Localisation, caractérisation et reconnaissance de voix chantées

Lise Regnier

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité

Acoustique, traitement du signal et informatique appliqués à la musique
(Ecole doctorale EDITE)

Présentée par

Mlle Lise Regnier

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Localization, characterization and recognition of singing voices

soutenue le 8 Mars 2012

devant le jury composé de : (préciser la qualité de chacun des membres).

Mr. Xavier Rodet	Directeur de thèse
Mr. Geoffroy Peeters	Encadrant de thèse
Mme. Emilia Gomez	Rapporteur
Mr. Sylvain Marchand	Rapporteur
Mr. Dan Ellis	Examinateur
Mr. Olivier Adam	Examinateur

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à MON : 01 42 34 68 41
Tél. pour les étudiants de MOO à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

Contents

Abstract

This dissertation is concerned with the problem of describing the singing voice within the audio signal of a song. This work is motivated by the fact that the lead vocal is the element that attracts the attention of most listeners. For this reason it is common for music listeners to organize and browse music collections using information related to the singing voice such as the singer name.

Our research concentrates on the three major problems of music information retrieval: the localization of the source to be described (i.e. the recognition of the elements corresponding to the singing voice in the signal of a mixture of instruments), the search of pertinent features to describe the singing voice, and finally the development of pattern recognition methods based on these features to identify the singer.

For this purpose we propose a set of novel features computed on the temporal variations of the fundamental frequency of the sung melody. These features, which aim to describe the vibrato and the portamento, are obtained with the aid of a dedicated model. In practice, these features are computed on the time-varying frequency of partials obtained using the sinusoidal model.

In the first experiment we show that partials corresponding to the singing voice can be accurately differentiated from the partials produced by other instruments using decisions based on the parameters of the vibrato and the portamento. Once the partials emitted by the singer are identified, the segments of the song containing singing can be directly localized. To improve the recognition of the partials emitted by the singer we propose to group partials that are related harmonically. Partials are clustered according to their degree of similarity. This similarity is computed using a set of CASA cues including their temporal frequency variations (i.e. the vibrato and the portamento). The clusters of harmonically related partials corresponding to the singing voice are identified using the vocal vibrato and the portamento parameters. Groups of vocal partials can then be re-synthesized to isolate the voice. The result of the partial grouping can also be used to transcribe the sung melody.

We then propose to go further with these features and study if the vibrato and portamento characteristics can be considered as a part of the singers' signature. Previous works on singer identification describe audio signals using features extracted on the short-term amplitude spectrum. The latter features aim to characterize the timbre of the sound, which, in the case of singing, is related to the vocal tract of the singer. The features we develop in this document capture long-term information related to the intonation of the singer, which is relevant to the style and the technique of the singer. We propose a method to combine these two complementary descriptions of the singing voice to increase the recognition rate of singer identification. In addition we evaluate the robustness of each type of feature against a set of variations. We show the singing voice is a highly variable instrument. To obtain a representative model of a singer's voice it is thus necessary to build models using a large set of examples

covering the full tessitura of a singer. In addition, we show that features extracted directly from the partials are more robust to the presence of an instrumental accompaniment than features derived from the amplitude spectrum.

Chapter 1

Introduction

Many auditors have a remarkable ability to identify the singer of a new song as long as they have already heard some songs performed by the same singer. When the singer is unfamiliar it is common for listeners to attempt to characterize the singer's voice by finding singers with similar vocal characteristics. The most remarkable thing is that these judgments of similarity seem to be consistent across a large number of auditors. This remark suggests that there are some characteristics of the singing voice, perceived by most people, which clearly define the identity and the type of voice of singers. Humans also have an incredible ability to distinguish different instruments playing simultaneously. Everyone is capable of detecting when a singing voice is present in a mixture with a high rate of precision. The capacity to separate the different sources of a mixture may be at the basis of this capacity of music listeners to recognize singers independently of the band with which they perform. We note that it is far more complex, even for experienced music listeners, to recognize a given instrumentalist (e.g. a guitarist, a bassist, a drummer, etc.) through different music bands. One of the elements that make the recognition of a singer in familiar and new contexts easier is the fact that the instrument and the performer are not separable in the case of the singing voice.

The purpose of our research is to study the characteristics of the singing voice that make the voice differ from other musical instruments and to study the characteristics of the singing voice that can be used to define the signature of a singer. The motivation of this research is to develop systems able to automatically detect the presence of a singing voice in the signal of a mixture of instruments and to automatically identify the singer of a given song.

In most studies conducted on the detection of the singing voice and on the recognition of singer, the singing voice is described by means of features extracted from the amplitude spectrum and its envelope. The envelope of the (short-term) amplitude spectrum conveys information related to the timbre. As explained by the source-filter model, the spectral envelope estimated on signals of speech and singing gives an estimation of the transfer function of the vocal tract, which filters the sounds created by the vibration of the vocal folds and is assumed to be responsible for the vowel quality and the overall color of the sound produced. Audio features derived from the spectral envelope, such as the Mel Frequency Cepstral Coefficients have been successfully used in tasks of speaker recognition. It is thus reasonable to assume that the same features can be used to describe the "instrument" of the singer; its vocal tract. This statement remains valid as long as the "timbral" features are extracted from the signals of sounds produced by a unique vocal source.

When these features are extracted from the signal of a song (i.e. a singing voice accompanied by other instruments) they describe the overall spectral characteristics of the whole mixture and it is almost impossible to derive information directly related to the singing voice from this analysis. However, if one wants to build an accurate singer identification system, it is necessary to work with information that describes only the singing voice. Otherwise the system is trained to recognize the artists rather than the singer themselves. In other words, such a system cannot recognize singers independently of the band with which they perform.

In this document we propose a novel approach to describe the singing voice, which extracts features directly related to the singing voice from the signal of a mixture. This approach is based on the interpretation of musical signals given by the harmonic sinusoidal model. This model describes the harmonic sounds as a sum of sine waves evolving slowly over time. Ideally, each sine wave, or partial, corresponds to one harmonic of one tone played by a single instrument and is described by a support (onset/offset), a time-varying frequency, a time-varying amplitude and a phase. This signal representation allows a separation of the spectral components played by the different instruments of a mixture. In this dissertation, we propose a method to distinguish partials corresponding to the singing voice among partials of other instruments using two characteristics of the singing voice: the vocal vibrato, which occurs naturally on sustained sung tone and the portamento, which occurs when two successive notes with distinct pitches are sung without interruption. These elements are known to be features of the singing voice that add richness and expression to the musical content. In our research we propose to model these temporal variations of frequency and use the parameters of the vibrato and portamento as new features to describe the singing voice. We show though different experiments that these features capture information related to the style and the technique of the singer and can thus be used to describe a part of the signature of a singer in tasks of singer identification. This novel approach to describe the singing voice provides information complementary to the information related to the timbre of the voice. Finally, we propose a method to combine this complementary information to improve the recognition rate of singer identification.

1 Structure of the document

The document is organized as follows:

Chapter 2 - **Reminder of classification techniques involving combination of information**. Most of the experiments conducted in this dissertation rely on supervised machine learning techniques. To lighten the description of our experiments we present the theoretical backgrounds and the machine learning algorithms that will be used in the rest of this document in a separated chapter. Beside the presentation of the classifiers, this chapter discusses the following points of supervised classification: the selection of the best set of features for a given problem, the choice of appropriate measures to report classification performances, and the use of information combination to improve classification performances.

Chapter 3 - **Singing voice: production, models and features**. The goal of this chapter is to introduce the features used in this document to describe the singing signals. This chapter starts with a general presentation of the mechanisms involved in the production of sung sounds and presents some specific aspects of singing, such as the vocal vibrato and the portamento. Then, the two major models

for sound representation, the source-filter model and the sinusoidal model, are presented. We propose to take advantage of these two models to derive two complementary types of features to describe the singing voice. The source-filter model, which describes the vocal production as an excitation produced by the vibration of the vocal folds filtered by the vocal tract, is used to obtain information related to the timbre of the singer. The sinusoidal model, which interprets the harmonic sounds as a sum of partials whose frequency and amplitude vary slowly over time, is used to extract information related to the intonation of the singer that is assumed to be linked with the style and the technique of the singer. The latter features are obtained with the aid of a dedicated model described and evaluated in this chapter.

Chapter 4 - **Singing voice localization and tracking in polyphonic context**. This chapter presents a set of methods to distinguish the partials corresponding to the singing voice among the other partials of a polyphonic mixture. First we present a method to directly identify the partials emitted by the singing voice. This method differentiates the vocal partials from the partials emitted by the other instruments using a set of decisions made by evaluating the characteristics of the vocal vibrato and portamento. The result of this simple method is used to localize the segments of a song containing a singing voice. We propose to improve the recognition rate of the vocal partials by exploiting the harmonic nature of the singing voice. In this chapter we present a method to group the partials emitted at the same instant by the same instrument based on some CASA cues (common onsets, harmonicity and common variations of the frequency based on the parameters of the vibrato and portamento). The groups of harmonically related partials corresponding to the singing voice are then identified using the vibrato and portamento criteria. Once the groups of vocal partials are identified they can be used to isolate the voice from the instrumental accompaniment and/or to transcribe the melody performed by the singer.

Chapter 5 - **Singer identification** In this chapter, we propose to evaluate the capacity of "timbral" and "intonative" features to capture information related to the signature of a singer. This evaluation is conducted through a series of singer identification experiments performed on lyric and pop-rock singers. The underlying problem of singer identification is to find an invariant "voiceprint" in the signals of song of a given singer, characterizing the voice of said singer. This task is challenging because the voice is a highly variable instrument that can produce an extremely large variety of sounds. We propose in this chapter to evaluate if the features related to the timbre and the intonative features are robust against the variations of pitch and loudness. Since the voice is usually accompanied by other instruments we also evaluate the capacity of these features to capture information related to the singing voice when they are extracted directly from a mixture. Finally, we present a method to combine the information conveyed by timbral and intonative features to increase the recognition rate of singer identification.

Chapter 6 - **Conclusion**. The last chapter discusses the results of this dissertation, draws some conclusions and suggests some directions for future works.

2 Outlines

The interest of the work presented in this document relies mainly on the development and the utilization of novel features describing some intonative and/or expressive attributes of the singing voice.

1. The model developed to extract the intonative features is presented in Chap.3: Sec.2.3. This model describes the time-varying frequency of a partial as the sum of a continuous variation, representing the portamento, and a quasi-sinusoidal modulation, representing the vibrato. One specificity of the vocal vibrato, the presence of an amplitude modulation which occurs passively in presence of a frequency modulation is presented in Chap.3: Sec.2.4. The characteristics of the temporal variations of amplitude and frequency constitute the set of the intonative features as explained in Chap.3: Sec.3.2.

2. The proposed method to localize the vocal segments of a song using the direct recognition of partials emitted by the singing voice is presented in Chap.4:Sec.3.

3. In Chap.4:Sec.4.1 we propose to improve the recognition of vocal partials by grouping the partials emitted at the same instant by the same instrument. The proposed approach is based on the definition of a similarity between partials, which is computed using the parameters of the vibrato and portamento. This method is presented as a novel approach for the extraction of the singing voice.

4. Finally, the proposed method developed to exploit the synergy offered by the intonative and timbral feature for the singer identification task is presented in Chap.5: Sec.3.1.2. This method uses the intonative features to refine the decisions made on the timbral features to increase the recognition rate of singer identification.

Chapter 2

Reminder of classification techniques involving combination of information

The past two decades have been characterized by the development of the internet, resulting in rapidly growing collections of data, including musical data. It became evident that it was necessary to develop systems able to classify music collections automatically. The research on the automatic organization of music collection is often referred to as Music Information Retrieval (MIR). MIR research has two main goals: to develop systems able to transcribe the score of a given song (e.g. transcribe the melody, the chords, the rhythm, etc.) and generate playlists of music automatically (e.g. music recommendations based on user specifications: music genre, music mood, tempo, decade of production, etc.).

In Pop-rock music most of the musical production is made of songs. A song can be defined as a lead vocal, carrying the main melody and the lyrics, accompanied by a set of musical instruments. The vocal lead and by extension the singer's voice is the element that attracts the attention of many listeners. Naturally, it has been proposed to organize, browse, and retrieve popular music recordings using information related to the singer's voice (e.g. the singer name, the type of voice, the sex of the singer, the lyrics of the song, etc.). However, the quasi-systematic presence of an instrumental accompaniment makes the extraction of information relative to the singer's voice rather difficult.

Any system developed to extract information on the sung melody or on the identity of the singer requires locating the portions of the song where the singing voice is present. The problem of singing voice localization (Chap.4) is generally viewed as a problem of classification with two classes: the vocal and the purely instrumental classes. The problem of singer identification (Chap.5) is also regarded as a problem of classification where each class represents one singer. In this document we deal with these two problems using supervised machine learning techniques.

The goal of this chapter is to introduce the theoretical backgrounds and the different methods that will be used in the different experimentations of this document. The details of the steps involved in supervised machine learning techniques are described in Sec.1. In the present work we attempt to describe the characteristics of singing voice using different types of information. To take into consideration the maximum aspects of the singing voice, it is necessary to combine this information at a stage of the class training process. We present in Sec.2 different techniques to combine information for classification problems.

1 Supervised pattern recognition

The goal of classification is to assign unlabeled patterns to a number of known categories. A system of classification is based on both an appropriate description of the patterns and on a statistical algorithm trained to learn the specificities of each pattern for the given problem. A system of supervised classification can be summarized as shown in Fig.2.1.

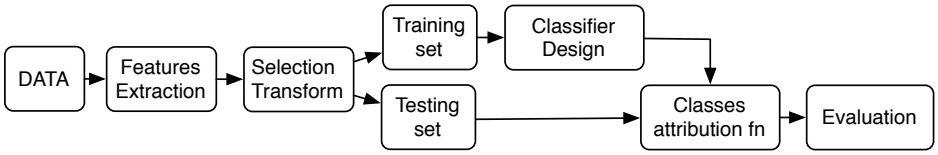


Figure 2.1: Basic steps of supervised classification

The problem starts with a set of data and a set of known classes. We assume in the following that each item of the data set belongs to one and only one class. The data set is described by a set of features, which put the important data points for the problem into a concise form. Features can be selected and transformed to increase the performance of the classification. This point is discussed in Sec.1.1. The set of data is split into training examples and testing samples [1]. A supervised learning algorithm analyzes the training examples and produces an inferred function, which is called a classifier. Different types of classifiers are presented in Sec.1.2. Once trained, the classifier is then used to predict the class of the examples of the testing set. In practice, the true classes of the testing patterns are known. The performance of the system of classification (features + classifier) is measured by comparing the predicted class of the testing examples with their true classes. Different measures to evaluate and compare the performance of classification systems are presented in Sec.1.3. The same classification problem can be solved with an infinite number of systems of classification (either by changing the features or the classifier). Each system has its own strengths and weaknesses leading to a bounded performance. In Sec.2 we present some methods to combine different systems of classification. The aim of these methods is to increase the classification performance by combining the strengths of different systems of classification without accumulating their weaknesses.

1.1 Features transformation and selection

We assume in the following that the features have already been extracted from the original data. Then the data is represented by a feature space, which is a compact representation of the salient aspects of the data for the given problem. Because of the curse of dimensionality it is often desirable to reduce the size of features in the feature space. In addition, it is necessary to delete non relevant features because they can deteriorate the classification performance significantly. To increase the performance of classification systems numerous methods have been developed to find the minimal set of informative and non-redundant features.

Feature selection methods are grouped into three categories [MBN02]. *Filter* methods select

[1] It is also possible to have a third part referred to as development set on which internal parameters of the classifiers can be optimized. In this presentation we assume that this optimization is realized with a cross-validation on the training set of samples.

features independently of the classifier using statistical tests or criteria of class separability. *Embedded* methods perform feature selection during the training process. They are specific to some classifiers. *Wrapped* methods select the best subset of features for a given classification task by minimizing the prediction error on the training data set. The subset of features returned by these methods is specific to a given classifier. Filter methods, applied as a pre-processing step, have the advantage of being very fast. Wrapped methods can be computationally expensive, however they offer the advantage of measuring the relevance of each feature (and each combination of features) for the given problem. These measures can be used backwards to ascertain the most discriminative aspects of the pattern studied, which can be very useful if the original features have a clear physical or perceptual meaning. A detailed overview of feature selection methods is presented in Fig.2.2 extracted from [SIL07].

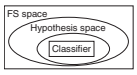
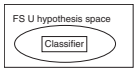
Model search	Advantages	Disadvantages	Examples
Filter	Univariate		
	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	χ^2 Euclidean distance *i*-test Information gain, Gain ratio (Ben-Bassat, 1982)
	Multivariate		
	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation-based feature selection (CFS) (Hall, 1999) Markov blanket filter (MBF) (Koller and Sahami, 1996) Fast correlation-based feature selection (FCBF) (Yu and Liu, 2004)
Wrapper	Deterministic		
	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) (Kittler, 1978) Sequential backward elimination (SBE) (Kittler, 1978) Plus q take-away r (Ferri *et al.*, 1994) Beam search (Siedelecky and Sklansky, 1988)
	Randomized		
	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing (Skalak, 1994) Genetic algorithms (Holland, 1975) Estimation of distribution algorithms (Inza *et al.*, 2000)
Embedded	Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes (Duda *et al.*, 2001) Feature selection using the weight vector of SVM (Guyon *et al.*, 2002; Weston *et al.*, 2003)
			

Figure 2.2: A taxonomy of feature selection methods from [SIL07]

The first idea to reduce the size of a feature space is to eliminate the redundancies that can exist between the original features. Methods like the Principal Component Analysis [Jol02] project the

original features space into a space in which the features are not correlated. The uncorrelated features are ordered in decreasing order of variance and the features with the highest variance are selected to form a new feature space. This type of method compresses the information of the original features space into a smaller set of features. Nothing ensures that the new feature space is more adapted for the classification.

To optimize the classification performance, features can be selected according to their capacity to separate the classes of the problem. This idea is exploited in the Linear Discriminant Analysis [DHS01] that computes the linear transformation that maximizes the ratio of the between-class to the within-class variance. The LDA returns a feature space whose dimension cannot be higher than the number of classes minus one. Thus, it cannot be applied when the number of informative features is higher than the number of classes. The IRMSFP (Interna Ratio Maximization using Feature Space Projection) algorithm [Pee03] proposes another method to compute a criterion of class separability for each feature of the original feature space. Then the features that best separate the classes are selected by fixing a threshold on the values of the separability criterion. The LDA and IRMSFP algorithms offer the advantage of preserving the original values of the features selected, which can be very useful for the interpretation of the model. We note that LDA and IRMFSP return the k best features of the feature space. Nothing ensures that the k best individual features form the best subset of k features for the problem.

As far as we know, the only way to find the best subspace of features is to use an embedded feature selection method. To find the best combination of k features of an original space of N features there are C_N^k cases to evaluate, which is in most cases not feasible. Numerous alternatives have been proposed to accelerate the search. According to [JZ02], the most effective feature selection technique is the sequential floating search methods. There are two main approaches of floating search methods: the Sequential Forward Search (SFFS), which starts with an empty set and at each step one feature is added to the current feature set as well as the Sequential Backward Search (SFBS), which starts with the full set of features, and discards one feature at each step. At each step, features are selected by minimizing a criterion function (e.g. the classification error rate). SFFS and SFBS algorithms proceed without backtracking: when one feature is added (or discarded) it cannot be removed, (or re-added) even if at a highest level in the search this feature would minimize the cost function. A possible improvement is to add and/or discard more than one feature at each step as done by the "plus-l-take-away-r" algorithm. In this method, l features are added one at a time and r features are removed from the expanded set. When $l > r$, the algorithm is a forward search, otherwise it is a backward search. The results obtained with this last approach are closer to the optimal solution than the results of the SFFS or SFBS algorithms. In the worst case they return the same subset. A complete description of these methods is given in [SPN$^+$99]. These embedded methods can be applied on the output of any classifier.

1.2 Classifiers architecture

There are a very large collection of classifiers. In this section we present three classifiers (GMM, SVM and kNN) that will be used in the experiments in the rest of this document. These classifiers are representative of the three main approaches for supervised classification: (a) the generative approach that learns the properties of each class, (b) the discriminative approach that learns the boundaries

between the classes and (c) the instance-based approach that uses the training examples as references. We present in table 2.1 an example of these three categories of classifiers accompanied by a brief description. A complete description of these classifiers can be found in [DHS01] or [Bis06].

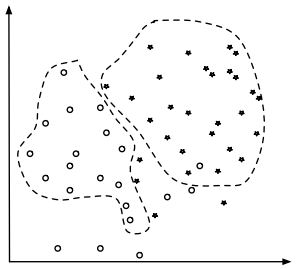
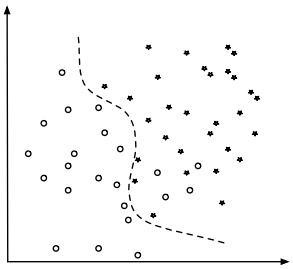
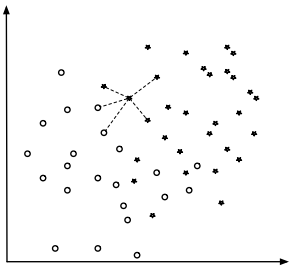
Type	Generative	Discriminative	Instance-based
General Idea			
Principles	Build one model per class	Learn the boundaries between the classes	Compare item
Example	Gaussian Mixture Model **GMM**	Support Vector Machine **SVM**	k-Nearest Neighbors **kNN**
Classif.	Likelihood for each class	Distances to the margins	Distance to the closest neighbors
Output	Posterior probability	Distance	Distance

Table 2.1: Examples of (a) generative (b) discriminative and (c) instance-based approaches for supervised classification

1.2.1 Generative approach: GMM

Gaussian Mixture Models have been widely used in audio classification tasks. They have been proven to be particularly efficient for the problem of speaker recognition [C$^+$97] and instruments identification [BHM01]. In the GMM classifier, the conditional probability density function (pdf) of the feature vector with respect to the different classes is modeled as a linear combination of multivariate Gaussian distributions. Each distribution p_i models a specific region of the feature space (or a different cluster). This is why GMMs are particularly suited to model data whose distribution cannot be modeled by a single cluster. The GMM for the class ω_i and a D dimensional feature vector x, can be written as:

$$p(x|\omega_i) = \sum_{i=1}^{M} c_i p_i(x) \tag{2.1}$$

where M is the number of components and c_i are non-negative weights that sum up to unity ($\sum_{i=1}^{M} c_i = 1$). In the case of a mixture of gaussians, each $p_i(x)$ is a multi-dimensional gaussian defined by a mean vector μ_i of length D and a D-by-D covariance matrix Σ_i as shown in Eq.(2.2):

$$p_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{2.2}$$

Thus, $p(x|\omega_i)$ is entirely defined with the parameters: (c_i, μ_i, Σ_i) for $i = 1 \dots M$. These parameters can be learned on a large set of training examples belonging to ω_i with the Expectation-Maximization algorithm [DLR$^+$77]. This algorithm iteratively estimates the GMM parameters to increase monotonically the likelihood of the estimated model for the observed feature vectors (i.e. to maximize $p(x|\omega_i)$). The covariance matrix Σ_i can be chosen to be full, diagonal or also spherical. In practice, the full covariance matrix leads to better results, but is constrained by a conditioning problem encountered by the matrix inversion of Eq.(2.2). Full covariance matrix requires a large amount of data to train efficiently. Using a diagonal covariance matrix clearly simplifies the matrix inversion problem. The precision of a density modeling obtained with an M-component GMM with a full covariance matrix can be equaled by a GMM with a diagonal covariance matrix with a higher order (M' > M).

An unknown pattern, represented by a feature vector x, is assigned to the class that maximizes $p(\omega_i|x)$. We have:

$$\hat{w} = \arg\max_{\omega_i \in \Omega} p(\omega_i|x) \tag{2.3}$$

If the classes have the same prior probability, using Bayes rule we have:

$$\hat{w} = \arg\max_{\omega_i \in \Omega} \frac{p(x|\omega_i)P(\omega_i)}{p(x)} \tag{2.4}$$

$$\hat{w} = \arg\max_{\omega_i \in \Omega} p(x|\omega_i) \tag{2.5}$$

Finally, an unknown pattern is assigned to the class that maximizes the likelihood $p(x|\omega_i)$. With GMM, we can compute a (pseudo) posterior probability for each class of the problem. We will discuss the interest of this point in the section dedicated to the fusion of classification.

- Advantages of GMM:
  - GMM is based on a well understood statistical model
  - They also present the advantage of being computationally inexpensive
- Disadvantages of GMM:
  - The tuning of GMMs is not straightforward: the number of mixtures, the training method and the choice of the covariance matrix type can be optimized using a cross-validation process on the training data set (but it increases the risk considerably to obtain over-fitted models).

1.2.2 Discriminative approach: SVM

The SVM classifier has also been widely used in MIR related classification tasks. In general, the SVM classifier is chosen to solve classification problems where the boundaries between the classes are complex. The goal of SVM is to find the best separator to discriminate two classes as shown in Tab.2.1.

In its simplest version, SVM uses a linear separating hyperplane to create a classifier between two classes linearly separable. If the data is not linearly separable the SVM classifier maps the original feature space into a higher dimensional space where a linear separating hyperplane can be found.

The mapping is done by the mean of a kernel function. In the linear case, the solution is given in terms of inner products. In the non-linear case, there is no need to compute the explicit mapping. The only requirement is to find a kernel function that is an inner product in the mapped space. This requirement can be tested with the condition of Mercer. The radial basis function (RBF) is commonly used as kernel. As shown in [HCL$^+$03], the RBF performs well in a wide variety of circumstances. A complete description of the most common kernel can be found in [HCL$^+$03].

In the definition of the SVM, the hyperplane is considered optimal if it maximizes the margin, which is given by the sum of the minimum distances between the decision boundary and the instances of each class. These instances, known as the *support vectors*, are given by the samples that are the closest to the boundary. The maximization of the margin can be formulated as a convex optimization problem and can be solved using standard methods. However, in many circumstances the classes are not perfectly separable and some samples of the training set lie on the "wrong" side of the boundary. In this case the best separation is found by minimizing a cost function that takes into consideration the margin and the classification error rate on the training data.

Though the basic formulation of SVM uses two classes, SVM can be generalized to problems involving multiple classes by finding the mean of a set of binary classifiers [Bur98]. The "one-versus-one" method trains a classifier for each pair of classes resulting in N(N-1)/2 binary classifiers for a problem with N classes. Then, the classification is done by max-wins voting strategy. Each binary classifier assigns the query sample to one of its two possible classes and its decision is considered as a vote for one class. In the end, the query sample is assigned to the class with the maximum number of votes. The other common method for multi-class SVM is the "one-versus-all" approach. In this case N classifiers are trained to discriminate one class against all other classes. During the test, the query sample is assigned to the class that has obtained the maximum output value within the set of binary classifications. The "one-versus-one" method is computationally more expensive but its performance is significantly better [SG96].

These versions of the multi-classes SVM provides a "hard" label. To compute pseudo posterior probability for each class of the problem the methods proposed, [WLW04] and [GTA08] can be applied.

- Advantages of SVM:
  - SVM indicates which examples are particularly important to classification through their support vectors.
  - SVM have high capacity and can typically generalize well to new data.
  - SVM seems to be less sensitive to other classification methods to the choice of the feature space (selection and transformation of the feature set).
- Disadvantages of SVM
  - Although SVMs have good generalization performance, they can be abysmally slow in the test phase.
  - SVM has a high algorithmic complexity and requires an extensive memory.
  - Perhaps the biggest limitation of the support vector approach lies in the search of an appropriate kernel function for a particular application and particular set of features.

1.2.3 Instance-based approach: k-NN

The k-nearest neighbor algorithm is the simplest example of an instance-based classification algorithm. It is a simple, intuitive and elegant algorithm that belongs to the category of non-parametric classifiers. Contrary to generative and discriminative classifiers, instance-based classifiers do not try to make any generalization on the training data. Instead, training examples are considered prototypes of the classes to which they belong. An unlabeled example is assigned to a class by measuring the similarity of these examples to the prototypes. Typically, the similarity is defined by the mean of a distance computed in the feature space. For this reason, the most similar prototype is called the closest neighbor.

In it simplest version, 1-NN, a query example is labeled with the class of its closest neighbor. For a number of neighbors $k > 1$, the k nearest neighbors are retrieved from the training samples and the query sample is assigned to the class that is the most represented among the k labels. In the weighed version of k-NN the distances between the query example and its k closest neighbors are used as weights to determine the closest set of prototypes. In all cases, k-NN returns a unique class label for the tested examples. As previously mentioned, in many circumstances it is necessary to have an estimation of the posterior probability for each class of the problem. These measurements can be obtained with the probabilistic version of the k-NN.

In the probabilistic version of k-NN, we suppose that the distance between examples is given by a Euclidean distance. We denote by R the hypersphere of volume V_R containing exactly k training examples. The unconditional probability density function (pdf) of a query sample represented by x can be approximated by:

$$p(x) \approx \frac{k}{NV_R} \tag{2.6}$$

where N is the number of samples in the training set. If k_i denotes the number of samples in R that belong to the class ω_i, and N_i the number of elements from class ω_i in the training set, then the class-conditional pdf in R for the class ω_i is given by:

$$p(x|\omega_i) \approx \frac{k_i}{N_i V_R} \tag{2.7}$$

If the prior probability of each class is given by the percentage of elements from class ω_i in Ω ($P(\omega_i) = \frac{N_i}{N}$), the posterior probabilities can be computed, using the Bayes rule, as follows:

$$p(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)} = \frac{k_i}{k} \tag{2.8}$$

In this probabilistic version of the k-NN, the region R and the volume V_R depends on x and on the training set. However, the k-NN classification rule, which assigns the class label using the labels of the neighbors, does not depend on volumeV_R. The k-NN is optimal (i.e. minimizes the Bayes error) when $N \to \infty$ and $V_R \to 0$ which is equivalent to $k \to \infty$ and $k/N \to 0$.

Different metrics define different regions hypersphere R. But for any metric, the k-NN rule is applied in the same way and the choice of the metric does not affect the property of the k-NN classifier because the shape of R is not fixed.

- Advantages of k-NN

  - The main advantage of the k-NN algorithm is it simplicity of interpretation and implementation
  - kNN is well adapted to problems where the boundary between the classes

- Disadvantages of k-NN

  - The results of k-NN highly dependent on the choice of the distance. Some distances are highly influenced by the values of the features. So that, it is necessary to adapt the feature space to the chosen distance.
  - The computational cost grows with the dimension of the feature space and the size of the training set. Some techniques, like feature vector pruning, can be applied to limit storage and computation.
  - This algorithm is greatly sensitive to the local presence of noise in the training-data.
  - The choice of the number of neighbor k is not straightforward. Choosing a large k reduces the effect of noise on the classification but makes the boundary between the classes less distinct. A good value of k for a given problem can be selected using a cross-validation evaluation on the training set. It is generally recommended to chose an odd k such as $k < \sqrt{N}$.

