The Flamenco *Cante*:

Automatic Characterization of Flamenco Singing by Analyzing Audio Recordings

Nadine Kroher

MASTER THESIS UPF / 2013

Master in Sound and Music Computing

Master thesis supervisors:

Dr. Emilia Gómez and Dr. Rafael Ramírez

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

UNIVERSITAT POMPEU FABRA

The Flamenco *Cante*: Automatic Characterization of Flamenco Singing by Analyzing Audio Recordings

Nadine Kroher

Music Technology Group
Universitat Pompeu Fabra
Tanger, 122-140, 3rd Floor
08018 Barcelona, SPAIN.

Master's thesis

Abstract. Flamenco singing is a highly expressive improvisational artform characterized by its deviation from the Western tonal system, freedom in rhythmic interpretation and a high amount of melodic ornamentation. Consequently, a singing performance represents a fusion of style-related constraints and the individual spontaneous interpretation. This study focuses on the description of the characteristics of a particular singer. In order to find suitable feature sets, a genre-specific automatic singer identification is implemented. For Western classical and popular music, related approaches have mainly relied on the extraction of timbre-based features to automatically recognize a singer by analyzing audio recordings. However, a classification solely based on spectral descriptors is prone to errors for low quality audio recordings. In order to obtain a more robust approach, low-level timbre features are combined with vibrato- and performance-related descriptors. Furthermore, differences among interpretations within a style are analyzed: Versions of the same a cappella *cante* have a common melodic skeleton which is subject to strong, individually determined melodic and rhythmic modifications. Similarity among such performances is modeled by applying dynamic time-warping to align automatic transcriptions and extracting performance-related descriptors. Resulting distances are evaluated by analyzing their correlation to human ratings.

Computing Reviews (1998) Categories and Subject Descriptors:

H Information Systems
H.5 Information Interfaces and Presentation
H.5.5 Sound and Music Computing

Acknowledgments

I would like to thank my supervisors Emilia Gómez and Rafael Ramírez for their guidance and support throughout this year and Xavier Serra for giving me the opportunity of being part in the SMC Master. I would also like to thank the members of the MTG and the Flamenco expert Jose Miguel Diáz-Bañez for being open to my questions and taking the time to give valuable feedback and support. I thank my fellow students for the great time we had during this year and for sharing jelly shots, Irish stew, Chinese wine, Risotto and much more. Of course I thank Cyril for his patience, understanding and graphic design support. Last but not least, I would like to thank Jose Manuel Diáz for sharing his passion for the fascinating world of Flamenco.

Content

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Research in music information retrieval is of great importance for managing large music databases and providing users with recommendations and automatically generated meta-data. Despite the great variety of music traditions and corresponding communities of enthusiasts, most algorithms are designed and tested on databases containing mainly Western commercial music. Studies have shown that applying state of the art content description technologies, such as automatic singer identification [Rao et al., 2011], to non-Western music databases leads to comparatively inferior results. The musical content of a piece of audio is described by combining low-level features directly extracted from the signal with generalizations based on musical knowledge. Consequently, algorithm performance can increase when genre-specific characteristics, such as the underlying tuning system, commonly used rhythmic patterns or typical instrumentation, are considered during the design stage.

Flamenco singing as an art form is of special interest for this type of studies due to its expressive improvisational character and deviations from the Western tonal system. Furthermore, traditional singing styles are performed in free rhythm and without or with very sparse musical accompaniment. These genre-specific properties together with the limited availability of recordings and the often suboptimal audio quality represent the main challenges of computationally description of this style.

Automatic singer identification is a suitable task to study the description of individual properties of a singer. Related approaches identify an unknown singer mainly by analyzing voice timbre descriptors. Consequently, spectral distortion in old or low quality recordings can cause classification errors. Also, the limited range of instrumentation, the acoustical properties of the Flamenco singing voice, such as the extensive use of vocal vibrato and the strong expressive elements as well as the spontaneous melodic interpretations provide useful generalizations for developing a more robust singer identification

algorithm for this genre. Songs of the same monophonic singing style, as for example *Martinetes*, have a common melodic skeleton which is strongly modified and ornamented during performance. A way of computationally understanding and measuring perceived similarity between different performances provides a useful tool for musicological studies, performance modeling, singing assessment and education.

1.2 Goals

This study aims to improve the accuracy of automatic singer identification for monophonic and polyphonic Flamenco recordings and to find a suitable model to characterize melodic similarity among a set of performances of the same *cante*.

The methodology to this approach comprises the following steps:

1. A state of the art review of automatic singer identification algorithms, performance similarity analysis and recent computational approaches to Flamenco description.

2. Collection of suitable datasets of monophonic and polyphonic Flamenco singing.

3. The implementation of a timbre-based singer identification algorithm in order to evaluate its performance for monophonic Flamenco singing.

4. Extraction of vibrato and performance-related features and their application in automatic singer identification for monophonic and polyphonic Flamenco music.

5. Implementation of several approaches to computational performance similarity analysis and comparison to human ratings for monophonic Flamenco *cantes*.

1.3 Structure of the thesis

In chapter two, the acoustic properties of the singing voice and underlying concepts of automatic classification are reviewed. Furthermore, related state of the art approaches to automatic singer identification and melodic similarity modeling are discussed. Chapter three introduces the methodology applied in this research, including the collection of databases and the approaches for feature extraction, classification and similarity analysis as well as evaluation

methods. In chapter four, results are presented and discussed. Chapter five provides a summary of the contributions made in the scope of this research and gives suggestions for future work.

Chapter 2

State of the art

This section reviews relevant concepts and related research work in computational analysis of Flamenco music, automatic singer identification and melodic similarity analysis. The first subsection briefly describes the Flamenco music tradition, summarizes the properties of Flamenco singing and gives an overview of computational approaches to Flamenco analysis. In the second part, related work in the area of singer identification is reviewed and categorized. The last sections provides an overview of related research dealing with melodic and timbre similarity.

2.1 Flamenco

Flamenco is a rich improvisational music tradition and its influences are as diverse as the roots of the population of its geographic origin, Andalusia. Even though the ancient history of the genre is not clearly determined, it is speculated to have Arabic, Jewish, Northern African and Hispanic influences ([Pohren, 2005]). Including various sub-styles and genres *(palos)*, Flamenco music is characterized by the singing voice *(cante)* which carries the main melody and the dance *(baile)* as its central elements, accompanied by guitar playing *(toque)* and accentuated hand clapping *(palmas)*. As an oral tradition with strong improvisational aspects, skills and compositions have been passed from generation to generation and only very rare manual transcriptions exist.

2.1.1 Flamenco singing

In Flamenco music the singing voice does not only carry the main melody, it is also thought to be the most expressive element of the genre. The *toná*, one of the oldest and purest styles, is mainly sung a cappella and only sometimes, depending on the sub-style, accompanied by rather sparse percussive elements. As described in [Pohren, 2005], Flamenco singing can be classified by its style and complexity into three categories: *Cante jondo*, *intermedio* and *chico*. The oldest and most expressive style, *jondo*, has its origin in religious songs and is thought the be the most complex form interpretation. It is usually sung in a *voz*

afillá, an extremely rough and airy shaping of the voice. The *cante intermedio* is less expressive and more ornamental and mainly sung by clearer, high-pitched voices, while the *cante chico* is associated with a brighter mood and a strong alignment to an underlying rhythm. Furthermore, traditional Flamenco voices are categorized into *gitano*, a very hoarse and noisy voice, and *bien,* a clearer and smoother voice. [Merchán Higuera, 2008] carried out interviews with three Flamenco experts, who characterized Flamenco singing in general as unstable in pitch, timbre and dynamics. Furthermore the typical voice is described as matt, containing few high frequency harmonic components and usually lacking the singer's formant.

2.1.2 Computational description and analysis

While a large amount of research in Flamenco deals with ethnomusicological and anthropological aspects, computational description and analysis is a comparatively young field. In a first study to assess acoustic properties of the Flamenco singing voice, [Merchán Higuera, 2008] extracted low-level timbre descriptors and melodic contours of popular songs to compare to alternative Flamenco versions recorded by the same singer under similar conditions. Results indicate that the Flamenco voice tends to be noisier and formants are shifted to higher frequency regions. Furthermore, timbre and pitch are less stable over time and vibrato tends to be irregular. These observations largely correspond to a qualitative assessment through interviews with experts carried out in the same study. However, the broad spectrum of voice qualities and singing styles impedes a strict differentiation and a clear characterization of the typical Flamenco voice.

In order to facilitate an analysis of mid- and high-level musical features without the presence of a score, a main focus of recent research in the field has been the development of an automatic transcription tool. In a first approach ([Gómez and Bonada, 2008]), a melody extraction algorithm based on fundamental frequency estimation and iterative note consolidation and tuning refinement has been implemented. Even though preliminary results for melodic similarity tasks are promising, limitations of the tool become obvious when comparing the unprocessed output to manual transcriptions. Errors mainly occur as incorrect note segmentation due to smooth note attacks and pitch estimation errors originated from unstable tuning. In a later approach ([Gómez and Bonada, 2013]), the incorporation of a state of the art predominant frequency algorithm allows the extension of automatic transcription to polyphonic styles. Accuracies for both, monophonic and polyphonic material range around 80%. It should be mentioned that manual annotations, which are used as a ground truth, tend to be prone to inter-subjective disagreements ([Gómez and Bonada, 2013]). Furthermore, due to the lack of scores in the Flamenco tradition and the presence of micro-tonal material, a consistent

notation format has not been established so far. The described system therefore outputs a detailed representation displaying fundamental frequency contour as well as a more abstract and score-oriented layer showing quantized pitch and energy envelopes.

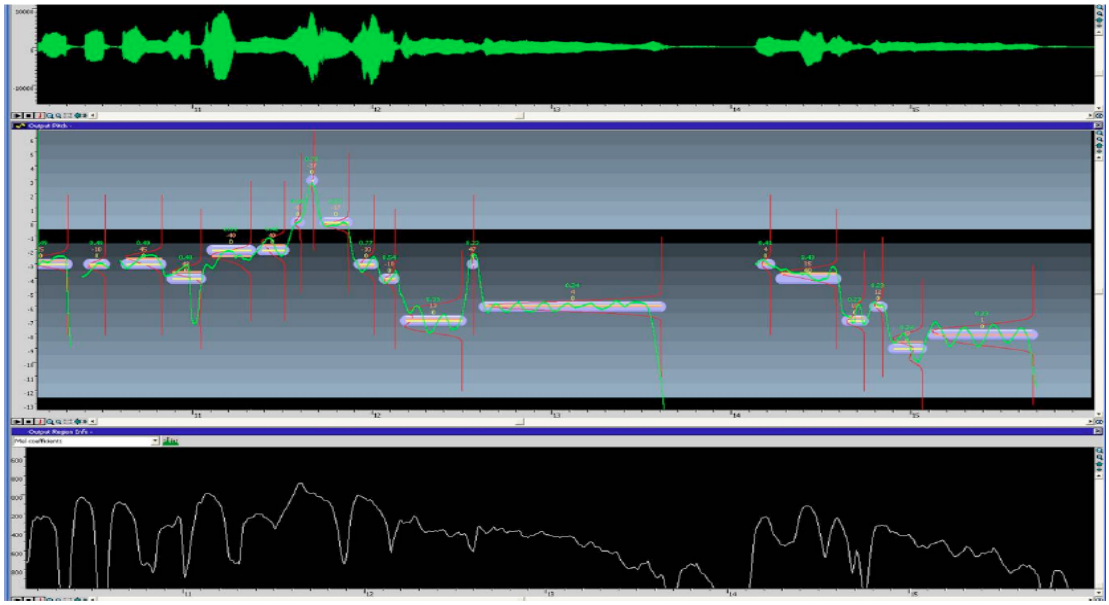

***Figure 2.1**. Screenshot of the existing transcription tool ([Gómez and Bonada, 2013], with permission of the authors) displaying the waveform, symbolic description and temporal energy evolution.*

A further area of interest is the inter- and intra-style analysis of sub-genres. A cappella singing styles are characterized by a common melodic skeleton which is subject to strong improvisational ornamentations, performed in free rhythm. However, manually recognizing these underlying melodic figures and determining the corresponding style is not a trivial task for untrained listeners but more obvious to Flamenco enthusiasts: A recent study on perceived melodic similarity ([Gómez et al., 2012c]) shows that Flamenco experts tend to group melodies by their underlying figure while novice listener rely their judgement on other features, such as the pitch range or the actual contour. Furthermore, experiments with synthesized melodies indicate that perceived similarity is influenced by the voice timbre. In a computational approach to characterize and study melodic similarity among various styles ([Cabrera et al. 2008]), the aforementioned computer-assisted transcription tool is used to extract the melodic contour. This way, after removing ornamentations, the melodic skeleton can be estimated. Calculating distance measures among songs from styles based on the underlying melodic figures, a phylogenetic graph is created and analyzed regarding style organization and historic evolution. A more detailed description is obtained in [Mora et al., 2010] based on manually extracted style-specific features defined by experts in the field. Clustering based

on these features showed a stronger grouping of different styles, but intra-style differences in the extracted features are vanishingly low. In a different approach ([Gómez at al., 2011]), an algorithm automatically detects previously defined ornamentations. Analyzing the results among styles reveals a tendency of certain ornamentation techniques to occur in some styles more than in others, which makes them a suitable feature to incorporate into style classification.

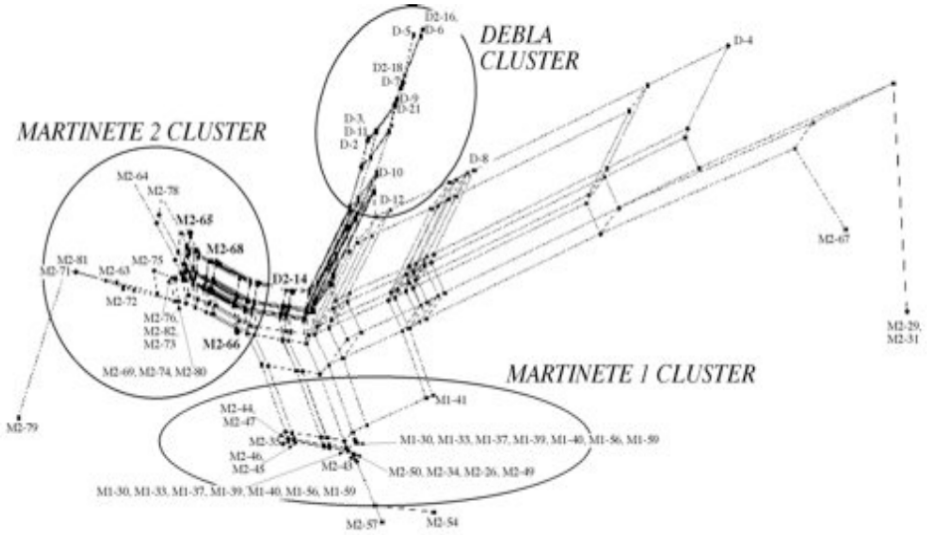

***Figure 2.2**. Graphic representation of melodic style similarity. Taken from [Mora et al., 2010] with permission of the authors.*

Detecting Flamenco as a genre among other music styles and traditions is a comparatively simple task: Satisfying results are obtained when combining global melodic features with instantaneous spectral and vibrato descriptors in [Salamon et al., 2012].

