skewness Vibrato Rate Big Length
Mean Vibrato Extent Big Length
Mean Vibrato Coverage Big Length
Std Dev Vibrato Coverage Big Length

Observing these descriptors, we can notice the relevance of vibrato descriptors, especially the ones that are length-based as well. This reassures our belief that the presence and the type of vibrato is one of the most important characteristics that leads us to distinguish between singing styles.

We now turn to examine the classification results in Chart 1.

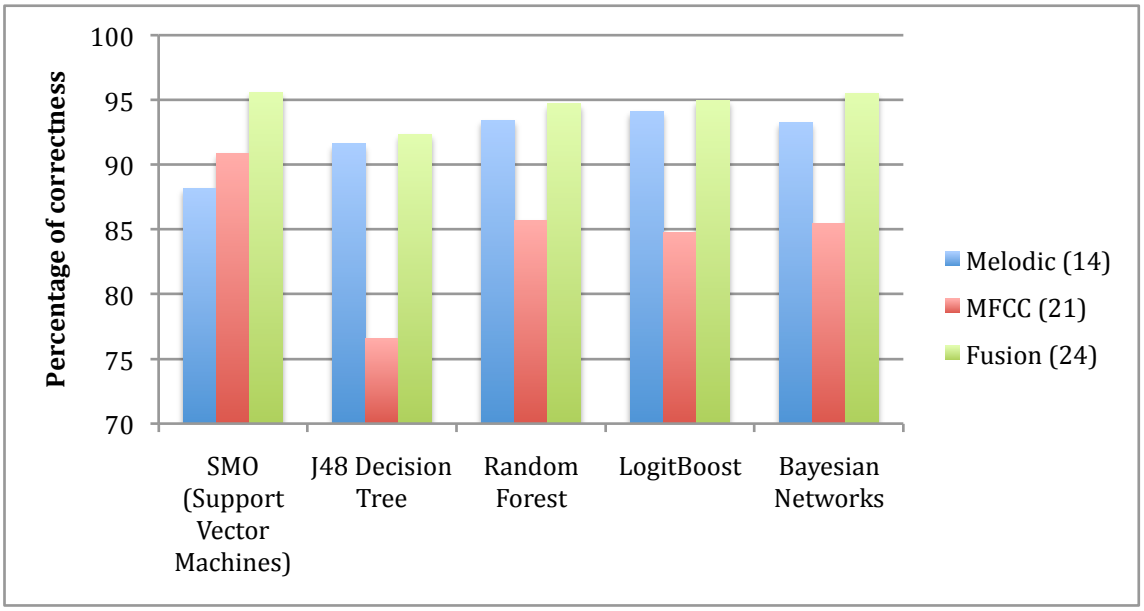

Chart 1: Results for the initial dataset using *CfsSubsetEval+BestFirst* as the attribute selection method

Looking at these results, we can observe our system achieves a performance of over 90% for almost all classifiers, while the baseline approach performs slightly worse in all classifiers except SMO. It is also recognizable a slight improvement in the results when we combine both types of features into a single vector, reaching 95% in the best case.

Regarding the confusion matrices in Figure 7 we can again take some interesting considerations. It was mentioned before (section 3.1) that vocal jazz and pop singing styles have much in common and can be easily mistaken one for the other even by humans. We can confirm that for most of the classifiers using the melodic features this confusion is present - see for example the first melodic confusion matrix, in which 11 vocal jazz excerpts are classified as pop, while 10 pop excerpts are labeled as vocal jazz. When we fuse the melodic and MFCC features, this error is reduced and the overall classification accuracy improves.