1.3 Performance of classification

The goal of any classification problem is to retrieve the class of each pattern composing the testing-set. The performance of a system of classification is always measured by comparing the predicted classes with the real classes of the testing samples. However, multiple points can be taken into consideration when evaluating the performance of a classification. If the problem is presented as an identification problem, i.e. the test samples have to be assigned to one of the N possible classes and the N classes are of equal importance, then the measure of performance should be computed as a combination of the performances for each class. If the problem is presented as a detection problem, then the measure of performance should focus on the performance of the class to be detected. The detection case can always be considered a binary classification.

Here we present some of these alternatives to measure the performance of the recognition problem that will be used in the rest of this document.

For a problem with two classes (Positive and Negative) the result of a classification process can be stored in a confusion matrix as shown on Fig.2.3.

		ESTIMATION	
		Positive	Negative
REF	Positive	*tp*	*fp*
	Negative	*fn*	*tn*

Figure 2.3: Confusion matrix for two classes

1.3.1 Measure of performance

Accuracy The most intuitive measure to evaluate the performance of a classification is the measure of *accuracy*, which is given by the percentage of sample of the testing set assigned to their true class. For a problem with N samples, we denote by N_{corr} the number of sample correctly classified (i.e the sum of the element on the diagonal of the confusion matrix), the accuracy is given by:

$$acc = \frac{N_{corr}}{N} \tag{2.9}$$

For a problem with two classes, Eq.(2.9) can be written as:

$$acc = \frac{tp + tn}{tp + fp + fn + tn} \tag{2.10}$$

We note that the accuracy does not take into consideration the distribution of the classes of the problem. Then, this measure gives a good indication of the performance of a classification if the classes of the problem occur with the same probability.

Recall, Precision and F-measure When the classes of the testing-set are not equilibrated, the classification performance can be assed by measuring the recall, the precision and/or the F-measure. These measures are computed independently for each class as follows. For a given class, letÕs consider the positive class, the *recall* indicates the percentage of the samples of that class that are successfully retrieved:

$$recall = \frac{tp}{tp + fn} \tag{2.11}$$

The *precision* indicates the percentage of the samples assigned to the given class that really belong to the class:

$$prec = \frac{tp}{tp + fp} \tag{2.12}$$

Finally, the *F-measure* combines the recall and the precision to give an overall measure of the classification performance:

$$F_{measure} = \frac{2.recall.prec}{recall + prec} = \frac{2.tp}{2.tp + fn + fp} \tag{2.13}$$

For a problem involving two classes, the $recall$, $prec$ and $F_{measure}$ of the "negative" class can be deduced from the $recall$ and $prec$ of the "positive" class if, and only if, the distribution of the classes is known.

In practice, none of these measure should ever be given without the prior probability of the class studied. On Fig.2.4 we plot the $F_{measure}$ associated with the positive class of pseudo random classifiers in function of the proportion of the positive class. The pseudo random classifiers correspond to classifiers that randomly generate binary classifications with q% of samples belonging to the positive class. As shown in this plot, the $F_{measure}$ does not vary linearly with the proportion p of the positive class. If the classes are well balanced ($p = 0.5$), a random classifier that assigns half of the samples to the positive class ($q = 0.5$) has a $F_{measure}$ equal to 0.5, which is what everyone expects. The same classifier ($q = 0.5$) applied on a data set where $p = 0.8$ obtains a $F_{measure}$ equal to 0.63. For a data set where the positive class is over-represented ($p > 0.5$) a dummy classifier that assigns most elements to the positive class ($q > 0.5$) will always obtain a very high $F_{measure}$. For instance, on a data set

where $p = 0.8$, a classifier that assigns all samples to the positive class ($q = 1$) obtained a $F_{measure}$ of 90%. The same classifier on the same data set will have a very low $F_{measure}$ for the negative class. However, the mean of the $F_{measure}$ of both class is not equal to 0.5 because the $F_{measure}$ is not linear.

From our experiments, the best measure to report the overall performance of a classifier on a problem with multiple classes that is not well balanced is the mean of the recalls of each class. This measure is definitely the best measure to take the recognition of rare events into account. Intuitively, if the recall of each class is weighted by the proportion of the class then the mean of the weighted recalls is equal to the accuracy presented in the previous paragraph.

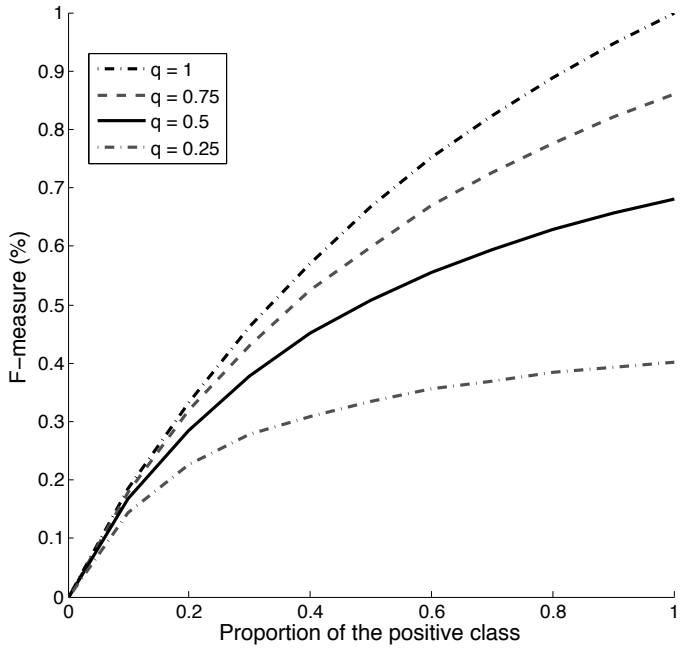


Figure 2.4: $F_{measure}$ in function of the proportion of the positive class

PR-curves For the case of binary classification, if all samples are assigned to the "positive" class, then the recall for this class is maximized and the precision is equal to the proportion of samples belonging to this class. On the other hand, if a very few samples are assigned to this class the precision will be maximized at the expense of the recall. If the decision is based on a threshold, it is possible to draw on a graph a point p = (recall, precision) for each threshold tested and then to plot the curve passing through all the operating points. The curves obtained are usually named *Recall-Precision curve* (PR-curve). The best configuration for the classifier (i.e. the best threshold) is given by the point that is closer to the top-right corner of the figure.

The PR-curves can be plotted for the multi-classes problems using the mean recall and the mean precision through classes.

ROC curves A similar idea, applied when two classes are well balanced, is used in the *Receiver-Operator Curve* (ROC). A full explanation on the ROC curves can be found in [Faw04]. This representation shows how the number of samples of the positive class correctly classified (tp) varies with

the number of miss-classified samples from the negative class (fp). In practice, the fp rate, called *specificity* in ROC terminology, is plotted on the abscissa and the tp rate, called *sensibility*, is plotted on the ordinate. The ROC curve of a random classifier is the diagonal y = x if the two classes are well balanced. ROC curves are very convenient to compare the performances of different classifiers (or different configurations of one classifier). The classifier whose curve is closer to the top-left corner of the plot has a better performance. The relation between the PR and ROC curves is fully explained in [DG06].

In many situations, it is preferable to report the performance of a classification using a single value. It is common to characterize the performance of a classifier using the area under the curve (AUC) as explained in [Bra97]. No realistic classifier should have an AUC less than 0.5.

DET curves The ROC and PR curves introduced above are a mean of representing and comparing the performance of classifiers for tasks of identification. In recognition problems there is also another task closely related to the identification task, referred to as *verification*. In a verification problem, each test samples come with a “claimed" class and the task is to verify if the sample really belong to the class or not. The verification paradigm, widely used in problems of speech recognition, is fully detailed in [vLMPB06].

In verification tasks, two errors are considered: *False acceptation FA* which correspond to a False Positive (fp), i.e accept the class of the test sample when the claimed class is not correct and *Miss detection MD* i.e reject a valid class (fn).

In general, to evaluate the performance of a verification system, a certain number of trials are given as inputs to the system. The system has to distinguish between true-trials (the sample is from the claimed class) and false-trial (the sample is not from the claimed class). The acceptation or rejection of the claimed class is based on a threshold τ. For a given threshold τ, the numbers (or rates) of FA and MD are coupled to form an operating point p = (FA, MD). By varying τ, we obtain a set of points that can be plotted to draw a Detection Error Tradeoff (DET) curve. DET curves are plotted using a non-linear axis which spreads out the plot which improves the readability (compared to ROC curves).

To compare performances of verification systems, a DET curve for each system is plotted and the curve that is closer to the bottom-left corner of the plot corresponds to the best system. In general, to report the performance of any task, a single value is preferred. The performance of verification systems is expressed in terms of Minimum Detection Cost (MDC) or in terms of Equal Error Rate (EER). This last term corresponds to the threshold τ that obtains an equal number of MD and FA. The Minimum Detection Cost is found by minimizing a detection cost function specified in terms of cost of misses and the cost of false alarms as well as the prior probability for the target hypothesis. Unlike ERR, the minimum detection cost depends on the particular application-dependent parameters of the cost function.

1.4 Comparison of classification performances

For each system of classification there is one configuration (feature selection and transformation + choice of classifier parameters) that yields the best performance. Depending on the nature of the classifier and on the size of the feature space, it may be computationally expensive to find the optimal configuration. Many factors can influence the performance including the choice of training and testing

set, the internal randomness of the training algorithm and random classifier errors [Die98]. We also remark that it is not possible to decorrelate the contribution of the feature space from the contribution of the classifier in any measure of performance presented above.

Before comparing the elements (features + classifier) of two systems of classification based on the same training and testing data sets, one may be sure that a pseudo optimal configuration has been reached. As mentioned in Sec. 1.2, any classifier has some internal parameters that can be optimized using a cross-validation on the training set of data and the feature space can be optimized using the methods presented in 1.1 .

Once these pseudo optimal configurations have been found, the classification systems can be compared using any of the curves presented in the previous paragraph (PR, DET or ROC) depending on the problem stated and the application (choice of a cost function). When several classifications lead to equivalent accuracies, tests such as McNemar (for 2 classifiers) and Cochran's Q (for $K > 2$ classifiers) can be applied to test if the difference between their accuracies is significant [KHDM02] .

As mentioned in this section, it is possible to solve the same problem of classification with an infinite number of approaches. Each system has its own strengths and weakness that vary with the problem to be solved. In addition, ensuring the superiority of one approach over others is not straightforward. In many cases, combining the results of different systems of classification leads to better performance than using a single classification. In the next section we introduce some of the most common strategies developed to combine different information relative to the same problem.

2 Combination of information for classification problems

The ultimate goal in classification combination is to combine the strengths of the different approaches without combining their weaknesses. The classification combination problem (also referred to as fusion of information) has received considerable attention in recent decades and is now considered a new direction for the development of highly reliable recognition systems.

There are an infinite number of ways to combine information for a given classification task. First, the combination of information can be performed at every stage of the classification process. The different levels of fusion are presented in Sec.2.1. Howeve, in many situations, it is only possible to combine the decisions of the classifiers. The two main schemes of decision combination (**parallel** and **sequential** combination) are presented in Sec. 2.2. Most combination methods are based on predefined rules but it is also possible to learn the combination using the outputs of classifiers as new features. The difference between trainable and non-trainable combiners is presented in 2.3.

Notations

The problem of classification combination generally involves multiple descriptions of the same pattern and multiple classifiers used to learn the specificities of the class for the given problem. In this overview of methods, for the combination of information for classification, we use the following notations:

- Each pattern z is assigned to one of the N possible **classes**: $\Omega = \{\omega_1 \dots \omega_N\}$.

- Each pattern can be described using different sets of features, the set of all available **descriptions** D_i is denoted by $\mathscr{D} = \{D_1 \dots D_L\}$.
- The set of **classifiers** is denoted by $\mathscr{C} = \{C_1 \dots C_M\}$.
- A **system of classification** is composed of one description and one classifier: $S^{(k)} = \left(D^{(k)}, C^{(k)}\right)$ where $D^{(k)} \in \mathscr{D}$ and $C^{(k)} \in \mathscr{C}$.

2.1 Levels of combination

To improve the classification accuracy using combining of information it is necessary to have complementary information on the same problem. For a given classification problem different information can be obtained by choosing different data sets, different descriptions of the data, or different classifiers. The combination can thus be performed at different stages of the classification process. Fig.**??** summarizes the different levels of combination. Each level of combination is further described in the following paragraphs.

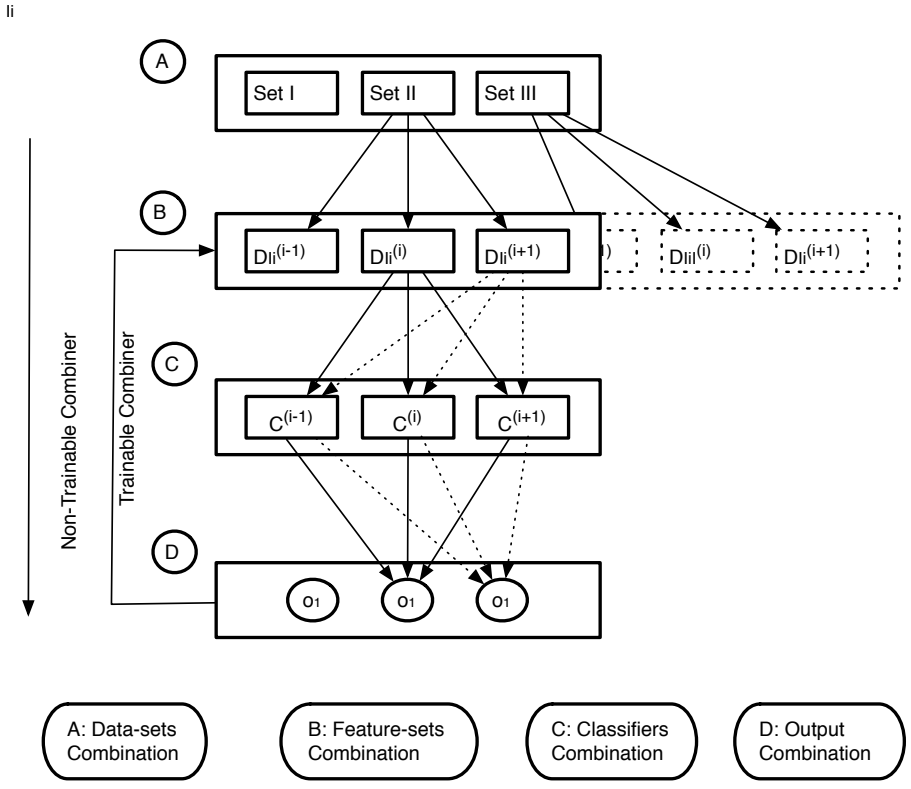


Figure 2.5: Level of combination

A. Data-sets combination Different sets of data can be used to obtain different knowledge pertaining to the same problem. Bagging [NH91] can be viewed as data-set combination. In general, if the same description D_i is used to train the same classifier C_j on different data -sets, the information conveyed by the different $S^{(k)}$ may be different, but not complementary.

Working with complementary (but not contradictory) information is the necessary requirement to improve the classification performance by combination. Maybe, the best way to ensure the synergy is to use complementary descriptions of the same data set (generally descriptions with different physical meanings).

B. Features combination Combination of descriptions, or "feature fusion", consist of combining different descriptions of the same pattern to form a unique description: $D(z) = \cup_{l=1}^{M} D_l(z)$.

This type of combination can be applied only if all descriptions to be fused have the same dimensions. If each description of an instance of the data set is given by a single feature vector, then the different vectors can be concatenated into a matrix to form a single, and more complete, description of the instance. In many cases, the pattern to be described is sampled into small portions and a feature vector is computed on each portion. As a result the pattern is described by a feature matrix. A major effect of this kind of sampling is the reduction of the noise that can be created by errors in the description. If two descriptions of the same pattern are not based on the same sampling, the different feature matrices have different dimensions and it is not feasible to concatenate them without adding redundant information or deleting a part of the information. Moreover, even if the dimensions of all descriptions to be combined are equals, the varieties in forms (scales of the values, continuous/discrete variables, binary/multi variables) and meanings of the features can create some problems for the feature space conditioning. Unless care is taken, the unique description obtained by concatenating the different descriptions can be sparse or ill conditioned. For these reasons, it is often preferable to train different classifiers, each of them using its own feature space, and combine their decisions afterwards.

C. Classifiers combination Boosting [FS95] is the most popular example of classifier combination. Weak classifiers (i.e. classifiers with low accuracies) are used to build a final " strong" classifier by learning the different strengths and weakness of each weak classifier. The process is iterative and weak classifiers are added to learn the specificities of samples that previous weak classifiers misclassified.

D. Decisions combination The combination of classifier outputs, also referred to as *decisions combination*, is the most popular scheme of combination because it can be applied to any problem. A classifier can return as much or as little information as you want about the class membership of the instance to be classified, as shown previously in the presentation of classifiers. The minimum information returned by a classifier is the predicted class. Additional information can be obtained by returning a ranked list of the possible classes. The maximum information a classifier can return is a measure of membership for each class of the problem. In any case the ranked list of classes can be deduced from the membership values and the predicted class can be deduced from the rank list. For more details on the output types of classifiers, we refer the reader to [XKS02]. In the next paragraph we present methods to combine the outputs of different classifiers for each type of output.

2.2 Schemes for combination of classifiers decisions

Combination methods based on the output of classifiers can be grouped into two categories. The first category, parallel combination, combines the outputs of classifiers dealing with the same problem (i.e the same original data set to be classified into the same known categories). The second category

combines the outputs of classifier runs one after another, where the set of classes for each classifier is determined by the previous classifier. This type of combination is reffered to as sequential combination in the following calculations.

2.2.1 Parallel combination

Parallel combination method deals with K systems of classification trained to solve the same problem. Then for an instance to be classified, we need to combine the K outputs that can be single label, ordered list of classes, or a vector of membership measurements for each class.

A: Single class For classifiers whose outputs are the predicted class, the combination can be performed using *voting methods* or the *Knowledge-Behavior Space* approach [HS93]. When the classifiers to be combined have different levels of performance, the decision of the classifiers can be weighted to obtain a better combination. Combining classifiers with single class output is limited, especially when $K = 2$. Often it is necessary to work with classifiers that can handle additional information about potential alternatives.

B: Ranked list Instead of returning a single decision, some classifiers are developed to return a ranked list of the possible classes. There are several methods to combine K ranked lists. The goal of these methods is to re-order the K ranked lists so that the true class is ranked higher than any other class. From these methods, we retain the *Highest rank, the Borda count* and the *Logistic regression* detailed in [HHS02b].

The *highest rank* analyzes the K lists to derive possible sub-sets of classes. Then, the union of these subsets is re-ordered according to their old ranks in each subset. The *Borda count* is a generalization of the majority vote. For each class candidate different scores are given according to its position on each list. The class that obtains the highest score is designated as the predicted class. This method does treat all classifier outputs equally. Some variations of this approach are described in [VES00]. The *Logistic regression* is a modification of the *Borda count* where weights are given to scores produced by each classifier. Additional information can be considered by using membership measurements instead of list of classes.

C: Vector of membership measurements Some classifiers, such as GMM, return a membership measure for each possible class. As we discussed in Sec.1.2 numerous methods have been developed to derive membership values for each class of the problem for classifiers that are originally developed to return a single class (such as the methods discussed for the SVM and kNN). If we consider that the K classifiers to be combined return a vector and membership measurements, then these vectors can be stored in a single decision matrix called *decision profile matrix* [Kun04]. Each column of the matrix contains the measurement vector of one classifier. As shown in section 1.2, these membership measurements can be (pseudo)-posterior probability (GMM) or a distance (SVM, kNN).

The variety in the form of the outputs can be problematic for the combination. Some requirements have to be carefully considered before applying the combination rule on classifier outputs. According to [KHDM02] there are two scenarios of combination.

1. **Scenario I :** All systems use the same feature space and the same classifier. The different decisions are obtained by varying the parameters of the classifier (e.g. the number k or the distance for kNN, the kernel for SVM, or the number of Gaussians for GMM, etc.). In this case, each classifier produces an estimate of the same membership measurements and they can be directly combined with no risk.

2. **Scenario II:** The different systems of classification have different feature spaces or different classifiers. The main interest of this scenario is the possibility of integrating complementary information into the data. In this case, it is not longer possible to combine the outputs directly, since the classification is done on different feature spaces, or the output returned by the classifier ranges in different intervals.

   To solve the problem of combination for scenario II, there are many transformations that can be applied to the outputs of the classifiers. We can consider, without lost of generality, that all these values are in the interval $[0, 1]$ and that the membership values are pseudo-posterior probability. The transformation of classifier outputs into pseudo-posterior probability can be done using the *softmax* method proposed in [DHS01]. The *bin method* has also been proposed to calibrate the output of classifiers such as the decision tree, naive Bayesian classifiers, [ZE01] and SVM [Dri01]. A detailed description of the classifier outputs transformation methods can be found in [Liu05].

When membership measurements have comparable scales they can be combined using classical rules such as: *sum-rule, product-rule, max-rule, min-rule or median-rule*. The *sum-rule* and the *median-rule* generally perform better than the three other ones if the classifiers to be combined are not fully independent [KHDM02]. The *Dempster-Shafer Theory* [MS88] can also been used as an alternative for the combination of decisions.

The difficulties encountered in combining decisions increases with the amount of information handled by the outputs. The single label-based combination is straightforward, while the combination on class membership measurements requires considerable precautions. Ranked list of classes is a good compromise. From these lists we can derive subset of possible classes and train new classifiers on these subsets to simplify the task. This idea is at the basis of sequential combination methods.

2.2.2 Sequential combination

We define sequential decision combination classification schemes, where the classification systems are applied one after another, and where the problem treated at one stage is defined by the previous system. The decision of a multi-stage classifier can be conveniently described by the mean of a decision tree as demonstrated in [Kur88]. The terminal nodes of the tree (the leaves), represent the classes and every interior node is connected with an appropriate set of classes accessible from that node. The root represents the entire set of classes into which a pattern may be classified. Then the classification process traverses the path of the tree starting at the root. At each non-terminal node encountered one aught to make a decision about the continued path in the tree until a terminal node is reached. To deal with decision trees, two components have to be defined beforehand: (a) The skeleton of the tree (i.e. the set of possible alternative at every stage) and (b) the decision rules at

every interior node. Depending on the choice of (a) and (b), we can group the multi-stage classifier into the following categories.

A: Hierarchical combination A hierarchical classification involves a set of classes organized with a predefined taxonomy. At each step a classifier has to choose the best family of classes. A specific features set can be used by each classifier at each step of the process. In general, classifiers are trained to drive the instance to be classified towards the most specific classifier that makes the final decision [GN02]. The successive choices form a set of decisions that can be interpreted as an automatic indexing system with different granularity levels. Real life data often has a hierarchical or clustered structure. For such data, using a hierarchical system of recognition can improve the performance and the rapidity of the classification process especially for problems involving a large number of classes [SBC$^+$02]. Hierarchical classification can be regarded as a sub-class recognition problem. To define the hierarchical organization of the classes, some clustering methods can be applied.

B: Cascade classification At each stage of a cascade classification a new classifier is used to either accept or reject the input instance. When the classifier accepts the instance is assigned directly to one class. Otherwise, when the instance is rejected, it is given to the next classifier for further processing. Each classifier is trained with a feature set defined beforehand. These methods have been proven to improve the classification performance and the recognition of rare events [ZBS07].

C: Multi-stage classification We refer to this as multi-stage[2] classification, which is a system of classification that, at each stage, reduces the set of possible classes and does not require any specific organization of the data. The goal of class set reduction is to derive a subset from the original class set such that is as small as possible and contains the true class. In general, class-set reduction methods attempt to derive a threshold on the ranks according to the worst-case ranks of the true classes [HHS02a]. The two main class-set reduction methods (intersection of large neighborhood or union of small neighborhood) use the rankings of all classifiers obtained on the training data set to determine the thresholds. This threshold is determined by the worst position of the true class. This kind of approach can suffer from over-fitting problems. The final decision is made when one only class remains possible or is given by the last classifier available.

2.3 Trainable .vs. non-trainable combiners

Most of the methods mentioned previously use a fixed rule, (parallel combination) or a set of rules (sequential combination) to combine the outputs of classifiers. It is possible to find a better combination by learning the behavior of the different classifiers. We refer to *non-trainable combiner* as schemes of combination that have no extra parameter that needs to be trained. The combination can be performed directly on the decisions of the classifiers. These are opposed to *trainable combiner*, which are classifiers that use the decisions of classifiers as new features to train a new classifier (see fig.). This new classifier, or meta-classifier, is trained to learn the best combination rule. These

[2]The term multi-layer is also found to refer to this type of combination.

trained-rules show very good accuracy, but they introduce a new problem regarding the separation of the data into training and testing data, as shown on Fig.2.6.

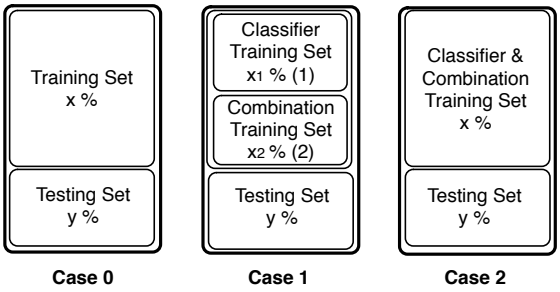


Figure 2.6: Possible data set reorganization to train class models and combination rules

The **case 0** illustrates a non-trainable combination. The two options for trainable combination are illustrated in case 1 and 2. In **case 1**, the training set is split into two parts. The first part (1) is used to learn information about the different classes. The resting part (2) is temporarily used as testing set. The membership measurements obtained from part (2) are used to learn the combination rule tested on part (3). To be accurate, such a system requires a very large amount of data. In **case 2**, the same data set is used to acquire information about the classes and about the combination rule. However, we know that membership measurements obtained on the training set are not necessarily comparable to the ones obtained on new testing data. This strategy can suffer from over-fitting consequences.

Non-trainable parallel combination schemes are simple and efficient to combine classifier decisions when the systems to be combined have comparable accuracies. Even a system with a lower accuracy can handle complementary information on the problem, but to be incorporated into the decision-making process, special care must be taken. When the classification systems have different degrees of performance they can be combined using a meta-classifier. Otherwise sequential combination can be used to combine different descriptions of the data. If the studied data has taxonomy, hierarchical classification can be applied. Otherwise, classification can be performed in cascade.

3 Conclusions

Supervised classification methods are powerful tools to solve problems involving the recognition of predefined patterns. In this document we use these methods to detect the singing voice (Chap.4) and to identify singers (Chap.5). For both problems we propose to describe the signals of singing voice with two independent sets of features that describe two distinct aspects of the singing voice. These features are described in Chap.3. For each experiment, the best features (within their respective feature set) are selected with the IRMFSP or the “plus-l-take-away-r” algorithms presented in Sec.1.1. The results of the feature selection methods are interpreted to evaluate the relevance of each feature for the given problem. All experiments are conducted with the classifiers presented in Sec.1.2 where the internal parameters of each classifier are chosen so that they minimize the classification error on the training set using a cross validation on this set. To avoid bias in the performances of our experiments we chose to work with data sets with well balanced classes. For the task of singer identification, we propose

to improve the recognition performance by combining the information conveyed by the two sets of features. For this combination we develop a specific method, mingling the sequential and parallel combinations of decisions.

Chapter 3

Singing voice: production, models and features

The goal of this chapter is to introduce the audio features used in the rest of this document to describe signals containing the singing voice. As mentioned in the previous chapter, problems involving pattern recognition (e.g. tasks of detection or identification) require representing the pertinent information of the data for a given problem in terms of a compact set of parameters. These parameters, referred to as audio features, should have a clear physical or perceptual meaning to allow an interpretation of the class models obtained with the statistical tools presented in the previous chapter.

In this work we describe signals containing signing voice using features derived from two of the major models developed for sound representation: the source-filter and the sinusoidal models. The source-filter model is used to extract information on the vocal tract of the singers, by means of a compact representation of the spectral envelope. This information is related to the timbre of the singer's voice. On the other hand, the sinusoidal model is used to extract information on the style and the technique of the singer by analyzing the temporal variations of the frequency and amplitude of partials.

We present in Sec.1 the basics of vocal production on which rely the interpretability of the source-filter and sinusoidal model. In this section we also present some elements that make singing differ from speech. Special attention is paid to the vocal vibrato. Then, in Sec.2 we give the details of the source-filter and the sinusoidal models and present some methods to estimate the parameters of these models. In this section we also present a model to interpret the temporal variation of the frequency (and amplitude) of the partials obtained using the sinusoidal model. Different methods to estimate the parameters of this model, called "intonative model", are presented and compared. Finally, Sec.3 reviews the features used to describe signals of the singing voice.

1 Singing voice production

A large number of studies have been conducted to understand the mechanisms involved in the production of speech. Some of these researches have been motivated by the desire to create synthetic voices. The level of current understanding allows the synthesis of intelligible and natural speech. The understanding of speech production is at the basis of the understanding of the singing production since

speech and singing production involve similar mechanisms. The similarity is such that the boundary between singing and speech is very fuzzy. However, in many cases, signals of the singing voice are heard as clearly different from speech. There is clearly a difference in the prosodic (i.e. the duration of the sounds, the rhythm, etc.) and intonative (i.e. the pitch movements) attributes of the vocal production between speech and singing. Beyond this melodic aspect of singing, some peculiar aspects of singing, such as the capacity of singing to pass over a loud orchestra, have attracted researchers' attention. Some of the elements that make singing so different from speech are presented in Sec.1.2 after a presentation, in Sec.1.1, of the basics of the vocal production involved in the production of speech and singing.

1.1 Vocal production

The voice organ is generally composed of three structures: the respiratory system, the vocal folds and the vocal and nasal tracts. The voice sounds originate from an airstream formed in the lungs, which is possibly interrupted by the vocal folds and then filtered by the vocal and nasal tracts.

During the phonation, the respiratory system (the lungs and the diaphragm muscle) acts as a compressor and controls the pressure under the glottis. When a certain amount of pressure is reached the vocal folds (located within the larynx at the top of the trachea) are triggered to vibrate and act as an oscillator. This phenomenon is called the phonatory process, or voicing. The oscillation of the vocal folds modulates the pressure and the flow of the air through the larynx. The resulting flow of air is a pulsating airstream named the voice source (or glottal source). The spectrum of the voice source is a harmonic spectrum whose fundamental frequency (f_0) is determined by the frequency of the vocal folds oscillation. Each harmonic component has a frequency located at an integer multiple of the f_0.