2.2 Singer identification

The task of identifying a particular singer solely based on features extracted from the raw audio signal is a challenging task for monophonic recordings and gains in complexity for polyphonic signals where additional pre-processing steps are required to reduce the influence of the accompaniment sounds. In the training phase, a suitable set of audio features, usually low-level timbre descriptors, is extracted from a database containing labeled audio recordings. Using the extracted information, a machine learning model can be estimated for each singer. The learning task is defined as finding a classifier F of the form:

(2.2.1.) $$F(MusicFragment) \rightarrow Performers$$

where *MusicFragment* is the set of music fragments and *Performers* is the set of possible singers to be identified. Each music fragment is characterized by different subsets of the extracted descriptors. To classify an unknown audio recording among the set of candidates in the training database, the same features are extracted. The classifier then assigns a class, in this case a singer, to the unlabeled audio file.

This section presents several approaches and compares their feature selection and classification algorithms. Pre-processing strategies for singer identification in polyphonic music are summarized and evaluation strategies and results of related work is presented.

2.2.1 Feature extraction

From an acoustic standpoint of view, the singing voice can be described as a transformation of aerodynamic energy from the breathing apparatus into sound waves ([Kob et al., 2011]). The fundamental frequency is determined by the vibrato frequency of the vocal folds. Vocal tract resonances shape the spectrum and contribute to a large extend to our timbre perception.

Most timbre features used in singer identification are based on the concept of separating the voice signal into a source signal and a filter transfer function. This concept is often referred to as source-filter modeling: The source corresponds to the excitation signal from the glottis and the transfer function describes the shaping of the excitation spectrum in the vocal tract. The source spectrum contains equally spaced frequency components with a decreasing amplitude of 12dB per octave. The shape of the filter function shows characteristic peaks at the resonant frequencies of the vocal tract.

The *Mel-frequency cepstral coefficients* (MFCCs) represent a commonly used feature set based on the source-filter model ([Tsai and Lin, 2012], [Nwe and Li, 2008], [Tsai and Lin, 2008] and [Lagrange et al., 2012]). The computation of the MFCCs can be summarized as follows:

1. Frame-wise computation of the short term Fourier transform
2. Logarithm of the magnitude spectrum
3. Bin reduction using overlapping triangular Mel-spaced filters
4. Cepstrum or discrete cosine transform (DCT) of the spectrum

Taking the logarithm of the magnitude spectrum adjusts the energy dynamics to the perception of loudness. To obtain a compact spectrum description, several bins need to be summarized. The bandwidth of Mel-filters increases with the center frequency and therefore the result has a higher resolution for low frequency components. Hereby, the properties of human frequency perception

are represented. The discrete cosine transform can be seen as the "spectrum of the spectrum": Slow magnitude changes with increasing frequencies will be represented by low components in the cepstrum. In a similar manner, low-pass filtering of time signals is applied to obtain an amplitude envelope. Higher cepstrum components represent the fine structure of the spectrum. Consequently, MFCCs with a low index represent the vocal tract transfer function and higher coefficients are related to the glottal source spectrum. Usually the DCT is used instead of a Fourier transform, since it has similar properties when applied to the magnitude spectrum and can be computed in a more efficient way.

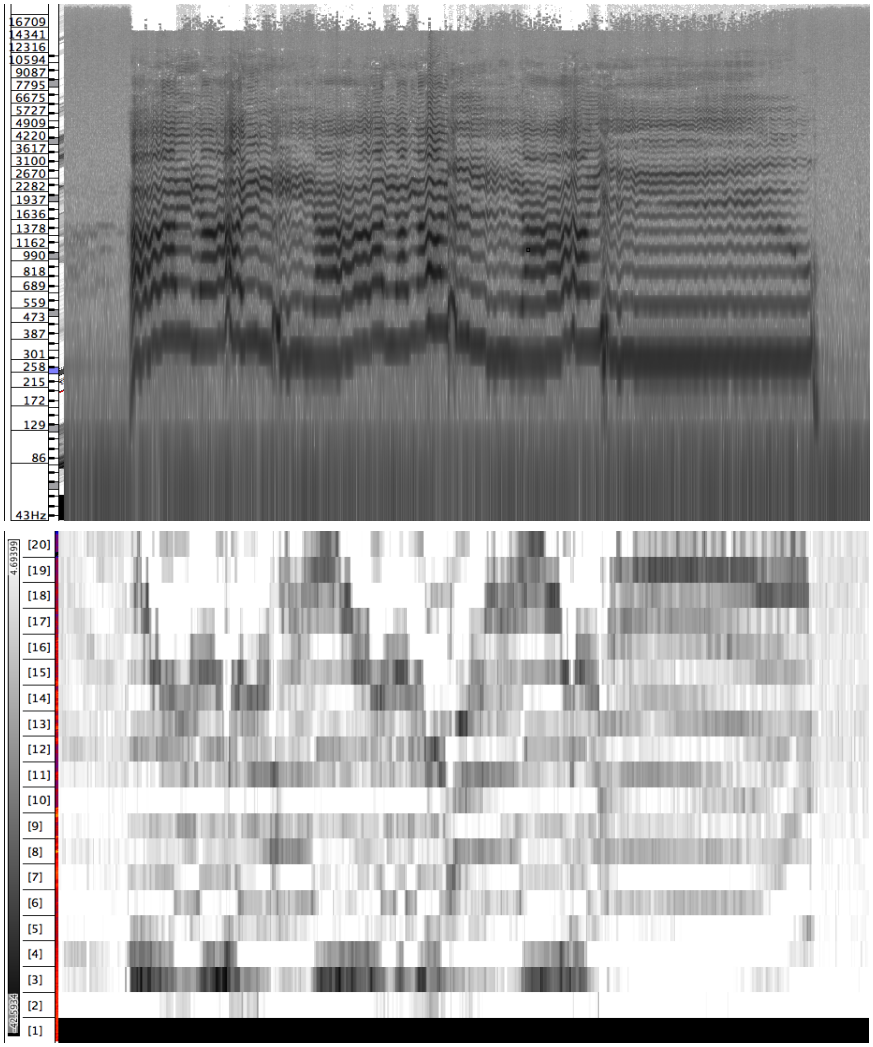

***Figure 2.3**. Spectrogram (top) and MFCCs (bottom) of a sung phrase.*

[Mesaros and Astola, 2005] compare low-order and high-order MFCCs and report that both equally contribute to singer identification tasks, while for speech recognition, using only low-order MFCCs to represent the spectral envelope, is usually sufficient. [Nwe and Li, 2006] compare MFCCs with a related feature set, the *linear frequency cepstral coefficients (LFCC)*: Following the same concept described above, equally spaced filters with a constant

bandwidth are used instead of the Mel-filterbank. The authors furthermore apply 96 triangular filters with center frequencies coinciding with musical notes of the Western tuning system to describe vibrato in the singing voice. In a cascaded filterbank containing 5 filters for each output of the previous stage, vibrato direction and depths can be estimated. Passing the output through nine octave-spaced filters and transforming to the cepstrum, they obtain the *vibrato octave frequency coefficients (OFCCs-VIB)*. Replacing the first two stages by a filterbank containing trapezoid filters centered around the fundamental frequencies of the musical tuning system with a bandwidth of +/- 1.5 semitones, harmonic properties of the signal are modeled. The obtained coefficients are referred to as *harmonic octave frequency coefficients (OFCCs-HARM)*. In [Nwe and Li, 2008] the fusion of these feature sets are studies. [Cai et al., 2011] extracted *Gamma-tone cepstral coefficients (GTCCs)* in addition to MFCCs and LPMCCs (described later) by applying a Gamma-tone filterbank, which approximates the filtering in the human cochlea, before passing to the cepstral domain. [Fujihara et al., 2010] use derivatives of the estimated fundamental frequencies as an extension of the MFCC model. A first genre-specific approach is carried out by [Sridhar and Geetha, 2008], using 22 filters based on the tuning system used in Carnatic music. They labeled the obtained coefficients as *Carnatic interval cepstral coefficients (CICCs)*.

A similar technique often used in speech processing is *linear predictive coding (LPC)*, where the vocal tract is modeled as a time-variant digital filter applied to the source signal, which is either white noise (for unvoiced speech) or an impulse train (voiced speech). To obtain the optimal filter coefficients (LPCs), an adaptive linear model compares the output predicted by the filter to the signal and minimizes the error by adjusting the coefficients. A detailed description of the method can be found in [Ó Cinnéide, A, 2008]. In contrast to the spectral based MFCC description, LPC models both the excitation signal and the filtering process in the time domain.

[Shen et al., 2009] combine LPCs with pitch histograms and genre estimation and describe instrumental sections using MFCCs. They report up to 10% improvement when additionally considering non-vocal segments. [Kim and Whitman, 2002] compute *warped LPCs* by using a Bark-scaled frequency axis. [Fujihara et al., 2010], [Zhang et al., 2003] and [Cai et al., 2011] extract LPMCCs, the Mel-frequency coefficients of the LPC spectrum. [Fujihara et al., 2010] report a significant improvement for various implementations of their algorithm when using LPMCCs instead of MFCCs. For their proposed system, the accuracy improves by 30%. [Cai et al., 2011] found that a combination of MFCCs, LPMCCs and GTCCs leads to an improvement in accuracy of 18% compared to an implementation in which only MFCCs are extracted.

Based on a modified source-filter model, [Bartsch and Wakefield, 2004] represent the signal by its *composite transfer function(CTF)*: The source signal

is modeled as sum of equally-space sinusoids with constant amplitude. The corresponding transfer function representing the vocal tract is estimated by minimizing a cost function. As a main drawback, the instantaneous fundamental frequency needs to be determined in order to model the source signal.

2.2.2 Classification

Once a set of suitable features has been found, a classifier can be built from a training database to identify the singer of a unknown audio recordings. Among the multitude of machine learning algorithms, *Gaussian Mixture Models (GMM)* have been frequently used in singer identification tasks ([Fujihara et al., 2010], [Shen et al., 2009], [Cai et al., 2011], [Sidhar and Geetha, 2008], [Tsai and Lin, 2010], [Lagrange et al., 2012] and [Zhang, 2003]). As a main advantage over many other algorithms, GMMs provide a probability for each class instead of a 1/0 classification and can therefore be used to estimate the reliability of the decision. Adding a new class at a later stage can simply be achieved by building an additional GMM without having to re-train the the entire data base. Furthermore, class-depended probabilities can be used to estimate perceived similarities to other singers.

The probability of occurrence of a single continuous variable can be modeled as Gaussian distribution $N(x \mid \mu, \sigma^2)$ with a mean value μ and a standard deviation σ^2 (Eq. 2.3.1).

(2.2.2)
$$N(x/\mu, \sigma^2) = \frac{1}{(2\pi\sigma)^2} * e^{\left(\frac{-x-\mu}{2\sigma^2}\right)}$$

In a toy example, x corresponds to the hight of a person, *N to* the number of people in a database with this hight and μ represents the average hight among the instances. The shape of the curve is determined by the parameter σ^2: The higher the value, the broader the peak. Given a distribution of one-dimensional data $\mathbf{x}=[x1,x2,..., x_N]$, the goal is to find the parameters μ and σ^2 which best approximate the data distribution as a Gaussian function. The probability of occurrence of data **x** is calculated as the product of probabilities of single data points:

(2.2.3)
$$P(\boldsymbol{x}/\mu, \sigma^2) = \prod_{i=1}^{N} (N(x_i), \mu, \sigma^2)$$

Since the data **x** is known and the distribution parameters are unknown, the probability distribution can be modeled as a likelihood *l* of the data **x** given the

distribution parameters μ and σ^2. The optimal parameters can be found by maximizing the likelihood, which corresponds to setting the partial derivatives zero and solving for the respective parameters.

(2.2.4) $$\underset{\mu,\sigma^2}{argmax}[l(\mu,\sigma^2/\boldsymbol{x})]$$

Most tasks require the integration of various feature variables. For a d-dimensional dataset, the Gaussian probability distribution is given by:

(2.2.5) $$N(\vec{x}/\vec{\mu},\Sigma^2)=\frac{1}{(2\pi\sigma)^{(d/2)}(|\Sigma|)^{(1/2)}}*e^{(-1/2(\vec{x}-\vec{\mu^2})^T\Sigma^{-1}(\vec{x}-\vec{\mu}))}$$

The corresponding mean is a d-dimensional vector and the covariance is represented by a d-dimensional matrix. Modeling a given dataset as a Gaussian distribution assumes that the present distribution can actually be approximated by Gaussian parameters. In many scenarios, this might not be the case: A distribution might have a more complex shape, including multiple maxima with varying amplitude. Such distributions can be approximated by a weighted sum of K Gaussian models. The resulting density function of a one-dimensional GMM combining K distributions is calculated as follows:

(2.2.6) $$P(x/\mu,\sigma^2)=\sum_{k=1}^{K}\pi_k N(x/\mu_k,\sigma_k^2)$$

where π_k are the weights for each distribution with

(2.2.7) $$\sum_{k=1}^{K}(\pi_k)=1 \quad for \quad 0<\pi_k<1 \quad .$$

Consequently, the coice of K and the resulting complexity are important classification parameter. While [Fujihara et al., 2010] used a combination of 64 Gaussian distributions, [Zhang, 2003] classified based on a combination of only 3 distributions and [Shen et al., 2009] determined in a parameter tuning experiment 2 to 7 Gaussian components as suitable. [Shen et al., 2009] built such a model for each of their feature groups: timbre, pitch and genre. They trained a combination function for a joined weighted classification using logistic regression.