a	b	c	d	e
47	0	0	3	0
0	49	0	1	0
0	1	49	0	0
0	3	1	36	10
0	2	0	11	37

a	b	c	d	e
41	0	0	5	4
1	47	0	0	2
0	0	48	0	2
0	0	0	46	4
2	1	0	2	45

a	b	c	d	e
46	1	0	2	1
0	49	0	0	1
0	1	49	0	0
0	1	0	47	2
0	2	0	0	48

a	b	c	d	e
49	0	0	0	1
3	46	0	0	1
2	0	47	0	1
0	1	2	46	1
0	0	0	7	43

a	b	c	d	e
42	0	3	2	3
1	41	1	7	0
6	1	40	2	1
2	3	0	41	4
5	1	4	4	36

a	b	c	d	e
49	1	0	0	0
2	46	0	1	1
2	1	47	0	0
0	2	2	44	2
0	0	0	5	45

a	b	c	d	e
48	0	1	0	1
2	47	0	0	1
1	0	48	1	0
0	1	1	48	0
0	0	0	7	43

a	b	c	d	e
37	1	4	3	5
2	46	1	1	0
3	0	47	0	0
3	0	0	43	4
5	1	1	7	36

a	b	c	d	e
47	1	0	0	2
1	48	0	0	1
1	1	47	0	1
0	1	0	49	0
0	0	1	2	47

a	b	c	d	e
50	0	0	0	0
1	49	0	0	0
2	0	47	1	0
0	1	0	48	1
0	0	0	6	44

a	b	c	d	e
39	1	3	4	3
0	46	1	0	3
1	2	46	0	1
4	1	0	40	5
4	2	1	3	40

a	b	c	d	e
49	0	0	0	1
1	49	0	0	0
1	1	48	0	0
0	1	0	47	2
1	0	0	2	47

a	b	c	d	e
50	0	0	0	0
2	48	0	0	0
0	1	49	0	0
0	3	1	46	0
0	3	0	4	43

a	b	c	d	e
40	0	2	5	3
1	45	0	2	2
2	0	46	0	2
2	0	0	40	8
3	1	0	6	40

a	b	c	d	e
48	1	0	1	0
2	48	0	0	0
0	1	49	0	0
0	1	0	46	3
0	0	1	4	45

Figure 7: Examples of confusion matrices. Left: melodic; center: MFCC; right: fusion. Machine learning algorithms from top to bottom: SMO, J48, RandomForest, LogitBoost, BayesNet. Music genres indicated by the letters: a) flamenco; b) instrumental jazz; c) opera; d) pop; e) vocal jazz

It is also intriguing to notice that a simple algorithm such as a decision tree conveys valuable results from a small 14-dimension vector, in the case of melodic features alone. Figure 8 demonstrates the relevance of these features, from which a small tree that delivers remarkable results can be obtained. Figure 9 reveals that employing the correct machine learning algorithm (K-Nearest Neighbours *KStar*) it is possible to obtain relevant results (91%) with only two descriptors: *Mean Vibrato Rate Big Length* and *Mean Vibrato Coverage Big Length*.

```
MeanVibratoCoverageBigLength <= 0.29744
|   MeanVibratoRateBigLength <= 6.1147
|   |   MeanVibratoExtentBigLength <= 63.7504
|   |   |   SC4 <= 0.13636: instrjazz (50.0/1.0)
|   |   |   SC4 > 0.13636
|   |   |   |   SC1 <= 0.10714
|   |   |   |   |   SC1 <= 0.073171: instrjazz (2.0/1.0)
|   |   |   |   |   SC1 > 0.073171: vocaljazz (3.0)
|   |   |   |   SC1 > 0.10714: pop (6.0/1.0)
|   |   MeanVibratoExtentBigLength > 63.7504
|   |   |   StdDevVibratoCoverage <= 0.29924: flamenco (50.0)
|   |   |   StdDevVibratoCoverage > 0.29924: opera (2.0)
|   MeanVibratoRateBigLength > 6.1147
|   |   MeanVibratoRateBigLength <= 6.4224: vocaljazz (43.0)
|   |   MeanVibratoRateBigLength > 6.4224
|   |   |   MeanPitchRange <= 464.2857: pop (44.0/1.0)
|   |   |   MeanPitchRange > 464.2857: vocaljazz (2.0)
MeanVibratoCoverageBigLength > 0.29744: opera (48.0/1.0)
```