If the vocal folds are sufficiently close to each other, or if they are under too much or too little tension and pressure, the vocal folds do not oscillate. The sound produced in this case is qualified as voiceless (or unvoiced). For instance, in whispering, vocals folds are too tense to vibrate normally, but they form a narrow passage that makes the airstream become turbulent and generate an audible noise. By definition, unvoiced sounds have no pitch. During the creation of unvoiced sounds, different parts of the vocal organ work as sound sources.

In both cases, the sounds created are filtered by the vocal tract, which is composed of everything located above the vocal folds: the supra-glottic larynx, the pharynx, the mouth, the vocal cavities, etc. The vocal tract acts as a resonator by enhancing certain frequencies. These resonances, and the corresponding enhanced frequency regions of the speech short-time spectrum, are commonly called formants. These resonances are created by the cavities of the vocal tract. Thus, by modifying the shape of the vocal tract, i.e. by changing the configuration of the articulators (the tongue, the soft palate, the jaw), the frequencies of the formants are modified. Each configuration of the vocal tract creates a different acoustic filter. In particular, this leads to the different vowels, and gives a special color to the overall sound. In practice, each vowel is essentially characterized by the frequency of the two first formants. The filtered sound is then radiated by the two outputs of the vocal tract: the nose and the lips.

This description of voice production is at the basis of the source-filter model developed by Fant [Fan81]. We give a description of this model in Sec.2.1. This model, at the foundation of the majority of the research in the speech processing area, can also been applied to describe sung sounds

since singing also involves the production of a voice source-filtered by the vocal tracts to produce the phonemes of the lyrics. However, in the case of singing, this model is sometimes altered. Some reasons for these modifications are presented below.

A comprehensive description of each step of the singing production can be found in the numerous works of Sundberg [SR90] and Titze [TM98].

1.2 Some specificities of the singing production

The first noticeable difference between speech and singing is the proportion of voiced sounds. According to [Coo90] 90% of the singing production is voiced compared to 60% in speech.

We review some elements that have been shown to differentiate singing from speech on voiced sounds. Most of these elements are supposed to help the voice to be powerful, intelligible and stand out from the instrumental accompaniment. First we present the formant tuning and the singing formant that result from special modifications of the vocal tract. Then, we present the vocal vibrato and the portamento that can be observed on the pitch contours of singing.

1.2.1 Formants tuning

The source-filter model supposes the vocal tract and the source are independent of one another. In other words, it assumes that the frequencies of the formants are independent of the pitch.

In singing, singers are required to produce specific vowels on specific pitches. At the same time, the sound produced has to be beautiful and audible even in the presence of a loud orchestra. When the first formant is lower than the pitch, it has been hypothesized [MS90] that singers adjust their formants such that they coincide with the harmonic partials. With this technique the overall sound level is increased without raising much vocal effort. This formant tuning, also called vowel modification, is produced by altering the jaw and/or lips openings to change to shape of the vocal tract [RMW90]. Formant tuning is not limited to high pitches. Male singers, like the baritone studied in [MS90], take advantage of the formant tuning on the higher partials even when the formant adjustments are not required by a low fundamental.

The formant tuning offers the advantage of increasing the level of the sound but it has as the main inconvenience of decreasing the vowel intelligibility. As found in [Sun94], [SSWR83] and [SMS95], the vowel intelligibility in lyric singers voices (soprano especially) decreases when the pitch increases. Therefore, formant tuning may not be as commonly used as expected. Some singing teachers prefer the fixed formant scale, suggesting that the gain in energy associated with the formant tuning is not sufficient to overcome the intelligibility problems associated with moving the formant [CS92].

1.2.2 Singing formant

The singing formant is an additional formant, around 3000 Hz, noted in the spectrum of many professional lyric singers. This formant was first noticed in "good" male voice by Bartholomew [Bar34]. It was latter named singing formant (or singer's formant) by Sundberg who proposed that the formant is a clustering of the third, fourth and fifth formants [Sun74]. He also suggests that singers produce this formant by lowering the larynx and by narrowing the vocal tract just above the glottis. It was later shown that the singing formant was also present in high voices. In [BP86], Bloothooft explains the

singing formant in terms of energy in one third of an octave centered around 2500 and 3160 Hz for female and male singers respectively. In both cases this formant is located in the range of frequency where the ear is very sensitive. In addition, orchestras have little power in this range. This suggests that the presence of a singing formant helps the singing voice to be heard above a large orchestra.

According to the results of various studies, this formant appears more often in low voices and is weaker in soprano voices [WBM+01]. Soprano singers have less need of this formant because their pitches fall in the range of sensitive human hearing. In addition, high voices have harmonics that are widely spaced and it is very likely that the singing formant falls between two harmonics. Finally, high voices can use formant tuning more effectively than other singers, and may therefore have less need of a singer formant.

The singing formant is considered a desirable property of lyric singers and may not exist in voices using microphone amplification (typically, pop-rock voices).

1.2.3 Vocal vibrato

The vibrato is considered an expressive attribute of music that adds richness, warmth and emotion to musical performance[Met32]. It is also considered a feature that helps the voice be heard even in the presence of a loud orchestra. In [Ras78], Rasch shows that notes with vibrato pass over a set of non-vibrated tones even if the latter have higher amplitudes or masking components. It has also been shown that vibrato makes vowels more prominent allowing them to be more easily separated from background sounds [MM91] [McA84] . An acoustical explanation of this phenomenon is given in [Mel92].

Utilization of vocal vibrato over history In early music vibrato was mainly used by singers on sustained notes to reinforce their presence in a musical ensemble and to emphasize some musical expressions [Tof96]. The continuous use of vibrato began during the 19^{th} century amongst the singers and was later copied by other musical instruments to imitate the expressive aspect of the singing voice [Bur01]. Today, vibrato is used 95% of the time on voiced vocal segments [Sea31] performed by lyric singers. Vibrato is also used by string and wind instruments on sustained tones. The use of vibrato has changed throughout the periods of music history, along with musical styles and musical traditions.

Definition of vibrato If the term vibrato evokes a common idea for musicians, it remains particularly difficult to give a precise definition because it covers a number of concepts and techniques. The first definition was formulated by Seashore ([Sea31], [Sea36]) as follows : " A good vibrato is a pulsation of pitch, usually accompanied by synchronous pulsations of loudness and timbre, of such extent and rate as to give a pleasing flexibility, tenderness, and richness to the tone." The rate and extent of the periodical modulation are illustrated in Fig.3.1. Physically, a separation can be made between pitch, amplitude and timbre modulations, but they are often fused into one perception of vibrato movement because their separation is not feasible perceptually [DHAT99].Theoretically, the term vibrato should refer to the modulation of frequency only (FM) and the modulation of amplitude (AM) should be referred to as *tremolo*.

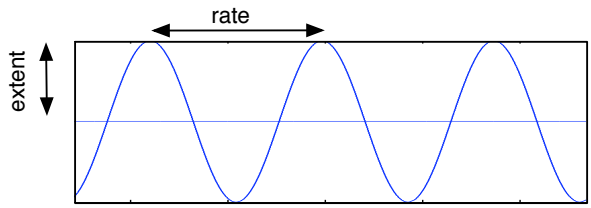


Figure 3.1: Rate and Extent of sinusoidal modulation

Production of vocal vibrato It is relatively easy to understand the interaction between the physical control of string players and the production of vibrato. The mechanisms involved in the production of vocal vibrato (and vibrato of wind instruments) still remain unclear. Numerous explanations have been given which suggest that vibrato enters vocal production as a relaxant principle. It is generally assumed that vocal vibrato result from neuromuscular excitation of the laryngeal mechanism [BZSJ09]. "Vibrato is the result of a balance between muscle systems in antagonistic relation to each other during phonation. When this balance occurs, the antagonistic muscle systems develop an alternating pulse that is a reflection of the continued energy level required of them to maintain equilibrium and muscular health. In other words, the muscles of the larynx begin to pulse rhythmically in response to tension and sub-glottic pressure, which produces the characteristic vibrato sound. It occurs naturally in order to protect the vocal folds." [1]

This explanation justifies the presence of the frequency modulation in the definition given by Seashore. It has been shown in many studies that the modulation of amplitude occurs passively when the fundamental frequency of the sung tone (and its harmonics) coincides with the upward and downward slopes of the vocal tract transfer function of the singer. This phenomenon is known as "resonance harmonic interaction" [TSSL02], [HH88]. The third part of the modulation, the modulation of timbre, is the result of the combination of the amplitude modulations of each harmonic partials independently [SM91] [Leb99].

Parameters of vocal vibrato As all (quasi) sinusoidal modulation, the vibrato can be described with the following parameters: extent (or amplitude), rate (or frequency), waveform, and regularity. Because the vibrato is not instantaneous, the following elements can also be considered: the time delay until the vibrato onset and the percentage of vibrato presence over the duration of a sustained note.

Most studies on the vocal vibrato have focused on the characterization of the modulation of frequency of lyric singers. In the following paragraphs we present the results of these studies.

- **FM Extent**
  The extent of the frequency modulation is given by the distance of the fluctuation of pitch above and below the mean pitch as shown in Fig.3.1. The extent is generally measured in cents (1 semi-tone = 100 cents). In Tab.3.1 we report the values of the pitch vibrato extent found in various studies. On average, the vibrato extent ranges from $\pm$ 50 et $\pm$ 100 cents. According to a perceptual evaluation of vibrato [Cas93], this value should not exceed a semitone, otherwise it would be perceived as an ornamentation (vocal trill). The extent value under

[1] http://www.singwise.com/cgi-bin/main.pl?section=articles&doc=Vibrato

which the modulation is not perceived may exist but it has never been measured. It is likely that the perception of vibrato varies according to the listeners.

The extent of the FM is much larger for lyric voices than pop-rock voices. It has been shown in [BS03] that the extent FM value increases when the loudness increases.

- **FM Rate**
  The vibrato rate, i.e. the number of modulation cycles per second, is measured in Hertz. The studies on vibrato rate measurement show that the rate of the vocal vibrato varies between 5 and 8 Hz. The rate is quasi constant per singer [Sun95] and on average the value is higher for female singers ([DHS95], [SLS80]). The vibrato rate is not always constant along the duration of the sung tone. According to [VGD05] the vibrato rate increases towards the end of notes. In many studies it is found that the vibrato rate varies with the pitch but remains invariant against changes in loudness.

  The values found across various studies are reported in Tab.3.2

- **FM Waveform**
  The frequency modulation of the vocal vibrato is a quasi sinusoidal modulation with small fluctuations [Sea31]. In [Hor89], four classical waveforms are suggested for the vocal vibrato: sinusoidal, triangular, trapezoidal and non-identifiable.

- **FM Regularity**
  The regularity between the different cycles of a vibrated tone can be used as an indication on the level of performance of the singer. According to [Sun94] singers with a high level of practice have a more regular vibrato than novice singers. A measure of the vibrato regularity can be computed using a correlation coefficient measured between two adjacent cycles, as suggested in [SCB81].

The amplitude modulation, which occurs passively in the presence of a frequency modulation, has been the subject of fewer studies than the frequency modulation. According to [Sea31] the amplitude modulation is present at 70% of the time when there is a frequency modulation.

Control of vibrato parameters In [SR90], Sundberg shows that singers can not modify their vibrato voluntarily, suggesting that the vibrato parameters remain constant for a given singer. Ssingers and singing teachers alike admit that singers develop their own vibrato naturally and subconsciously when their vocal technique is appropriated [SSH84]. When student singers force their vibrato it often results in a too slow (called wobble) or an overly fast (called bleat) vibrato that sounds more like errors in pitch than expressive effects. In particular, a wobble is picked up by the human ear as a succession of separate pitches and is perceived as unpleasant.

In several other studies, it is mentioned that singers adapt their vibrato with the tempo [Ven67]. In [MB90], it is suggested that the control of the vibrato's parameters is a good indication of the singer's skill and flexibility. In an extended study on the vibrato and note's transition [Mah08], it is suggested that the vibrato rate is adapted to facilitate the note's transition (i.e. the last vibrato cycle is at its bottom-most point when the next tone has a lower pitch, and vice versa).

Ref	min extent	max extent	instru	Notes
[Sea31]	±10	± 160	± 50	The extent value does not vary with the loudness, the pitch, the sex or the musical mode. Children have a lower extent. The extent is constant for a given singer.
[Win53]				The value change with the phonation and the loudness
[SLS80]	± 80	± 200		mean: ± 120 for lyric singers
[SCB81]				varies with the loudness
[Mil86]	± 66	± 100		
[Hor89]	± 50	± 200		
[SR90]		± 100		
[Cas93]	± 100	± 150		
[Sun95]	± 10	± 100	± 50	choir singers: much lower extent ± 10 cents.
[Pra97]	± 50	± 150		vary with the pitch and the loudness
[TD00]	± 60	± 200	± 20 - ± 35	The value for instru corresponds to string instruments

Table 3.1: Extent values (in cents) for vocal FM found across different studies

Ref	min rate	max rate	Instru	Notes
[Sea31]	6	7		for 50% of singers: there is no relation between the vibrato rate the pitch, the loudness, the vowel or the sex of the singer.
[Sea38]	≈ 6.6	≈ 6.6		Bellow 4.5 Hz the modulation is not perceived by the auditors, but this value depends on the listeners.
[Win74]	5.5	7.7		median = 6.9 Hz for vocal vibrato
[SLS80]	male: 4.7 female: 4.9	male 6.3, female 6.6		the rate varies with the age of the singers, their sex, their emotional implication and the vowel. For 10 opera singer there is not relation between the pitch and the rate
[Ben81]				The rate varies with the pitch.
[Pra94]	≈ 6	≈ 6		The vibrato rate increases at the end of the sustained notes. (variation ± 8%).
[Sun95]	5	7		The rate is constant per singer.
[TD00]	5.5	8		
[DHAT99]	6	7	de 4 - 12	increases towards the end of the note.

Table 3.2: Vibrato rate values across different studies

We now present another feature of the singing voice, still occurring on the pitch contour, that helps to distinguish the voice from the instrumental background during the transitions between two notes.

1.2.4 Portamento and Legato

In singing, when two distant tones are sung in the same breath, it is common to glide smoothly and continuously from one pitch to another. This technique is called "portamento". The same effect can be used by other musical instruments and is referred to as a glissando. When the two notes are spaced apart by an interval smaller than a third this effect is referred to as "legato". Musicians also use the term legato to refer to smooth repetitions of the same note. In [Pol02], Pollastri defines the latter as *spikes*. Spikes are characterized as a monotonically increasing sequence of pitches followed by a monotonically decreasing one. They appear on note repetitions and are sometimes used as an ornamentation. In our representation and modeling of singing pitch contours, all of these effets (portamento, legato, spikes) are modeled as portamento.

We note that the use of portamento is much more popular in singing than for any other instrument. It is considered an attribute of the bel canto technique and is used less today than in the past [Pot06].

2 Models for voiced sounds

In this section we present several models to describe the voiced sounds of singing. During the production of pitched sounds the periodicity of the voice source is never constant due to the fact the voice organ is a complex system with many variables that evolve constantly through time. Considering these points, a voiced utterance can be interpreted as:

1. a temporal succession of voice pulses modified by a time-varying configuration of the vocal tract or as
2. a set of time-varying frequencies (with harmonic ratio to the fundamental) whose amplitudes vary over time.

These two interpretations correspond respectively to the source-filter model and the sinusoidal model. These dual representations of voiced sounds are described in the next sections. For each model, we also present some methods to estimate their parameters.

2.1 Source-filter model

2.1.1 Model description

The source-filter model, introduced by Fant [Fan81], describes speech production as a two stage process involving the generation of a voice source (with its own spectral shape and spectral fine structure) which is filtered by the resonant properties of the vocal tract. If the voice source (also called excitation signal) is denoted by $e(t)$ and the impulse response (or transfer function) of the filter is denoted by $h(t)$, then the voiced signal resulting from $e(t)$ is passed through the filter whose impulse response $h(t)$ is given by:

$$x(t) = e(t) * h(t) \tag{3.1}$$

In the frequency domain Eq.3.1 becomes:

$$X(\omega) = E(\omega)H(\omega) \tag{3.2}$$

The source-filter model is compelling because it reflects the physical mechanism of vocal production. It has been used as an underlying representation of signals in numerous studies on speech synthesis, speech recognition and speech transformation. These studies consider the voice source to be related to the qualities of the vocal folds (length, mass, tension), and they suggest that the filter is related to the physical shape of the vocal tract. The contribution of the source and the filter are generally separated using inverse filter techniques where $h(t)$ is given by the estimation of the spectral envelope of $|X(\omega)|$ the amplitude spectrum of the signal. The source-filter model as presented here is a simplified model since it does not consider the radiation given off by the nose and the lips.

This model supposes the source and the filter clearly independent, in practice there are always non-linear interactions between the two components [AF82]. It is possible that the interaction between the source-tract interaction is more complex in singing than in speech [HdD01]. It is also noted in [Hen01] that the source-filter model is not appropriate for singing voice signals. However, these statements are not demonstrated and the source-filter model has been used successfully to synthesize singing voice in [Lu02]. We note that the inverse filtering techniques yield worse results as the fundamental frequency increases. For high-pitch sounds the envelope estimation can be critical because the harmonics are widely spaced and they excite the vocal tract on a reduced number of frequency bands. As a result, the spectral envelope can be properly estimated on a reduced number of frequencies. In [AC04] it is suggested that the source-filter model (as all other non-interactive models) could model the singing voice and the reason for the failures of inverse filtering techniques found in [Hen01] is due to the high pitches achieved by female singers.

2.1.2 Estimation of the vocal transfer function

There are two main approaches to estimate the transfer function of the vocal tract: the first approach finds the position of the formants directly by mean of linear prediction [MG76] and the second set of methods are based on the real cepstrum whose aim is to estimate the spectral envelope of sounds [OSB$^+$89].

Linear prediction analysis Linear prediction estimates a signal at a given instant n using a linear combination of its p previous samples. We have:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e(n) \tag{3.3}$$

The transfer function of the filter, in the frequency domain, is deduced from Eq.(3.3) and Eq.(3.2) as follows:

$$H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{3.4}$$

where $H(z)$ is the z-transform of the vocal tract filter $h(n)$. This function is an all-pole filter as shown in Eq.(3.4).

The goal of the linear prediction is to determine the set of coefficients a_k that provides the best estimation of the samples of the signal. These coefficients are estimated by minimizing the error between $x(n)$ and the values predicted from the p past samples. The two most common methods to solve this problem of minimization, the covariance matrix and the auto-correlation, are described in [Mak75]. The values of $A(z)$ can be factored to determine the location of the poles, which correspond to the formant location. This is one of the greatest advantages offered by the LP analysis: it estimates coefficients that are directly related to the physical properties of the person who emitted the sound.

The selection of the order of the model (the number of poles) is quite critical: if p increases, the estimated envelope starts to follow the peaks of the harmonic signal instead of its overall shape. An appropriate model order can be chosen using the physical properties of the vocal tract filter as explained in [Osh87]. Even though, the all-pole model suffers from systematic errors for high-pitched signals.

These errors have been addressed in [EJM91] where the discrete all-pole (DAP) model is presented as an alternative for these cases. The idea of the DAP model is to fit the all-pole model using a finite set of spectral locations that are related to the harmonic position of the fundamental frequency. This method requires the estimation of the f_0.

Cepstral analysis The second set of spectral envelope estimation methods, based on the cepstrum, is also inherited from the source-filter model. The cepstrum, originally presented in [BHT63], is given by the inverse Fourier transformation of the log-amplitude spectrum of the signal. The main advantage of the cepstral domain is that it transforms the convolution of two signals into the addition of their cepstra. Thus, the cepstrum of a vocal signal is the addition of the cepstrum of the vocal tract response and the cepstrum of the excitation signal. If F^{-1} denotes the inverse fourier fransformation, and the real cepstrum of $x(t)$ is given by:

$$C(\omega) = F^{-1}(\log(|E(\omega)H(\omega)|)) = F^{-1}(\log(|E(\omega)|)) + F^{-1}(\log(|H(\omega)|)) \tag{3.5}$$

The spectral envelope is considered a smooth version of the amplitude spectrum. The simplest mean to estimate this envelope is to set all high frequency components of the cepstrum to zero and to retain only the low frequency components. This filtered version of the cepstrum creates an envelope that follows the mean of the amplitude spectrum but not the contour of the spectral peaks.

To preserve the original values of the harmonic's amplitudes, Galas [GR91] and then Cappe [CM96] propose a new method named the discrete cepstrum. This method is more robust than the traditional cepstrum, but it is very complex, computationally expensive and requires a fundamental frequency analysis to preselect the spectral peaks.

There is also another method based on the cepstrum used to estimate accurately the peak contours: *the true envelope* estimation method. This method has been introduced in [IA79] and then improved in [Röb05]. The true envelope estimation is based on iterative cepstral smoothing of the log-amplitude spectrum. Let $X(\omega)$ be the spectrum of the signal and $V_i(\omega)$ be the cepstral representation at step i (i.e the Fourier transformation of the filtered cepstrum). The algorithm starts with $A_0(\omega) = log(|X(\omega)|)$ and $V_0(\omega) = -\inf$ and iteratively updates the smoothed input spectrum using the maximum of the original spectrum and the current cepstral representation.

$$A_i(\omega) = \max(A_{i-1}(\omega), V_{i-1}(\omega)) \tag{3.6}$$

The algorithm stops if for all bin, the relation $A_i(\omega) < C_i(\omega) + \tau$ is satisfied where τ is a fixed threshold. With this procedure, the valley between the peaks will be filled by the filtered cepstral representation and the estimated envelope will gradually grow until the peaks are covered.

2.2 Harmonic sinusoidal model

2.2.1 Model description

The harmonic sinusoidal model is another model that is widely used to interpret harmonic signals. In the harmonic sinusoidal model, the glottal source is represented as a sum of sine waves that, when applied to a time-varying vocal-tract filter, leads to the desired sinusoidal representation of speech or singing waveform.

The sinusoidal models have their origin in the vocoder developed by Dudley [Dud39]. The first harmonic sinusoidal model is presented by McAulay and Quatieri in [MQ86] where they describe speech sounds as the sum of sine waves, with amplitude and frequency evolving slowly over time. This model was then improved in [GL$^+$88]. As a next step, the harmonic sinusoidal model was extended by means of a dedicated noise model [SS90]. Thus, the sinusoidal model given by Eq.(3.7) was completed.

$$x(t) = \sum_{h=0}^{H(t)} p_h(t) + b(t) \text{ where} \tag{3.7}$$

$$p_h(t) = a_h(t)\cos(2\pi f_h(t) + \phi_{h,0}) \tag{3.8}$$

In Eq.(3.8), $p_h(t)$ is a sinusoid with a slow time-varying amplitude $a_h(t)$, a slow time-varying frequency $f_h(t)$ and a phase at the origin $\phi_{h,0}$. In Eq.(3.7), $x(t)$ is the signal that consists of a sum of sinusoidal components with slow time-varying amplitude and frequency (also called partials) plus an additive noise. If $x(t)$ is the signal of a tone produced by a monotonic instrument, such as the singing voice, the H components are harmonics and their frequencies are in harmonic ratio to the fundamental. In practice, the number of components varies over the duration of the sustained tone.

The sinusoidal components of the sinusoidal model can vary in amplitude and frequency independently. Thus, the model can be used to represent signal modulations such as the vibrato or the portamento.

To represent a signal with the sinusoidal model it is necessary to estimate the parameters $a_h(t)$, $f_h(t)$ and $\phi_{h,0}$ for each instant of the signal. In practice, the signal is discretized and we assume the parameters constant on each discrete interval. In others words, we make the assumption that the amplitude and the frequency vary slowly over time.

2.2.2 Sinusoidal model parameters estimation

To estimate the parameters of the sinusoidal model, the values $f_h(n)$, $a_h(n)$ and $\phi_{0,h}(n)$ are estimated on each discrete portion of the signal on which the signal is supposed to be stationary. The number of sinusoidal components $H(n)$ on each frame has to be determined. Then, the trajectories of each partial's frequency and amplitude are determined using a continuation algorithm.

We describe in the next paragraph a method to estimate these parameters. We start with the estimation of the parameters for a single sinusoid ($H = 1$), which is the basic case. This simple estimation method is then extended to the case of signals with multiple components.

Simplified case: monochromatic stationary signal We consider the simplest case of sinusoidal model, i.e. a stationary signal with a single component. Such a signal is written as:

$$x(t) = a.cos(2\pi ft + \phi_0) \tag{3.9}$$

where f, a and ϕ_0 are the frequency, the amplitude and the phase of the modulation we need to estimate. In practice, if f is known, the other parameters of the modulation are estimated by minimizing the modeling error.

- **Frequency estimation**
  The frequency f is usually estimated by analyzing the spectrum $|X(f)|$ of the signal $x(t)$. The value of f is given by the position of the highest peak of the spectrum.

  If we consider the discrete version of x, defined on a finite number of sample N, its spectrum is approximated by the Discrete Fourier Transform whose equation is given in Eq.(3.10). This function is evaluated on a finite set of frequencies: $f = k/N$ for $k \in \{0, \ldots, N-1\}$.

$$X\left(\frac{k}{N}\right) = \sum_{n=0}^{N-1} x(n) \cdot e^{-2i\pi \frac{k}{N}} \tag{3.10}$$

  The estimation of f is given by the value of the bin where the spectrum is maximal. For a signal of size N sampled at frequency f_s we have:

$$\hat{k} = \arg\max_k(|X(k/N)|) \tag{3.11}$$

$$\hat{f} = \frac{\hat{k}}{N} \cdot f_s \tag{3.12}$$

  In practice, the DFT is computed on a centered version of the signal to remove the constant component ($X(0) = \sum_{n=0}^{N-1} x(n)$). Otherwise, the highest peak could lead to an estimation of the constant. The original signal is weighted with a window (Rectangular, Triangular, Hamming,Hanning, Blackman, Kaiser, Harris, etc.[Har78], [Nut81]) to reduce the side-lobes in the spectrum.

- **Amplitude and phase estimation**
  The amplitude associated to the modulation of frequency $\hat{f}$ is given by the amplitude value of the (normalized) spectrum at bin $\hat{k}$. The estimation of a is given by:

$$\hat{a} = \frac{1}{H(0)} \left| X\left(\frac{\hat{k}}{N}\right) \right| \tag{3.13}$$

  where $H(\omega)$ is the Fourier Transform of the window used to segment the original signal into a set of pseudo stationary signals with finite duration.

Finally, the phase of the sinusoidal modulation is given by the angle of the DFT at this point:

$$\hat{\phi}_0 = \angle X\left(\frac{\hat{k}}{N}\right) \tag{3.14}$$

The estimation of the sinusoidal parameters relies entirely on the good estimation of $\hat{k}$, which is constrained by the limitation of the Fourier analysis: the precision and the resolution.

Improvement of parameters estimation:

- The spectral precision is limited by the number of points used to compute the DFT. For a DTF computed on M points, only the frequency multiples of $\frac{f_s}{M}$ can be estimated exactly. For the other frequencies, their estimation is given with a precision equal to f_s/M. The spectral precision can be increased by using a DFT computed with a larger number of points: $M > N$. This technique, called *zero padding*, is equivalent to a quasi perfect interpolation in the frequency domain. However, this technique has a very limited effect in the sense that it is necessary to add an incredibly large number of points (which is computationally very expensive) to obtain a weak improvement of the spectral precision.

- The spectral resolution indicates the ability to discriminate two close frequencies. The spectral resolution is limited by the width of the main lobe of the smoothing window. The ability to distinguish two close frequency components increases as the main lobe of the window narrows. The main-lobe bandwidth reduces as the length of the window increases. For each classical widow the bandwidth of the main-lobe is given by: $B_\omega = C_\omega/L$ where C_ω is a parameter defined for each window and L is the size of the window. To estimate the sinusoidal parameters of a monochromatic signal the resolution has no influence, but it can become a serious issue when the signal is composed of several sinusoidal components.

Numerous alternatives have been developed to overcome the problem of spectral precision and resolution. There are methods based on techniques of interpolation ([SS87], [AKZ99]), regression methods [MD92] and techniques of spectral reassignments ([KGV78], [Aug95]).

Another set of techniques, not based on the amelioration of the Fourier transformation, has been developed: the High Resolution (HR) methods. As their name indicates, these methods offer a much better spectral resolution than any method based on the Fourier transformation. HR methods found their origins in the early works of Prony [Pro95] and Pisarenko [Pis73] which aim to estimate the parameters of a sum of exponentials with techniques of linear prediction. When the signal is not corrupted by any noise the HR methods have a virtually infinite resolution and precision (even when the analyzed signal has a very few points). The first HR methods, proposed in [Pro95] and [Pis73], were not robust to the presence of additive noise. Some alternatives to the high resolution parametric subspace methods, such as MUSIC ([Sch86]) and ESPRIT ([RPK86]) have been proposed to increase the performance in the presence of noise. The underlying idea of these methods is to separate the data into signal (sinusoidal) and noise subspaces via eigenvalue decomposition of the covariance matrix or the singular value decomposition of the raw data matrix. The main advantage of the HR methods is that they are not constrained by the time-frequency compromise. The main drawback is that the

number of sinusoidal components has to be pre-estimated and these methods can be computationally expensive since they require the inversion of matrices.

Multiple sine waves stationary signal When multiple sinusoids are present on a frame of the signal, the method presented for monochromatic signal is applied iteratively on the highest peaks of the signal. Once the parameters of the sinusoid corresponding to the highest peak of the spectrum have been estimated, the process is iterated with the second highest peak and so on until H peaks are found.

The problem of defining H for each frame remains. Peaks can be selected with an amplitude-based threshold. A global threshold generally leads to non-uniform results: the difficulty is to reject spurious peaks in the high amplitude region without removing valid peaks in the lower amplitude region. Thresholds based on local amplitude produce better results because the number of tracks can vary considerably from frame to frame.

Estimation of the time-varying frequency and amplitude Once the parameters of the sinusoidal model are estimated for each frame they have to be connected to form coherent sine wave frequency and amplitude trajectories. Several strategies have been explored to realize this task.

In [MQ86], a simple sinusoidal continuation algorithm is proposed: each spectral peak is connected to its closest peak in the next frame. This algorithm may create unreasonable jumps and is often unable to track frequency trajectories with a significant frequency modulation without introducing numerous false connections.