Singer identification is a multi-class problem and consequently the training phase consists of building a GMM for each singer based on the extracted features. The *expectation maximization* (EM) algorithm is a commonly used implementation to estimate multivariate Gaussian parameters. To classify a new

test instance, the same features are extracted and the probabilities of the sample belonging to each class is estimated using the GMMs. In a *maximum likelihood classification*, the class with the highest probability is assigned to the new instance:

$$\lambda_{(\hat{S})} = \underset{s \in S}{argmax}\, P(x_{(unknown)} / \lambda_s) \tag{2.2.8}$$

[Zhang, 2003] furthermore only considered frames above a fixed likelihood threshold for the decision. As an alternative to the maximum likelihood classification, [Cai et al., 2011] assigns the class with the *maximum a posteriori estimation* (MAP):

$$\lambda_{(\hat{S})} = \underset{s \in S}{argmax}\, P(\lambda_s / x_{unknown}) = \underset{s \in S}{argmax}\, \frac{P(x_{unknown} / \lambda_s) * P(\lambda_s)}{\sum_{(s \in S)} P(x_{unknown} / \lambda_s) * P(\lambda_s)} \tag{2.2.9}$$

[Kim and Whitman, 2002] compare GMM and support vector machine (SVM) based classification for singer identification considering various scenarios and find that in most cases SVMs performed slightly better (39.6% vs. 32.1% correctly classified songs using LPCs extracted from polyphonic signals). In alternative approaches, [Nwe and Li, 2008] implement a maximum likelihood estimation based on Hidden Markov Models (HMMs), [Mesaros and Astola, 2005] use artificial neural networks (ANNs) and [Bartsch and Wakefield, 2004] apply standard quadratic classifiers.

2.2.3 Voice detection and accompaniment sound reduction

Since in most musical genres a cappella singing is the exception, various attempts have been made to adapt singer identification techniques to polyphonic recordings. [Tsai and Lee, 2012] train a model for each singer based on spoken sound samples but the rather poor results rule out this option. However, the authors show that for a very small amount of available training data results can be slightly improved when spoken samples are included in the singer model.

Several systems therefore try to detect the vocal segments of each song to at least avoid a classification based on features extracted from instrumental sections. This process is often referred to as *vocal identification* or *reliable frame selection*. Based on the assumption that the singing voice energy is mainly distributed between 200 and 2000Hz, [Kim and Whitman, 2002] filter this range and use a threshold in the harmonicity measure to distinguish between voice segments and drums. The obtained accuracy for voice detection is rather low with values around 55% depending on the threshold. [Zhang, 2003] analyses only the first part of each song and detects the start of the singing voice

using a predetermined threshold for low-level features such as the average zero-crossing rate and spectral flux. From the detected starting point onwards, only a fixed amount of frames are considered. Based on frame-wise extracted MFCCs, [Cai et al., 2008] apply a sparse representation based classification and subsequent low-pass filtering of the classifier output. In a similar approach, [Shen et al., 2009] train SVMs on MFCCs from vocal and non-vocal sections and obtained up to 92.4% accuracy. They observed that about 90% of all false positives contained unusually loud instrumental sections or strong noise. [Cai et al., 2011] also report loud background music as the main source of misclassification. Using manually annotated data, [Nwe and Li, 2011] categorize the training database by structural sections (intro, chorus, etc.), gender of the singer, tempo and loudness and construct a HMM for each possible combination. This multi-model approach leads to a slightly improved accuracy of 82.7% compared to 81.1% for a single HMM implementation. [Lagrange et al., 2012] adapt an approach from speech recognition in noisy environment: They built GMMs based on a source-filter model for the voice segments and non-negative matrix factorization for the accompaniment. The obtained likelihood for the analyzed segment containing singing voice is propagated to the singer identification step and integrated as an uncertainty measure to weight the segments' influence in the classification. Using machine learning to model vocal and non-vocal segments assumes that non-vocal segments have similar characteristics among different songs. This might not be the case in general, but within styles with a limited set of possible instrumentations, as it is the case for Flamenco music, this strategy might lead to better results.

A different strategy, which aims to isolate the singing voice before extracting acoustic features, is often referred to as *accompaniment sound reduction*: [Sridhar and Geetha, 2008] evaluate different source separation algorithms for Carnatic music and apply the best performing system in a pre-processing step.

Combining both approaches, [Fujihara et al., 2010] estimate the instantaneous fundamental frequency and extracted the harmonic structure of the predominant melody to re-synthesize the voice. The drawback of this process is the fact that harmonics of voice and accompaniment might coincide and furthermore non-harmonic components of the voice are neglected. To exclude non-vocal sections, GMMs are built from extracted LPMCCs for both vocal and non-vocal training samples. The authors report an improvement in accuracy of 45% compared to the case when no pre-processing is applied. They even obtain better results when the automatic reliable frame selection is used compared to manual labeling, since the algorithm also rejects unreliable vocal frames. [Rao et al., 2011] extend this algorithm performing the vocal detection based on static and dynamic timbre low-level features as well estimated instabilities in the fundamental frequency typical for the singing voice. They obtain an overall accuracy of 85.9% for a database containing 45 songs from western and non-western genres. Furthermore, the achieved accuracy is strongly genre

dependent and for some styles reduced feature sets performed better than the proposed approach. Analyzing the performance of single features shows that for example in Bollywood music, where the vocal timbre varies strongly compared to the rather constant instrumental timbre, dynamic timbre features perform best. [Tsai and Lin, 2010] apply a background music removal based on estimating the cepstral transfer function between solo voice and the mixed signal. They combine this technique with singing voice detection in a pre-processing step.

2.2.4 Results

Testing the proposed monophonic singer identification algorithm based on MFCC extraction and GMM classification, [Tsai and Lee, 2012] obtain up to 95% accuracy for a database containing 30 a cappella passages of Mandarin pop music divided into 15 training and 15 test samples. They show that the resulting accuracy strongly increases with the number of training samples (65% for 5 training passages vs. 95% for 15 training passages). In an earlier approach based on spectral envelope estimation using CTF, [Bartsch and Wakefield, 2004] obtain 99.5% accuracy. It has though to be mentioned that the test set contains single notes sung by classically trained singers and is therefore not representative for commonly available music databases.

In a first approach towards polyphonic singer identification based on linear and warped LPCs, [Kim and Whitman, 2002] detect only 45.3% of a database containing 250 songs by 20 different artists correctly. A crucial factor might be the rather poor performance of the voice detection. Detecting the starting point of the voice and extracting LPMCCs, [Zahng, 2003] obtains an accuracy of 82% testing on a small database containing a total of 45 songs by 8 different singers. [Shen et al., 2009] test their algorithm, which combines voice, instrumental and genre features, on two databases containing 230 songs by 23 artists and 8000 songs by 90 artists, respectively. They furthermore test on the databases *USPOP* and *Magnatune* which were used in the 2005 MIREX competition. Obtained accuracies range between 76.1% and 84.2%. Similar results are reported by [Nwe and Li, 2006] when applying their vibrato-based multimodal HMM classification to a small database containing 84 songs: Best performance is achieved using features extracted with cascaded vibrato filters (83.8%) which corresponds to an improvement of 5% over the use of MFCCs. Analyzing only the verse parts lead to a further increase in accuracy of 3%. The authors furthermore investigated the so-called a*lbum effect*: When the models are trained on songs contained in the same album, the accuracy when testing on the entire database significantly decreased (61.3%). The is suspected that in this case the homogeneous sound of the album determined by the production, mixing and mastering processes is modeled instead of the characteristics of the singing voice. Studying the fusion of the described features ([Nwe and Li,

2008]), a combination of MFCCs, LFCCs, harmonic and vibrato features gives better results than using single feature sets. Combining LFCCs with harmonic and vibrato features gives the best result with an accuracy of 86.7%. Based on a database containing 40 pop-songs by 10 different singers, [Fujihara et al, 2010] obtain an accuracy of 95% using 4-fold cross-validation for the method combining accompaniment sound reduction and reliable frame selection. A similar result is achieved for a larger commercial database containing 246 songs by 20 different singers. Using voice-enhancement as a pre-processing step while passing uncertainties to the final classification step, [Lagrange et al., 2012] obtaine up to 94% accuracy. The test is performed on the same database as in [Fujihara et al., 2012]. It is to mention that the authors built their classifiers solely based on MFCCs, while [Fujihara et al., 2010] additionally extracted F0-trajectories. [Cai and al., 2011] report a slightly lower accuracy of 90% for the combination of MFCCs, GTCCs and LPMCCs, tested on a smaller database containing a total of 100 songs. [Tsai and Lin, 2010] obtain 92.7% accuracy for their approach combining background music removal based on cepstral transformation and singing voice detection. They test on a larger database containing 300 pop music tracks. For 56 Carnatic music samples, [Sridhar and Geetha, 2008] improve the accuracy by using the Carnatic-specific CICC features to 87.5% compared to 30% using MFCCs.

Summarizing the results, it can be said that the state of the art algorithm by [Tsai and Lee, 2012] can reliably identify an unknown singer in monophonic signals. For polyphonic recordings, both the selected features as well as the treatment of the background music is crucial. Combining various reliable features and performing both, background removal and voice detection, leads to promising results above 90% for representative databases. Furthermore, various approaches have shown that specific genres might require adapted feature extraction and therefore testing and improving existing singer identification algorithms for non-Western music traditions is of great importance.

2.3 Similarity characterization

The similarity between pieces of music is a complex multi-dimensional problem with great importance for music recommendation systems. This review will only focus on melodic and voice timbre similarity measures, since these are the areas relevant to research task presented here.

2.3.1 Voice timbre similarity

Recent approaches have mainly focused on the overall timbre of polyphonic music pieces, which is to a great extend shaped by its instrumentation. Comparatively little work has been carried out in order to characterize timbre similarity among singers and its relation to perceived melodic similarity. As mentioned above, [Gómez et al., 2012c] report significant differences between human melodic similarity judgements of real and synthesized melodies and suggest to include timbral features when modeling perceived melodic similarity.

Similar to their singer identification approach, [Fujihara and Goto, 2007] used the LPMFCCs of the re-synthesized voice to determine the mutual information content of two GMMs as a similarity measure for the corresponding tracks. The proposed algorithm leads to a better agreement with human voice similarity judgements than a baseline approach where MFCCs are extracted from the untreated music signal. [Garnier et al., 2007] study voice quality in a top-down approach by qualitatively assessing voice quality from verbal descriptions of singing teachers and define corresponding acoustic descriptors based on low-level feature extraction.

2.3.2 Melodic similarity

Melodic similarity has been extensively studied in the past. A complete review would be out of scope, but a detailed summary of similarity measures and ground truth gathering can be found in [Marseden, 2012]. Here, only a few, representative concepts will be reviewed.

Based on previous experiments which indicate that the frequency of events is correlated to their contribution to perceived melodic similarity, [Eerola et al., 2001] use statistical properties of folk melodies, such as the distribution of notes and intervals, to predict similarity. Comparing to human ratings, they obtain a predictive power of 39% which is significantly increased to 52% when the instantaneous features are weighted by the note duration and the corresponding metrical hierarchy. When score-based musical descriptors, i.e. consonance or tonal stability, are used, a higher predictive power of 62% can be obtained. [Orio and Rodá, 2009] construct graph representations of melodies based on musicological studies: Elements of the melody are organized in a hierarchical tree structure, based on their harmonic and rhythmic significance. Similarity is then determined by comparing the respective graphs. Various recent approaches are based on note-based alignment: Melodies are treated as sequences of symbols and a cost function is used to relate two sequences in the best possible way. Hereby, large distances between melodies due to temporal deviations, i.e. caused by performers, do not influence the measure. A scoring function evaluates the resulting similarity of the aligned melodies. A review of

commonly used symbol alignment strategies and scoring function can be found in [Kranenburg et al., 2009]. Instead of a note-based symbolic representation, [Zhu and Kankanhalli, 2002] construct continuous melody contours from automatic transcriptions. Silent frames are filled by prolonging the previous note and melodic slopes are defined between local maxima and minima of non-overlapping time-windows. After aligning the slopes of two tracks, similarity is calculated as the distance between the resulting sequences. Further recent approaches also use continuous time-pitch representations: [Urbano et al., 2010] interpolate a pitch contour from discrete note values. After applying a local alignment algorithm, similarity is determined by the distance between the first derivatives of the shapes.

When dealing with raw audio material for which no score is available, it seems natural to compare estimated f0-trajectories directly instead of note symbols obtained from an automatic transcription. This is of specific relevance when analyzing the singing voice, since onsets are often vague and note segmentation is specially prone to errors. With the goal of implementing a singing assessment system, [Molina et al., 2013] use *dynamic time warping* to align f0-contours extracted from the audio signal: A cost matrix M describes the squared frequency deviation between all possible combinations of samples of both curves in the observed time frame:

(2.3.1) $$M_{ij} = min((f0_1(i) - f0_2(j))^2, \alpha)$$

The constant α limits the contribution to the cost function in case of extreme values due to errors in the f0-estimation or silent frames. The resulting optimal alignment corresponds to the optimal path among the matrix M. The total cost of the optimal path serves as a measure for the pitch deviation while the slope of the path characterizes rhythmic similarity.

Chapter 3

Methodology

In order to study the suitability of various feature sets for singer identification in Flamenco Cantes, several approaches are compared regarding their performance: A baseline approach for monophonic audio material based on frame-wise extracted MFCCs and GMM classification is implemented and evaluated for three datasets containing different music material. In an extension, MFCCs are combined with vibrato and note-based features obtained from an automatic transcription algorithm. Both monophonic and polyphonic material is classified using these feature sets in the WEKA machine learning environment ([Hall et. al., 2009]). In combination with temporal and pitch distances obtained from melodic alignment, the same attributes described above serve to study melodic similarity among a small data collection of monophonic *Martinetes*.

3.1 Dataset collection

For the automatic singer identification task, five datasets are used to evaluate the performance of the algorithms: For monophonic material, performances are compared for a cappella cantes, monophonic excerpts from opera singing and a monophonic Pop dataset. The extended feature set is furthermore tested for two datasets, the first containing polyphonic Flamenco recordings and the second classical opera singing with instrumental accompaniment.

3.1.1 Monophonic Pop database

As a reference for the baseline approach, a subset of the *MIR-1K* dataset ([Hsu, 2010]) was used. The audio material consists of short clips of amateurs singing Mandarin pop songs. Voice and accompaniment music are provided on separate channels. The subset used for the experiment contains a total of 228 clips sung by 5 different male singers. Based on the class distribution, the baseline accuracy of this database for song-wise classification corresponds to 22.08% correctly classified instances.