Figure 8: Example of a J48 decision tree that delivers 92.4% accuracy

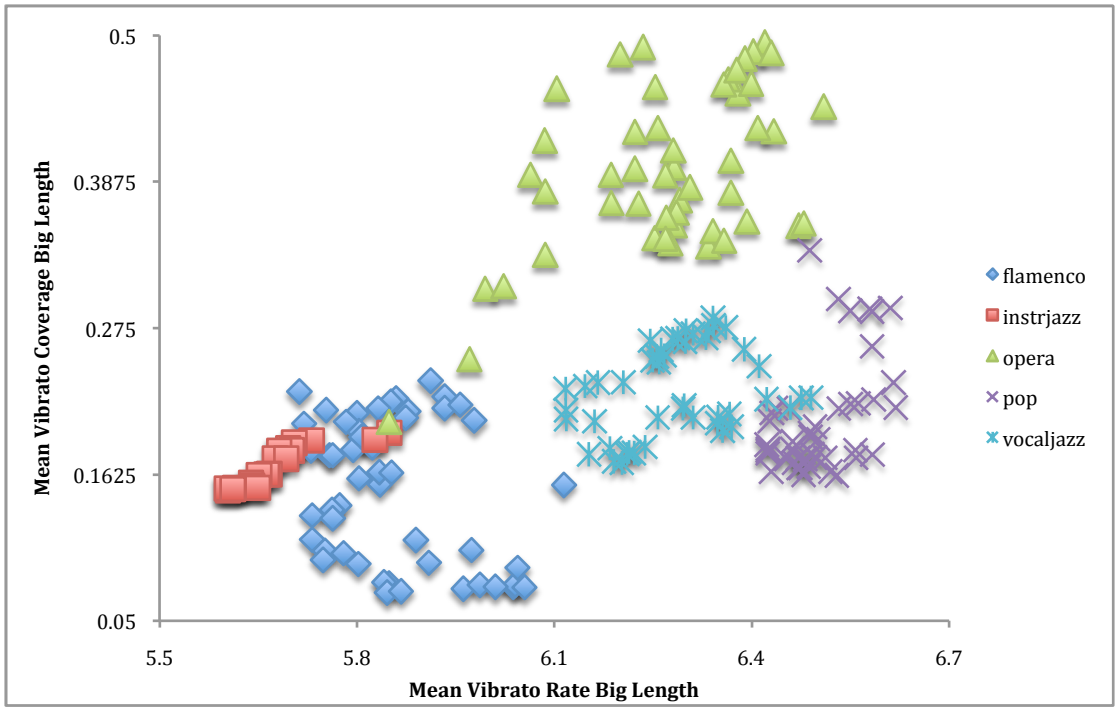

Figure 9: Mean Vibrato Rate Big Length vs Mean Vibrato Coverage Big Length

The first expected conclusion we can take is that vibrato coverage helps us to classify opera. In almost all of the opera excerpts, the biggest exctracted melodic contours (which should correspond to the bigger phrases) show a mean vibrato coverage above 30%. We can also observe that the mean vibrato rate in vocal jazz and pop is usually greater than in instrumental jazz and flamenco, making it a useful feature to isolate these two styles. On the other hand, to distinguish between them, one useful feature is the mean pitch range, meaning that the distance between the lowest and highest pitch

in vocal jazz melodies is generally larger than in pop melodies. Shape class descriptors prove to be useful in this tree, as instrumental jazz can be discriminated using the fourth of these descriptors (for more information, see section 3.3.3).

4.1.2 Attribute Selection: *SVMAttributeEval + Ranker*

The results exhibited in Chart 2 were obtained using the same approaches and classifiers as the ones explained before for Chart 1. The only difference relies on the attribute selection, which in this case uses *Ranker* as the search method and *SVMAttributeEval* as the evaluator. *Ranker* allows the user to define a maximum number of attributes to be kept. We chose a set of ten descriptors to have significantly less features than instances, avoiding overfitting. It also makes it possible to compare if the results vary significantly when we use different attribute selection methods and less descriptors. The list of features is shown in Table 4.

Rank	Melodic	MFCC	Fusion
1	Skewness Vibrato Rate Big Length	Mean MFCC5	Skewness Vibrato Rate Big Length
2	Mean Pitch Range	Variance MFCC5	Mean Pitch Range
3	Mean Vibrato Coverage Big Length	Mean MFCC1	Mean Vibrato Coverage Big Length
4	Mean Lowest Big Length	Mean MFCC3	Mean MFCC1
5	Kurtosis Vibrato Coverage	Mean MFCC15	Mean Lowest Big Length
6	Mean Pitch Std Deviation	Variance MFCC3	Kurtosis Vibrato Coverage
7	Mean Vibrato Extent	Mean MFCC10	Variance MFCC5
8	Std Dev Pitch Range	Mean MFCC4	Std Dev Pitch Range
9	Std Dev Vibrato Coverage Big Length	Variance MFCC6	Std Dev Vibrato Coverage Big Length
10	Shape Class 3	Mean MFCC6	Variance MFCC20