This algorithm was improved upon in [SS90] by using a set of frequency guides to create sinusoidal trajectories. The values of the frequency guides are obtained from the peak value and their context, such as surrounding peaks and fundamental frequency. If the sound to be modeled is known to be harmonic, the frequency guides are initiated according to the harmonic series corresponding to the estimated fundamental frequency.

Another strategy, proposed in [ABLS02] compares amplitude and frequency difference for the candidates to connect. This algorithm connects only the peaks that do not exceed a minimum variation for both parameters. Unconnected peaks belong to dying partials. Peaks with no connection may represent a newborn sinusoid. Some other methods, based on Hidden Markov Models or the Viterbi algorithm have also been proposed. These models are told to be very efficient to find the trajectories of partials in polyphonic mixtures and inharmonic signals.

Alternative approaches for sinusoidal parameters estimation Some models have been proposed to directly estimate the time-varying frequency $f(t)$ and time-varying amplitude $a(t)$ on a non stationary signal. In the method proposed in [Lar89], the amplitude of the signal is described by a polynomial with complex coefficients. This polynomial, chosen with an order equal to 2 [Pee01], is sufficient to describe the variations in frequency and amplitude independently. This method, however, assumes the frequency of the modulation known is a priori and the coefficients of the polynomial are estimated by a least-squares error minimization. In some simple cases, when the variations of amplitude and frequency are linear, the coefficients of the linear variations can be estimated by a polynomial interpolation of the instantaneous amplitude and frequency respectively (the order of the polynomial is equal to 1).

At the end of the sinusoidal parameter estimation process each partial h present in the signal is described with:

- an onset and an offset
- a function of frequency $f_h(t)$
- a function of amplitude and $a_h(t)$
- a phase Φ_0

The greatest advantage of the sinusoidal method is that it describes the temporal variations of each sinusoidal component independently. It also supposes that the amplitude and frequency are independent components. These temporal variations of frequency (and amplitude) can be interpreted as prosodic elements in speech and as intonative elements in singing. Next we present a parametric model to estimate the parameters of vibrato and/or portamento from a time-varying function $f_h(t)$. The model can also be applied to estimate the parameters of a tremolo from an amplitude function $a_h(t)$.

2.3 Intonative model

In the description of the source-filter model presented in Sec.2.1 we have shown that the spectral envelope of speech and singing sounds is commonly used to extract information on the speaker/singer. In the same way, we believe that an appropriate model to extract information from the partials obtained with the sinusoidal model can be used to extract additional information on the singer. The characteristics extracted are related to the style and the technique of the singer.

2.3.1 Model description

We suppose that $f(t)$ is the time-varying frequency of a partial covering a sustained note or a continuous transition between two notes. As shown in Fig.3.2, the fundamental frequency of a sung melody can be decomposed into these two types of events.

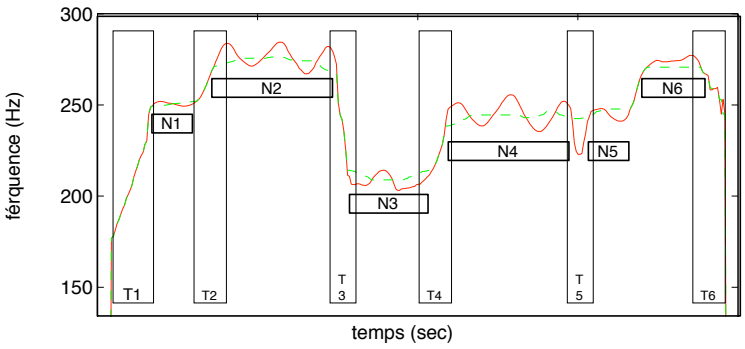


Figure 3.2: Fundamental frequency of a sung melody: Segmentation into sustained notes (N) and transitions (T)

On $f(t)$, we can expect two types of variation: a periodic modulation associated to the vibrato and a continuous (and slow) variation associated with the portamento. Typically, the vibrato occurs on sustained notes and the portamento appears during the transition between two notes. However, it is also common to observe a slow continuous variation of frequency over the duration of a sustained tone.

Considering these points, we propose to model the frequency trajectory of a partial as the sum of the periodic modulation $x_f(t)$ plus a slow continuous variation $d_f(t)$. An additional error $\epsilon(t)$ is also comprised in the model. With $\bar{f}$ denoting the mean frequency (the perceived pitch), $f(t)$ can be written as:

$$f(t) = \bar{f} \cdot (d_f(t) + x_f(t)) + \epsilon(t) \tag{3.15}$$

In Eq.(3.15), the two frequency variations are given in terms of relative variations so that these variations are similar for two partials in harmonic ratio to the same fundamental.

If we assume a vibrato rate of r and extent a [2]constants over the duration of the tone, the periodic modulation $x_f(t)$ can be written as:

$$x_f(t) = a.\cos(2\pi r t + \phi_0) \tag{3.16}$$

In practice, the rate of the vibrato remains quasi constant over the duration of the tone, but the amplitude can vary significantly. For some singers, the amplitude increases consistently at the beginning of the notes and for some other singers the amplitude decreases towards the end of the notes. To allow a more flexible model of vibrato we propose to describe the amplitude of the modulation with an exponential. Thus the vibrato model is given by:

$$x_f(t) = a_0.e^{a_1.t}.\cos(2\pi r t + \phi_0) \tag{3.17}$$

If a_1 is positive, Eq.(3.17) represents vibrato whose extent decreases over the duration of the note. Conversely, if a_1 is negative, Eq.(3.17) describes a vibrato whose amplitude increases over time.

2.3.2 Model parameters estimation

- It is first necessary to compute $\bar{f}$ the **mean frequency**, which corresponds to the perceived pitch. This value is simply given to be the average value of $f(t)$ over the time. The other parameters are estimated on $f(t)/\bar{f}$. This is so that when two partials are harmonics of the same fundamental, the parameters of the model (except $\bar{f}$) have the same magnitude.

- The **relative frequency deviation** $d_f(t)$ is given by the overall shape of $f(t)$. In practice, it is given by the low-pass filtered version of $f(t)/\bar{f}$. To preserve the modulation associated with the vibrato, whose minimal rate is around 5Hz, we chose a cutoff frequency of $f_c = 4$ Hz. If $f(t)$ is the fundamental frequency of a sustained note with vibrato only, then the frequency

[2]To avoid the confusion between the meanings of amplitude and frequency, used both to refer to the frequency modulation (FM) and amplitude modulation (AM) as well as to qualify the speed and the width of a sinusoidal modulation. The parameters of the vibrato $x_f(t)$ are named rate (frequency) and extent (amplitude)

deviation is null. If $f(t)$ covers a note transition, then the frequency deviation has up to two points of inflections. For these reasons, we chose to model the frequency deviation with a third order polynomial (with real coefficients).

- The **sinusoidal part** of the $f(t)$ is estimated by subtracting the relative frequency deviation d_f from the normalized version of the frequency trajectory:

$$\hat{x}_f(t) = \frac{f(t)}{\bar{f}} - \hat{d}_f(t) \tag{3.18}$$

$x_f(t)$ is a sinusoidal modulation, which can be written with Eq.(3.16). This equation is similar to Eq.(3.9). Thus, the parameters of the modulation (rate, extent and phase) can be estimated with any method presented in Sec.2.2.2.

The same model can be applied on amplitude trajectory $a(t)$. Contrary to the frequency modulation, the amplitude modulation of two harmonically related partials do not have any predefined relation because each harmonic partial is modulated according to its frequency position in the vocal tract.

The function of amplitude can be written as:

$$a(t) = \bar{a}.(d_a(t) + x_a(t)) + \epsilon \tag{3.19}$$

where $\bar{a}$ is related to the global dynamic of the sound ({***p***,***mf***,***f***, ... }), $d_a(t)$ transcribes a possible variation of dynamic (for instance a *crescendo*) and $x_a(t)$ represents the amplitude modulation. The parameters of Eq.(3.19) can be estimated with the methods presented above.

2.3.3 Model evaluation

We evaluate the ability of the proposed model to fit the time-varying frequency function of partials. The performance of the method relies on two elements: the estimation of the parameters of the sinusoidal modulation and the estimation of the continuous variation of frequency.

Position of the problem The parameters of the sinusoidal modulation can be estimated with any method presented in Sec.2.2.2, where the number of sinusoidal components is fixed to one. Numerous studies have been conducted on the comparison of sinusoidal parameter estimation methods. However, the time-frequency resolution constraints encountered by these methods to track partials or to extract the parameters of partials are different. In the case of partial tracking, the parameter estimation is conducted on signals with a rather high sampling rate (usually between 11 and 44 kHz) in order to estimate frequencies produced by the instruments of the mixture (whose frequencies range from 60 to 5000 Hz). When these methods are applied to estimate the intonative model parameters, the signal under analysis has a low sampling rate (whose value is determined by the hop size of the STFT used in the partial tracking) and the frequency to be estimated is between 4 and 8 Hz (these values correspond to the lower and upper rate of vibrato). Since we want to estimate frequency ranging from 4 to 8 Hz, it is very important to have an extremely precise estimation (contrary to the frequency estimation of partial tracks an error of 1 Hz will be very important).

Considering these points, we propose to compare the classical sinusoidal parameters estimation methods for signals corresponding to the frequency of a vibrated note. We chose to compare the following methods:

- M0: Basic method based on the Fourier transformation of the signal described in Sec.2.2.2, where the Fourier transformation is computed with a zero padding factor equal to 4.
- M1: QIFFT method proposed in [SS87]. This method uses a quadratic interpolation of the three samples surrounding a spectral peak to refine the estimation of the position of the spectral peak with the maximal amplitude. This method is applied on a spectrum obtained with a zero padding factor equal to 4, and the phase is given by a linear interpolation.
- M2: The high resolution subspace method ESPRIT proposed in [BDR06].
- M3: The method based on complex polynomials proposed in [Lar89]. This method requires an estimation of the frequency of the modulation as input. In our case, this estimation is given by the results of the QIFFT method.

M0, M1 and M2 have the assumption that the rate and extent are constant along the duration of the modulation. The method ESPRIT supposes the rate constant and the extent exponentially damped. M3 considers the rate and the extent linearly variable. We summarize the settings of the methods in Tab.3.3.

Method	Base	Window	Rate	Extent
M0: TF	TF , padding 4	hanning	f	a_0
M1: QIFFT	TF, padding 4	hanning	f	a_0
M2: ESPRIT	Sub-space, order: N/2		f	$a_0 \cdot e^{\delta t}$
M3: POLY	QIFFT	Gaussian	$r_0 + f_1.t$	$a_0 + a_1.t$

Table 3.3: Methods and parameters used in the evaluation

Data We evaluate these methods on three distinct sets of signals representing the fundamental frequency $f(t)$ of vibrated sung tones. The first set, denoted by S_1, is composed of sinusoidal modulation with a constant rate and extent. S_2 is composed with sinusoidal modulation whose extent and rate vary linearly over the time. The third set is composed of a fundamental frequency estimated from "real" a cappella recordings with the YIN algorithm [dCK02] and checked manually. The synthetic signals of S_1 and S_2 are generated with the parameters presented in Tab.3.4.

The parameters of Tab.3.4 are chosen for the following reasons:

- 1. The vocal vibrato extent is chosen between 1/5 and 1 semi tone
- 2. The vocal vibrato rate is comprised between 4 and 8 Hz ([VGD05])
- 3. The phase is randomly chosen between $-\pi$ and π.
- 4 and 5. The values chosen correspond to the linear variation coefficients of extent and rate measured in spoken sounds [AS05]. These values may not be larger for sung sounds but we chose the values measured in speech utterances in order to have reasonable values. The linear variation of rate and amplitude are set equal to zero in the signals of S_1.

Indice	Name	Parameter	Distribution	Unit	Intervalle
1	extent	a_0	Uniforme	cents	$[10, 100]$
2	rate	r_0	Uniforme	Hz	$[4, 8]$
3	phase	ϕ_0	Uniforme	rad	$[-\pi, \pi]$
4	damped factor	a_1	Gaussian	$cents/sec^{-1}$	$N(0, 10)$
5	linear factor	f_1	Gaussian	rad/sec^2	$N(0, 100)$
6	length	$N * Fs$	Uniforme	sec	[0.3, 1]

Table 3.4: Values of parameters for synthetic vibrato

- 6. The sampling rate is chosen equal to 1000 Hz, which is the sampling rate of the signals of S_3. The duration of the signal is comprised between 0.3 and 1 sec. These values correspond to the duration of a quaver and a minim at an *allegreto* tempo (120 bpm).

We evaluate the performance of the method for the estimation of the vibrato parameters on S_1 and S_2. The performance of the estimation of all parameters of the model is evaluated on the signals of S_3.

Results - Sinusoidal modulation estimation - S_1 and S_2 We evaluate the performance of the methods presented in Tab.3.3 for the signals of S_1 and S_2 in the presence of an additive noise. The experiments are conducted for a signal-to-noise ratio (SNR) varying from -10 to 80 decibels.

For all evaluations we compare the Mean-Square Error (MSE) of the estimation with the Cramer-Rao Bound (CBR) [VT68]. The CRB is an indicator of quality that gives the lower bound on the error variance of unbiased estimator. If $CRB(\hat{P})$ denotes the Cramer Rao Bound of the estimation of P and σ_P^2 denotes the variance of an unbiased estimator then we have:

$$CRB(\hat{P}) \leq \sigma_P^2$$

In practice, most estimators are biased. Their comparison with the CBR remains possible if the bound is compared to the mean square error. The MSE can be decomposed into square bias and variance:

$$MSE(\hat{P}) = \mathbb{E}\left[(\hat{P} - P)^2\right] = \frac{1}{N}\sum_{n=0}^{N}\left(\hat{P} - \mathbb{E}(\hat{P}) + \mathbb{E}(\hat{P}) - P\right)^2 \quad (3.20)$$

$$= \frac{1}{N}\sum_{n=0}^{N}\left(\hat{P} - \mathbb{E}(\hat{P}) + B_P\right)^2 = \sigma_P^2 + B_P^2 \quad (3.21)$$

where B_P^2 is an indicator for systematic errors while σ_P^2 is an indicator for noise sensitivity. For biased and unbiased estimators, the minimal value of the MSE is given by the CRB. The theoretical computation of the CRB for the sinusoidal parameters can be found in [Kay88]. These values for a sinusoidal modulation of length N and amplitude A in the presence of an additive Gaussian noise with variance σ_B^2 are given in Tab.3.5.

We evaluate the quality of the estimators on signal of S_1 with various durations (3, 1 and 0.3 second). The performance of the frequency estimators are plotted in Fig.3.3 - 3.3 and 3.5. The

Amplitude	$CRB(\hat{a}) = \frac{2\sigma_b^2}{N} + O\left(\frac{1}{N^2}\right)$
Frequency	$CRB(\hat{f}) = \frac{6\sigma_b^2}{4\pi^2 N^3 A^2} + O\left(\frac{1}{N^4}\right)$
Phase	$CRB(\hat{\phi}_0) = \frac{2\sigma_b^2}{NA^2} + O\left(\frac{1}{N^2}\right)$
Noise	$CRB(\hat{\sigma}_b) = \frac{2\sigma_b^2}{4N} + O\left(\frac{1}{N^2}\right)$

Table 3.5: Theoretical CRB for sinusoidal parameters estimation (stationary case)

performances for the amplitude estimation are plotted on Fig.3.6 - 3.7 and 3.8. Finally the performance for the phase estimation are given in Fig.3.9 - 3.10 and 3.11.

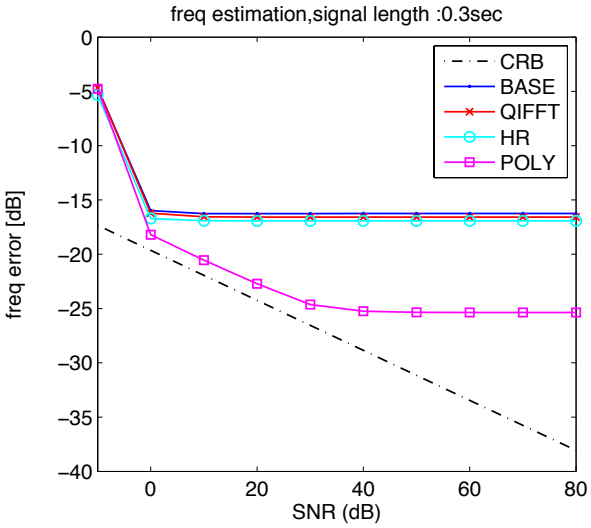


Figure 3.3: CRB : S1- frequency - (0.3sec)

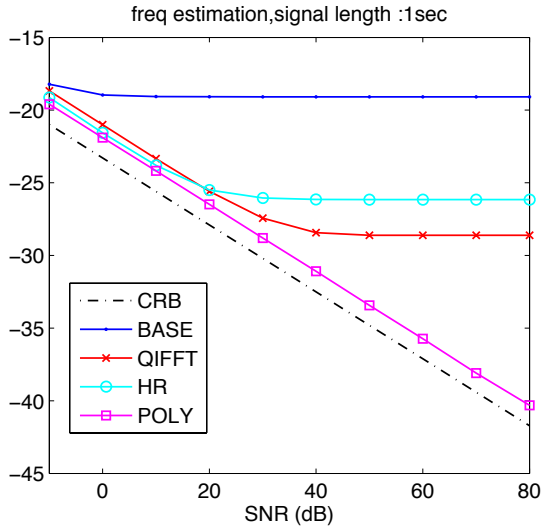


Figure 3.4: CRB : S1 - frequency - (1sec)

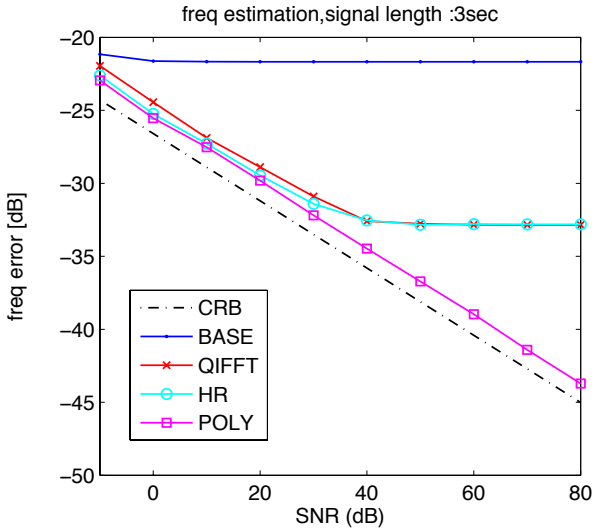


Figure 3.5: CRB : S1 - frequency - (3sec)

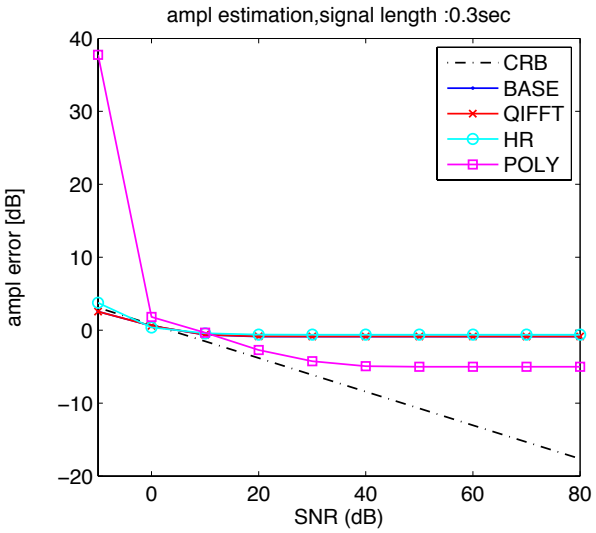


Figure 3.6: CRB : S1 - amplitude - (0.3sec)

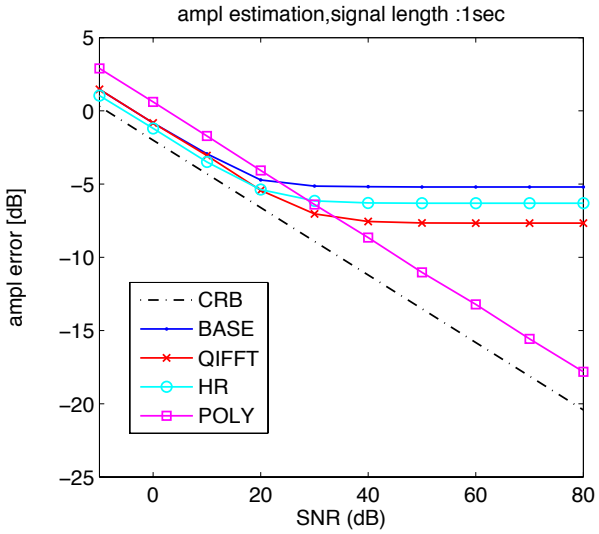


Figure 3.7: CRB : S1 - amplitude - (1sec)

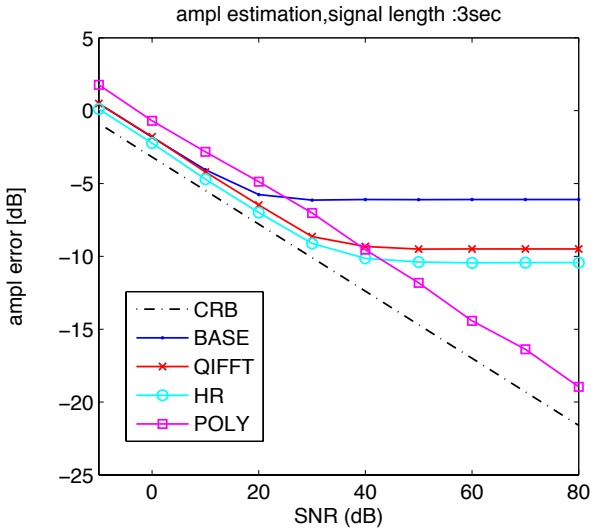


Figure 3.8: CRB : S1- amplitude - (3sec)

We conduct the same evaluation for the signal of S_2, i.e. sinusoidal modulation with rate and extent varying linearly. The performance is evaluated on signal with length 1 and 0.3 seconds.

The closer the curve to the CRB, the better the estimator. When the curve associated with the estimator follows the CRB, the error is dominated by the variance. We observe that after a given SNR threshold, the variance of the estimator errors saturates at a fixed level given by the estimator bias.

The frequency estimation is the central part of the algorithm. Phase and amplitude use the frequency to determine their estimate. The frequency estimate is influenced more by the noise so that it shows the largest sensitivity to noise.

For both sets of signals, we can see that the length of the signals has a high influence on the performance of the estimation. On the signals of S_1, where the modulation is stationary, the method based on the complex polynomials proposed by Laroche (POLY) can be considered an unbiased estimator on signals of 3 and 1 sec. On these signals, the basic method (M0) has a much higher bias than the

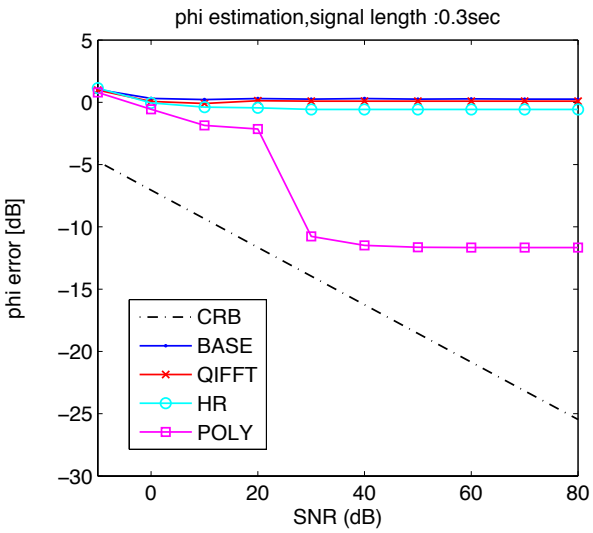


Figure 3.9: CRB : S1 - ϕ_0 - (0.3sec)

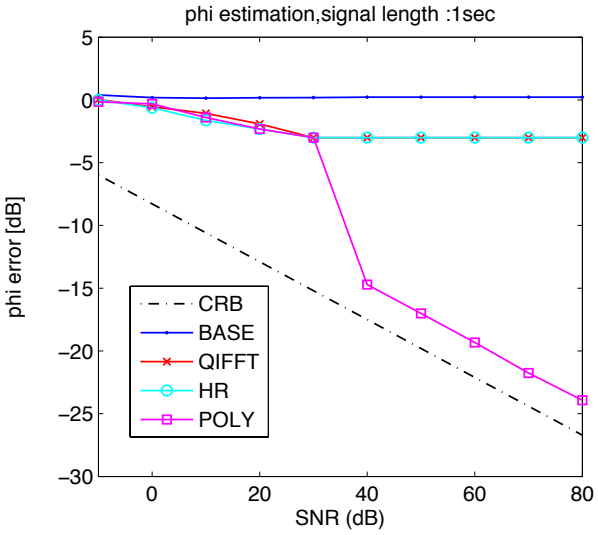


Figure 3.10: CRB : S1 - ϕ_0 - (1sec)

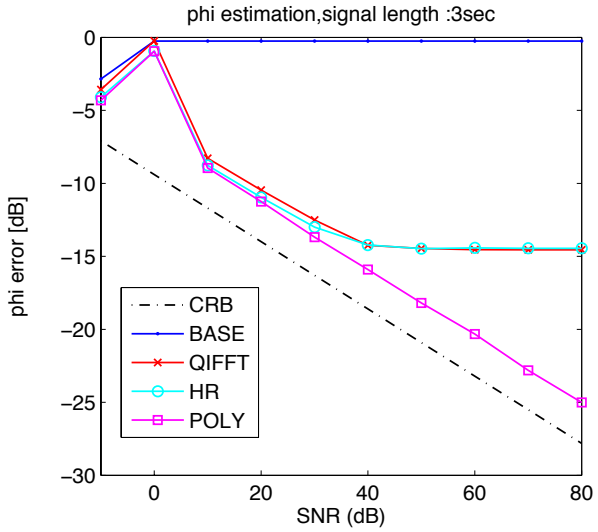


Figure 3.11: CRB : S1 - ϕ_0 - (3sec)

QIFFT algorithm. On these "long" signals, the QIFFT method performs better than HR and for a low SAR this method performs better than POLY. On shorter signals (0.3 sec), POLY is biased for SNR higher than 40 dB. On these short signals, the HR method performs better than the basic and QIFFT method. We note that the amelioration given by the interpolation in the QIFFT method on "long" signals disappears on "short" signals. The relative performances of the methods evaluated remain more or less similar on signals of vibrato with varying rate and extent. In general, POLY provides the best estimation. We note that the bias of the amplitude estimation is much larger for the signals of S_2 than the signals of S_1.

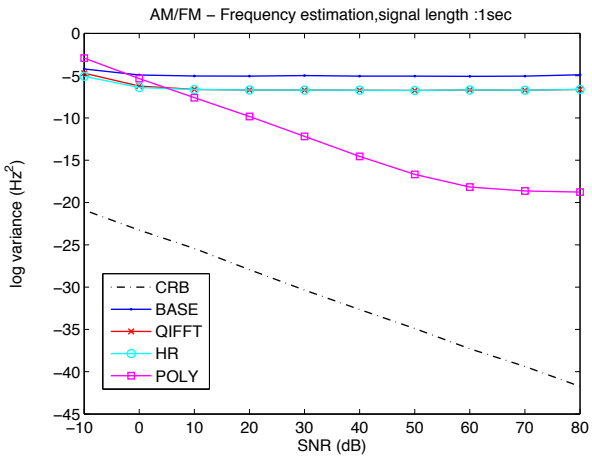


Figure 3.12: CRB - S2 - frequency - (1sec)

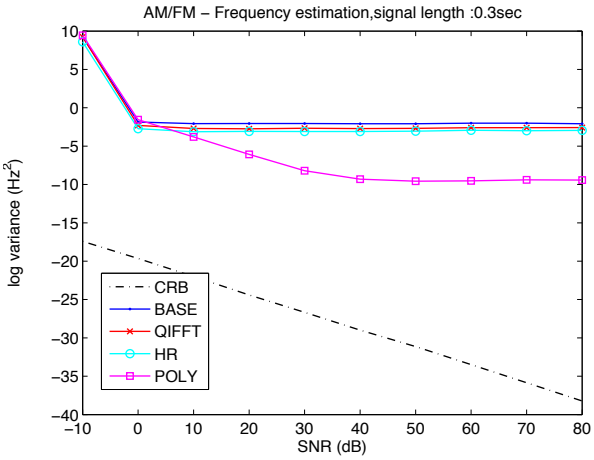


Figure 3.13: CRB - S2 - frequency - (0.3sec)

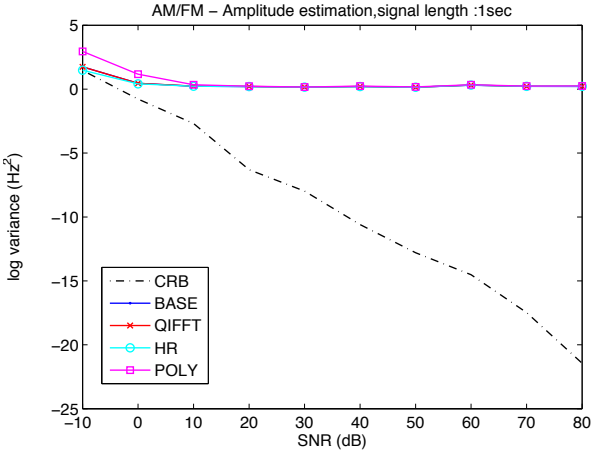


Figure 3.14: CRB - S2 - amplitude - (1sec)

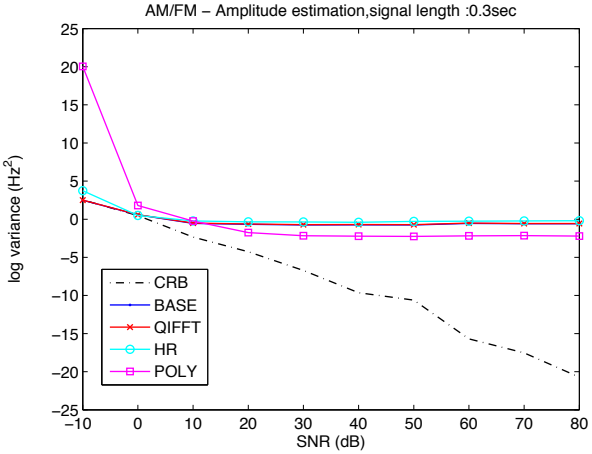


Figure 3.15: CRB - S2 - amplitude - (0.3sec)

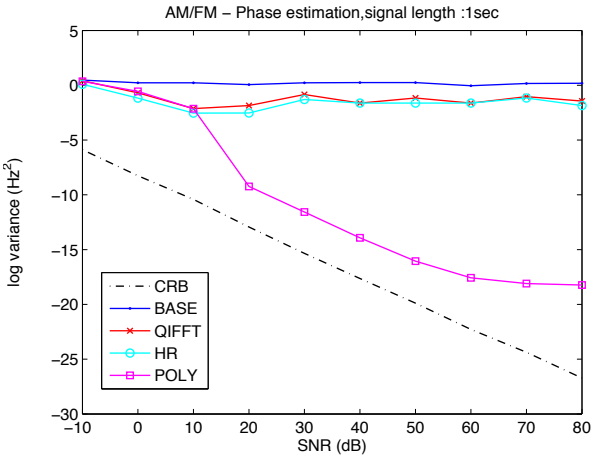


Figure 3.16: CRB - S2 - ϕ_0 - (1sec)

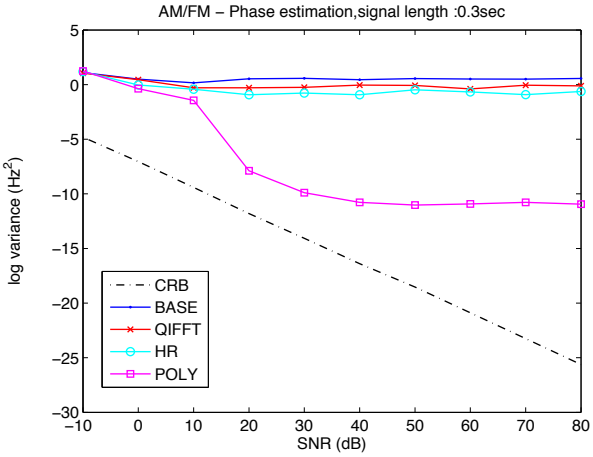


Figure 3.17: CRB -S2 - ϕ_0 - (0.3sec)

Results - Entire note model estimation - S_3 The performance on real signals (S_3) cannot be evaluated with the CRB since the parameters of the sinusoidal modulation are not known *a priori*. In this case, we measure the performance with the RMS error computed between the original $f(n)$ and the estimate $\hat{f}(n)$:

$$RMS_f = \sqrt{\frac{1}{N-1}\sum_{n=0}^{N}|f(n)-\hat{f}(n)|^2} \tag{3.22}$$

We illustrate the different steps of the estimation (estimation of the deviation d_f, then estimation of the sinusoidal part $\hat{x}$ and finally the estimation of the parameters of $\hat{x}$) on a vibrated note and a portamento note transition in Fig.3.18 and Fig.3.19.