Singer	No. of clips	Total length |hh/mm/ss]
“Abjones”	40	0:04:48
“Bobon”	50	0:06:02
“Bug”	38	0:04:57
“Davidson”	52	0:06:33
“Fdps”	48	0:06:15

Table 3.1: *MIR-1K subset*

3.1.2 Monophonic Flamenco database

Obtaining a suitable amount of monophonic Flamenco voice recordings has shown to be problematic: Since a cappella singing styles have lost in popularity during the past decades and are rarely being recorded, a significant amount of available recordings are more than 25 years old and the sound quality varies strongly among songs. The dataset was collected from commercially available recordings, excerpts of TV documentaries and the *TONAS* Dataset ([Mora et al., 2010]). The set includes a total of 65 songs by five male singers. The baseline accuracy is 28.79%.

Singer	No. of songs	Total length |hh/mm/ss]
Antonio Mairena	19	0:54:38
“El Chocolate”	10	0:19:49
Juan Talega	9	0:27:36
Manuel “Agujetas”	16	0:34:05
Rafael Romero	10	0:07:28

Table 3.2: Monophonic Flamenco dataset

3.1.3 Monophonic opera singing excerpts

Singer	No. of songs	Total length [hh/mm/ss]
José Carreras	8	00:06:45
Plácido Domingo	7	00:15:28

Table 3.3: Monophonic classical dataset

A total of 15 excerpts are extracted from commercially available CDs which contain no or very low instrumental accompaniment. The resulting baseline accuracy is 53.55%.

3.1.4 Polyphonic Flamenco database

A polyphonic Flamenco dataset has been collected from commercially available recordings and song excerpts including different styles and albums. All audio examples contain a similar instrumentation: singing voice accompanied by one or two guitars and in some cases hand-clapping (*palmas*). A total of 150 songs sung by three male and two female artists are included. Classes are equally distributed with a resulting baseline accuracy of 20.00%.

Singer	No. of songs	Total length |hh/mm/ss]
Antonio Mairena	30	2:30:55
"El Camarón"	30	1:49:12
Carmen Linares	30	2:40:32
Manuel "Agujetas"	30	1:42:32
Niña de los Peines	30	1:14:28

Table 3.4: Polyphonic Flamenco dataset

3.1.5 Polyphonic classical database

Opera recordings containing singing voice and instrumental accompaniment and interludes have been collected from a variety of commercially available CDs and compilations. The resulting baseline accuracy is 20.00%.

Singer	No. of tracks	Total length |hh/mm/ss]
Monserrat Caballé	30	02:32:21
Cecilia Bartoli	29	02:19:35
José Carreras	30	01:38:34
Luciano Pavarotti	30	01:43:18
Plácido Domingo	30	02:03:26

Table 3.5: Polyphonic classical dataset

3.1.6 Audio material for melodic similarity analysis

Melodic similarity is studied on a small audio collection of 12 monophonic recordings of *Martinetes* performed by different artists. For each track, only the the first verse was extracted. The dataset includes the most representative singers of this style and has been collected after consulting with Flamenco experts. As mentioned above, all interpretations beloning to this style can be characterized by a common melodic skeleton which is subject to strong ornamentation and melismatic variation. Furthermore, *Martinetes* are performed in free rhythm, which leads to strong temporal deviations between the performances.

Singer	Lyrics	Duration [s]
Antonio Mairena	"A la puerta de la fragua"	11
Chano Lobato	"Ya no era aquel que era"	15
Chocolate	"Ay, ven acá mujer del mundo"	8
Jacinto Almadén	"Ay, ven acá mujer del mundo"	16
Jesús Heredia	"Ay, ven acá mujer del mundo"	20
Manuel Simón	"Ay, lo sacaban por la carraca"	30
Miguel Vargas	"To viene a chocá conmigo"	14
Naranjito	"Ay, estando yo en el altozano"	26
Paco De Lucia	"En la fragua de tio Ramón"	15
Talegón de Córdoba	Ay, yo no tengo la pena"	16
Tomás Pabón	"Ay, ven acá mujer del mundo"	11
Turronero	"A la puerta de la fragua"	10

Table 3.6: Collection of *Martinetes* for melodic similarity analysis

3.2 Monophonic MFCC-based singer identification

As baseline approach, a singer identification algorithm solely based on frame-wise extracted MFCCs using Gaussian mixture models for classification has been implemented in *MATLAB*. After splitting into training and test database for the current fold, the MFCCs 1 to 13 are extracted for each frame contained in the training set using the Auditory Toolbox [Slaney, 1993]. After grouping the frames by their corresponding class, a Gaussian Mixture Model is estimated for each singer using the *expectation maximization (EM)* algorithm. This iterative parameter adjustment process is based on the assumption, that every observation x_i is generated by a single component of the Gaussian mixture

model. Therefore, the *membership weight* is defined, which reflects the uncertainty that a specific datapoint is generated by a certain weight. Given a GMM with k components and the parameters Θ_k and component weights π_k are unknown,

(3.2.1) $$P(x/\Theta)=\sum_{k=1}^{K}\pi_k N(x/\Theta_k)$$

the membership weight for component k and observation i are calculated as:

(3.2.2) $$w(i,k)=\frac{\pi_k P(x_i/\Theta_k)}{\sum_{j=1}^{K}\pi_j P(x_i/\Theta i)}$$

In the EM-algorithm, the parameter values are initialized using random values or a given parameter set. In the so-called *E-step*, the membership weights are calculated for each datapoint k and each component weight k. In the *M-step*, parameter values are updated using the sum of the membership weights N_k for component k over all instances and the number of instances N:

(3.2.3) $$\pi_{(k,new)}=\frac{N_k}{N}$$

(3.2.4) $$\mu_{(k,new)}=\frac{1}{N_k}\sum_{i=1}^{N} w_{ik}*x_i$$

(3.2.5) $$\Sigma_{(k,new)}=\frac{1}{N_k}\sum_{i=1}^{N} w_{ik}*(x_i-\mu_{(k,new)})(x_i-\mu_{(k,new)})^T$$

Hereafter, the log-likelihood is computed as the logarithm of equation (3.2.1). Its behavior is preserved while the product in the formula simplifies to a sum. The process is repeated until either the log-likelihood or the parameters converge. Once the optimal parameters for each GMM are estimated, given an unknown instance all class likelihoods are computed frame-wise and frames with a maximum likelihood below an adjustable reliability threshold are discarded. The class is finally estimated by an unweighted majority vote among all frames. The EM-based parameter estimation as well as the Baysian class likelihood estimation was implemented using the *GMMBAYES* Toolbox [Paalanen, 2004].

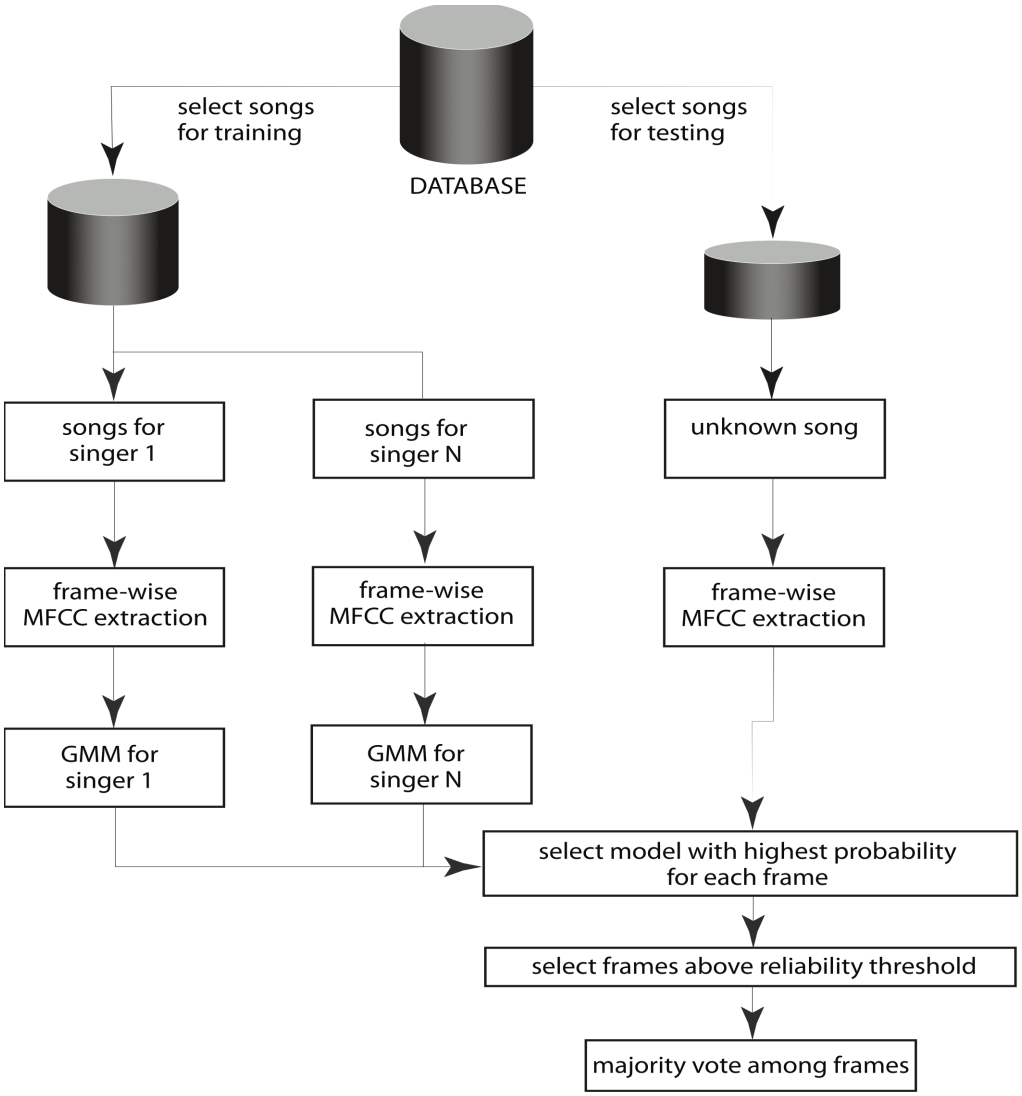

Figure 3.1: *MFCC-based singer identification algorithm*

3.3 Singer identification using timbre, vibrato and note features

In an extended approach, a singer identification based on SVM-classification in WEKA using global features extracted from the audio material was implemented. The features can be divided into three groups: Timbre, vibrato and note-based attributes.

3.3.1 Extraction of timbre features

Based on the assumption, that the sound quality present audio material is rather poor, at least for the monophonic dataset, and the focus is set to obtain a system robust to timbre distortion, the timbre-based feature extraction was

limited to the average values of the ***MFCCs 1-13***. For the monophonic database, silent frames were estimated from the energy envelope and excluded from the feature extraction process. For polyphonic material, voiced sections are estimated from the pitch salience function described in the following section.

3.3.2 Extraction of vibrato-motivated features

In [Nwe and Li, 2009] it is shown that including vibrato features improves the accuracy of a timbre-based singer identification algorithm. Vocal vibrato is furthermore a key feature of Flamenco singing and extensively used, especially in A Cappella styles. While [Nwe and Li, 2009] extracted vibrato descriptors directly from the audio signal using cascaded filterbanks, here vibrato rate, depth and amount is estimated from predominant fundamental frequency trajectories.

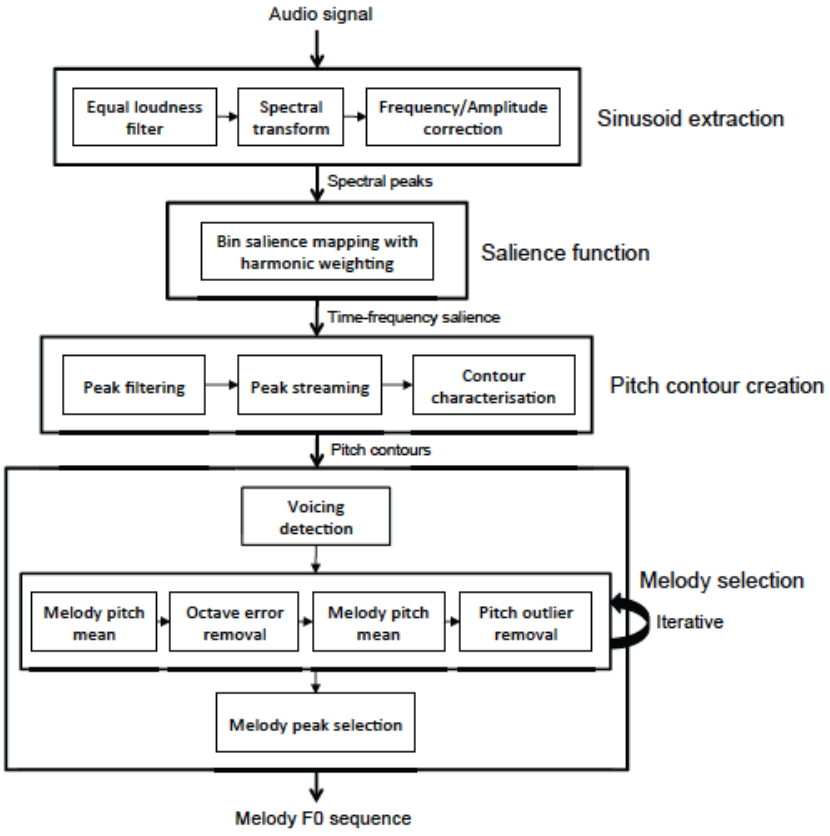

Figure 3.2: *predominant fo estimation [Salamon and Gómez, 2012]*

The pitch contour was obtained using the MELODIA vamp plugin[1] which implements a state-of-the-art predominant f0-estimation algorithm for monophonic and polyphonic audio material [Salamon and Gómez, 2012]. A block diagram of the algorithm (reproduced from [Salamon and Gómez, 2012]

[1] http://mtg.upf.edu/technologies/melodia

with permission from the authors) is provided in Figure 3.2: After applying a perceptually-based filter, peaks are extracted from the short-term spectrum and used to compute a salience value based on harmonic summation for all possible f0 values between 55-1760Hz. The final f0-curve is estimated by contour tracking based on auditory streaming principles and evaluating contour characteristics against thresholds determined from distributions of melodic contour features.