Table 4: List of selected features by order of relevance

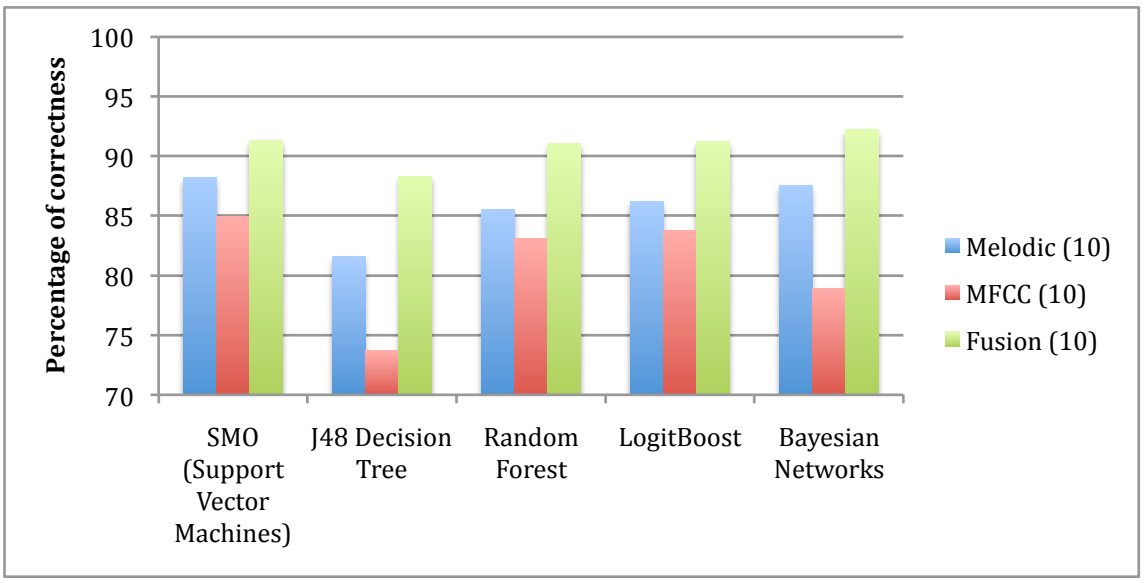

Chart 2: Results for the initial dataset using *SVMAttributeEval+Ranker* as the attribute selection method

We can detect immediately a significant decrease in the accuracy of our approach for all classifiers except SMO, which maintains the same level of accuracy. This may mean that this classifier is more resilient to changes in the number of descriptors or that the attribute selector favours Support Vector Machines classifiers. The accuracy also drops for the baseline approach and for the fusion of both approaches, although by a smaller margin, keeping the level above 90% for all classifiers except J48 tree.

4.2 Extended Dataset

4.2.1 Attribute Selection: *CfsSubsetEval + BestFirst*

The same approaches, classifiers and attribute selection method as the ones explained for Chart 1 were used for Chart 3. However, the dataset is an expanded version of the first, adding 250 snippets to achieve a total of 500 excerpts.