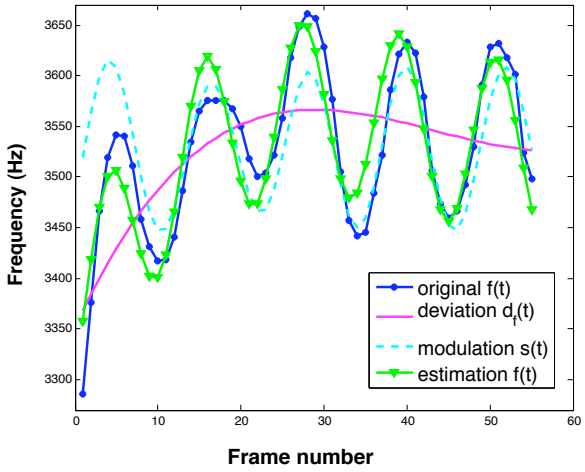


Figure 3.18: Estimation on a vibrated tone

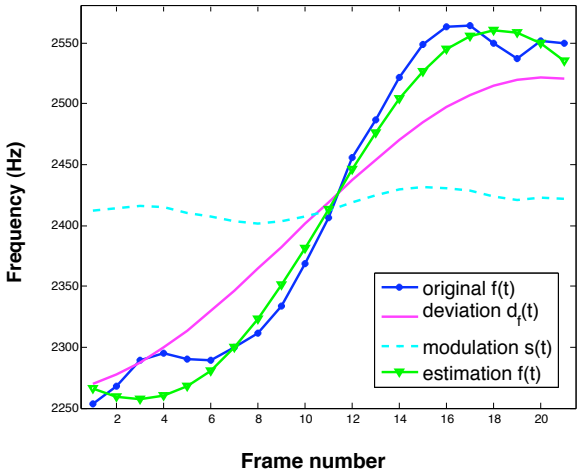


Figure 3.19: Estimation on a portamento (note transition)

The model assumes that the input $f(t)$ covers a sustained tone or a note transition. In practice, the partial tracking algorithm returns partials whose onset and offset do not meet these requirements. To obtain partials satisfying these constraints we propose to segment partials using the BIC criterion [Sch78]. The BIC criterion is applied to detect significant changes in the frequency function $f(t)$.

For $f(n)$ defined on N points $f(n) = \{x_1, \ldots, x_T\}$ the BIC criterion (for a problem in dimension 1) is computed for each n as follows:

$$BIC(n) = R(n) - \lambda log(T) \tag{3.23}$$

where λ is the penalty term and $R(n)$ is the value of the likelihood function for the estimated model. In dimension 1, R is given by:

$$R(n) = T\log(\sigma_f^2) - (n\log(\sigma_l^2) + (N-n)\log(\sigma_r^2)) \tag{3.24}$$

where $l(n)$ and $r(n)$ corresponds to the left and right sides of f split at instant n: $l(n) = \{x_0, \ldots, x_{n-1}\}$ and $r(n) = \{x_n, \ldots, x_{N-1}\}$. The terms σ_f^2, σ_l^2 , σ_r^2 indicate the variances of f, l and r respectively. The term λ is introduced to avoid over-fitting problems. In our problem, we use a large λ ($\lambda = 5$) to detect significant changes in f. This value was found empirically.

If $BIC(n) > 0$ then f is better modeled using the separated parts l and r obtained by splitting f at sample n. Fig.3.20 and 3.21 show the enhancement provided by the BIC criterion for the modeling of time-varying frequencies. On these figures, we plot the estimation of $f(t)$ obtained using a single model and the estimation obtained by concatenation of two distinct models. The vertical line represents the point of separation returned by the BIC criterion.

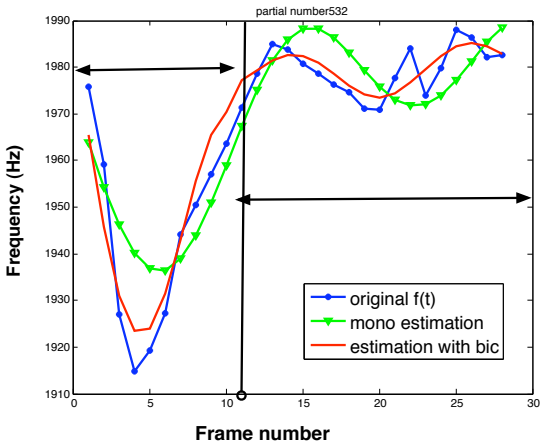


Figure 3.20: Two vibrated notes

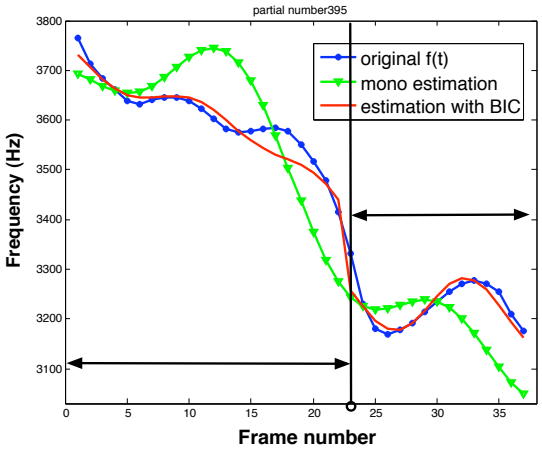


Figure 3.21: Vibrated notes linked + portamento

We report in Tab.3.6 the RMS error values for the 4 methods tested on a set of 1000 partials (S_3) segmented with the BIC criterion when it was necessary. The methods HR and POLY have been tested for fixed and varying amplitude and extent.

Method	Base	QIFFT	HR	HR	POLY	POLY
Model $A(t)$	a	a	a	$a_0.e^{a_1.t}$	a	$a_0 + a_1.t$
Model $F(t)$	f	f	f	f	f	$f_0 + f_1.t$
RMS (Hz)	4.11	3.81	3.48	3.01	6.85	6.79

Table 3.6: RMS errors for various estimation methods

The method based on the complex polynomial (POLY), that showed better results than any other method on S_1 and S_2 performs very poorly on real signals. On the synthetic signals, we showed that POLY is rather robust to the presence of additive noise and is very accurate to estimate the parameters of a sinusoidal modulation whose parameters vary linearly. We assume from the results obtained on S_3 that the POLY method is not robust to non-linear variations of the rate and extent of the sinusoidal modulation. On these signals, the HR method with an exponentially damped amplitude has the best performance. In average, the method has an RMS error of 3 Hz, which is relatively low. In the rest of the document, the parameters of the vibrato are estimated with the high-resolution method.

2.4 Relation between intonative and source-filter model

The vibrato is an interesting feature of the singing voice which has been presented by Arroabarren [AC04] as the link between the source-filter and the sinusoidal models. As mentioned in Sec.1.2.3, the frequency modulation creates an amplitude modulation on each harmonic partial. The amplitude of each partial is modulated according to its frequency in the vocal tract. Considering this point, by analyzing the amplitude modulation occurring along a few cycles of a frequency modulation, it is possible to deduce interesting information on the frequency of the formants.

To be reliable, this analysis has to be conducted on "ideal" recordings. First of all, it is important that the signal is recorded in a room with a low reverberation. Indeed, the presence of a reverberation modifies the values of partials' amplitude. In addition, the vocal tract must be invariant along the duration of the signal analyzed. Finally, the portion of the signal considered must have a quasi-constant loudness (i.e. there is no major change in dynamics, such as a crescendo). The last two assumptions are fairly reasonable since the analysis is performed on a very short portion of the signal.

In the following we propose a simple method to measure the position of the formants on an excerpt of singing. We start with some observations and then present a simple method to estimate the formants' frequencies.

2.4.1 Estimation of the formant position from a cycle of frequency modulation

Data We analyze a sound emitted by a female singer on an A5 (880 Hz) on the vowel /a/. We plot in Fig.3.22 the six first harmonic partials of this sound. We note that the theoretical formants value for the vowel /a/ pronounced by a female are : $f_1 = 600$ Hz and $f_2 = 1100$ Hz [SR90].

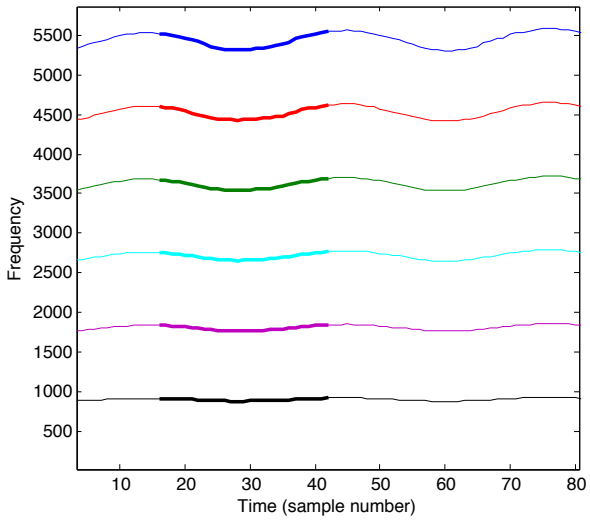


Figure 3.22: Frequency trajectories of the six first partials.

In the following, we study the amplitude modulation of one cycle of the frequency modulation corresponding to the bold part, i.e. for the samples between 16 and 42.

Illustration of the amplitude modulations We propose to illustrate the amplitude modulation against two variables: the frequency and the time. The variations of amplitude in function of the frequencies are plotted on the top plot of Fig.3.22. The bottom plot of Fig.3.22 represents the amplitude variations along the time axis. For both representations, the variations of the first six partials are presented on a third axis. The first half of the modulation is plotted with a plain line and the second half with dashed lines. On both plots, the amplitude of each partial is normalized so that the maximum amplitude of each partial is equal to one.

Observations From the plots of Fig.3.23 we can deduce that we have three categories of partials:

- The amplitude variations of partials 2, 4 and 5 demonstrate similar behaviors. As shown on the top plot, the amplitude increases while the frequency decreases and the maximum amplitude value is reached for the minimum frequency. This observation can also be deduced from the bottom plot where the amplitude of these partials is maximal in the middle of the modulation. From these plots we can deduce the presence of a formant close to the minimum frequency of partials 2, 4 and 5. Finally, we have 3 formants around 1750 Hz, 3500 Hz and 4400Hz.
- We can observe the opposite phenomenon on partials 3 and 6. The amplitude of these partials increases when the frequency increases and the amplitude is minimal in the middle of the cycle. For both partials we can deduce that there is a formant located just above the maximum frequency covered by each partial. Finally, we can deduce a formant slightly above 2800 Hz and a formant slightly above 5500 Hz.
- The amplitude of the first formant does not vary along the duration of the modulation cycle. In this case, the mean frequency of the partial coincides with a formant whose bandwidth is larger than the frequency range covered by the partial. For this reason, there is no significant amplitude modulation. We can assume As explained previously, the presence of a formant around 880 Hz is the result of the formant tuning process.

If the bandwidth of the formant is narrower than the range of frequency covered by the partial, then the amplitude modulation appears and its rate is twice the rate of the frequency modulation as illustrated in Fig.3.24. On this plot the frequency and amplitude of a single partial are normalized, scaled and superimposed to highlight the relationship between the rates of the modulations.

The rates of FM and AM, computed with the intonative model on all partials of a sung sound, can thus be used to give a rough estimate of the formant frequencies of a singer. In Chap.5 we will further show that this information, i.e. the correlation between the AM and FM rate for a given frequency, can be use to identify singers without computing the exact positions of the formants. In a study on the singing synthesis [MB90], it has also been proven that the correlation between the two modulations has an important effect on the natural aspect of singing synthesis.

Method to estimate the position of each formant Like the problem of envelope estimation, the estimation of formant frequency is rather difficult on high-pitched signals because of the width space between each of the harmonics. On Fig.3.25, we plot the amplitude against the frequency of all partials composing a G4 sung by a female singer on the vowel /a/. Clearly, this plot shows a sampled version of the spectral envelope on a finite number of frequencies.

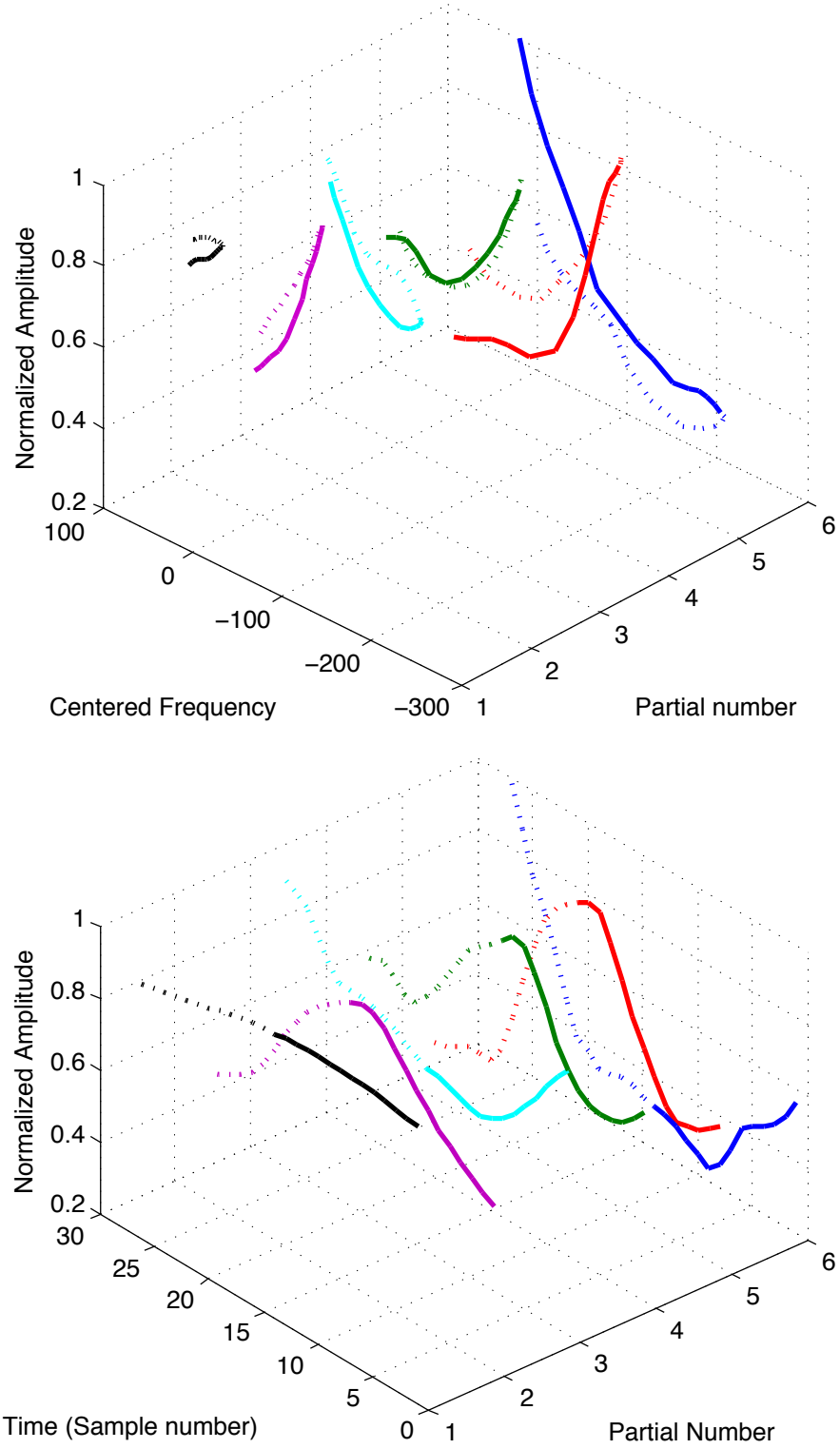


Figure 3.23: Evolution of the amplitude of six partials:
Amplitude .vs. Frequency (top), Amplitude .vs. Time (bottom)

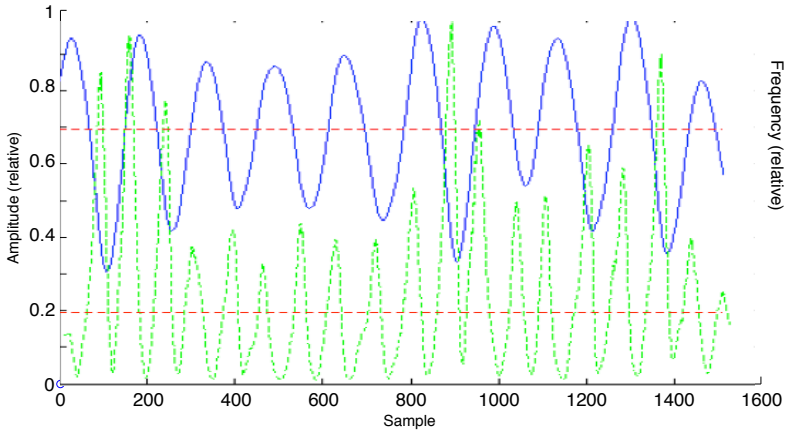


Figure 3.24: Correlation between AM and FM rate. Green dashed line: normalized amplitude; Blue plain line: normalized frequency

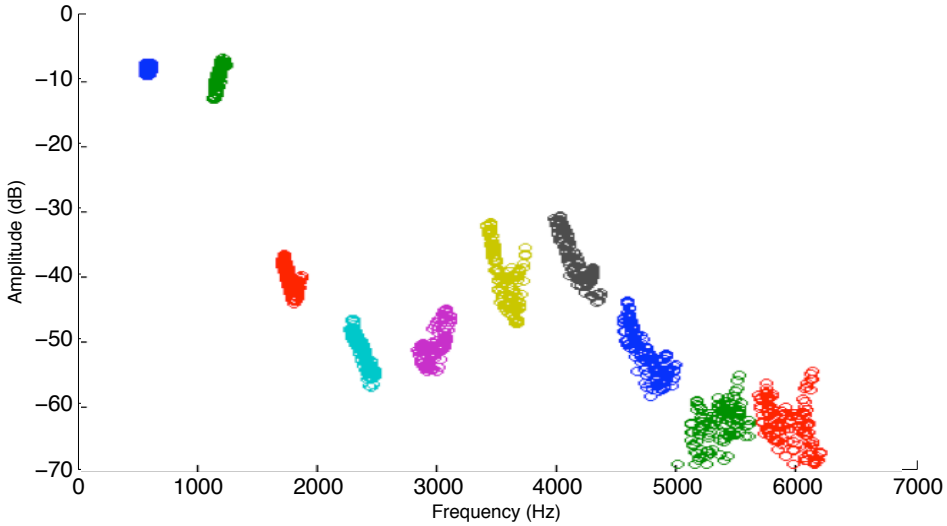


Figure 3.25: Samples of the vocal tract transfer function

With minimum information on the vocal production, the frequency and the bandwidth of the formants can be easily deduced from such a plot. From our experiments, a voice can be re-synthesized accurately using a formant-based synthesizer (such as CHANT [RPB84]) which requires the frequency and the bandwidth of formants as input. As observed before, there are two main situations for the formant estimation:

- **The partial's frequencies crosses the formant:** In this case, the frequency and the bandwidth of the formant can be easily deduced as illustrated in Fig.3.26. The frequency of the formant is simply given by the position of the peak with the maximum amplitude. The bandwidth is measured by first estimating the slope of the formant, and then by finding the point whose amplitude is equal to the amplitude of the formant minus three decibels.

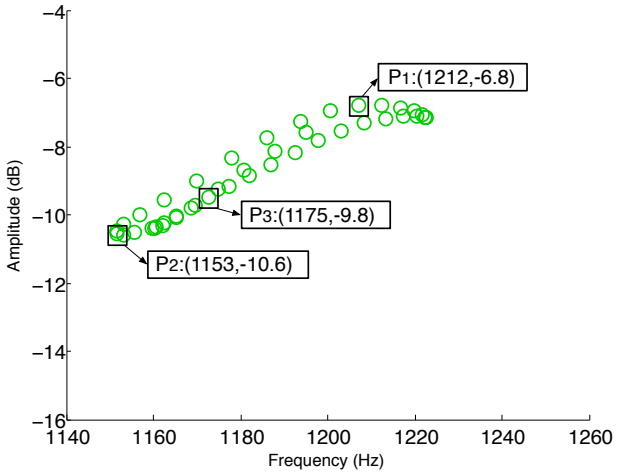


Figure 3.26: Measure of the bandwidth of the formant for the second partial, $\Delta f = 74$ Hz

On Fig.3.26 we have a formant at $P_1 = (1212, -6.8)$ and we can deduce that the bandwidth of this formant is equal to 74 Hz. The bandwidth is equal to twice the distance between the frequency of P_1 and P_3.

- **The formant is missing:** When a formant is comprised between two partials, in most situations the frequency of the formant can be deduced from the slopes of the amplitudes of the partial surrounding the formant as illustrated in Fig.3.27. It is however not possible to measure the bandwidth of these formants accurately. In this case the bandwidth can be set with theoretical values.

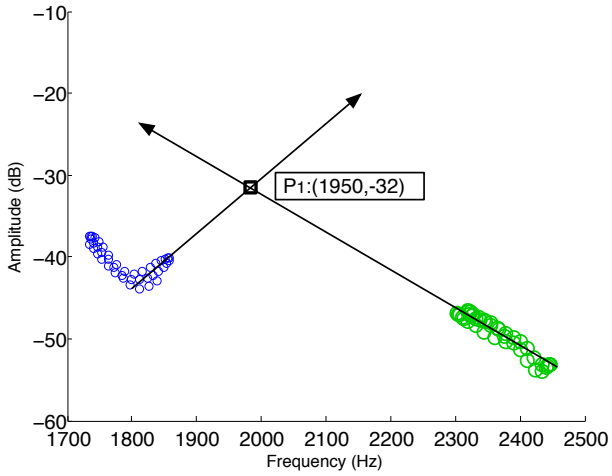


Figure 3.27: Computation of the location of a missing formant

We report the frequency and bandwidth values of the firsts 8 formants for the analysis of the Fig.3.25 in Tab.3.7.

Formant Number	1	2	3	4	5	6	7	8
Frequency (Hz)	595	1212	1950	3462	4035	4430	5535	6153
Amplitude (dB)	-7.85	-9.26	-32	-32	-30.62	-35	-55.15	-54.5

Table 3.7: Formants values for the sound sample illustrated in Fig.3.25

The idea of modeling the vocal tract of a singer using the instantaneous frequency and amplitude of partials has been further exploited in the Composite Transfer Function. This method was first presented in [Mel01] and then improved in [Bar04]. The studies performed by Arroabarren [AC04] go even further and suggest that the vibrato can be used to de-correlate the glottal source and the vocal tract transfer function. In [AC04], the amplitude modulation is presented as the element solving the problem of filtering an inversion encounter on high-pitched signals. This study concludes that the instantaneous amplitude and frequency obtained by sinusoidal modeling provide more, as well as complementary, information about the vocal tract function than known analysis methods.

3 Features for singing voice

In this section we present several features, deduced from the source-filter model and the intonative model, to describe the signals of singing voice. Features derived from the spectral envelope of sounds are referred to as "timbral" features because their purpose is to characterize the timbre of the sound. Features extracted from the time-varying frequency of partials are named "intonative" features.

3.1 Timbral features

In Sec.2.1.2 we presented a set of methods to estimate the spectral envelope. Methods based on linear prediction techniques represent the spectral envelope with a reduced set of coefficients (the poles of the filter) that can be used directly as features. Estimation methods based on the cepstrum lead to an envelope which has the same number of coefficients as the original spectrum. In this section we present different methods that have been proposed to increase the performance of the linear prediction coefficients or to compress the information conveyed by the envelope estimated with cepstrum-based approaches.

Linear Predictive Coefficients and their variants Linear predictive coefficients (LPC) and their variants have been widely used in problems involving speech and singing processing. The main disadvantage of the traditional LPC is that it treats all frequencies equally on a linear scale. Numerous methods have been proposed to compute these coefficients on a frequency scale closer to human perception.

The **Wrapped Linear Predictive Coefficient**, computed on a Bark scale, have been used in [KW02] for a task of singing detection. Similar coefficients have been used for the same task in [BE01] where they are named Perceptual Linear Predictive coefficients (PLPC).

It has also been proposed to derive Cepstral Coefficients from the LPC using a recursion formula [RS78]. The coefficients obtained are named **Linear Predictive Cepstral Coefficients (LPCC)**. This

set of coefficients have been used to detect the singing voice in [Zha03] and to model singer identity in [NSW04]. The LPCC are supposed to more be robust to noise than the cepstral coefficients obtained directly from the spectrum.

Mel Frequency Cepstral Coefficients As explained in Sec.2.1.2 the other major idea we explored to estimate the spectral envelope is based on the cepstrum. In the cepstral representation of a sound (see Eq.3.5) the envelope is estimated using the low cefrencies of the cepstrum. In general, the Discrete Cosine Transform (DCT) is applied to the cepstrum in order to concentrate the most useful information into a set of few coefficients. Thus, the estimated envelope is parameterized by the first coefficients of the cepstrum DCT. Similarly to the LPC, it has been suggested that, by changing the frequency scale, one could enhance the envelope coding. The most popular envelope coding is certainly the **Mel Frequency Cepstral Coefficients**, which are coefficients obtained from the DCT of the cepstrum computed on the Mel scale. This scale is, as the Bark scale used in the WLPC, used to approximate the response of the human auditory system.

The majority of studies in speech and singing processing describe the audio contents (on pseudo stationary segment) with the MFCC or the LPC. It is also common to complete the description by using the first and second derivatives of these coefficients.

Cesptral Coefficients derived from the True Envelope From the last envelope estimation method presented in Sec.2.1.2, the True Envelope estimator, we propose a new set of coefficients to describe the timbre of a sound. As done for the MFCC, the essential information of the true envelope is compacted into a few coefficients by applying the DCT.

Timbral features, in particular the MFCC, have been widely used for speech and musical signal representation. In the case of speech or singing signals, the spectral envelope is supposed to convey information about the vocal tract. Most generally, when the signal is composed of a single source, the spectral envelope conveys information related to the timbre of the instrument. A contrario, is rather difficult to give a clear meaning of spectral envelopes estimated on sounds created by a mixture of instruments. The post-processing effects can modify the overall spectral characteristics of a sound significantly. As a results, two sounds with clearly different timbres, but post-processed with the same effects can have rather similar timbral features. This problem, known as the “album effect” or “produced effect”, is further described in Chap.5.

3.2 Intonative features

The vibrato and the portamento are among the most important features of singing voice. For each sustained note or note transition, the time varying frequency and amplitude are modeled with the method presented in Sec.2.3. This method can be applied on segments of the fundamental frequency covering either a transition between two notes or a sustained note.

On each portion of the fundamental frequency we obtain the following set of coefficients:

- The mean frequency
- The mean amplitude
- 4 real coefficients for the three order polynomial representing the slow frequency deviation

- 4 real coefficients for the three order polynomial representing the slow amplitude deviation
- 3 coefficients (extent, amplitude at the origin and amplitude decay) to describe the sinusoidal frequency modulation of type SEM
- 3 coefficients to describe the sinusoidal amplitude modulation of type SEM

Finally, a set of 16 coefficients is obtained to characterize the varying frequency and amplitude associated with a portion of f_0. Per our analysis, not all these features are relevant in regards to characterizing the voice. The most important criteria are the vibrato (FM) rate and the vibrato extent. The rate of the amplitude modulation is also important. In the evaluation conducted in the next chapter, we will show that the simultaneous presence of the frequency and amplitude modulation is characteristic of singing. Other musical instruments can produce tones with a vibrato (string instruments, some wind instruments) or a tremolo (some wind instruments) but they do not have the two modulations occurring simultaneously. The vibrato and tremolo rate (given with the mean frequency) are features that are singer specific. They depend upon the technique (vibrato rate at a given mean frequency) and the vocal tract (ratio of AM and FM rate for a given mean frequency) of the singer. In Chap.5, we will show that these features can be used to identify singers.