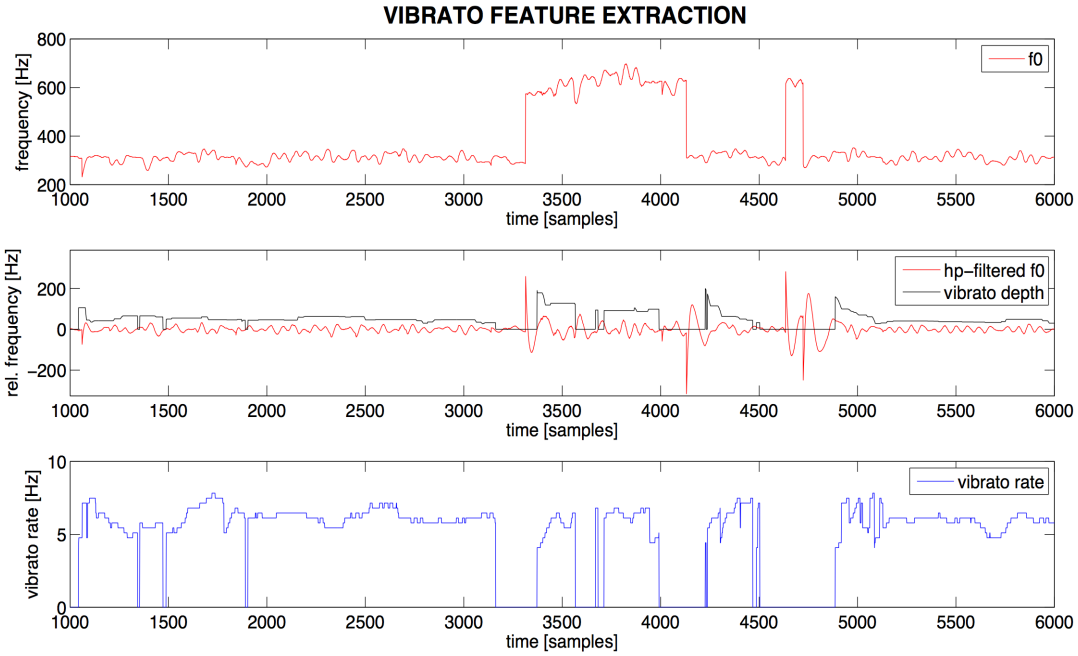

Figure 3.3: *Top: F0-contour; middle: pitch fluctuation curve and estimated vibrato depth; bottom: estimated vibrato rate*

Vocal vibrato can be characterized as an oscillation of the fundamental frequency within a range of 4 to 8 Hz and a depth of less than 200 cents. Therefore, the f0-curve was treated as a time series and high-pass filtered at 2Hz in order to obtain a zero-centered signal which preserves only fast pitch fluctuation. If vibrato is present, the spectrum of this signal shows a peak in considered frequency range. The instantaneous ***vibrato rate*** corresponds to the peak frequency in the spectrum and the ***vibrato depth*** can be estimated as the magnitude of the pitch fluctuation in the current frame. The total number of voiced frames divided by number of frames in which vibrato is detected is extracted as the ***vibrato ratio***.

3.3.3 Extraction of note-based features

Flamenco singing is to a great extend characterized by the way of employing ornamentation and melisma to the melodic corpus of a song. It seems therefore

natural to include global note-based descriptors as attributes in the classification process.

The transcription system described in [Gómez and Bonada, 2013] estimates a note representation from the predominant frequency trajectories mentioned in the previous section: First, an adaptive algorithm segments the f0 contour into short notes by minimizing a cost function which describes the melodic paths. The so obtained transcription is refined in an iterative approach: First, a global tuning frequency is estimated from a histogram containing the deviation of the frame-based f0 values with respect to the equal-tempered scale. A dynamic programming method segments the f0 trajectories into single notes considering energy, timbre, voicing and f0 cues. In an iterative refinement process, consecutive short notes with the same pitch and a smooth transition are consolidated, the estimated tuning frequency is refined according to the obtained notes, and the note nominal pitch is re-estimated based on the new tuning. Frequency. The system outputs a symbolic note representation of the analyzed audio including duration, energy and a pitch value quantized to an equal-tempered scale relative to the locally estimated tuning frequency. Furthermore, the tool provides a non-quantized, absolute pitch notation, allowing the estimation of the instantaneous ***detuning amount*** and its standard deviation, the ***tuning fluctuation***. Comparing the frame-wise note representation to the f0-contour, the ***amount of ornamentation*** is defined as the average absolute difference between both curves per frame.

A key characteristic of Flamenco singing is a singer's ability to stretch phrases with melisma and ornamentation without breaking the note flow by breathing or pausing. Silences between phrases are chosen consciously and serve as an expressive factor. Consequently, the ***amount of silence*** estimated from the relative number of unvoiced frames and the ***maximum phrase length*** have been selected as corresponding descriptors.

The melodic content is described by the following attributes obtained directly from the note transcription:

- ***pitch range*** in cents
- ***pitch fluctuation*** (standard deviation of the normalized pitch values)
- ***lowest pitch*** and ***highest pitch***
- ***onset rate*** (average number of notes per second)

3.3.4 Evaluation

Classification is performed using the WEKA [Hall et. Al, 2009] machine learning environment. For both monophonic and polyphonic material, the

singer identification is applied using all features as well as combinations of single feature sets to analyze their suitability for this task. To avoid misinterpretation of the results due to overfitting, all experiments are being conducted in a 10-fold cross-validation. In a further step, an automatic feature selection using a SVM classifier based subset evaluation is applied.

Three classifiers are compared regarding their performance:

- Support Vector Machines (SVM). SVM ([Cristiani and Shawe-Taylor, 2000]) take advantage of using a non linear attribute mapping that allows them to be able to predict non-linear models (though they remain linear in a higher dimension space). The classification accuracy of SVM largely depends on the choice of the kernel evaluation function and the parameters which control the amount to which deviations are tolerated (denoted by ε). Here, SVM have been explored with empirically determined paramenters: complexity c=15, tolerance=0.001 and $\varepsilon = 1^{15}$, using feature set normalization and a linear kernel.

- Artificial Neural Networks (ANN). ANN learning methods provide a robust approach to approximating a target function. In this study, a gradient descent back propagation algorithm ([Chauvin and Rumelhart, 1995]) is applied to tune the neural network parameters to best fit the training set. The momentum applied to the weights during updating is set to 0.2 and the learning rate (the amount the weights are updated) to 0.3. A fully-connected multi layer neural network with one hidden layer (one input neuron for each attribute and one output neuron for each class) is used.

- k-NN. k-NN are based on the notion of lazy learning which subsumes a family of algorithms that store the complete set of given (classified) examples of an underlying example language and delay all further calculations until requests for classifying yet unseen instances are received. k=8 was chosen for the monophonic and k=12 for the polyphonic database. These values were picked according to the rule of thumb relating the recommended number of observed neighbors k to the number of instances n:

(3.3.1) $$k = \sqrt{n}$$

Algorithm performance is evaluated by calculating the percentage of correctly classified instances, averaged over N=10 folds:

(3.3.2) $$CCI = \sum_{i=1}^{N} \frac{number\ of\ correctly\ classified\ instances}{total\ number\ of\ instances} * \frac{1}{N} * 100$$

3.3.5 Summary of extracted attributes

Table 3.5 gives an overview of the timbre, vibrato and note-based attributes extracted from the audio recordings.

Feature set	Attribute	Details
Timbre	MFCC 1 - 13	Frame-wise extracted mel-frequency cepstral coefficients 1 to 13 averaged over each track.
Vibrato	vibrato rate	Average vibrato frequency in *Hz*.
	vibrato rate standard deviation	Standard deviation of the detected vibrato frequencies.
	average vibrato depth	Average vibrato extend in *cents* calculated from the relative pitch difference of the zero-centered f0-contour.
	vibrato depth standard deviation	Standard deviation of the detected vibrato depth.
	Vibrato ratio	Amount of detected vibrato in each track: $\frac{number\ of\ \ frames\ containing\ vibrato}{total\ number\ of\ \ frames}$
Note-based	detuning amount	Maximum change of the tuning frequency in *cents*.
	tuning fluctuation	Standard deviation of the local detuning.
	ornamentation amount	For the current frame *n* and the total number of frames *N*: $\sum_{n=1}^{N} \frac{f0-value[n]}{note-value[n]} * \frac{1}{N}$
	silence ration	Ratio of unvoiced frames: $\frac{number\ of\ unvoiced}{total\ number\ of\ \ frames}$
	maximum phrase length	Length of the longest frame within the song in *s*.
	pitch range	Difference between highest and lowest detected pitch in MIDI notes.
	pitch fluctuation	Standard deviation of the MIDI pitch
	lowest pitch	Lowest detected MIDI pitch.
	onset rate	Average number of notes per second.
	Average volume	Average MIDI velocity of the normalized transcription
	Dynamic fluctuation	Standard deviation of the MIDI velocity.
	Average duration	Average note duration in *s*.
	Duration fluctuation	Standard deviation of the note duration.

Table 3.7: Extracted features

3.4 Melodic similarity

3.4.1. Human ratings

Human similarity ratings are taken over from [Gómez, 2012c]: 19 novice listeners and 3 Flamenco experts were asked to form groups of similar melodies, based on both, real recordings and melodies synthesized from the estimated f0-contour. Distance matrices were constructed based on the number of participants which grouped a pair of melodies together. While agreement among naïve listeners was rather low (μ=0.0824 and σ=0.2109), the small group of experts gave more homogeneous ratings (μ=0.1891 and σ=0.1170). Even though participants were specifically asked to judge melodic similarity, a qualitative analysis in form of an open questionnaire showed that specially naïve listeners took different characteristics, i.e. vibrato and dynamics, into consideration. As mentioned above, significant variations between perceived similarity of real recordings and synthesized melodies were observed for both, novice listeners and experts.

3.4.2. Melodic alignment using dynamic time warping

Similar to [Molina et al., 2013], dynamic time warping is used to align two melodies and estimate their temporal and pitch similarity. The vocal vibrato and short ornamentations have a strong influence on the cost matrix defined in eq. (2.3.1) and f0-contours are consequently prone to misalignments. Therefore, a continuous pitch curve of quantized note values is used instead. Silent frames correspond to negative pitch values. The dynamic time warping algorithm, which finds the optimal path along the cost matrix, is based on the implementation in [Ellis, 2003] and includes several restrictions for a meaningful alignment.

For a cost matrix M, the and the indice vector of the optimal path p, the temporal deviation among two melodies can be estimated as the squared distance to the diagonal path p_{diag}, normalized by the path length N:

(3.4.1)
$$\Delta_{temp} = \frac{\sum_{i=1}^{N} (p(i) - p_{diag}(i))^2}{N}$$

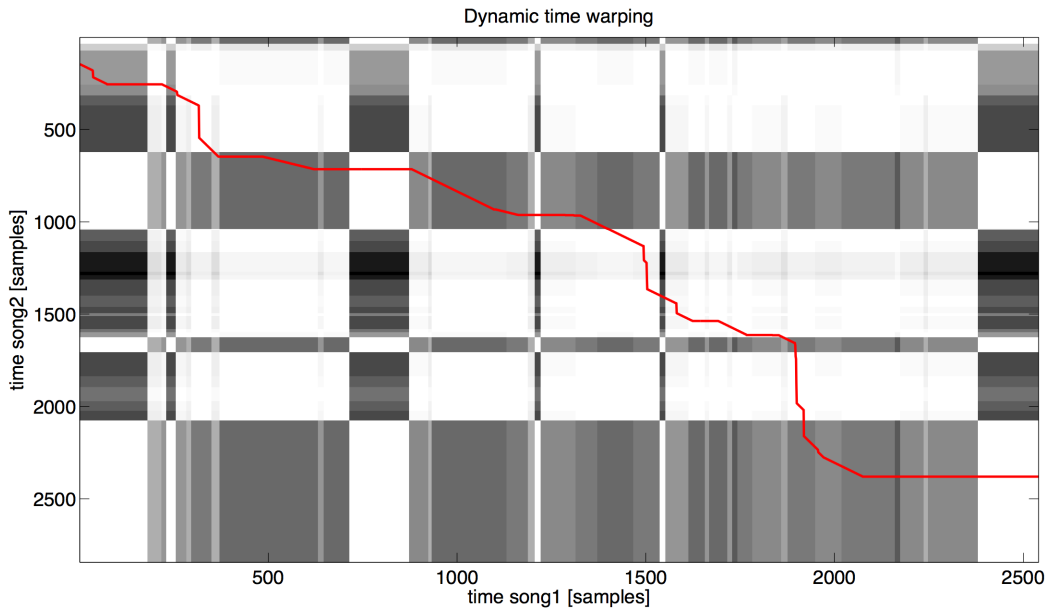

Figure 3.4: *Cost matrix and ideal path*

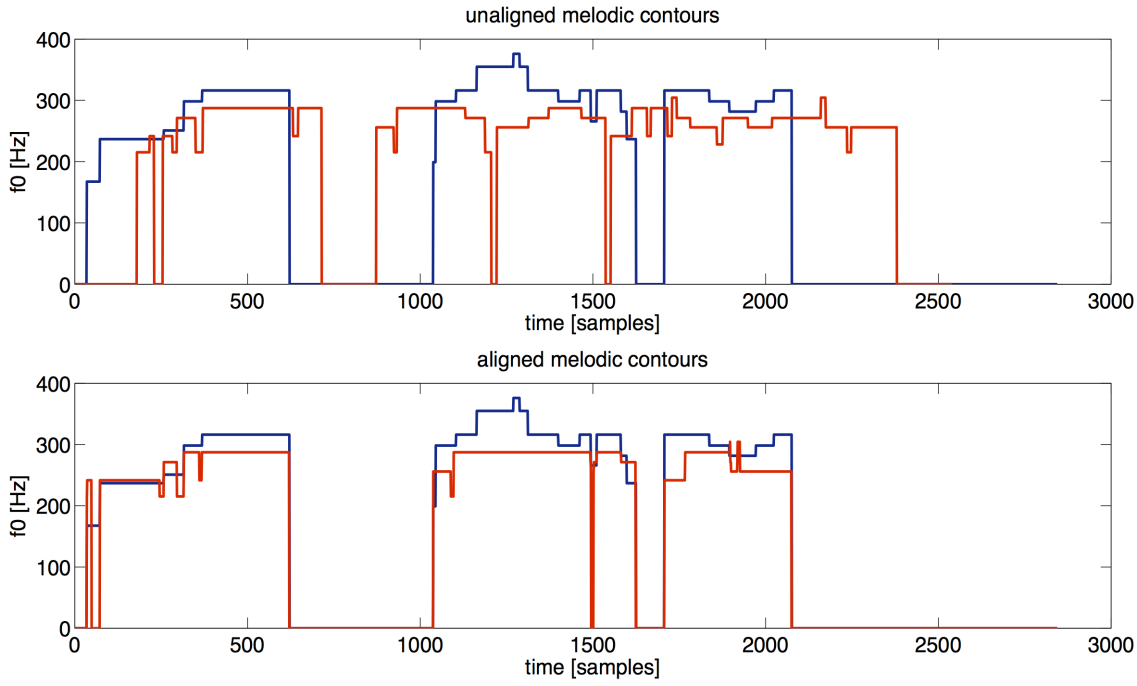

Figure 3.5: *Melodic alignment using dynamic time warping*

The pitch deviation is defined as the sum of the elements of the cost matrix over the optimal path p:

(3.4.2)
$$\Delta\, pitch = \frac{\sum_{i=1}^{N} (M(p(i)))}{N}$$

This value corresponds to the distance between the melodic contours normalized by their length.