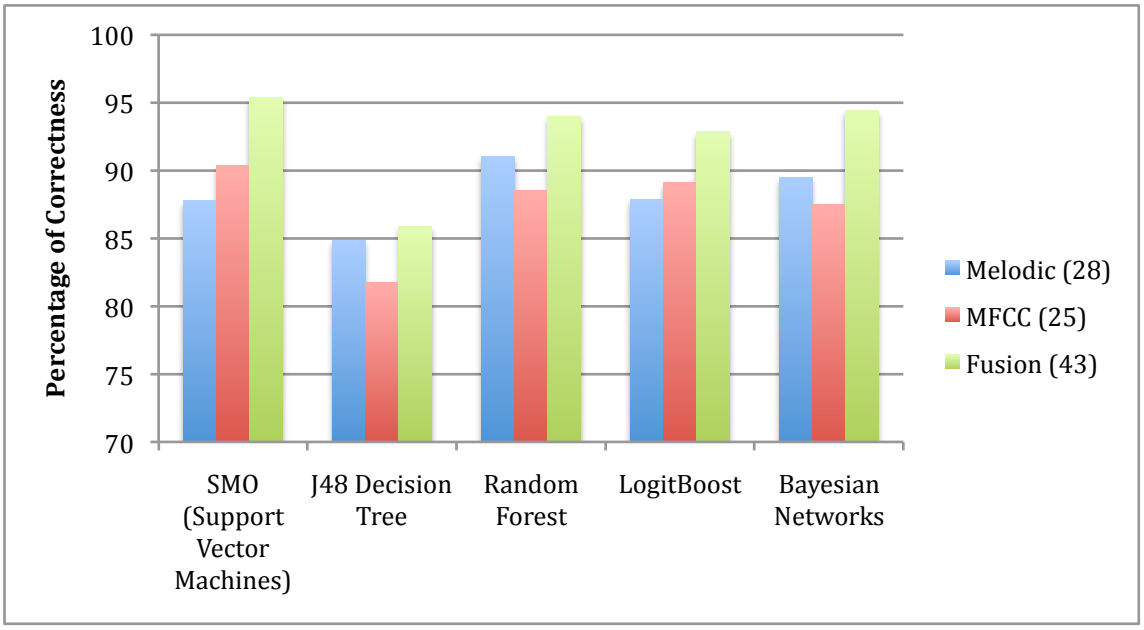

Chart 3: Results for the expanded dataset using *CfsSubsetEval+BestFirst* as the attribute selection method

Several observations can be made from these results:

1. For melodic features, accuracy decreases by an average of 4%, nevertheless staying close to the 90% mark;
2. For MFCC features, accuracy increases by an average of 2%, reaching the same level as the melodic features approach;
3. With the exception of J48, there is no significant decrease in accuracy when using a single vector containing both types of features.

The first statement was expected, as accuracy tends to decrease when we increase the dataset. The second evidence may be explained as the result of more training, which may allow for a proper stabilisation of the MFCC means and variances. Concerning the third consideration, maybe the accuracy is kept because the number of selected features is high.

4.2.2 Attribute Selection: *SVMAttributeEval + Ranker*

To avoid overfitting, once again we tried this attribute selection method that allows to sort the features by order of relevance and select a small portion of them, in this case ten (Table 5).

Rank	Melodic	MFCC	Fusion
1	Kurtosis Vibrato Rate Big Length	Mean MFCC5	Kurtosis Vibrato Rate Big Length
2	Mean Pitch Range	Variance MFCC5	Mean Pitch Range
3	Mean Vibrato Coverage Big Length	Mean MFCC1	Mean Vibrato Coverage Big Length
4	Skewness Vibrato Rate Big Length	Variance MFCC4	Mean MFCC1
5	Mean Lowest Big Length	Variance MFCC7	Mean Lowest Big Length
6	Std Dev Vibrato Rate Big Length	Mean MFCC7	Mean MFCC5
7	Mean Pitch Std Deviation	Mean MFCC10	Mean Pitch Std Deviation
8	Mean Vibrato Extent Big Length	Mean MFCC4	Mean MFCC10
9	Std Dev Pitch Range	Mean MFCC11	Mean MFCC2
10	Kurtosis Vibrato Coverage	Variance MFCC20	Variance MFCC5