4 Summary and Conclusions

The vocal production can be interpreted as a temporal succession of voice pulses, created by an airflow produced by the lungs and processed by the vocal folds, modified by an instantaneous configuration of the vocal tract. The vocal tract filters the voice source to give a particular vowel quality with a particular color which depends upon the physical characteristics of the singer. The source-filter model formalizes this interpretation of vocal production. The transfer function of the vocal tract is given by the envelope of the amplitude spectrum computed on stationary portions of the signal. There are two main approaches to estimate the spectral envelope: the linear predictive approach and the cepstrum-based approaches. The coefficients given by the linear prediction can be used directly as a compact representation of the vocal tract of the singer. The envelopes estimated via the cepstrum can be compacted using compression algorithms such as the DCT to concentrate the most informative elements into a set of few coefficients. To lead to a representation of sounds closer to what humans hear, the spectral envelope can be computed on logarithmic frequency scales, such as the Mel or the Bark scales. This idea is at the origin of the most commonly used features for sound descriptions: the MFCC. We proposed in this chapter to compute similar coefficients on an envelope computed with the true envelope estimation method. This method has been proven to be more efficient than any other method to estimate the envelope of sounds with high pitches that occur frequently for alto and soprano voices.

The vocal production can also be interpreted as a set of frequencies varying slowly over time whose amplitudes also vary over time. This interpretation is formally given by the sinusoidal model, which decomposes the signal into a sum of partials. Each partial is defined on a support (onset, offset) by a frequency function $f(t)$, an amplitude function $a(t)$ and a phase. If the partial covers a sustained note or a transition between two notes, then the time-varying frequency $f(t)$ can be modeled by the sum of a slow, continuous variation and a periodic modulation. This model allows the interpretation of each $f(t)$ in terms of portamento and vibrato, which are characteristics of singing known to add

expression and help the voice stand out from the musical accompaniment. The function of the vocal tract implies that some frequencies are naturally enhanced. So when $f(t)$ follows a periodic modulation, the associated amplitude function $a(t)$ also follows a regular modulation. The conjoint analysis of $f(t)$ and $a(t)$ applied on all harmonically related partials of a sung sound can be used to obtain information on the vocal tract of the singer. The parameters of the continuous variation and the periodic modulation can be used directly as features to describe some intonative and expressive components of sung sounds.

Finally, for each singing signal we propose the extraction of two types of features. These are features related to the vocal tract of the singer, capturing information on the timbre of the sound. Features obtained on the analysis of the temporal variation of the sinusoidal components of harmonic sounds capturing information related to the style and the technique of the singer.

In the rest of the documents we will use the vibrato as criterion to detect the voice. Since the vibrato also exists in other instruments, we will detect vocal vibrato by analyzing the frequency and amplitude modulations of each sinusoidal component of a sound. Since the vibrato is supposed to be a natural attribute of singing that can hardly be modified voluntarily by the singer, we will examine if the vibrato is a relevant feature to classify singers.

Chapter 4

Singing voice localization and tracking in polyphonic context

A very large portion of the music produced today is composed of songs. A song can be defined as a lead vocal accompanied by a set of instruments. The lead vocal is the element that attracts the attention of most of the listeners. First of all, it carries the main melody line, which is clearly the most memorable element of a song. In addition, the lead vocal, along with the lyrics, conveys the message of the song.

Because of the importance of the lead vocal, numerous investigations have been conducted to develop systems able to extract information related to the singing voice within a song. The most typical examples of these studies involve the following: the extraction of the singing voice (i.e. isolate the voice from the musical accompaniment), the transcription of the sung melody (i.e. write the score performed by the singer) and the identification of the singer. The problem of singer identification is addressed in the next chapter (Chap.5). The present chapter focuses on the general problem of discriminating within the signal of a song the elements emitted by the singing voice from the elements produced by the other instruments. In the rest of this chapter we refer to this problem as *singing voice tracking*. Tracking elements produced by the voice within a song is the first step for the transcription of the sung melody or the singing voice separation. In general, the singing voice is not present throughout the song. Typically, the introduction of the song, the bridges between the verses and chorus and the coda of the song are purely instrumental. So that, much research conducted on the topic of singing voice uses as a first step a system to locate the portions of the song where the voice is present. This task is generally referred to as *singing voice detection* or *vocal segments localization*.

The research presented in this chapter is performed on a simplified scheme of a song. We consider that only one singer is present in the song. In other words, we work with songs where there are no back vocals in the accompaniment and where only one singer performs the lead vocal (with no doubling).

In this chapter we investigate the two following points: the localization of the vocal segments and the tracking of the content produced by the singing voice. First, the two tasks are defined in Sec.1. Then, in Sec.2 we present a set of works related to these two problems. In Sec.3 we present and evaluate a novel approach to locate the vocal portions of a song. The method proposed here is based on the identification of partials likely produced by the singing voice using criterion of vocal vibrato and portamento. This simple method is next extended to develop a system to track all partials

produced by the singing voice and to group partials harmonically related. This method, presented in Sec.4.1, is then used to improve the localization of vocal segments and to transcribe the sung melody when one other instrument accompanies the voice.

1 Problems statement

In the present chapter we conduct investigations on the following points: 1) locate the vocal portions of a song and 2) track the content produced by the singing voice within the signal of a song. The details of these tasks are presented below.

1.1 Singing voice detection

In a song, the singing voice is usually accompanied by other musical instruments. However, there are moments in the song where the voice is not present. Much research that attempts to extract information on the sung melody or on the singer requires knowing precisely when the voice is present or not. This is the overarching goal of the singing voice localization task. In the literature, the terms *singing voice detection*, *vocal segments localization* or also *singing voice localization* are used independently to refer to the same task: segment a song into vocal and non-vocal portions. Vocal segments are defined as portions of a song where the voice is present, whereas non-vocal portions encompass the purely instrumental and silent portions of a song.

Theoretically, the localization and the detection are slightly different tasks. Given a song, a singing voice detection system returns for each subunit (e.g. a frame) an indication of the presence or absence of a singing voice. A system of localization returns the starting and ending instants of each vocal portion of the song. The localization is generally performed after the detection task by smoothing the result of the detection stage.

The problem of singing voice detection is commonly viewed as a problem of classification with two classes: “vocal” and “non-vocal”. There are two main approaches to solving this problem, both based on supervised learning techniques:

1. The singing voice detection can be considered a problem of binary classification. “Vocal” and “non-vocal” classes are modeled using salient features extracted on labeled data and a classifier is trained to recognize and/or discriminate between the two classes. Then, unlabeled songs are automatically segmented into vocal and non-vocal portions by the trained classifier.

2. The singing voice detection can also be viewed as a specific case of musical instrument recognition in a polyphonic mixture. In this case, the system has to discriminate the singing voice among all other possible instruments. The recognition of the individual instrument has no importance and the two classes remain more or less the same as in the previous approach: “singing voice” and “non-singing voice”. Once the singing voice is identified, the vocal portions can be easily deduced.

Methods developed using these approaches are described in details in the section dedicated to the related works (see Sec.2.1 and Sec.2.2). They are considered the most straightforward approaches to solving this problem. It is also possible to deduce the position of the vocal segments from the separated

sources or from the transcription of the sung melody line. To obtain an accurate transcription or extraction of the lead vocal, it is necessary to retrieve within the signals the elements that are produced by the voice. We refer to this problem as "singing voice tracking".

1.2 Singing voice tracking

We define singing voice tracking [1] as the process of finding and following within the signal of a song all elements produced by the singing voice. In our approach to this problem, this is done by finding within the set of all partials extracted from the polyphonic signal the ones that are produced by the voice.

The result of the singing voice tracking can be directly used to re-synthesize the voice (with additive synthesis) or used as a first step for the transcription of the sung melody. In our approach, the harmonic content of the voice is tracked within the signal without using any information on the localization of the singing voice. *A contrario*, we suggest using the result of the singing voice tracking to locate the vocal portions.

The singing voice tracking task, as defined here, can be viewed as a step of a source separation method where two sources ("singing voice" and "instrumental accompaniment") are considered. The discrimination of the different sources in the "instrumental accompaniment" has no importance for the task. A large collection of works has been proposed to solve the problem of source separation. We present in Sec.2.3 a brief review of the source separation methods related to the Blind Source Separation (BSS) and the Computational Audio Scene Analysis (CASA) approaches.

When the singing voice has been isolated, the sung melody can be easily transcribed since it is admitted that techniques developed to find the fundamental frequency of a monophonic source are highly reliable. However, some other approaches have been proposed to transcribe the sung melody (or the main melody line) without separating the sources beforehand. These methods usually track the most prominent melody and suppose that the voice, when present, carries the dominant melody. We note that, in the case of dominant melody transcription, the results cannot be used directly to locate the voice. However, it is still possible to derive from the main melody line a system to detect the presence of voice (e.g. based on the characteristics of the pitch contour of the voice). We review in Sec.2.4 a set of methods proposed to transcribe the sung melody.

2 Related works

In this section we present different approaches that can be used to locate the singing voice or to track the elements of the signal produced by the singing voice. We considered that these two problems are related and can be solved using one or a combination of the following; a vocal/non-vocal detector, a system of instrument recognition, a source separation approach or a system for dominant melody transcription. The result of one of these approaches can be used by another approach to reach the final goal, which is to differentiate between the singing voice and elements produced by other instruments from within the signal elements.

[1] The term tracking in this definition is inherited from the general problem of partial tracking (and has no correlation with real time tracking).

Each of the four problems mentioned above are complex and each of them could be the subject of a dissertation. Therefore, it is clear that the reviews proposed here are not exhaustive and are given in order to provide a general idea of the problems related to the extraction of information from the singing voice of a song.

2.1 Singing voice localization

The first research on singing voice detection was done with tools originally developed in the field of speech processing. In [BE01], Berenzweig proposes to detect the vocal segments with a system originally trained to recognize the phonemes of the English language. Although the singing voice differs from the natural speech, the study is based on the hypothesis that an acoustic model trained on speech would respond in a different manner to singing than to any other musical instruments. For each frame of the signal, a speech recognizer estimates the posterior probability for each possible phoneme. The posterior probabilities are then used as new features. The method supposes that when the singing voice is present and pronounces one of the phonemes, then one posterior probability is high and all other posterior probabilities are relatively low. Conversely, when the voice is not present, all posterior probabilities are more or less similar. The difference between the two situations is modeled with an HMM (Hidden Markov Model) with two states. On a test-set composed of 245 audio clips with length of 15 sec, they obtain 81.2% accuracy for vocal segments. This method is, as far as we know, the only one based on a speech-processing tool.

The general approach to detecting the voice is to train a classifier to recognize the specificities of the "vocal" and "non-vocal" classes. In most cases, the classes are modeled using features computed on the amplitude spectrum of stationary portions of the signal (frame). The underlying idea of this approach is based on the observation that there is a significant difference in the spectral distribution of vocal and non-vocal portions. This approach has been successfully employed in many problems of audio classification including speech recognition, speaker identification and has also proven to be well suited for numerous MIR tasks. This approach is a frame-by-frame singing voice detection system. In most cases it provides a rapidly alternating output (detection function) that is meaningless for the application. To obtain coherent vocal segments, the detection function can be smoothed out using filtering methods. It is also possible to first decompose the signal into portions with common spectral characteristics and assign each portion to one class using voting rules on the class labels of the frames within the portions.

We present here a set of works conducted on the singing voice detection task. We first present the methods based on "timbral" features (i.e. features derived from the spectral envelope of stationary audio segments). Then we present some methods that have proposed features specific to the singing voice: the variation of energy and the harmonicity of the voice. The methods listed below are evaluated from different data sets and use different measures to report the performance (in general the F-measure of the vocal class or the global accuracy). Therefore, these methods cannot be compared with one another directly.

The most common approach for audio classification is to model the classes with GMM trained on MFCC and then classify each frame of the song using a maximum likelihood criterion. This idea has been explored by Li in [LW07] where the singing voice localization is processed as a first step for the singing voice separation. In this approach, the signal is first partitioned into spectrally homogenous

portions by detecting significant spectral changes. Then each portions is classified as vocal or non-vocal, based on the overall likelihood given by GMM trained on MFCC. On a test set of 10 songs, the proposed approach reaches 86.65% accuracy.

This combination of MFCC and GMM has also been used by Lukashevich [LGD07]. In this study, the frame-by-frame classification is filtered with an ARMA filter. The idea is based on the fact that the presence of voice tends to be continuous over consecutive frames, so that the presence of voice in frame i can be partially determined by the presence of voice in the k previous frames (AR model). In addition, the short-term outliers are removed by smoothing the decision function with a moving average filter (MA). On a set of 10 songs, the method reaches an accuracy of 72.7% before filtering. The accuracy is improved to 82% after ARMA filtering .

GMM trained on MFCC have also been used by Tsai [TRW04] where the singing detection is performed as a first step for singer identification. On a test-set composed with 416 mandarin pop music songs, the best accuracy achieved is 79.8%.

In the approach proposed by Berenzweig [BEL02], where the singing detection is performed as a first step for singer identification, the discrimination between the two classes is done with a multi-layer perceptron fed with a variant of the LPC (the perceptual linear predictive coefficients). On a test-set of 80 clips of 15 second length the proposed method obtain a F-measure of 40% for the vocal class.

In a comparative study [RH07] of the performance of features and classifier for the task of singing detection, Rocamora finds that the best performance is obtained with the SVM classifier trained on MFCC (78.5% vocal detection accuracy for 46 songs). The study shows that the variation of performance is highly influenced by the choice of the features.

An SVM classifier trained on spectral features has also been used by Ramona [RR08]. In this study, three post-processing methods to smooth the singing voice detection function are compared: Median filter, HMM trained on the duration of vocal and non-vocal segments and an heuristic rule based on a fixed length for segments. On a set of 90 songs, the best accuracy (82.2%) is obtained with the HMM based post-processing, however there is not a large difference with the median filter method.

In [Zha03] Zhang proposes a method to detect the starting point of the voice using adapted features: energy, spectral-flux, harmonicity and zero-crossing rate. Energy and spectral-flux are used to detect the high variations in the signal created when the voice starts. The average zero-crossing rate (ZCR) is used to detect consonants. Onsets are detected by comparing the values of these features to a set of predefined thresholds.

This harmonicity criterion used in [Zha03] was first developed by Chou [CG01] for the problem of speech and music discrimination. According to [CG01], the higher harmonics are stronger for the singing voice than for any other instruments. They define a harmonicity coefficient, given by the maximum value of a linear combination of the temporal and the spectral auto-correlations, which is supposed to have higher values when the singing voice is present.

Numerous studies have investigated the harmonic character of the voice for its detection. In the study proposed by Kim [KW02] the voice detection is performed as a first step of a system of singer identification. For this application, the recall of the method can be neglected in favor of the precision. In this study, the signal is band-pass filtered to conserve only the frequencies covered by the vocal

tessitura ([200-2000 Hz]). This is achieved with an IIR digital filter. To remove the frequency components related to the other instruments present in the filtered signal an inverse comb-filter is applied to find the frequency at which the signal is most attenuated. A harmonicity coefficient is then computed on each frame as the ratio of the signal's energy to the energy of the most attenuated signal. The assignment of each frame into a vocal or non-vocal class is done with a simple threshold. In the best case 55.4% of the vocal portions of the song are correctly classified on a test-set composed of 20 songs.

Another technique, named twice-iterated composite Fourier transform technique (TIFCT), has been proposed by Maddage [MWXW04] to detect the singing voice. The idea of the TIFCT is also related to the harmonic character of the singing voice. They show that the cumulative energy in the lower coefficients of the TIFCT is capable of differentiating the harmonic structure of vocal and instrumental music in higher octaves. Finally, a set of thresholds is applied on these coefficients to detect the frames where the singing voice is present.

Shenoy [SWW05], also proposes a harmonic coefficient. Like [KW02], the frequencies related to a given fundamental (in this case, the tonality of the song) are filtered with a series of inverse comb-filters to remove the majority of the components created by the pitched instruments of the song. They hypothesize that the energy of the harmonic components corresponding to the singing voice is not entirely removed because the fundamental frequency associated with the voice always varies through time because of the vibrato and intonation. Finally, the regions where the energy of the filtered signal is higher than a certain threshold are classified as vocal. The process is applied on 4 distinct frequency bands of the signal and the decision is made by considering the decision on all bands. Finally, on a set of 10 songs they obtain a Recall and a Precision of 89.44% and 77.37% for the vocal class.

2.2 Instrument identification

The detection of the singing voice in a polyphonic mixture can also been considered a special case of instrument detection (or instrument identification). The problem of instrument identification has been widely investigated in the Music Information Retrieval field. Early methods of musical instrument identification focus on isolated notes identification performed on solo instrument recordings. More recent studies have proposed some methods to perform instrument recognition in polyphonic recording.

In the following paragraphs we present a brief overview of methods developed for the recognition of instruments in solo (Sec.2.2.1) polyphonic recordings (Sec.2.2.2). There are very few studies that consider the voice as one of the instruments to be identified. However, methods and features used for solo and multiple instruments recognition can be adapted for signals containing a singing voice.

2.2.1 Solo instrument identification

The most common approach to recognizing an instrument in a monophonic signal is based on supervised learning methods. During a training phase, each instrument of the data set is modeled and a classifier is trained to learn the specificities of each class. Numerous methods, such as the methods developed for the singing detection use features derived from the amplitude spectrum and the spectral envelope.

Marques and Moreno [MM99] evaluate the performance of GMM and SVM classifiers trained on LPC, linear spaced cesptral coefficients and MFCC for solo instrument identification. MFCC features yield the best performances and LPC the worst performances. In general, SVM based classifications obtained better performance than GMM-based classification. The best performance, on a set of 8 instruments, is found to be 70%. It is shown that the choice of the features has a higher influence on the overall accuracy than the choice of the classifier.

The importance of features, and the feature dependence, has also been studied by Brown *et al.* [BHM01]. They examine the performance of cepstral features based on constant-Q coefficients (previously used in [Osh87]), spectral smoothness (derived from the constant-Q coefficients), spectral centroid, average energy and LPC for a kNN classifier. On a set of 4 woodwinds instrument, the spectral smoothness obtains the best performance (84% of correct classification). The authors suggest that this result is due to the fact that the instruments studied have distinct formant structures.

Martin and Kim [MK98] have conducted an evaluation on a large set of orchestral instruments. In their approach, the instrument identification is based on a hierarchic classification inherited from the taxonomy of musical instruments. The instrument's family is first identified, and then an instrument is identified within its family. The classification at each stage is performed by a kNN. Numerous features are examined included the parameters of vibrato, tremolo, pitch variations, spectral centroid and harmonic skew. Some of these features require the transcription of the pitch beforehand. The study shows that the parameters of vibrato and tremolo are salient information for the discrimination of the family of an instrument. For a set of 37 orchestral instruments, they obtain 71.6% accuracy for instrument recognition (with 90% of correct family recognition).

The use of the taxonomy of musical instrument has also been examined by Eronen and Klapuri [EK00] on the same data set as [MK98]. In this study, a very large number of features, independent of the pitch, are assessed. Contrary to [MK98], the best classification accuracy (80%), is obtained with a direct classification (i.e. without the taxonomy).

A comprehensive review of the features and the statistical classifiers used for the recognition of musical instrument as been proposed by Herrera [HBPD03].

The next step for instrument identification is the extension of this technique to polyphonic recordings. The recognition of a specific instrument in a polyphonic mixture is close to the task of singing voice detection.

2.2.2 Multiple instruments recognition in polyphonic recordings

The problem of simultaneous recognition of instruments is much more complex than the solo instrument identification. For instance, on a mono source signal, information obtained from the spectral envelope has a rather clear meaning. In contrast, it is difficult to interpret the information derived from the spectral envelope of a mixture of instruments. In the case of tonal harmonic music, which represents a large portion of the musical production, the different instruments play in harmonic accordance with the tonality of the musical piece. As the result, the harmonics of the tones played simultaneously coincide on multiple frequency regions. The values of the amplitude spectrum on these regions are then corrupted by the interfering sounds. It is rather complex to de-correlate the contribution of each instrument on these frequency bins and obtain an estimation of the spectrum associated to each source. Next we present some alternatives we propose to solve this problem.

The first method for multi-instrument recognition, proposed by Kashino and Murase [KM99] relies on a set of templates computed in the time domain. During a training phase, a template waveform for each note of each possible instrument is computed. Then, for a mixture composed of two instruments, the tone and the identity of the most prominent instrument are estimated by comparing the mixture to the predefined templates. The energy of the corresponding waveform is then subtracted and the process is iterated with the second instrument. When the correct f_0 of the two instruments is given as input to the system, the method achieves a 68% correct classification rate. The study is limited to the 3 following instruments: flute, violin and piano. The recognition accuracy is improved to 88% when higher musical knowledge (e.g. voice leading rule) is considered.

A spectral approach has been proposed by Kinoshita *et al.* [KST99]. In this study, the fundamental frequencies of the two instruments of the mixture are estimated to determine the frequency regions where partials of the two instruments coincide. Features related to the sharpness of onset and the spectral distribution of partials are computed on the mixture and the values corresponding to the overlapped frequency regions are ignored in the recognition process. It is also proposed to subtract the values corresponding to the first instrument and keep the remaining values for the recognition of the second instrument. The classification performance of the proposed approach ranges from 66% to 75%. The performance varies with the interval between the two tones.

The suggestion to discard or modify the features corresponding to frequency regions where harmonic partials of concurrent sounds coincide has also been examined by Eggink and Brown [EB03]. The proposed approach is based on the missing feature theory that had been successfully applied in the field of speech and speaker recognition [CGJV01]. Like [KST99], the features obtained on frequency regions where interfering coinciding tones are marked as unreliable and are excluded in the classification process. The study shows that cepstral features computed on a non-linear scale are not compatible with the missing feature theory since frequency regions rarely have clear counterparts in the cepstral domain. The evaluation is performed on real recordings. The performance of the method varies with the instrument: the flute is recognized with 73% accuracy while the accuracy for the clarinet falls to 47% accuracy.

A completely different approach has been proposed by Essid *and al.* [ERD06]. Contrary to most methods presented above, this system does not rely on note model or pitch estimation. The system is trained to recognize the combination of instruments directly. This idea is closer to the approach presented in Sec. 2 for the singing voice detection. The system is trained using temporal features (Auto-correlation, zero-crossing rate, amplitude modulation), cepstral features and spectral features (spectral width, spectral asymmetry, . . . , spectral slope, frequency derivative of the constant-Q coefficients). Since the number of combinations is infinite, they start by generating taxonomy of instrument combinations using a hierarchical clustering. A set of SVM classifiers are then trained on each node of the taxonomy, each of them is trained with a specific set of features determined by a pair-wise feature selection algorithm. The average accuracy for all combinations of instruments present in is found to be 53%. This study includes the percussion instruments and the singing voice in the classes of the problem and deal with the recognition of instrument of jazzy orchestration ranging from a solo to a quartet.

Other methods for multi-instruments recognition are based on source separation methods. Vincent [VR04] proposes a method based on an adaptation of the Independent Subspace Analysis (ISA) to

identify instruments of an excerpt of a duet. In ISA, the short-term spectrum of a polyphonic sound is represented as a weighted linear combination of "note spectra" plus a background noise. Note spectra, which are templates of notes in the frequency domain, are learned on solo instrument recordings for each possible note and each possible instrument. The goal of the ISA method it to span the best subspace for each original source and decompose the signal into a base of note spectra. The procedure uses the models of note spectra to perform the identification and the transcription of each source simultaneously. The study does not report quantitative results on multi-instrument recognition, but the authors claim that the proposed approach has a performance comparable to existing methods on solo excerpts and is robust to degradation under reverberant conditions. A similar idea has been explored by Jincahitra [Jin04].

Martin *and al.* [MBTL07] also proposed a method for instrument recognition based on a source separation approach which does not requires the multi-pitch transcription. In this approach the source separation detects spectral peaks and groups them into clusters using cues inspired by CASA (these cues will later be explained in Sec 2.3). Finally each cluster is matched to a collection of timbre models, which are compact descriptions of the spectral envelope (and its temporal evolution) of each instrument. These models are trained on solo recordings. The proposed approach is evaluated on 54 audio files with up to 4 notes played simultaneously by different instruments. The average F-measure for all instruments through experiments with 2, 3 and 4 notes is equal to 60%. The performance decreases with the number of notes: 77% for 2 notes, 50% for 4 notes.

Heittola *and al.* [HKV09] propose a method for multi-instruments recognition based on source-filter models and non-negative matrix factorization to separate the sources of a polyphonic mixture. The mixture is decomposed into a sum of spectral bases modeled as a product of excitation and filters. The instrument recognition is then performed on solo source using MFCC as features and GMM as the classifier. The method is evaluated on polyphonic signals randomly generated from 19 instruments classes. For signals composed with 6 simultaneous sources the recognition rate reaches 56%.

In this section, we have seen that the recognition of instruments in polyphonic context requires either the pitch transcription of each source present in the signal or the separation of the mixture into sources. In the next sections we investigate these problems in greater detail in the specific case of singing voice. In Sec.2.3, we present a series of works conducted on the separation of the voice from the musical accompaniment. The problem of singing pitch transcription and main melody transcription are presented in Sec.2.4.

2.3 Singing voice extraction using source separation approach

The music source separation is the problem of extracting each instrument from a polyphonic mixture. This problem has been widely investigated in the last decades because it is the key technique in applications such as automatic music transcription, lyrics recognition and alignments, remixing and audio content analysis. In the case of signals of tonal music, the different instruments play tones in a harmonic ratio that makes the separation task more difficult for the same reasons as the ones mentioned in the recognition of multiple instruments. The specific case of singing voice separation assumes that the signal is composed of two sources only: the singing voice and the instrumental accompaniment.

We present three categories of singing voice separation methods applied on monaural (single channel) recordings:

1. The Blind Source Separation methods, whose aim is to be data-driven and to adapt thanks to a criterion of the independence of the different sources.

2. The statistical modeling methods, which are based on the adaptation of singing voice and accompaniment models.

3. The methods inherited from the Computational Auditory Scene Analysis (CASA) that are based on signal processing and psycho-acoustic elements. These methods attempt to separate the sources using a set of simple and coarse cues. A comprehensive review of these methods, in the general case of polyphonic signal source separation, can be found in [LL09].

2.3.1 Blind Source Separation (BBS) approaches

The most popular approaches for BBS are the Independent Component Analysis (ICA) and the Non-negative Matrix Factorization (NMF) methods that were originally developed to decompose the source of stationary signals. Numerous attempts have been made to adapt these methods to non-stationary signal, such as the signal of singing voice.

Vembu and Baumann [VB05] propose a method for separation of vocals based on ICA for monaural recordings (i.e. signals with a single channel). The method first detects the vocal segments, then for each frame of the vocal segments, the redundancy information of the time-frequency representation of the signal is reduced by PCA. Then, ICA is applied to decompose the remaining information into a matrix of spectral bases (that are elements that occur many times detected by the PCA stage) and a matrix of time-varying gains (where each spectral base corresponds to one gain). The spectral bases are then clustered into vocal and non-vocal classes using a set of features known to differentiate between these classes. The stage performed with the ICA can be replaced by a NMF.

The NMF method, first used for audio source separation in [WP05], has been shown to give good results without any prior information about each source. The NMF technique has been used in the case of vocal separation by Chanrungutai *and al.* [CR08]. In their approach, the STFT of the signal is computed and analyzed to initialize the NMF input (i.e. the non-negative matrix). Then the signal is decomposed with NMF to obtain a matrix of spectral bases and a matrix of gains. This stage is followed by a manual selection of the spectral bases belonging to the singing voice. The selected elements are then used to recompose a spectrogram containing components related to the singing voice only. The inverse DFT is computed on each frame of the obtained spectrogram to re-synthesis the voice.

The system proposed by Virtanen *and al.* [VMR08] is also based on the NMF. The proposed approach relies on the singing voice pitch extraction obtained with the method presented in [RK06]. On vocal segments, the harmonic components related to the singing voice melody are masked to obtain an approximation of the musical accompaniment in vocal segments. NFM is applied on the results to obtain a model for the musical accompaniment. Like [CR08], the elements corresponding to the accompaniment are removed and the isolated voice is obtained using spectrogram inversion re-synthesis technique.

The NFM approach has also been used by Durrieu to extract the vocal line. In [DRDF10], the NMF technique is combined with a source-filter approach. The singing source, which is supposed to be dominant in the mixture, is characterized through a source-filter model that allows a first estimation

of the pitch contour. This estimation of the melody is then processed by the Viterbi algorithm to improve the quality of the melody tracking. The results of the Viterbi algorithm can be used to re-estimate the singing source as proposed in [DRD09]. The results of this method are very promising, as proven by the results of the MIREX'08 evaluation.

2.3.2 Statistical Modeling approaches

Tsai has proposed a method based on statistical modeling. In [TW06] a model of a solo singing voice is deduced from two GMM models: a music GMM and accompanied-vocals GMM. Both models are trained respectively on non-vocals and vocal segments estimated in a previous step. The solo voice model is obtained by attenuating the musical background of the accompanied-vocals model. This step is realized by subtracting the purely instrumental model to the accompanied-vocals model. It supposes that the instrumental part is more or less similar during purely instrumental and vocal sections of the song.

A similar approach has been developed by Ozerov *and al* [OPGB05]. Their system requires the use of training data consisting of solo vocal recordings to train a Bayesian model for the singing voice. Similarly, a set of instrumental tracks is used to train a model of instrumental accompaniment. For a given song, the non-vocal portions are used to adapt the accompaniment model to better fit the actual instruments present in the input signal. The obtained model, in conjunction with the singing model, is then used to separate the vocals.

The method proposed by Raj and Smaragdis [RSSS07] is also based on pure statistical modeling, but uses an approach close to the spectrogram factorization. The signal is decomposed into time-frequency components and each component is assigned to either a vocal or accompaniment class using models trained on both solo voice and purely instrumental recordings.

The three methods mentioned above ([TW06], [OPGB05], [RSSS07]) are based on the hypothesis that, on a given song, the accompaniment is similar on vocal and non-vocal portions.

2.3.3 Computational Auditory Scene Analysis (CASA) approaches

CASA methods find their origins in the perceptual works of Bregman [Bre90]. One of the Bregman's motivations was to build systems, based on humans' auditory models, able to hear like humans. Compared to other sound separation approaches, CASA makes minimal assumptions about concurrent sounds. Instead it relies on the intrinsic properties of sounds and therefore has greater potential in regards to singing voice separation on monaural recordings.