3.4.3. Global note, vibrato and timbre features

The global timbre, note-based and vibrato attributes extracted from the audio recordings correspond to the full feature set used for singer identification, as described in the previous section. Vibrato and timbre descriptors are grouped together to form one feature set. Distances between tracks are estimated as the euclidean distance between their corresponding feature vectors.

3.4.4. Evaluation

Once distance matrices based on computational analysis are found, their correlation to distances obtained from human judgement is of special interest. A common method to evaluate a possible relation between to distance matrices is the *Mantel* test [Bonnet and Van de Peer, 2002]. First, the linear correlation between two matrices A and B is measured with the Pearson correlation coefficient, which gives a value r between -1 and 1 for matrices with N values in the upper triangle:

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{A_{(i,j)} - \bar{A}}{std(A)} * \frac{B_{(i,j)} - \bar{B}}{std(B)} \quad , \tag{3.4.3}$$

where $\bar{A}$ is the mean and *std(A)* the standard deviation of the elements in matrix *A*. A strong correlation is indicated by a value significantly different from zero. To verify that a relation exists, the value is compared to correlations to per-mutated versions of one of the matrices. Here, 10000 random permutations are performed. The confidence value *p* corresponds to the proportion of permutations which gave a higher correlation than the original matrix. Consequently, a *p*-value close to zero confirms an existing relation.

Chapter 4

Results and discussion

This chapter gives and overview of quantitative results for the singer identification task obtained using the baseline approach in comparison with the extended feature set described in the previous section. Furthermore, results of the computational melodic similarity approach are given as their correlation with human ratings.

4.1 Monophonic MFCC-based singer identification

The baseline approach based on spectral descriptors has been applied all three monophonic databases.

4.1.1. Classification results

The GMM classification based on frame-wise extracted MFCCs gives convincing results for the Pop music database given the relatively small amount of audio material. This result is similar to the accuracy reported by [Tsai and Lee, 2012]. For the classical dataset, which includes only two possible classes, the singer identification was performed error-free. This confirms that for a suitable audio quality, timbre based singer identification gives reliable results even for small databases. By contrast, the accuracy is significantly lower for the monophonic Flamenco dataset, even though more audio material is available and the number of classes is equal to the Pop dataset.

Database	Mean CCI [%] among folds
Monophonic Pop (MIR 1-K subset)	97.14
Monophonic Classical	100.00
Monophonic Flamenco	65.00

Table 4.1: Classification results for MFCC-based singer identification

Since the classification is solely based on timbre features, the varying audio quality of the Flamenco recordings and the resulting distortion of the spectrum are probably the reason for inferior performance rather than the musical material. All clips from the MIR-1K database were recorded under similar circumstances and with unchanged technical setups, while the monophonic Flamenco database was gathered from different media and time spans. The classical dataset contains high-quality audio recordings, with no significant overall timbre variation, even tough collected from different sources. The experiment hence demonstrates the low robustness of MFCC-based singer identification towards spectral distortion. It consequently seems convincing to include melodic features in the singer identification algorithm, even tough humans seem to recognize voices mainly by their timbre.

4.1.2. Parameter optimization

The described system allows the adjustment of two parameters: The reliability threshold and the complexity of the Gaussian Mixture Model.

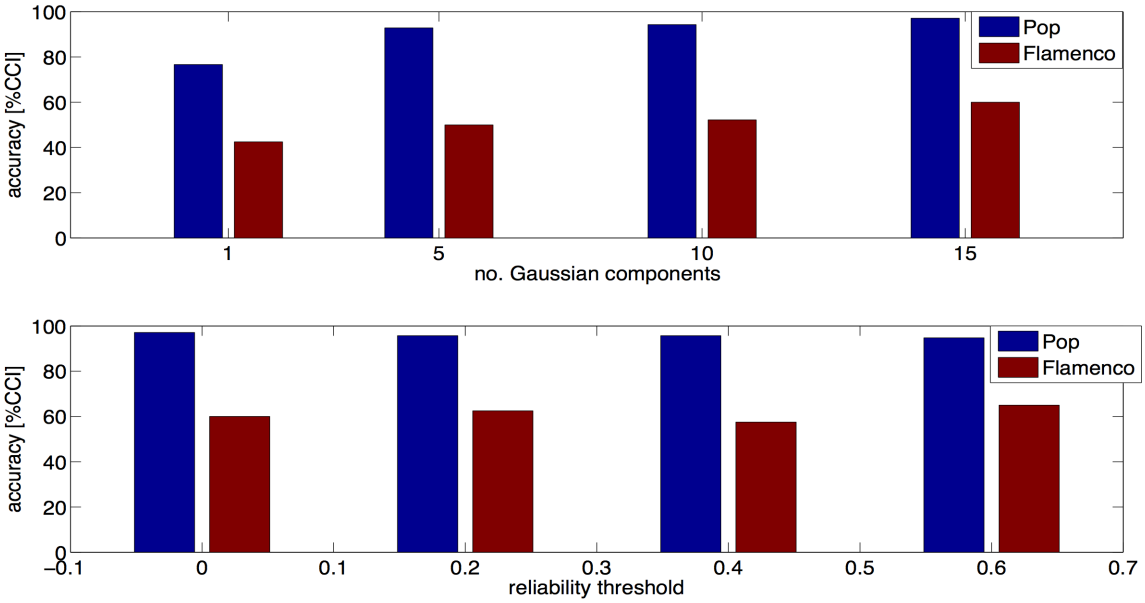

Figure 4.1: *GMM parameter optimization*

An empirical optimization has shown that the reliability threshold does not seem to improve classification results significantly. For both databases the variation is below one percent. However, the a higher complexity increases the performance and for 15 Gaussian components overfitting still does not seem to occur.

4.2 Singer identification based on timbre, vibrato and note features

The extended approach has been applied to the Monophonic and Polyphonic Flamenco dataset and furthermore to the polyphonic classical dataset. The method is not suitable for Pop singing, since in most tracks no vibrato is present. The monophonic Classical database has not been used, since the number of instances is too low for a meaningful classification with global descriptors.

4.2.1. Monophonic Flamenco dataset

For the monophonic Flamenco dataset, the singer identification based on averaged MFCCs only obtained slightly lower results than the GMM-based classification based on frame-wise extracted MFCCs (65.00%).

Attribute set	% Correctly Classified Instances		
	SVM	**8-NN**	**ANN**
All	80.00	80.00	84.62
Timbre	60.00	47.70	56.92
Vibrato	72.31	63.07	64.62
Note	63.08	70.67	69.23
SVM attribute selection	83.08	83.08	**84.62**

Table 4.2: Classification results: Monophonic Flamenco database

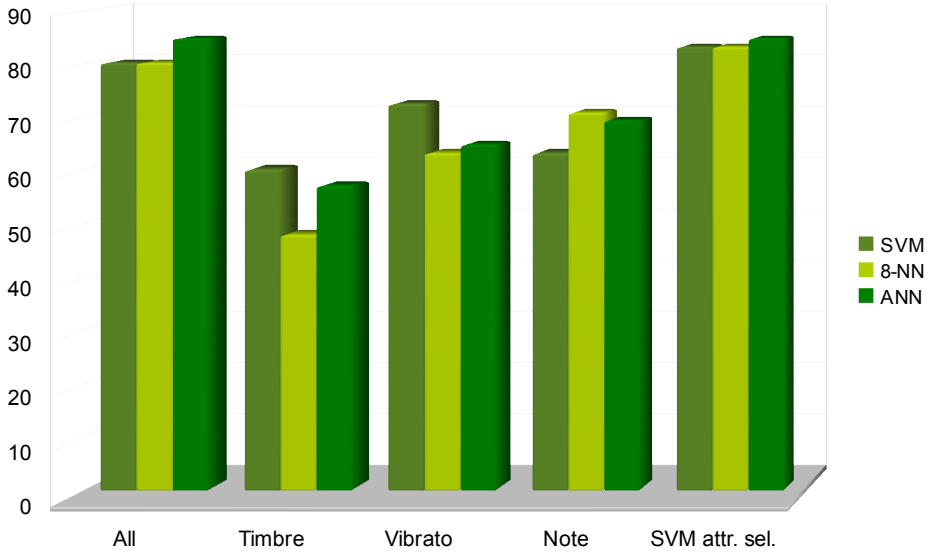

Figure 4.2: *Classification results [%CCI]: Monophonic Flamenco database*

Independent of the used classifier, vibrato and note-based features obtain higher accuracies than timbre descriptors and overall highest performance is achieved when all feature sets are combined. An SVM-based attribute selection filter ranks the features by their accuracy when used independently. When discarding the five features with the lowest rank (**MFCC12, *vibrato ratio, detuning amount, lowest pitch and std vibrato depth***), a further increase in accuracy can be obtained for the SVM and k-NN classification. Among the classifiers, results slightly vary. The k-nearest-neighbor algorithm which has been successfully used for complex music classification tasks such as audio tag classification [Sordo, et al., 2011] gives acceptable results, given its simplicity and short runtime.

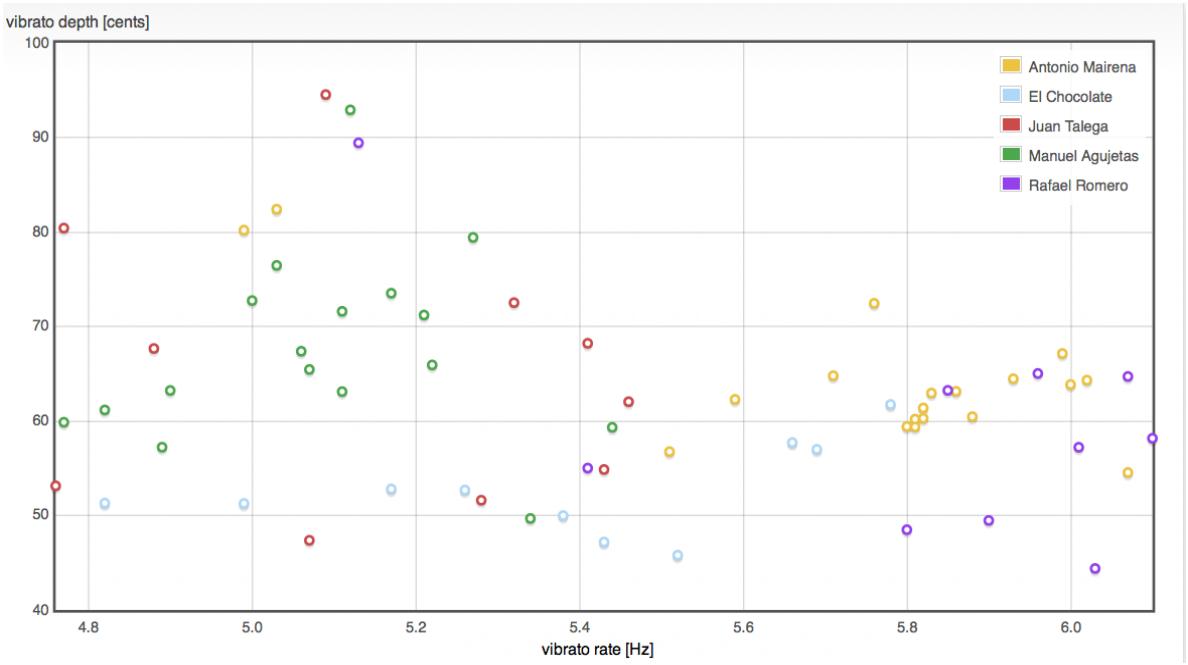

Figure 4.3. *vibrato rate and depth for the monophonic Flamenco database*

Analyzing the confusion matrix and the detailed classification results gives no indication for any singer being more prone to classification errors.

classified as → Singer	A. Mairena	Chocolate	J. Talega	Agujetas	R. Romero
A. Mairena	17	0	1	1	0
Chocolate	0	7	1	1	0
J. Talega	0	1	8	0	0
Agujetas	0	1	0	16	0
R. Romero	2	0	0	1	7

Table 4.3: Confusion matrix: Monophonic database, all attributes, multilayer perceptron

True positive rate	False positive rate	precision	recall	F-measure	Singer
0.895	0.043	0.895	0.895	0.895	Antonio Mairena
0.7	0.036	0.778	0.7	0.737	"El Chocolate"
0.889	0.036	0.8	0.889	0.842	Juan Talega
0.941	0.063	0.842	0.941	0.889	Manuel "Agujetas"
0.7	0.018	0.875	0.7	0.778	Rafael Romero

Table 4.4: Classification result analysis: Monophonic database

4.2.2. Polyphonic Flamenco dataset

Conducting the MFCC-based singer identification on the polyphonic flamenco dataset results in a significantly higher accuracy compared to the monophonic dataset. A possible reason is the lower amount of spectral distortion due to a better and more consistent audio quality. Furthermore, the *album effect* mentioned above has a strong influence as shown in the following section.