Table 5: List of selected features by order of relevance

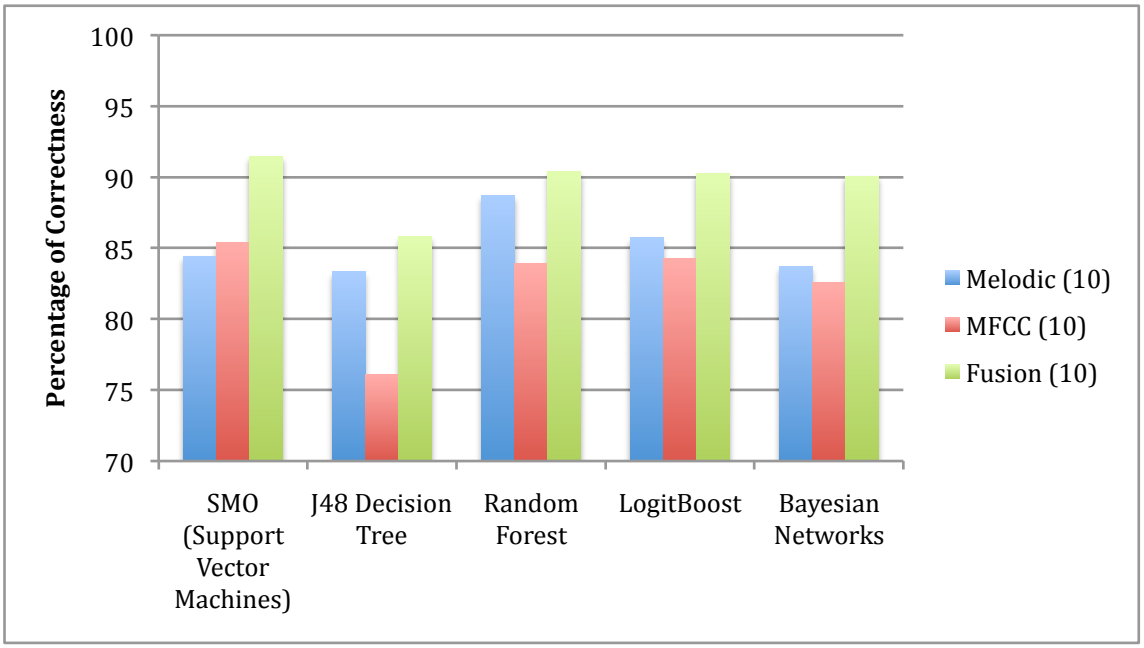

Chart 4: Results for the expanded dataset using *SVMAttributeEval+Ranker* as the attribute selection method

Comparing the results shown in Chart 4 with the ones exhibited in Chart 3, we can see that, especially for the fusion vector, using less features to perform the classification does not lead to a substantial decline in the accuracy of the system.
Comparing to Chart 2, we can take the same conclusions as we took in the previous section (4.2.1).

4.3 GTZAN Genre Collection

As this is not a prepared dataset, we were expecting the system's accuracy to drop considerably. In this collection, a great part of the samples have low quality and no clear melody, which leads to a poor melody extraction, hence weak classification. Chart 5 displays the results for the three approaches and five classifiers that were adopted before, using *CfsSubsetEval + BestFirst* as the attribute selection method.

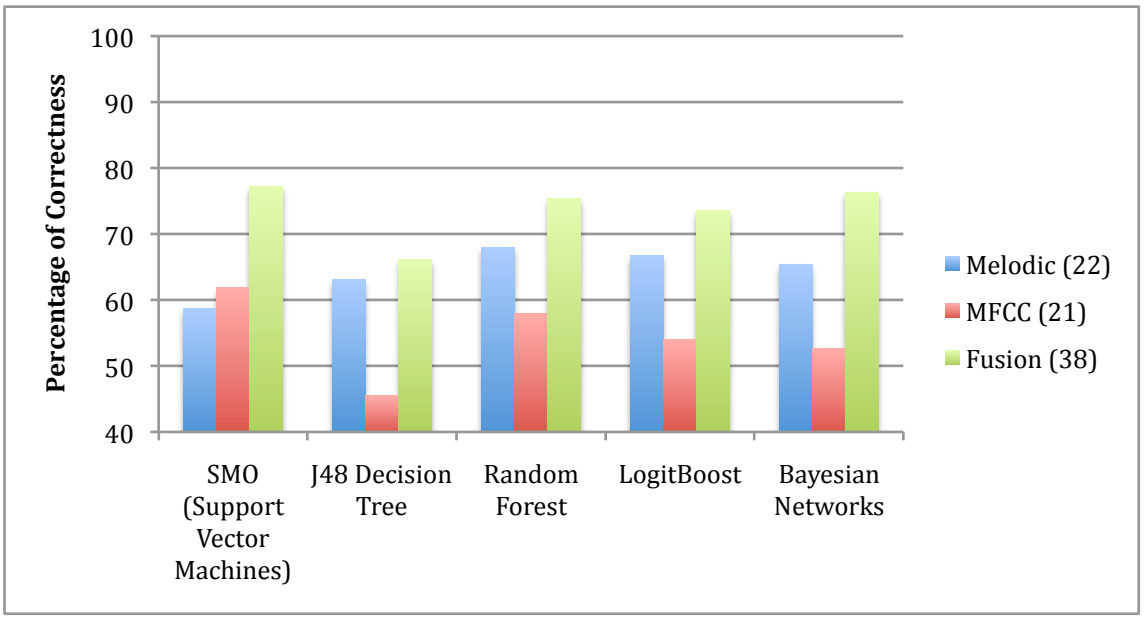

Chart 5: Results for the GTZAN dataset using *CfsSubsetEval + BestFirst* as the attribute selection method

As expected, accuracy is lower for all classifiers with both our and the baseline approaches. However, SMO, Random Forest and Bayesian Networks attain more than 75% precision when fusing both kinds of features, lower than state-of-the-art performance of 92% achieved by Panagakis, Kotropoulos, and Arce (2009). Nevertheless, it is interesting to note that by fusion we can significantly improve the results. This suggests that complementing low level features with high level melody features leads to promising results.