Approaches inspired by CASA can be summarized as follows. 1) The signal is segmented into units, which likely originate from a single source. These units can be time-frequency components of the spectrogram (or cochleagram), spectral peaks obtained on each frame of the signal, or directly partials (or *strands*) obtained with the harmonic sinusoidal model. 2) These units are grouped, using ASA principles, to form "auditory events" produced by a single source. 3) Then, events are followed through time to form "source events". The most common rules for CASA are listed bellow:

- Harmonic concordance
- Spectral proximity (closeness in time and frequency)
- Spatial proximity

- Synchronous changes of the following components
  - Common onsets and offsets
  - Common frequency variations
  - Common amplitude variations
  - Equi-directional movement of spectrum

These elements have been described by Bregman as *cues* for simultaneous integration (because they are short-time cues that can be thought of as slices in the same short interval). They have been proven to be significant for "events" formation in the psychoacoustic and neurophysiologic area. [1]

A complete description of evidence for each cue can be found in the detailed work of Mellinger [Mel92]. The system proposed by Mellinger groups partials produced by the same instrument using common onset and common frequency variation as cues.

Li and Wang [LW07], propose an approach for singing voice separation based on the CASA framework: segmentation and grouping. The vocal portions of the song are decomposed into time-frequency (T-F) units (time frame and frequency channels given by an auditory filter bank). The estimated pitch contours of the singing voice are used as a cue to label each unit as singing-dominant or accompaniment-dominant. The T-F units are merged into segments using correlations of features extracted on units. Finally, the segments where the majority of units are labeled as singing-dominant are grouped to form the stream associated with the singing voice. This stream is then re-synthesized to obtain a signal of a solo singing voice.

This method, which is able to separate the voiced sounds of the singing voice, has been improved by Hsu and Jang [HJ10a] by adding the separation of unvoiced segments of singing. The signal of a song is on one hand decomposed into accompaniment, singing voiced and singing unvoiced segments using an HMM. On the other hand, the spectrogram of the signal is decomposed into T-F units using the same settings as [LW07]. Each T-F unit is labeled as accompaniment, voiced or unvoiced dominant using jointly the result of the segmentation stage and the contours of the singing pitch. Voiced and unvoiced units are re-synthesized independently and finally combined to form the separated singing voice stream. We remark that the method does not use specific rules to group T-F units.

Lagrange *and al.* [LMMT08] also propose a method for dominant source extraction based on ASA principles. In this approach, the problem of grouping elements of the spectrogram per source is casted into a graph cut formulation using the normalized cut criterion. The method uses spectral peaks as underlying representation of the input signal. The partial tracking and the source formation are jointly optimized by defining a measure of similarity between peaks based on CASA principles (amplitude and frequency proximity plus harmonicity). Groups of peaks related to the same source are automatically formed by applying the normalized cut criterion on the peak similarity matrix. The overall peak similarity within a cluster is maximized and the similarity between clusters is minimized. The clusters associated with the most dominant peaks are then re-synthesized to obtain the isolated voice.

[1] Psychoacoustic studies have shown that these features are an important part of the scene analysis process. Neurophysiological evidences shows that there exist neurons in the auditory system that respond to a certain feature. This evidence suggests that these features are filtered fairly early in the auditory pathway. Then they can be used as information for event and source formation.

Bartsch [Bar04] developed a system named PESCE to group partials produced by the same instrument in order to isolate the lead vocal. The method groups partials with common frequency modulation into sets of harmonically related partials named *harmonic complexes*. The method is based on a measure of similarity between two partials given by the normalized correlation of the time-varying frequencies of the partials computed in the log domain. The maximum score obtained between a given partial and each partial composing the harmonic complex gives the similarity of a harmonic complex and a partial. A greedy algorithm, applied on these measurements, is used to group harmonically related partials automatically.

The idea of grouping partials directly has also been investigated by Wang [Wan94]. In their approach, the fundamental frequency of the singing voice is given as a prior. Partials harmonically related to this f_0 are grouped using a technique named *harmonic-locked loop*. The system does not distinguish singing voice from other musical sounds, i.e. when the singing voice is absent the system incorrectly tracks partials that belong to another harmonic source. A comprehensive review of CASA methods for musical signal separation is given in [WB06].

When the vocals are isolated, the transcription of the sung melody can be easily performed since pitch transcription of monophonic signal is considered a solved problem. Reciprocally, numerous approaches for singing voice extraction rely on the estimation of the sung melody. Singing voice separation and sung melody transcription can be considered as "chicken-end-egg" problems. Numerous alternatives have been developed to transcribe the sung melody without separating the sources beforehand. These methods are presented in the next paragraph.

2.4 Singing melody transcription

The task of singing melody transcription has attracted a lot of attention because it is the first step of numerous applications such as query-by-humming recognition, detection of copyright, recognition of cover versions and extraction of singing voice.

In this section we review some methods proposed to transcribe the singing melody that are not based on source separation methods. In several studies, it is assumed that the most prominent pitch corresponds to the singing voice on vocal segments. Thus, the transcription of the singing melody can be realized with methods for dominant melody transcription applied on pre-determined vocal segments. It is also possible to transcribe the dominant melody on the whole song and then to identify the portions corresponding to the voice using characteristics of the pitch contours of singing.

A review and a comparison of numerous methods for singing pitch transcription have been proposed in [PEE$^+$07b]. This paper defines a set of measures to evaluate the performance of singing melody transcription. Goto [Got04] proposes to estimate the melody and bass line of pop and jazz music. The pitches of each melody are searched in limited frequency ranges. Using adaptive tone models within a multi-agent architecture, the proposed approach obtains about 88% of pitch accuracy for the singing melody line. In this work no attempt was made to distinguish between vocal and non-vocal sections.

Eggink and Brown [EB04a] propose a method to transcribe the melody performed by the soloist instrument of a musical piece. They take into consideration that fact that the solo instrument does not always produce the strongest f_0. This is so that they perform a multi-f_0 tracking and then apply a system for instrument recognition ([EB04b]) to determine which f_0 corresponds to the soloist's

instrument. The method supposes that the soloist instrument is known a priori. Information on the tessitura of this instrument is added for the f_0 selection. Additional high musical knowledge is used to find the most likely path between the f_0 candidates. Using two short excerpts of classical music, on a frame-by-frame evaluation, they obtain a pitch accuracy of 40% when the f_0 is given by the dominant melody. The performance is improved to 54% by adding the system for instrument recognition. We note that the improvement given by the knowledge of a priori of the tessitura can probably not be obtained for singing voice because the range of frequencies covered by the singing voice is very large.

The method proposed by Fujihara *and al.* [FKG+06] to track the singing melody is also based on a multi-pitch estimation. For each f_0 candidate, they evaluate the vocal probability of the harmonic structure. This is done by re-synthesizing the harmonic components related to each f_0. On the synthesis obtained they extract cepstral (MFCC and derivatives), spectral features (LPC and derivatives) and features related to the temporal variation of f_0. The vocal probability is then estimated by comparing the features to vocal and non-vocal GMM. Once the most likely vocal f_0 have been found, the vocal melody is tracked using the Viterbi algorithm. They compare the performance of their approach with the performance of the method proposed by Goto [Got04]. The evaluation is done on a set of 21 songs performed by 14 different singers (from the RWC Music Database). They obtain a pitch accuracy of 84.3% with their method, compared to 78.1% with the method of Goto.

The research conducted by Li in [LW05] is specially adapted to detect the pitch contours of the singing voice based on the assumption that the pitch contour of the singing voice tends to be relatively piece-wise constant (in the sense that notes are sustained on a rather long duration). The signal is first filtered by an auditory periphery and a correlogram is computed to extract the information of periodicity on each channel. Then, channels and peaks are selected using a statistical model to retain only the useful information on periodicity. Finally an HMM followed by the Viterbi algorithm are used to model the pitch generation process and track the most likely pitch track. In this study they compare the performance of their method with the multi-pitch approaches developed by Klapuri [Kla03] and Wu [WWB03]. They conduct the evaluation on 25 short excerpts from 9 songs and report an error rate of 16.2% for their method compared to 45.3 % and 44.3% for [Kla03] and [WWB03] respectively.

The method proposed by Cao *and al.* [CLLY07] takes into consideration the fact that the singing melody can be dominated on some frames by non-singing intrusions and thus local information obtained on frames is corrupted. For each local dominant f_0 they detect the associated harmonic structure. These structures are analyzed forward and backward along frames to determine which f_0 corresponds to a singing pitch. The method, evaluated on 13 songs, obtains a pitch accuracy of 79.39% compared to 74.12% obtained with the predominant pitch transcription.

The method proposed by Ryynanen and Klapuri [RK06] estimates the multi-f_0 and the note onset (accent detector) on each frame of the signal. These elements are given as input to an HMM to model "note events". Elements extracted on each frame are also used as input of a source separation algorithm. For each source obtained, cepstral features are computed to feed a GMM trained to recognize the presence/absence of a singing voice. The two elements (HMM and GMM) are used to model low-level acoustic information. They add another level to model higher-musical knowledge (tessitura of singing, musical key) to choose between note transition probabilities. Finally, information from the low-level acoustic model and high level musicological model are given as input of a Viterbi decoding system to find the optimal melody track. The melody obtained is transcribed into MIDI for

the evaluation. On a test-set of 96 songs (using a 3 fold cross validation) they found a recall (% of MIDI note correctly transcribed) equal to 63% when using low and high level information. Using only the low-level acoustical information they obtain a recall of 56%. The transcription into MIDI can be very adapted for applications such as "Query-by-humming". This representation has the advantage of being very compact, however it looses all information related to the intonation of the singer.

The method proposed by Salamon for singing melody transcription relies on a time-pitch salience function defined in [Sal08]. For each spectral peak of short-term spectrum, the salience is computed as the sum of the weighted energy of the spectral peak found at integer multiple (harmonics) of the given frequency. Peaks with the maximal salience are selected and connected across frames to form pitch contours. These contours are finally selected, using criterion on the trajectory of the frequency, to form the singing melody. The proposed approach obtains the best results (85%) at MIREX'11 [SG11].

The method proposed by Hsu and Jang [HJ10b] starts by removing the percussive elements of the song. On the remaining harmonic contents, they discriminate the singing partials from the partials produced by other instruments with the method we proposed in [RP09]. Once vocal partials have been identified, they evaluate the harmonic structure associated to the lowest partials of each frame to determine vocal f_0 candidates. On the other hand, a normalized sub-harmonic summation (NSHS) map is used to enhance the harmonic content of the associated spectrogram. By combining these elements, they propose an accurate method to estimate the singing pitch trend. This method has obtained the highest performance of the MIREX 10 contest. The average accuracy across all data sets tested reaches 82.72%.

We present in the next paragraph the method we have developed to identify partials emitted by the singing voice. This method has been originally proposed to locate the vocal portions of a song.

3 Proposed approach for singing voice detection

In this section we offer a very simple approach to detect the singing voice in a polyphonic mixture. Like the methods presented in Sec.2, the proposed method relies on the discrimination of two classes: vocal and non-vocal. Contrary to the methods presented in Sec.2, which discriminate the classes based on the distribution of their spectral content, the suggested approach performs the discrimination using the temporal variations of the partials frequency. Vibrato, tremolo and portamento parameters are computed for each partial, and a partial is assigned to the vocal class if the values of these parameters correspond to a vocal vibrato or a vocal portamento.

This approach can be viewed as a system of specific instrument recognition in a mixture. As shown in Sec.2.2 the vibrato has been successfully used to discriminate family of instruments. Contrary to instrument identification methods based on vibrato, which are able to recognize instrument on monophonic recordings [MK98], the proposed method detects the singing voice in mixtures directly. The detection is done by the recognition of partials emitted by the voice. Thus, the proposed method can also be viewed as variant of the CASA methods, since it searches in the signal components related to a given source. However, in our case, the problem is simplified. We do not attempt to group the partials to form note events.

Each step of the proposed approach is described in detail in Sec.3.1. We present in Sec.3.2 the

results obtained with the method for the task of singing voice detection.

3.1 Description of the proposed approach to localize vocal segments

The proposed approach for vocal segments localization goes through four steps. For each step we discuss the method utilized to choose the parameters.

3.1.1 Step 1 - Partial tracking and segmentation

The signal is given as input of a partial tracking algorithm. The problem of partial tracking, described in Sec.2.2, can be summarized as follows: on each frame, sinusoidal components are detected by selecting peaks of the magnitude spectrum given by the STFT of the signal. The detected peaks are connected across frames to form coherent partials' frequency and amplitude trajectories. Finally, a partial p is characterized by:

- a support given by the starting and ending frames: t_{beg} and t_{end}
- a vector of time-varying frequency: $f_p(t)$
- a vector of time-varying amplitude: $a_p(t)$
- and a vector of phase: $\phi_{0,p}(t)$ (representing the phase at the beginning of each frame)

In practice, we want the system to accurately track partials emitted by the singing voice. These partials may have large variations in frequency. Therefore, it is necessary to set the parameters of the tracking algorithm to allow the connection between peaks with a rather large difference in frequency. However, if this distance is too large, the system starts merging frequency trajectories corresponding to distinct notes coming from different instruments. In pop-rock music, the harmonic structure of a song follows simple harmonic rules. As a result, it occurs frequently that partials of different instruments overlap vertically (notes with common harmonic) and horizontally (a note played by an instrument is repeated by another instrument on the next beat).

In this work, partials are extracted using pm2 IRCAM's software with the parameters summarized in Tab.4.1. These parameters, found empirically, provide an accurate tracking of partials with large frequency variations.

For the next steps, it is necessary that each partial either covers a single note emitted by a unique instrument, or covers a note transition. This is so that partials are segmented with the methods presented in Chap.3, Sec.3.2. The rest of the method is then applied on segmented partials.

3.1.2 Step 2 - Extraction of features

For each partial we compute the intonative features using the model presented in Chap.3,Sec.3.2. We remind that the model decomposes the frequency (and amplitude) trajectory of a partial into the sum of a monotonic variation $d_x(t)$ and a sinusoidal modulation $s_x(t)$:

$$x(t) = \bar{x} \cdot (d_x(t) + s_x(t)) + \epsilon_x(t) \text{ with,} \quad (4.1)$$

$$s_x(t) = e_x.cos(2.\pi.r_x.t + \phi_{x,0}) \quad (4.2)$$

Parameters	Values	Parameters	Values
STFT parameters		Partial tracking parameters	
Sampling Rate	44.1 Khz	Relative frequency deviation (devFR)	20 cents
Window size (win_{size})	80 ms	Constant frequency deviation (devFC)	3 Hz
Window type	Blackman	Relative amplitude deviation (devA)	30%
NFFT	$win_{size} \times 4$	Source partials neighbors (devM)	4
Step Size	$win_{size}/4$	Target partials neighbors (devK)	4
Peak selection parameters		Partial connection parameters	
Amplitude threshold (m)	-60 dB	Time gap to connect partials over (Ct)	0.1 sec
Number max of partials on each frame (q)	40	Frequency gap to connect partials over (Cf)	50 cents
		Minimum partial length (L)	0.08 sec

Table 4.1: Parameters values for partial tracking with pm2 IRCAM's software

where $\bar{x}$ corresponds to the mean value of $x(t)$ over time and $\epsilon_x(t)$ denotes the modeling error. This model applied on the frequency trajectory ($x(t) = f_p(t)$) estimates the parameters of the portamento and vibrato. The same model applied on the amplitude trajectory ($x(t) = a_p(t)$) estimates the parameters of the tremolo.

At the end of this stage, for each partial we retain the following features:

- r_f: vibrato rate (Hz)
- e_f: vibrato extent (cents)
- r_a: tremolo rate (Hz)
- e_a: tremolo extent (relative measurement %)
- Δ_f: the maximum frequency deviation which corresponds to the interval (in cents) covered by $d_f(t)$ and
- ϵ_f: the time-varying frequency modeling error

Then, vocal partials (i.e. partials produced by the voice) are selected using a set of thresholds applied on the values of these features.

3.1.3 Step 3 - Selection of vocal partials

A partial is assigned to the vocal class if its vibrato has characteristics that correspond to vocal vibrato and and tremolo parameters correspond to the values of a vocal vibrato or if the partial corresponds to a portamento. We consider that a vibrato is a vocal vibrato if its vibrato rate is around 6 Hz and its vibrato and tremolo extents are sufficiently larges. To be considered as a portamento, the frequency deviation of the partial has to be sufficiently large. These values are deduced from the knowledge of singing production. We denote by P_{vib} the set of partials with vibrato and by P_{trem} the set of partials tremolo. Finally, the set of vocal partials, denoted by P_{voc}, is given by the union of P_{vib} and P_{trem}:

$$P_{voc} = P_{vib} \cup P_{trem}$$

The set of thresholds used for the vocal partial selection is summarized in Tab.4.2.

In practice, partials whose frequency modeling error (ϵ_f) is too high are automatically assigned to the non-vocal class. We suppose that these partials have a time-varying frequency $f_p(t)$ that cannot be modeled with Eq.(4.1) and thus they are likely not produced by the singing voice.

p has a vocal vibrato ($p \in P_{vib}$)	its vibrato rate r_f is around 6 Hz	$r_f \in [6 - \tau_{\Delta r}, 6 + \tau_{\Delta r}]$
	the extent of the vibrato is sufficiently large	$e_f > \tau_{e,f}$
	there is a tremolo with an extent sufficiently large	$e_a > \tau_{e,a}$
OR		
p is a note transition portamento ($p \in P_{port}$)	the frequency deviation $d_f(t)$ crosses a sufficiently large range of frequencies	$\Delta_{d_f} > \tau_{\Delta df}$

Table 4.2: Thresholds for vocal partials selection

The classification of a partial into the vocal and non-vocal classes is thus determined by the set of thresholds: $\{\tau_{\Delta r}, \tau_{e,f}, \tau_{e,a}, \tau_{\Delta df}\}$. The choice of the threshold values is highly dependent on the application of the method. We discuss the choice of these parameters for the task of singing detection in the evaluation proposed in Sec.3.2.

3.1.4 Step 4 - Formation of vocal segments

The positions of vocal segments are given by the supports of the partials of P_{voc}. Because the voice is a highly harmonic instrument, a vocal segment should be covered by several vocal partials that are harmonically related. We propose to assign a given frame to the vocal class if there is more than one vocal partial on this frame. The result of this step is a frame-by-frame vocal detection.

As explained in Sec.1.1, a frame-by-fame vocal detection is usually over-segmented. To obtain coherent vocal segments (i.e. vocal segments that would correspond to the manual annotation) it is necessary to smooth the frame-by-frame decision. By construction, the proposed method is unable to determine the vocal sections on which the singing voice produces non-voiced sounds. We can suppose that unvoiced sounds usually occur between two voiced sounds. Our proposal is to eliminate short non-vocal sections (with a duration lower than 1 sec) located between two vocal sections. We chose this value because the analysis of the manual annotation into vocal and non-vocal segments shows that non-vocal segments with duration lower than a second were not annotated as non-vocal. This smoothing stage removes non-vocal sections corresponding to short breaks in singing (such as a breathing break).

3.2 Evaluation

We evaluate the performance of the proposed method to locate the vocal segments of a song.

3.2.1 Data set

The method is evaluated on a set of 90 popular songs chosen for their variety in artists, languages, tempi and music genre. They constitute a representative sampling of commercial music. The data set has been originally used for the singing detection task in [RR08]. All songs of the data set have been manually annotated by the same person to provide a reference for the position of vocal segments. The whole set is well balanced: 50.3% of the frames belong to the vocal class and the 49.7% remaining frames are non-vocal. We note that all files contain singing and instrumental accompaniment.

In the following, 58 songs are used for the training and the evaluation is conducted on the 32 remaining songs. The choice of training and testing songs was given in [RR08].

3.2.2 Results

Evaluations are conducted using a ten folds cross-validation. The performance is given by the F-measure of the vocal class. This measure is, in the case of a balanced data set, equivalent to the accuracy of the classification.

Since the method is based on a set of thresholds it is possible to vary the tradeoff between recall and precision by varying the values of the thresholds. In Fig.4.1 we plot the recall and precision of the vocal class obtained by fixing all thresholds but one. This figure plots the PR curve obtained for vocal partials ($p \in P_{vib}$) whose vibrato rates range from 4 to 6 Hz (i.e. for $\Delta_r = 2$ Hz). The green curve is obtained by fixing all thresholds except for the vibrato extent. The blue curve is obtained by fixing all thresholds and varying the tremolo extent. The values expressed as percentages in the legend of the figure can be converted into cents with the formulae given in Eq.(4.3).

$$x_{cents} = 1200. \log 2(x_{\%} + 1) \tag{4.3}$$

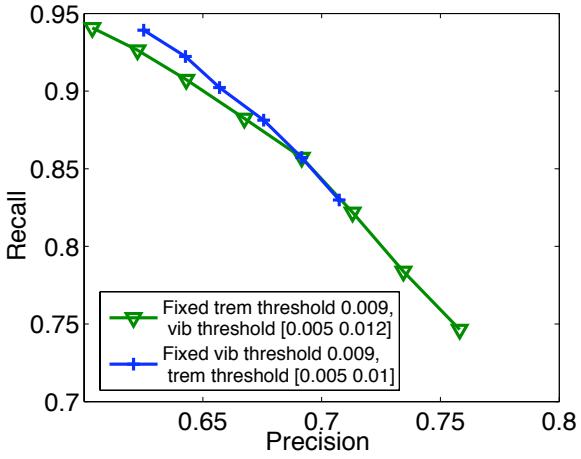


Figure 4.1: Examples of PR-curve obtained when varying one threshold

The thresholds leading to the best F-measure are learned on the training data. Once the partials have been extracted and parameterized (steps 1-2) the procedure of partials selection (step 3) and vocal segment localization (step 4) is extremely fast. Thus, an exhaustive search of the best combination of thresholds can be applied. The ranges of values tested for each threshold are given in Tab.4.3.

The last column of the table summarized the thresholds values leading to the best F-measure on the training-set.

Threshold	Range/set of values tested	Best set
τ_ϵ	10 Hz	10 Hz
τ_r	$[0, 4]$ Hz , step = 0.1 Hz	2 Hz
$\tau_{e,f}$	$[0, 100]$ cents, step = 5 cents	16 cents
$\tau_{e,a}$	$[1, 20]$ %, step = 0.5%	9 %
$\tau_{\Delta df}$	$[0, 500]$ cents, step = 5 cents	85 cents
τ_t	$\{0.5, 1, 1.5\}$ sec	1 sec

Table 4.3: Values tested for each threshold

The results obtained with these thresholds on the testing-set are reported in Tab.4.4. These results are compared to results obtained with a "classical machine learning approach" shown in Tab.4.5. These results are obtained using MFCC and their derivatives, SFM (Spectral Flatness Measure) and its derivatives as features [Pee07a]. These features are normalized by the inter-quantile-range (IQR). Features with an IQR equal to zero are discarded. The set of the 40 best features, selected using the IRMFSP ([Pee03]) are retained to train two GMM (vocal and non-vocal) with 8 components. For both approaches, the results are given before and after the filtering process (step 4). We note that the same strategy is applied for smoothing the frame-by-frame classification.

	Before filtering	After Filtering
Fmeasure	69.85%	76.52%
Recall	60.20%	71.09%
Precision	83.20%	85.45%

Table 4.4: Results of the proposed approach for the vocal class

	Before filtering	After Filtering
Fmeasure	75.81%	77.40%
Recall	75.80%	77.40%
Precision	75.81%	77.40%

Table 4.5: Results of the classical machine learning approach for the vocal class

On Fig.4.2 we plot an example of the segmentation obtained before and after smoothing the vocal detection on an excerpt of the song "aline.wav" of the test-set.

As shown by Tab.4.4 and Tab.4.5, the singing voice detection method proposed in this section achieves a performance close to the "classical" approach. The set of thresholds optimized on the training set to maximize the F-measure locates the vocal segments very precisely, however some vocal segments are not detected. We can conclude that vibrato, tremolo and portamento features are really efficient to detect the voice. However, vocal portions cannot always be detected with these features. A deeper analysis of the results of vocal partial identification shows that short vocal partials are not detected with the proposed method. This is due to the fact that it is difficult to obtain a reliable estimation of vibrato parameters on short segments. Also, partials with a low frequency can be misestimated due to problems of frequency resolution.

We have come up with a method that overcomes these problems. The aim of the method is to retrieve all partials corresponding to the singing voice in a mixture. The improvement relies essentially on a better formulation of the harmonic nature of the singing voice.

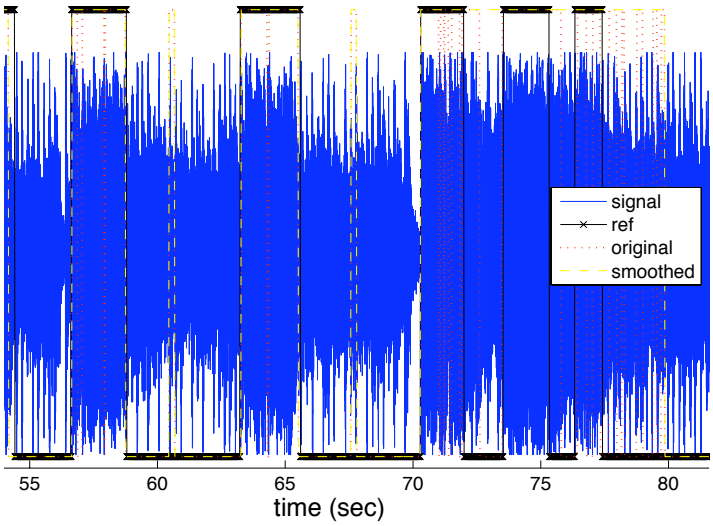


Figure 4.2: Vocal segments before and after smoothing on “aline.wav”

4 Proposed approach to track harmonic contents of the singing voice

We present in this section a method to group partials produced simultaneously by each instrument of a polyphonic audio mixture. The proposed method is adapted for pitched (harmonic) instruments and is specially designed to group vocal partials harmonically related. The primary goal of the method is to retrieve all partials produced by the voice in a mixture. The result of the method can be used directly to isolate the lead vocal. The groups of vocal partials can also be used to precisely determine the vocal portions of the song and to transcribe the singing melody.

The proposed approach is based on some CASA principles: common onsets, synchronous variations of frequency and harmonicity. Like the methods presented in Sec.2.3, the method uses partials as input. The main idea of the method is to define a similarity measure between partials based on the parameters of vibrato and portamento. Then the groups of partials are formed automatically using traditional clustering techniques.

The details of the proposed method are given in Sec.4.1. The ability of the method to group harmonically-related partials is evaluated in Sec.4.2. Then, we present two applications of this method. In Sec.4.3 we evaluate if groups of partials can be used to locate the vocal segment of a song. Finally, in Sec.4.4, we propose a method to deduce the fundamental frequency associated with a group of partials and then evaluate how the group of partials can be used to transcribe the vocal melody on a polyphonic excerpt.

4.1 Description of the method to group harmonically related partials

Before giving the details of the different step of the method, we first review the theoretical fundamentals on which the method is based.

4.1.1 Theoretical fundaments

To be fully characterized, a partial is described by a frequency and an amplitude functions plus an interval of time (i.e. an onset and an offset). Thus, a partial can be defined as follows:

$$p(t; a(t), f(t), t_{beg}, t_{end}) = \begin{cases} a(t)\cos(2\pi \int_0^t f(\tau)d\tau + \phi_0) & \text{if } t_{beg} \leq t \leq t_{end} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where $a(t)$ is the time-varying amplitude and $f(t)$ the time-varying frequency of the partial. Both functions are defined between the onset t_{beg} and the offset t_{end}. The support of the partial p is denoted $I_p = [t_{(p,beg)}, \ldots, t_{(p,end)}]$.

Based on this definition, the signal associated with a single harmonic source with H harmonics, for one "note event", can be written as:

$$s(t) = \sum_{h=0}^{H} p_h(t; a_h(t), f_h(t), I_h) + \epsilon(t) \tag{4.5}$$

where ϵ is the modeling error. This equation is equivalent to the harmonic sinusoidal model presented in Sec.2.2.

Theoretically, for any harmonic instrument, all time-varying frequency functions $f_h(t)$ follow the same trajectory. If $f_0(t)$ denotes the fundamental frequency of the sound, (i.e. the lowest harmonic of the sound) the relation between the f_h at each instant t is given by:

$$f_h(t) = (h+1).f_0(t) \ , \forall h \in H \tag{4.6}$$

This relation has motivated a number of CASA cues such as: *common frequency variations* and *harmonic concordance*. And theoretically, all harmonic partials associated to the same source have the same support:

$$I_h = I_0 \ , \forall h \in H. \tag{4.7}$$

This last relation is at the basis of the *common onset and offset* ASA cues. In practice, it is common to find partials (especially higher partials) starting with a small delay and stopping before the offset of the fundamental partial. We can assume that the relation given by Eq.(4.7) remains true and the amplitude of these partials is null at the beginning and the end of $a_h(t)$. There is no predefined relation between the $a_h(t)$ of harmonically related partials. This relation depends on the timbre of the instrument. In numerous source separation studies based on CASA, the *common amplitude variations* is taken as a cue for the segregation and integration of partials. However, in the case of the singing voice, such a cue cannot be considered as reliable since the amplitude of each partial is modulated according to its position of the frequency within the vocal tract.

A signal $s(t)$ composed with a singing voice source s_{voc} and instrumental background source s_{instru} is written as:

$$s(t) \quad = \quad s_{voc}(t) + s_{instru}(t) + \epsilon(t) \tag{4.8}$$

Where the singing voice source is written as the sum of N note events, each of them defined as a sum of H harmonic partials:

$$s_{voc}(t) = \sum_{n=1}^{N}\sum_{h=1}^{H} p_{n,h}(t; a_{n,h}(t), f_{n,h}(t), I_{n,h}) \tag{4.9}$$

Since we are not concerned with the separation of the instrumental accompaniment into note events, we simply write s_{instru} as the sum of M partials that are not necessarily harmonically related.

$$s_{instru}(t) = \sum_{m=1}^{M} p'_m(t; a'_m(t), f'_m(t), I_m) \tag{4.10}$$

Each time-varying frequency can be modeled with the intonative model presented in Sec.2.3 of the previous chapter. For the h^{th} harmonic we rewrite Eq.(4.2) as:

$$f_h(t) = \bar{f}_h \cdot (df_h(t) + e_{f,h}.\cos(2\pi r_{f,h}t + \phi_{0,h})) + \epsilon_{f,h}(t). \tag{4.11}$$

In this expression, the frequency variations are expressed in terms of relative variations. From Eq.(4.6) we can deduce that all harmonics related to a fundamental whose mean frequency is denoted by $\bar{f}_0$ follow:

$$\begin{cases} \bar{f}_h = (h+1).\bar{f}_0\ , \\ e_{f,h} = e_{f,0} \\ r_{f,h} = r_{f,0} \\ df_h(t) = df_0(t) \forall t \in I_p \end{cases} \tag{4.12}$$

To group harmonically related partials we propose to define a distance between two partials and then realize the grouping stage with traditional clustering tools. The different steps of the methods are described below.