Attribute set	% Correctly Classified Instances		
	SVM	12-NN	ANN
All	88.00	74.00	89.33
Timbre	86.67	68.67	85.33
Vibrato	63.67	67.33	68.67
Note	53.33	47.33	53.67
Timbre + Vibrato	90.00	76.67	91.33
Note + Vibrato	71.33	67.33	72.00
Note + Timbre	76.00	66.67	85.33
SVM attribute selection	86.00	80.00	**92.00**

Table 4.5: Classification results: Polyphonic Flamenco database

However, performance is increased to 90.0% when timbre and vibrato features are combined. Incorporating additionally note-based features leads to a small decrease in accuracy and it can furthermore be seen that the performance solely based on note-features is comparatively low. A possible reason is the fact that the expressiveness of the singing voice in accompanied styles and especially the rhythmic and melodic modifications are rather limited compared to monophonic styles. Also, vibrato is used less frequently. Besides that, the voice

tuning is aligned to the guitar and therefore less detuning occurs during a song. Based on the SVM attribute selection, the features ***average note duration***, ***detuning amount***, ***tuning fluctuation***, ***pitch range***, ***MFCC 13*** and ***dynamic fluctuation*** discarded and the resulting accuracy further improved to 92.0 %.

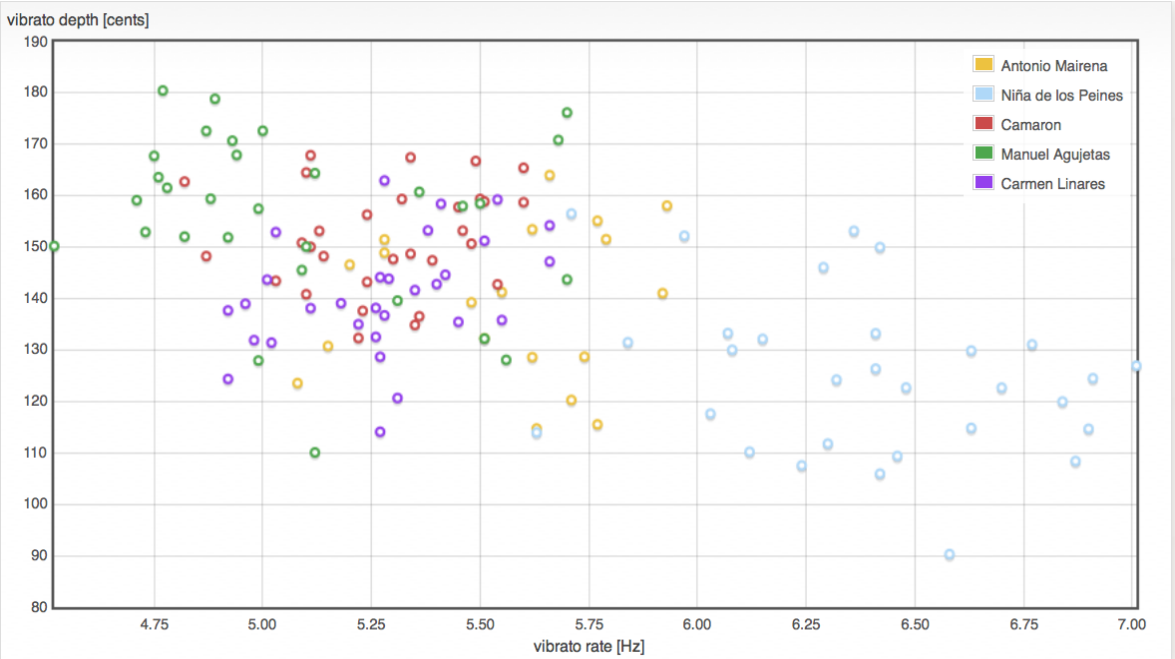

Figure 4.4. *vibrato rate and depth for the polyphonic Flamenco database.*

classified as → Singer	Antonio Mairena	Niña de los Peines	Camaron	Manuel Agujetas	Carmen Linares
Antonio Mairena	30 [25]	0 [1]	0 [0]	0 [4]	0 [0]
Niña de los Peines	0 [1]	27 [28]	1 [0]	2 [1]	0 [0]
Camaron	0 [1]	0 [0]	28 [26]	0 [1]	2 [0]
Manuel Agujetas	2 [7]	0 [1]	0 [0]	27 [22]	1 [0]
Carmen Linares	2 [2]	0 [0]	1 [0]	1 [1]	26 [27]

Table 4.6: Confusion matrix: Polyphonic Flamenco database, automatically selected features and [timbre features]

Analyzing the confusion matrices of both, the automatically selected feature set as well as the MFCC-based classification, it can be seen that interestingly enough there is no elevated confusion among singers of the same gender.

True positive rate	False positive rate	precision	recall	F-measure	Singer
1	0.033	0.882	1	0.938	Antonio Mairena
0.9	0	1	0.9	0.947	Niña de los Peines
0.933	0.017	0.933	0.933	0.933	Camaron
0.9	0.025	0.9	0.9	0.9	Manuel Agujetas
0.867	0.025	0.897	0.867	0.881	Carmen Linares

Table 4.7: Classification result analysis: Polyphonic Flamenco database, automatically selected features

4.2.3. Influence of the album effect

As shown in [Nwe and Li, 2006], it might occur that in timbre-based singer identification, not only the voice is modeled, but also the overall sound of an album. While the monophonic Flamenco database mainly consists of live recordings and compilations, the polyphonic dataset contains for some cases various songs from the same album. To investigate the album effect, a two-class classification was performed in two setups: First, the model is built on a database containing only songs from one album per singer and tested on all instances from other sources. This obviously represents a worst-case scenario , since a training database should be as diverse as possible. In a comparison, a test set of the same size is randomly selected. The experiment is repeated for the full attribute set as well as the timbre attributes only.

Attribute set	% Correctly Classified Instances	
	Album split	Random split
All	91.67	90.48
Timbre	41.67	95.24

Table 4.8: Influence of the album effect

The results clearly show a significant decrease in accuracy for the timbre-based classification when the model is trained on a single album per singer. When the full attribute set is used, both results are within the same range. This experiment clearly demonstrates a clear advantage of non timbre-based features for small training sets with little diversity.

4.2.4. Polyphonic classical dataset

Overall, the obtained accuracy is significantly lower for classical music tracks. As in the previous example, best performance is obtained using automatically selected attributes. Here, the features ***MFCC12, amount of silence, vibrato rate fluctuation, note duration fluctuation, lowest pitch*** and ***average note duration*** were discarded in the SVM attribute selection.

Attribute set	% Correctly Classified Instances		
	SVM	**12-NN**	**ANN**
All	66.44	72.48	75.17
Timbre	70.47	51.00	65.10
Vibrato	61.07	46.98	59.73
Note	57.04	45.64	48.99
Timbre + Vibrato	73.83	64.43	71.81
Note + Vibrato	63.09	57.05	60.40
Note + Timbre	71.81	55.03	63.09
SVM attribute selection	**76.51**	66.44	**72.48**

Table 4.9: Classification results polyphonic classical database

There are several possible reasons provoking a lower accuracy for this type of musical material: First, the accompaniment instrumentation in classical music does also cover main melodic lines during short interludes, causing a high pitch salience and consequently false fo-trajectories and errors in the note transcription. Furthermore, opera singing is strongly tied to a score and leaves comparatively little melodic interpretation to the performer. Accordingly, statistical note features are less informative, since they are to a large extend determined by the score.

True positive rate	False positive rate	precision	recall	F-measure	Singer
0.833	0.042	0.833	0.833	0.833	Monserrat Caballé
0.793	0.042	0.821	0.793	0.807	Cecilia Bartoli
0.733	0.050	0.786	0.733	0.759	José Carreras
0.667	0.134	0.556	0.667	0.606	Luciano Pavarotti
0.600	0.076	0.667	0.600	0.632	Plácido Domingo

Table 4.10: Classification result analysis: Polyphonic classical database, SVM selected features

classified as → **Singer**	Monserrat Caballé	Cecilia Bartoli	José Carreras	Luciano Pavarotti	Plácido Domingo
Monserrat Caballé	25 [23]	3 [1]	1 [4]	1 [0]	0 [2]
Cecilia Bartoli	2 [3]	23 [24]	3 [2]	1 [0]	0 [0]
José Carreras	2 [4]	2 [1]	22 [18]	4 [3]	0 [4]
Luciano Pavarotti	1 [1]	0 [0]	0 [4]	20 [21]	9 [4]
Plácido Domingo	0 [0]	0 [0]	2 [2]	10 [9]	18 [19]

Table 4.11: Confusion matrix: Polyphonic classical database, automatically selected features and [timbre features]

Analyzing the confusion matrix, a slight tendency towards an elevated number misclassifications among singers of the same gender can be observed.

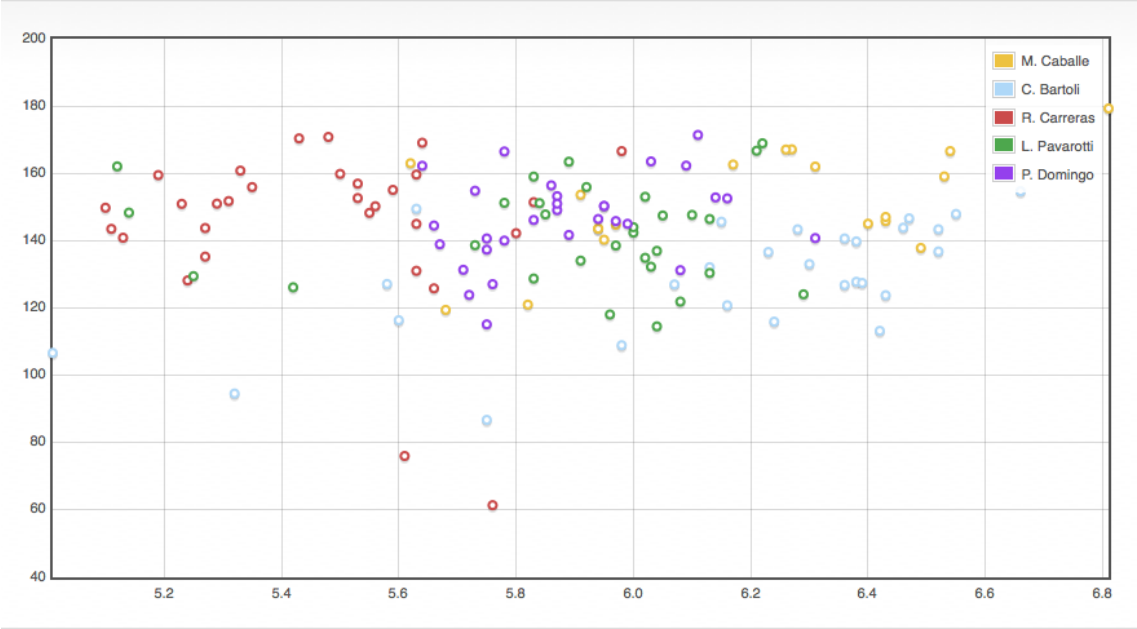

Figure 4.5. vibrato rate and depth for the polyphonic Classical database

4.3 Similarity

Analyzing the correlation between human ratings and the computationally generated similarities, the temporal deviation described by slope of the optimal path of alignment seems to model human perception best for both naïve and expert listeners. The pitch distances of the aligned sequences do not seem to be related to human judgement. Even though timbre seems to influence the experiment, no correlation has been found with the related timbre and vibrato descriptors. The similarity matrix constructed from statistical note features indicates a correlation to human judgements from real recordings.

Human judgement	Temporal deviation	Pitch deviation	Statistical note features	SVM-selected note features	Timbre and vibrato
Naïve listeners, real recordings	r=0.333 p=0.003	r=-0.130 p=0.198	r=0.219 p=0.053	r=0.308 p=0.0123	r=-0.085 p=0.335
Naïve listeners, synthetic melodies	r=0.245 p=0.123	r=-0.096 p=0.204	r=0.069 p=0.273	r=0.176 p=0.061	---
Expert listeners, real recordings	r=0.306 p=0.047	r=-0.118 p=0.256	r=0.236 p=0.083	r=0.431 p=0.011	r=0.050 p=0.358
Expert listeners, synthetic melodies	r=0.213 p=0.044	r=-0.094 p=0.224	r=0.102 p=0.192	r=0.207 p=0.044	---

Table 4.12: Correlation between computational similarity measures and human ratings

When the distance matrix of the expert ratings is represented as a phylogenetic graph[2], two clusters can be identified:

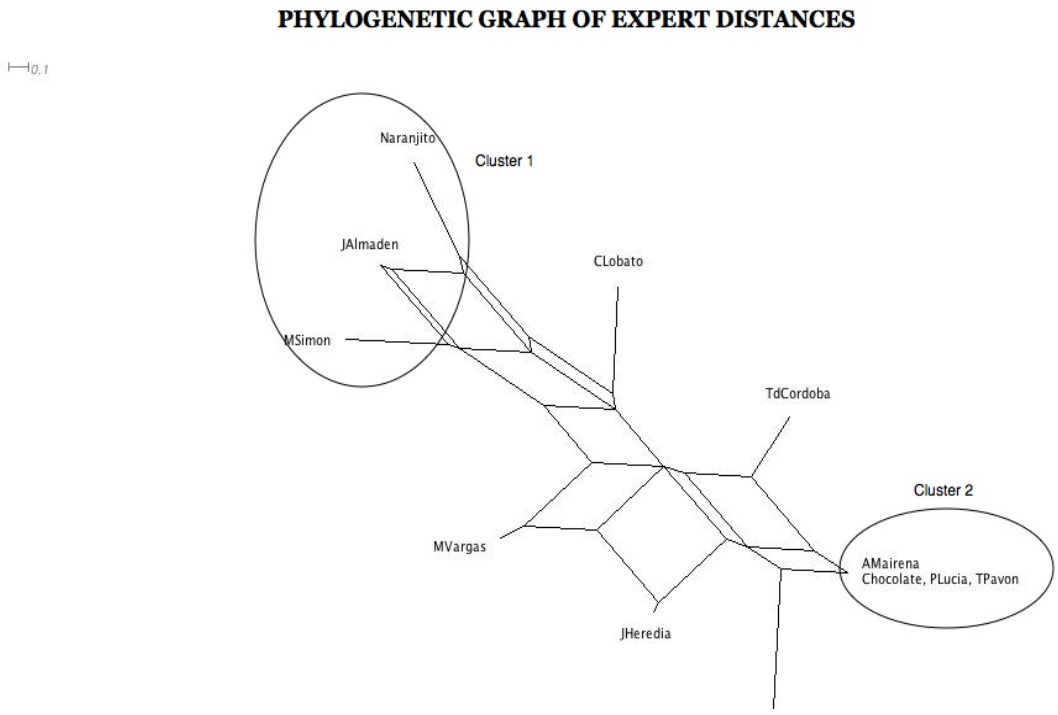

Figure 4.6. *phylogenetic graph of expert distances*

Using these two clusters as possible classes in a classification, an automatic SVM attribute selection can be used to identify the note-based features which are most suitable to distinguish between these two groups. Selecting the six best ranked attributes ***amount of silence, std duration, ornamentation***

[2] http://www.splitstree.org/

amount, average duration, average dynamics, and ***std dynamics***, a distance matrix can be calculated based on this subset (*SVM-selected note features*). The results show that this subset models the human ratings best among all selected approaches and a correlation can be found with all experimental setups. Interestingly enough, these features describe temporal and dynamic behavior and do not represent pitch content and melodic contours. This corresponds to the high correlation found with temporal distance matrix obtained from the dynamic time alignment. Consequently, human judgement of melodic similarity of monophonic Flamenco *cantes* seems to be based on temporal behavior of melodies rather than pitch content. Furthermore, the dynamic interpretation has an influence on the perceived similarity. This effect seems to be stronger for real performances than for synthesized melodies.