5. Conclusions

A final overview of the work carried out is provided in the last chapter of this thesis. First we present the goals achieved and contributions made and then make suggestions for future work.

5.1 Contributions

Looking at the goals we established in the introduction (chapter 1.2), we note that all of them have been met:

- A state-of-the-art review in the most relevant fields for this thesis was provided;
- A genre classifier using melody features was built;
- A set of reliable melody descriptors was achieved;
- A ground truth dataset for system evaluation was generated;
- Our method was evaluated employing different datasets and it was compared to other approaches;
- Low level and mid/high level descriptors were successfully fused, improving indeed the accuracy of the system;
- The evaluation results were presented and discussed.

Concentrating on the evaluation results, we can draw some final conclusions. The set of melody descriptors has proven to be robust, as the accuracy of the system did not fall substantially when doubling the dataset. We can also state that complementing low level timbre features with high level melody features is a promising direction for genre classification. Another conclusion we can take is that it is possible to achieve about 90% precision in genre classification with a melody description system that reaches about 70% accuracy, which is state-of-the-art level.

5.2 Future Work

The work developed throughout this thesis has given several interesting and promising results. Many of them can be extended and improved in several ways. We propose here some suggestions:

- The dataset should be expanded, through the addition of more excerpts and the introduction of other genres;
- Other low level features should be tested, in order to achieve a stronger set of descriptors;
- Different datasets should be tried, preferably ones which include melody annotation;
- Genre classification performance should be compared to the melody extraction accuracy, from which we could draw some interesting conclusions.

5.3 Final Words

As a personal conclusion, the main motivation for this project was fulfilled, as I had the possibility of combining my musical background with the technologies developed in the MTG. Throughout this work I have had the opportunity to learn from many people and would like to thank them all.

Bruno Rocha

References

C. Adams. Melodic contour typology. Ethnomusicology, 20: 179-215, 1976.

D. Arnold, N. Temperley, G. Norris, and P. Griffiths. Opera. In A. Latham (Ed.), *The Oxford Companion to Music, Oxford Music Online*, http://www.oxfordmusiconline.com/subscriber/article/opr/t114/e4847 (accessed September 4, 2011).

J. Bergstra, M. Mandel, and D. Eck. Scalable genre and tag prediction with spectral covariance. In *Procedings of the 11th International Symposium on Music Information Retrieval*, pages 507-512, Utrecht, The Netherlands, 2010.

J. Bonada and X. Serra. Synthesis of the singing voice by performance sampling and spectral models. In *IEEE Spectral Processing Magazine*, 24(2): 67-79, 2007.

M. Bunch. Dynamics of the singing voice (4th ed.). New York, Springer, 1997.

J. Durrieu, G. Richard, and B. David. Singer melody extraction in polyphonic signals using source separation methods. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 169–172, Las Vegas, USA, 2008.

D. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab. 2005. http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

D. Ellis. Classifying music audio with timbral and chroma features. In *Procedings of the 8th International Symposium on Music Information Retrieval*, pages 339–340, Vienna, Austria, 2007.

L. Fernandez. Flamenco music theory. Acordes Concert, 2004.

M. Genussov and I. Cohen. Musical genre classification of audio signals using geometric methods. In *Proceedings of the European Signal Processing Conference (EUSIPCO), pages 497-501, 2010*

E. Gómez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. Journal of New Music Research, 32: 23-40, 2003.

F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pages 196–204, London, UK, 2004.

E. Guaus. Audio content processing for automatic music genre classification: descriptors, databases, and classifiers. PhD Thesis. Barcelona, Universitat Pompeu Fabra, 2009.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning. 46: 389-422, 2002.

M. Hall. Correlation-based feature selection for machine learning. PhD Thesis. Hamilton, New Zealand, University of Waikato, 1999.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1, 2009.

C. Hsu, and J. Jang. Singing pitch extraction at MIREX 2010. In *6th Music Information Retrieval Evaluation eXchange (MIREX)*, extended abstract. Utrecht, The Netherlands, 2010.

T. Kako, Y. Ohishi, H. Kameoka, K. Kashino, K. Takeda. Automatic identification for singing styles based on sung melodic contour characterized in phase plane. In *Procedings of the 10th International Symposium on Music Information Retrieval*. Kobe, Japan, 2009.

I. Katz. Flamenco. In *Grove Music Online*. *Oxford Music Online*. http://www.oxfordmusiconline.com/subscriber/article/grove/music/09780 (accessed September 4, 2011).

A. Klapuri and M. Davy. Signal processing methods for music transcription. New York, Springer, 2006.

T. Langlois and G. Marques. Music classification method based on timbral features. In *Procedings of the 10th International Symposium on Music Information Retrieval*, Kobe, Japan, 2009.

C. McKay. Automatic genre classification of MIDI recordings. Master Thesis. Montreal, McGill University, Canada.

F. Merchán. Expressive characterization of flamenco singing. Master Thesis. Barcelona, Universitat Pompeu Fabra, 2008.

R. Middleton. Rock Singing. In J. Potter (Ed.), The Cambridge Companion to Singing. Cambridge, Cambridge University Press, 2000.

R. Middleton, D. Buckley, R. Walser, D. Laing, and P. Manuel. Pop. In *Grove Music Online*. *Oxford Music Online*. http://www.oxfordmusiconline.com/subscriber/article/grove/music/46845 (accessed September 4, 2011).

E. Pampalk, A. Flexer, and G.Widmer. Improvements of audio-based music similarity and genre classification. In *Procedings of the 6th International Symposium on Music Information Retrieval*, pages 628–633, London, UK, 2005.

Y. Panagakis, C. Kotropoulos, and G. Arce. Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In *Procedings of the 10th International Symposium on Music Information Retrieval*, pages 249-254, Kobe, Japan, 2009.

G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. In *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1247–1256, 2007.

J. Potter. Jazz singing. In J. Potter (Ed.), The Cambridge Companion to Singing. Cambridge, Cambridge University Press, 2000.

H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagama. Autoregressive MFCC Models for Genre Classification Improved by Harmonic-percussion Separation. In *Procedings of the 11th International Symposium on Music Information Retrieval*, pages 87-92, Utrecht, The Netherlands, 2010.

J. Salamon. Chroma-based predominant melody and bass line extraction from music audio signals. Master Thesis. Barcelona, Universitat Pompeu Fabra, 2008.

J. Salamon and E. Gómez. Melody extraction from polyphonic music audio. In *6th Music Information Retrieval Evaluation eXchange (MIREX)*, extended abstract. Utrecht, The Netherlands, 2010.

J. Salamon and E. Gómez. Melody extraction from polyphonic music: MIREX 2011. In *7th Music Information Retrieval Evaluation exchange (MIREX)*, extended abstract. Miami, USA, 2011.

N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2): 133-141, 2006.

C. Seashore. Psychology of music. New York, Dover, 1967.

G. Stefani. Melody: a popular perspective. Popular Music, 6(1): 21-35, 1987.

J. Stark. Bel canto: a history of vocal pedagogy. Toronto, University of Toronto Press, 1999.

J. Sundberg. Where does the sound come from? In J. Potter (Ed.), The Cambridge Companion to Singing. Cambridge, Cambridge University Press, 2000.

J. Sundberg. The science of the singing voice. Dekalb, Northern Illinois University Press, 1987.

J. Sundberg, and M. Thalén. Describing different styles of singing: a comparison of a female singer's voice source in "Classical", "Pop", "Jazz" and "Blues". Log Phon Vocol, 26: 82-93, 2001.

H. Tachibana, T. Ono, N. Ono, and S. Sagayama. In *6th Music Information Retrieval Evaluation eXchange (MIREX)*, extended abstract. Utrecht, The Netherlands, 2010.

M. Tucker and T. Jackson. Jazz. In *Grove Music Online*. *Oxford Music Online*. http://www.oxfordmusiconline.com/subscriber/article/grove/music/45011 (accessed September 4, 2011).

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, 10(5): 293-302, 2002.