4.4 General observations

Analyzing two-dimensional plots of the extracted vibrato features (Figure 4.3, 4.4 and 4.5, available online[3]) it can be seen that vocal vibrato is a suitable feature to describe the individual characteristics of a particular singer. Furthermore, the observation about genre-specific tendencies correspond to the findings in [Salamon et al, 2012]: Apart from a single exception (Niña de los Peines), the vibrato rate in Flamenco singing tends to be lower, between 4.5 and 6 Hz compared to Opera singing (5 to 7 Hz). Analyzing highest and lowest pitch does not give a clear distinction between male and female singers. This is probably due to (octave-) errors in the f0-estimation due to accompaniment sound.

[3] http://nadinekroher.wordpress.com/28-2/

Chapter 5

Conclusions and future work

The aim of this study is to explore computational analysis and characterization of Flamenco singing. The first focus is on the implementation of an automatic singer identification. Experiments show that an identification based on spectral descriptors gives reliable results for databases containing high-quality audio recordings, even if the amount of available material is limited. On the other hand, this approach not robust to spectral distortions and varying overall timbre due to inconsistent recoding situations and audio quality, as it is the case for the monophonic Flamenco database. Since vocal vibrato is extensively used in Flamenco singing, corresponding descriptors have been extracted from an estimated f0 envelope. Furthermore, since the Flamenco singing performances can be characterized by a large amount of spontaneous rhythmic, melodic and dynamic variation, statistical attributes describing temporal, melodic and dynamic behavior have been extracted from automatic transcriptions. The resulting method combining vibrato, note and timbre descriptors shows a significant increase in accuracy compared to the baseline approach. In a second step, various computational melodic similarity measures have been implemented: A dynamic time warping algorithm aligns two melodies and gives a measure for rhythmic similarity. Pitch similarity is calculated as the distance between the aligned sequences. Furthermore, vector distances between the attributes described above have been calculated. Analysis of the correlation between the resulting distance matrices and human ratings indicate that perceived melodic similarity can be modeled best by describing temporal and dynamic deviation compared to similarity measures for pitch content and melodic contour.

5.1 Contributions

According to the goals defined in section 1.2, in the scope of this study the following contributions have been made:

- A review of state of the art approaches to singer identification, melodic similarity analysis and computational analysis and description of Flamenco music.

- Collection of various datasets containing Flamenco, Pop and Classical music material, suitable for singer identification and analysis tasks.

- Implementation of a singer identification algorithm for Flamenco music based on genre-specific concepts, which is suitable for small datasets and low audio quality material.

- A method for reliably extracting vibrato features from a given fo-contour.

- Application of dynamic time warping for computational melodic similarity analysis of monophonic Flamenco *Cantes*.

- Analysis of correlation between computational models for melodic similarity and human ratings for monophonic Flamenco *Cantes*.

- General characterization of the properties of the Flamenco singing voice compared to other genres obtained from audio descriptors.

5.2 Future work

Based on the findings in this study, the following aspects are of interest for related studies:

- Improvement of the automatic transcription tool and incorporation of genre-specific knowledge such as detection and description of specific ornamentation. Definition of a symbolic notation system in cooperation with Flamenco experts.

- Experiments to obtain human ratings for artist and style similarity and implementation of corresponding measures based on the descriptors used for the singer identification algorithm.

- Development of a Flamenco singing assessment and classification tool for educational and musicological purposes.

- Development of a data exploration tool for low- and high-level descriptors of Flamenco music. Incorporation of Flamenco styles into existing music databases, such as *MusicBrainz*[4].
- Expressive performance modeling of Flamenco singing based on audio and image analysis.

[4] http://www.musicbrainz.org

Bibliography

Bartsch, M. A. and Wakefield G. H. (2004). Singing Voice Identification Using Spectral Envelope Estimation. *IEEE Transactions on speech and audio processing*, 12 (2): 100-110.

Bonnet, E. and Van de Peer, Y. (2002). zt: a software tool for simple and partial Mantel tests. *Journal of Statistical software*, 7 (10): 1-12.

Cristiani, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Chauvin, Y. and Rumelhart, D. E. (1995). *Backpropagation: theory, architectures and applications*. Lawrence Erlbaum Associates, Mawah, N.J..

Cabrera, J.J., Díaz-Báñez, J.M., Escobar, F.J., Gómez, E., Gómez, F., Mora, J. (2008). Comparative Melodic Analysis of A Cappella Flamenco Cantes. *Conference on Interdisciplinary Musicology*. Thessaloniki, Greece.

Cai, W., Quiang, L. And Guan, X. (2011). Automatic Singer Identification Based on Auditory Features. *Proceedings of the 7th International Conference on Natural Computation*, 1624-1628.

Delgado, M., Fajardo, W. and Molina-Solana, M. (2011). A state of the art on computational music performance. *Expert Systems with Applications*, 38: 155-160.

Eerola, T., Järvinen, T., Louhivuori, J. and Toivianen, P. (2001). Statistical Features and Perceived Similarity of Folk Melodies. *Music Perception*. Spring 2001, 18 (3), 275-296.

Ellis, D. (2003). Dynamic Time Warp (DTW) in Matlab. Web resource. http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/, retrieved: 15/5/2013.

Flemming, E. (2005). Course materials for 24.963 Linguistic Phonetics, Fall 2005. *MIT OpenCourseWare*, (http://ocw.mit.edu/), Massachusetts Institute of Technology. [Downloaded on 30/01/13].

Fujihara, H. and Goto, M. (2007). A Music Information Retrieval System Based on Singing Voice Timbre. *Proceedings of the 8th International Conference on Music Information Retrieval*, pp. 467-470.

Fujihara, H., Goto, M., Kitahara, T. and Okuno, H. G. (2010). A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval. *IEEE Transactions on audio, speech and language processing*, 18 : 638-648.

Garnier, M., Heinrich, N., Castellengo, M., Sotiropoulos, D. and Dubois, D. (2007). Characterisation of Voice Quality in Western Lyrical Singing. *Journal of interdisciplinary music studies*, 1 (2), pp.62.91.

Gómez, E. and Bonada, J. (2008). Automatic Melodic Transcription of Flamenco Singing. *Proceedings of the 4th Conference on Interdisciplinary Musicology*, 2-6 July, Thessaloniki, Greece.

Gómez, E. and Bonada, J. (2013). Towards computer-assisted Flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. *Computer Music Journal.* 37 (2): 73-90.

Gómez, E., Bonada, J. and Salamon, J. (2012a). Automatic Transcription of Flamenco Singing from Monophonic and Polyphonic Music Recordings. *Proceedings of the 3rd Interdisciplinary Conference on Flamenco Research and the 2nd International Workshop of Folk Music Analysis*, 19 April, Seville, Spain.

Gómez, E., Cañadas, F., Salamon, J., Bonada, J., Vera, P. and Cabañas, P. (2012b). Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing. *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 8-12 October, Porto, Portugal.

Gómez, E., Guastavino, C., Gómez, F. and Bonada, J. (2012c). Analyzing Melodic Similarity Judgements in Flamenco a Cappella Singing. *Proceedings of the joint 12th International Conference on Music Perception and Cognition and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music,* July 23-28, Thessaloniki, Greece.

Gómez, F., Pikrakis, A., Mora, J., Díaz-Báñez, J-M. and Gómez, E. (2011). Automatic Detection of Ornamentation in Flamenco Music. *4th International Workshop on Machine Learning and Music*, September 9, Athens, Greece.

Gómez, F., Pikrakis, A., Mora, J., Díaz-Báñez, J-M., Gómez, E., Escobar, F., Oramas, S. and Salamon, J. (2012). Automatic Detection of Melodic Patterns in Flamenco Singing by Analyzing Polyphonic Music Recordings. *2nd International Workshop of Folk Music Analysis,* April 19-20, Sevilla, Spain.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11 (1).

Hsu, C.-L. (2010). On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Trans. on Audio, Speech, and Language Processing*, 18 (2): 310-319.

Johnson, A. and Kempster, G. (2010). Classification of the Classical Male Singing Voice Using Long-Term Average Spectrum. *Journal of Voice*, 25 (5): 538-543.

Kim, Y. E. and Whitman, B. (2002). Singer Identification in Popular Music Recordings Using Voice Coding Features. 3rd *International Society for Music Information Retrieval Conference*, 13-17 October, Paris, France.

Kob, M., Heinrich, N., Herzel, H., Howard, D., Tokuda, I. and Wolfe, J. (2011). Analysing and Understanding the Singing Voice: Recent Progress and Open Questions. *Current Bioinformatics*, 6 (1): 362-374.

Kranenburg, van, P., Volk, A., Wiering, F. and Veltkamp, R. C. (2009). Musical Models for Folk-Song Melody Alignment. *10th International Society for Music Information Retrieval Conference*, pp. 507-512.

Lagrange, M., Ozerov, A. and Vincent, E. (2012). Robust Singer Identification in Polyphonic Music Using Melody Enhancement and Uncertainty-Based Learning. *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 8-12 October, Porto, Portugal.

Marsden, A. (2012). Interrogating Melodic Similarity: A Definite Pheonomenon or the Product of Interpretation? *Journal of New Music Research*, 41(4): 323-335.

Merchán Higuera, F. (2008). Expressive characterization of flamenco singing. *Master Thesis*, Universitat Pomeu Fabra, Barcelona, Spain.

Mesaros, A. and Astola, J. (2005). The mel-frequency cepstral coefficients in the context of singer identification. *Proceedings of the 6th International Conference on Music Information Retrieval*, 11-15 September, London, UK.

Mora, J., Gómez, F., Gómez, E., Escobar-Borrego, F.J., Diaz-Banez, J.M. (2010). Melodic Characterization and Similarity in A Cappella Flamenco Cantes. *11th International Society for Music Information Retrieval Conference ISMIR*.

Molina, E., Barbancho, I., Gómez, E., Barbancho, A. M. and Lorenzo, J. T. (2013). Fundamental Frequency Alignment vs. Note-based Melodic Similarity for Singing Voice Assessment. *Proceedings of the 8th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, Vancouver, Canada.

Nwe, T. and Li, H. (2007). Exploring Vibrato-Motivated Acoustic Features for Singer Identification. *IEEE Transactions on audio, speech and language processing*, 15 (2): 519-530.

Nwe, T. and Li, H. (2008). On Fusion of Timbre-Motivated Features for Singing Voice Detection and Singer Identification. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 30 March – 4 April, Las Vegas, Nevada, USA.

Ó Cinnéide, A. Linear Prediction. The Technique, Its Solution and Application to Speech. *Technical Report*. Dublin Institute of Technology. August 2008.

Orio, N. and Rodá, A. (2009). A Measure of Melodic Similarity Based on a Graph Representation of the Music Structure. *0th International Society for Music Information Retrieval Conference*, pp. 543-548.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO I.S.T. Project Report*.

Paalanen, P. (2004). Bayesian Classification Using Gaussian Mixture Models and EM Estimation: Implementations and Comparisons. *Information Technology Project*. Lappeenranta University of Technology.

Phoren, D.E. (2005). The Art of Flamenco. *Bold Strummer Ltd.,* 43rd edition.

Ramírez, R., Maestre, E. and Serra, X. (2010). Automatic performer identification in commercial monophonic Jazz performances. *Pattern Recognition Letters*, 43: 1514-1523.

Rao, V., Gupta, C. and Rao, O. (2011). Context-aware features for singing voice detection in polyphonic music. *Proceedings of the 9th International Workshop on Adaptive Multimedia Retrieval*, July 18-19, Barcelona, Spain.

Salamon, J., Rocha, B. and Gómez, E. (2012). Musical Genre Classification using Melody Features Extracted from Polyphonic Music Signals. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 25-30 March, Kyoto, Japan.

Salamon, J. and Gómez, E. (2012). Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, 20(6): 1759-1770.

Shen, J., Shepherd, J., Bin, C. and Tan, K.-L. (2009). A Novel Framework for Efficient Automated Singer Identification in Large Music Databases. *ACM Transactions on Information Systems*, 27 (3): Article 18.

Slaney, M. (1993). Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work. Apple Computer Technical Report #45.

Sordo, M., Celma,Ó. and Bogdanov, D. (2011). Audio Tag Classification Using Weighted-Vote Nearest-Neighbor Classification. *MIREX 2011 entry in Audio Tag Classification*.

Sridhar, R. and Geetha, T. V. (2008). Music Information Retrieval of Carnatic Songs Based on Carnatic Music Singer Identification. *Proceedings of the 2008 International Conference on Computer and Electrical Engineering*, 20-22 December, Phuket, Thailand.

Stamatatos, E. and Widmer, G. (2005). Automatic identification of music performers with learning ensembles. *Artificial Intelligence,* 165: 37-56.

Tsai, W.-H. and Lee, H.-C. (2012). Automatic Singer Identification Based on Speech-Derived Models. *Proceedings of the International Journal of Future Computer and Communication*, 1 (2): 94-96.

Tsai, W.-H. and Lin, H.-C. (2010). Popular Singer Identification Based On Cepstrum Transformation. *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo*, 19-23 July, Singapore.

Urbano, J., Lloréns, J., Morato, J. and Sánchez-Cuadrado, S. (2010). *Proceedings of the 7th International Symposium on Computer Music Modeling and Retrieval*, Málaga, Spain, pp. 338-355.

Zhang, T. (2003). System and Method for Automatic Singer Identification. *Proceedings of the IEEE International Conference on Multimedia and Expo*, 6-9 July, Baltimore, MD, USA.

Zhu, Y. and Kankanhalli, M. (2002). Similarity Matching of Continuous Melody Contours for Humming Query of Melody Databases. *Technical Report*. Laboratories for Information Technologies, Singapore.