4.1.2 Step 1 - Measure of similarity between partials

For two partials, denoted by p and q, we define a (dis)similarity measure. We want the measure to be close to zero if p and q are harmonically related and close to one otherwise. The measure is defined as a multi-criterion function. Partials are compared on a set of K criteria, analogous to CASA cues, with comparison functions denoted by $\phi_k(p,q)$ for $k = 1, \ldots, K$. The dissimilarity measure $m(p,q)$ is then given by the aggregation of the comparison functions:

$$m(p,q) = \psi(\phi_1(p,q), \ldots, \phi_K(p,q)) \tag{4.13}$$

As mentioned in the previous paragraph, partials associated with the same note event have similar frequency trajectories. They follow the relations given in Eq.(4.12). However, in a polyphonic mixture following traditional harmonic rules, the notes played by the different instruments have harmonics overlapping in frequency. This problem, which creates the main difficulty for multiple instrument recognition (see Sec.2.2), prevents the partial tracking from being perfect. As a result, a partial can be split into multiple segments where the different portions belong to different sources.

Considering this last point, we suggest to compare the temporal variation of partials on their common frames. If I_p and I_q denotes the support of partial p and q respectively, then the frequency

modulations of p and q are compared on their shared part: $I_{p,q} = I_p \cap I_q$. We denote by $|I_{p,q}|$ the length of $I_{p,q}$. On this interval, partials are compared with the following criteria.

- **Strict dissimilarity based on support of partials**
  First, partials that do not overlap in time can hardly be harmonically related. We define a first criterion based on the length of the common support of the two partials. If two partials have no common support then other criterion cannot be computed and the dissimilarity measure is set up to one (strict dissimilarity). Formally this is given by:

$$\phi_1(p,q) = 1 - \frac{|I_{p,q}|}{\max(|I_p|, |I_q|)} \tag{4.14}$$

  Since we want the measure to be bound to one, the denominator is given by the maximal length of I_p and I_q.

  The strict dissimilarity leads to:

$$\phi_1(p,q) = 1 \Rightarrow m(p,q) = 1 \tag{4.15}$$

  This first comparison is conducted to exclude the comparison of partials that have no common support, and to favor the partials that have (relatively) long common support. This criterion is based on the intuitive idea that partials related to the same sounds have common onsets and offsets. As mentioned by Hartmann [Har88] the time onset is the primary cue by which sounds from different sources are segregated in music listening.

- **Dissimilarity based on onset**
  However, the partials of a note do not start exactly at the same time and the slight asynchrony of partials' onsets is an important part of the timbre of the instrument as shown by Grey [Gre75]. Strong partials of a note start within a period of 27 to 49 ms [Ras79] and a large delay (> 30 ms) greatly reduces the degree to which two sounds are heard as a single tone [Bre90]. To take into consideration these points we define a criterion based on the length of the gap between the onsets of two partials as:

$$\phi_2(p,q) = \frac{|onset_p - onset_q|}{\max(|I_p|, |I_q|)} \tag{4.16}$$

  In all types of music, the different sources of the ensemble play in harmonic and rhythmic relation and they take great care to start with exact common onsets, even though the ear is still able to separate the different sounds. There are other evidences that make the partials produced by different instruments with common onset to be grouped per instrument. According to numerous studies on CASA principles, the common frequency variation is one of those. The term variation encompasses the periodic modulation (vibrato) and the continuous variations (portamento) that can appear on partial's frequency trajectories. We chose then to compare partials p and q with their features of vibrato and portamento computed on $f_p(t)$ and $f_q(t)$ for $t \in I_{p,q}$.

- **Dissimilarity based on vibrato**
  The vibratos of the two partials are compared using the rate and the extent values with the

following function:

$$\phi_3(p,q) = \frac{1}{2}\left(\frac{|e_{f,p} - e_{f,q}|}{\max(e_{f,p}, e_{f,q})} + \frac{|r_{f,p} - r_{f,q}|}{r_{max}}\right) \tag{4.17}$$

Where r_{max} is the maximum vibrato rate which depends on the sampling rate of f_p and f_q which depends on the hop size used in the partial tracking algorithm. If the sampling rate of f_q is equal to 100 Hz then $r_{max} = \frac{1}{2}.\frac{1}{0.01} = 50$ Hz.

- **Dissimilarity based on relative frequency variation**
  The comparison based on the portamento is done on the values of the third order polynomial used to model the slow monotonic variation $d_f(t)$). Since the polynomial is computed on the normalized version of $d_f(t)$) (which is monotonic) the poles of the polynomial lie between 0 and 1. For partial p, these coefficients are denoted by $c_{n,p}$ for $n = 1 \ldots 4$ and the comparison based on the continuous variation of frequency is given by :

$$\phi_4(p,q) = \frac{1}{4}\sum_{n=0}^{3} |c_{n,p} - c_{n,q}| \tag{4.18}$$

- **Dissimilarity based on harmonicity**
  Since all types of modulation cannot be modeled with a sum of the portamento and vibrato, we add a last criterion on the frequency modulation based on the idea that $f_p(t)$ and $f_q(t)$ are equal up to a constant multiplier on their shared part : $\frac{f_p(t)}{f_q(t)} = c.\mathbb{1}_{I_{p,q}}$. Then we define a measure of harmonicity based on the standard deviation of the ratio of the two frequencies as:

$$\phi_5(p,q) = \begin{cases} \sigma\left(\dfrac{f_p(t)}{f_q(t)}\right) & \text{if } \bar{f_p} > \bar{f_q} \\ \sigma\left(\dfrac{f_q(t)}{f_p(t)}\right) & \text{otherwise} \end{cases} \tag{4.19}$$

The dissimilarity measure between p and q is finally given by the aggregation of these comparison functions. The aggregation is done using the linear opinion pool [2] If the first comparison function is not equal to one (strict dissimilarity):

$$m(p,q) = \psi(\phi_1, \ldots, \phi_K) = \begin{cases} 1 & \text{if } \phi_1(p,q) = 1 \\ \displaystyle\sum_{i=1}^{K} w_i.\phi_i(p,q) & \text{otherwise.} \end{cases} \tag{4.20}$$

We chose the following weights: $w_i = [1/9, 1/9, 2/9, 2/9, 2/9]$. If we consider that the two criteria based on the support of the partials form a unique criterion, then distance is given by the mean of four equally weighted comparison functions (support, FM, frequency deviation and harmonicity).

[2]The log opinion pool, which is a weighted product of the $\phi_i(p,q)$ is an exclusive aggregation. If one $\phi_i(p,q)$ is equal to zero, the dissimilarity measure is also equal to zero. Since we want each piece of information to be considered, we chose the linear aggregation.

4.1.3 Step 2 - Distance matrix

In practice, dissimilarity is computed for all pairs of partials of a given segment of a song. A moving window of a few seconds length automatically determines these segments. On each segment we construct a distance matrix as follows:

For each pair of partials, the dissimilarity measurement is so that:

$$\begin{cases} 0 \leq m(p,q) \leq 1 \\ \quad m(p,q) = 0 \qquad p = q \end{cases} \tag{4.21}$$

We note that $m(p,q)$ is not a distance because $m(p,q) \neq m(q,p)$. The non-symmetry comes from the two first criteria: if $I_p \subset I_q$ then $m(p,q) < m(q,p)$. However, if $m(p,q)$ indicates that p and q are very similar there is no reason for q to not be similar to p. This is why we chose to define the distance between partials p and q as follows:

$$d(p,q) = \min\{m(p,q), m(p,q)\}. \tag{4.22}$$

Finally, the distance between the p^{th} and q^{th} is reported at the intersection of the p^{th} row and the q^{th} column of the similarity matrix M. This matrix is then given as input of the chosen clustering algorithm to group partials automatically.

4.1.4 Step 3 - Grouping harmonically related partials by clustering

Partials are grouped using an agglomerative hierarchical clustering algorithm based on the distance between partials defined above. The distance between two clusters is given by the average of all distances between pairs of partials (*average linkage method*), resulting in clusters with close variances. Given two clusters denoted by C_1 and C_2, such as C_1 is composed with $|C_1|$ partials p_i for $i = 1 \ldots |C_1|$ and C_2 is composed with partials q_j for $j = 1 \ldots |C_2|$. The distance between the two clusters is given by:

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i=1}^{|C_1|} \sum_{j=1}^{|C_2|} d(p_i, q_j) \tag{4.23}$$

If the task was to find all partials produced by one instrument throughout the song duration, the *single linkage method* would have been more appropriate. The *complete linkage method* produces very specific classes and does not fit our problem. The optimal cutoff for stopping the agglomerative process of the clustering is found using the VIF (Variance Inflation Factor) criterion [OSON05].

At the end of the clustering stage we obtain groups of harmonically related partials. Each group is defined by:

- a set of partials
- an onset and an offset, which are determined using the onset and offset of the longest partials of the cluster
- a fundamental frequency. The description of the fundamental frequency estimation for a group of harmonically related partials will be further described in Sec.4.4.

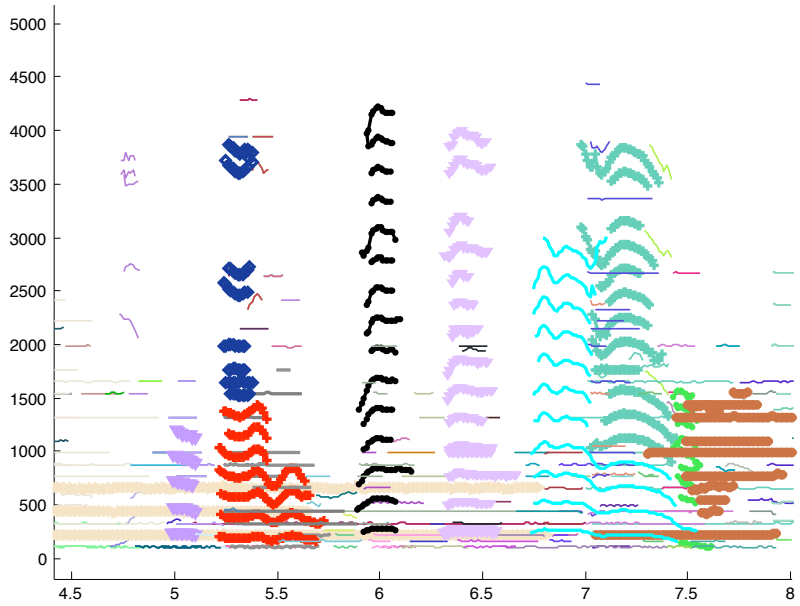


Figure 4.3: Example of partial clustering obtained on a real signal: voice + piano

In Figure 4.3, we plot the result of the proposed method computed on a real signal with three sources (voice and piano) where the partials of each source have been extracted on separated tracks. Each cluster is represented by a different line type.

Before explaining the details of the fundamental frequency estimation method and its application to multi-pitch estimation, we first evaluate the capacity of the method to group harmonically related partials. This evaluation is presented in Sec.4.2. Then, we apply the method to the task of singing detection in Sec.4.3.

4.2 Evaluation of clusters of harmonic partials

We evaluate if the clusters obtained with the proposed method are composed of partials corresponding to the same note coming from the same source. In practice, this evaluation is not feasible on “real" recordings because it is complex and time-consuming to obtain a ground thrust from real data. We propose to evaluate the quality of the clusters with a data set composed of multi-track recordings. Partials are extracted from separate tracks and labeled (to indicate from which source they belong) before being merged into a global set of partials. The method is applied on the global set of partials, and the reliability of each cluster is computed by regarding the labels of the partials composing the cluster.

4.2.1 Data set

The data set is generated from the multi-track recordings of 15 songs. From these songs we create the data set of partials as follows:

1. For each song we extract partials from the lead vocal track and at least one instrumental track of piano, guitar and/or bass. Partials are tracked independently on each mono-instrumental track.

2. To avoid bias in the evaluation we chose to equilibrate the partials of singing voice and the partials of all other instruments. For each song, one or two instrumental tracks are chosen so that there is the same number of instrumental partials as vocal partials.

3. Finally, the balanced sets of partials for each song are merged into a global set of partial. The distribution of partials per class of instruments is described in Tab. 4.6.

	Voice	Instruments		
	Voice	Guitar	Piano	Bass
Label	1	-1	-2	-3
Cardinality	56289	19092	25460	11737
Cardinality	56289	56289		

Table 4.6: Multi-track partials test-set description

4.2.2 Measure for cluster evaluation

We measure the performance of the clustering method using the measure of cluster purity. The cluster purity is given by the percentage of partials coming from the source that is the most represented within the cluster. If all partials composing a cluster come from the same source, the purity of this cluster is then equal to 1. For a cluster C_i (of size $|C_i|$) the cluster purity is given by:

$$purity(C_i) = \frac{1}{|C_i|} \max_j (|C_i|_{class=j}). \tag{4.24}$$

The overall purity of a clustering solution, with I clusters, is given by the weighted sum of individual cluster purities as shown in Eq.(4.25) where $|C|$ is the total number of partials (i.e. $|C| = \sum_{i=1}^{I} |C_i|$).

$$purity = \frac{1}{|C|} \sum_{i=1}^{I} |C_i| purity(C_i) \tag{4.25}$$

If all clusters are composed with a single partial, the overall purity is then equal to one. However, this clustering solution is of no interest to us. To avoid this problem, the evaluation is conducted on clusters composed with at least 3 partials.

4.2.3 Results

We can consider two cases for the evaluation: a) the signal is composed of two sources (vocal and instrumental) as described in Eq.(4.8) or b) each instrument in the signal is considered as an independent source (i.e. each source of the instrumental background is considered individually).

For the case with two sources, the overall purity of the clustering solution is equal to 0.8944. When all instruments are considered individually, the overall purity of the clustering solution is equal to 0.7802. In both cases, 82% of the partials have been assigned to one cluster.

These results show that the proposed method is clearly efficient to group partials coming from the same source. Naturally, the performance is better when considering only two sources. We note that

the instruments considered in the accompaniment are polyphonic instruments that make the task even more complicated. We also remark that the cluster purity measure does not take into consideration the problem of note events. However, the clusters of vocal partials cannot cover multiple note events because of the first comparison function given in Eq.(4.14). Nothing ensures that partials corresponding to the same note are not split into distinct clusters.

This evaluation is conducted as a first validation of the partial clustering method. The case analyzed here is not realistic since partials have been extracted on separate tracks. Thus, the problem encountered by partial tracking algorithms on the frequency regions where partials from different sources overlap is avoided. In the next sections, we apply the clustering method on partials extracted from a mixture of instruments.

4.3 Application to singing voice detection

We apply the partial clustering method to the task of singing detection. Clusters of partials are analyzed to identify “vocal clusters”, i.e. clusters composed of vocal partials. Then the vocal segments are deduced from the supports of vocal clusters. This problem can be viewed either as a specific case of instrument recognition in polyphonic context, or as a problem of source identification in a source separation method.

4.3.1 Data set

The evaluation is conducted on the same data set as in the previous evaluation of singing voice detection. The details of this data set are given in Sec.3.2.1.

4.3.2 Results

In the previous evaluation, Sec.3.2, we have shown that the singing voice detection based on the direct recognition of vocal partials has a very high precision but suffers from a relatively low recall. The low recall came from the bad estimation of intonative features on short partials (shorter than two cycles of vibrato) and on partials with a low mean frequency (probably due to the FFT resolution). We believe that these partials can be better interpreted with the cluster of partials. The hypothesis is that the groups of harmonically related partials carry more source information than the individual partials.

In this evaluation, a cluster is considered a “vocal cluster" if a certain number of partials within the partials are identified as vocal partials with the previous method. We plot in Fig.4.4 the partials of a given cluster. Partials identified as vocal partials with the previous method are plotted with dotted lines. Since the number of vocal partials is sufficiently large, the cluster is considered as a vocal cluster. This plot illustrates the improvement of vocal segment localization given by the cluster of partials: the previous method would detect a vocal segment between 2.55 sec to 2.8 sec. But the cluster of partials shows that the vocal segment is actually longer: from 2.55 sec to 3.15 sec.

We report in Tab.4.7 the results obtained with the clusters of partials on the testing set. We present the results obtained for X varying from 1 to 3 where X is the minimal number of vocal partials within a vocal cluster. These results have been obtained using the same set of thresholds as in the previous evaluation. The results obtained with the previous method are reported in the first line of the table.

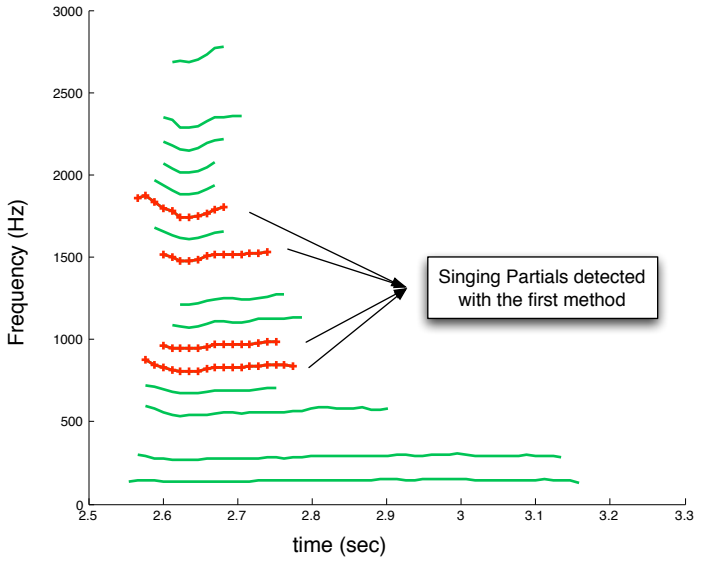


Figure 4.4: Comparison of the basic and improved method for vocal partial detection

X	F-measure	Precision	Recall
	76.52 %	85.45%	70.30%
≥ 3	77.91%	79.21%	76.65%
≥ 2	79.02 %	75.33%	83.08%
≥ 1	77.23 %	66.97%	91.19%

Table 4.7: Results of the voice detection for various X

As shown by these results, the clusters of partials increase the F-measure mainly because it improves the recall. Unfortunately, improving the recall leads to a degradation of the precision. In the ideal case, i.e. if the cluster purities were equal to 1 and the precision of the first method was equal to 100% the recall would be improved without affecting the precision.

As in the previous evaluation, the tradeoff between the recall and the precision can be adapted for the application. This tradeoff still depends on the values of the thresholds for vocal partial detection but also on the value of X. We note that a relatively good tradeoff is given for $X = 2$.

As shown in this section, the clusters of partials can be used to improve the localization of vocal segments. In addition, the clusters of harmonic partials are powerful for two other applications: the separation of the singing voice and the transcription of the vocal melody.

To isolate the singing voice, the partials of vocal clusters can be simply re-synthesized. This solution often leads to a very synthetic result, whose quality is affected by artifacts. In addition, all elements of the singing voice that are not voiced are missed. To obtain a better separation, we suggest re-synthesizing the signal after masking the components belonging to non-vocal cluster. The analysis of non-vocal and vocal clusters overlapping in time determines the bandwidth of the mask associated with each non-vocal partial.

We present in the next section a method to transcribe the vocal melody based on the analysis of the partials clusters.

4.4 Application to multi-pitch and sung melody transcription

4.4.1 Method to estimate the f_0 of a cluster

The fundamental frequency of the cluster plotted in Fig.4.4 can be deduced intuitively: it corresponds to the lowest partial. Theoretically, the fundamental frequency of a cluster can be deduced from the distances between consecutive harmonic partials. Unfortunately, during the formation of clusters, two types of errors can appear: introduction of impostor partials and suppression of correct partials. To estimate the f_0 of a cluster it is necessary to eliminate the impostor partials and to find the harmonic index for each remaining partial.

We describe a method to make this estimation for a cluster composed of a set of partials $\mathcal{P} = \{p_i\}$ whose frequency functions f_i are not defined outside the support of the cluster: $f_i(t) = \emptyset$ if $t \notin T = [t_{beg}, \ldots, t_{end}]$. The number of partials at each instant of the support is given by $P(t)$. The fundamental frequency of a cluster is a time-varying function denoted by $f_0(t)$. A first estimation of this function can be estimated by finding the value $\bar{f_0}$ (mean frequency over time of $f_0(t)$) and the set of harmonic index $\{h_i\}$ that minimize:

$$e = \frac{1}{|T|} \sum_{t=t_{beg}}^{t_{end}} e(t) \text{ with,} \tag{4.26}$$

$$e(t) = \frac{1}{P(t)} \sum_{i=1}^{P(t)} \left(\frac{f_i(t)}{h_i} - f_0(t) \right)^2 + \epsilon(t) \tag{4.27}$$

where the left part of Eq.(4.27) corresponds to the errors due to the variations of frequency around $\bar{f_0}$ and the term on the right side ($\epsilon(t)$) corresponds to the error introduced by the impostor partials. Thus, the error given in Eq.(4.26) is minimized by eliminating the impostor partials and by finding the set of harmonic index $\mathcal{H} = \{h_i\}$ and the value $\bar{f_0}$ that best explain the set of $f_i(t)$ of the partials considered as non-impostors. Theoretically, the conjoined estimation of $\mathcal{H}$ and $\bar{f_0}$ can be addressed as a mixed-integer quadratic problem (MIQP). But in practice this type of optimization is computationally too expensive to be applied.

We propose a simple and empirical method to solve this problem. The method starts with a rough estimation of $\bar{f_0}$ to give a first estimation of $\mathcal{H}$ and eliminates the impostor partials. The values of $\bar{f_0}$ and $\mathcal{H}$ can be re-estimated if the first estimation of $\bar{f_0}$ was given with an octave error. The steps of the methods are detailed below.

1. **Setting the data of the problem:** All partials within the cluster are arranged by increasing frequency. Keeping in mind that some harmonics may be missing, we number the partials with the following indices: p_1 corresponds to the partial with the lowest frequency and so on until the highest partial is numbered as p_K where $K = \max_t(P(t))$. In the ideal case, when all harmonics (and only the harmonics) are present for each i, p_i corresponds to h_i.

2. **Pre-estimation of $\bar{f_0}$:** The quantity $\bar{f_0}$ is roughly estimated by analyzing the frequency distances between the time-varying frequency of consecutive partials (i.e. partials close in frequency) .

- Given two consecutive partials p_i and p_{i+1} whose frequencies are denoted by f_i and f_{i+1}, the distance at instant t is given by:

$$y_i(t) = \begin{cases} f_{i+1}(t) - f_i(t) & \text{if } f_i(t) \neq \emptyset \text{ and } f_{i+1}(t) \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases} \tag{4.28}$$

- This quantity $y_i(t)$ is estimated for each pair of consecutive partials (i.e. for $i = 1, \ldots, K-1$) all along the support of the cluster (i.e. for $t \in [t_{beg}, \ldots, t_{end}]$).
- The mean frequency associated with the cluster is given by the analysis of the distribution of all $y_i \neq \emptyset$. The estimation of $\bar{f}_0$ is simply given by the position of the highest peak in the distribution.

3. **Estimation of the harmonic numbers:** Once $\bar{f}_0$ has been estimated the harmonic index of each partial p_i at each instant t is given by:

$$h_i(t) = \frac{f_i(t)}{\bar{f}_0} \tag{4.29}$$

For each partial, the continuation of $h_i(t)$ is analyzed to determine a single harmonic index for each partial. At the end of this step, for a cluster composed of K partials, we obtained a set of K harmonic indices $\mathcal{H} = \{h_1, \ldots, h_K\}$. The values of the series $\mathcal{H}$ give an indication on the probability that $\bar{f}_0$ was properly estimated. If the value of $\bar{f}_0$ is correctly estimated, and no impostor partials are present, then the values of $\mathcal{H}$ are natural integers. If $\bar{f}_0$ was estimated an octave lower, the series $\mathcal{H}$ contains even numbers only. Finally, if $\bar{f}_0$ was estimated an octave too high, then the series contains rational numbers. Typically, the series contains the following ratio: $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$ and so on. If one of the miss-estimation cases is encountered, we simply note it, but the procedure continues with the next steps.

4. **Detection of impostor partials:** The values of $\mathcal{H}$ are then rounded to the closest integer. In most situations, if an impostor partial is present, two partials are assigned to the same harmonic numbers. In this case, the impostor partial is determined by the values of the partials' similarity matrix presented in 4.1.3. The partial which is the less similar to the partial with the closest harmonic number is considered an impostor. At the end of this step, the partials composing the cluster are grouped into two sets: $\mathcal{P}_h$ the set of harmonic partials and $\mathcal{P}_i$ the set of impostor partials.

5. **Estimation of the error:** To each partial of $\mathcal{P}_h$ corresponds an index of harmonic number h_i: $\mathcal{P}_h = \{(f_1, h_1) \ldots (f_k, h_k)\}$. The error given in Eq.4.26 can finally be computed using the partials of $\mathcal{P}_h$. If the analysis of the series $\mathcal{H}$ at step 3 suggested that the $\bar{f}_0$ was estimated with an octave error then the analysis is done again (from step 2 to step 5) with the new estimate of $\bar{f}_0$. Finally, the solution $S = \{\bar{f}_0, \mathcal{H}_h\}$ yielding the lowest error is retained.

6. **Final estimation of the fundamental of the cluster** The time-varying frequency associated with the cluster is then computed using the partials of $\mathcal{P}_h$. The number of partials of $\mathcal{P}_h$ present at time t is denoted by $P_h(t)$ and each frequency denoted by f_j in the following equation corresponds to one f_i of the original set of partials composing the cluster.

$$f_0(t) = \frac{1}{P_h(t)} \sum_{j=1}^{P_h(t)} \frac{f_j(t)}{h_j} \tag{4.30}$$

This method can be applied to estimate the f_0 of each cluster returned by the method described in Sec.4. In this case, it produces a multi-pitch estimation. To estimate the vocal melody, the proposed approach is applied only on vocal clusters.

4.4.2 Data set

We evaluate the performance of the proposed approach to transcribe the vocal melody in accompanied vocal recordings. We work with a set of 200 excerpts of a cappella recordings from pop-rock and lyrical singers mixed with instrumental tracks of piano and guitar. Each excerpt is 2 seconds length and the voice is always present. For each a cappella recording we create 2 samples of accompanied vocal by mixing the solo voice track with an excerpt of piano or guitar. For each sample, the instrumental excerpt is chosen randomly within the instrumental track. Finally, the data set is composed with 600 excerpts: 200 a cappella, 200 voice + piano and 200 voice + guitar.

For each mix, the RMS energy of the two tracks to be mixed is normalized so that the singing-to-accompaniment ratio is equal to 0 dB. The fundamental frequency of the vocal melody is estimated on the a cappella recordings using the YIN algorithm. This estimation is considered the reference.

4.4.3 Measure

The performance of the proposed method for vocal melody transcription in polyphonic context, is evaluated with the measures proposed in [PEE+07b]. For each frequency estimated we measure the **raw pitch** and the **chroma pitch** accuracies. The raw pitch accuracy is given by the fraction of frames where the pitch has been estimated within one quarter of the tone of the true pitch (Eq.(4.31) and Eq.(4.32)). The chroma pitch accuracy is computed in the same way, except that the pitch is measured in term of chroma, or pitch class, a quantity derived from the pitch by wrapping the pitch into one octave.

$$score = 100 - \frac{1}{N} \sum_{t=1}^{N} err(t) \text{ where} \tag{4.31}$$

$$err(t) = \begin{cases} 100 & \text{if } |f^{ref}_{cents}(t) - f^{est}_{cents}(t)| \geq 100 \\ |f^{ref}_{cents}(t) - f^{est}_{cents}(t)| & \text{otherwise} \end{cases} \tag{4.32}$$

$$\tag{4.33}$$

where f^{ref} is the correct f_0 and f^{est} is the frequency estimated. The conversion of the fundamental frequency into cents is given by Eq.(4.3).

4.4.4 Results

The vocal melody is given by the fundamental frequency of the clusters classified as vocal clusters. In the evaluation on the localization of vocal portions (Sec.4.3), we have seen that the recognition of

vocal clusters was good but not perfect. Since we want to evaluate the performance of the vocal f_0 transcription which relies on the recognition of the vocal clusters, we propose to evaluate the following elements:

- **Presence of the vocal f_0 in the multi-pitch transcription:** For each frame of (f^{ref}) we peak in the multi-pitch transcription the frequency which is the closest of the reference frequency to give a pseudo f^{est}.
- **Accuracy of the vocal melody transcription:** The f^{est} is given by the f_0 of the clusters recognized as vocal clusters with the method presented in Sec.4.3
- **Accuracy of the dominant melody transcription** Additionally, we also evaluate the performance of the transcription when we consider that the dominant pitch at each instant is the vocal pitch

For each case, we measure the raw and chroma pitch accuracies. The results are reported in Tab.4.8 for each type of accompaniment.

	Accapela	
	Multi-pitch	Sung Melody
Chroma	96%	88%
Pitch	85%	74%
	Guitar	
	Multi-pitch	Sung Melody
Chroma	89%	71%
Pitch	79%	57%
	Piano	
	Multi-pitch	Sung Melody
Chroma	89%	74%
Pitch	78%	59%

Table 4.8: Results of melody transcription in polyphonic context per type of accompaniment

As shown by these results, the chroma pitch accuracy yields a much better performance than the pitch accuracy, which indicates that the proposed method to estimate the f_0 of a given cluster suffers from octave errors. The chroma pitch accuracy obtained for the sung melody is more or less equivalent to the performance of the method mentioned in Sec.2.4.

The high performances obtained with the multi-pitch experiments show that the method proposed to group harmonically related partials is efficient, even when partials have been tracked directly on the mixture of instruments.

Additionally, we can conclude that the method proposed to identify vocal clusters is also rather efficient since the performance of “multi-pitch" and “vocal melody" are close. This results show that the vocal pitch found by identifying vocal clusters was in most cases the correct pitch.

The performance obtained using vocal pitch as the dominant pitch is much lower than the performance obtained with the identification of vocal clusters. However, the accompanied vocals were synthesized so that the singing-to-accompaniment ratio was equal to 0 dB. In “real recordings", it is likely that the voice is louder than the instrumental accompaniment in vocal portions.

The performances for vocals accompanied by guitar and piano are equivalent. It should be mentioned that these two instruments do not have vibrato and portamento to aid the identification of vocal clusters.

5 Summary and conclusions

In this chapter we have investigated the general problem of finding the elements produced by the singing voice in the signal of a song (vocal + accompaniment). The main idea developed here is to discriminate the vocal partials from instrumental partials using some characteristics of the singing voice: the vibrato, the portamento and the harmonicity. These elements can be detected by analyzing the variations of the time-varying frequency of partials obtained using the sinusoidal model as the underlying representation of the audio signal.

In a first part (Sec.3), we developed a method to localize the vocal segments of a song based on the direct recognition of vocal partials using a set of thresholds on the values of intonative features. These features are given by the intonative model applied on time-varying frequency and amplitude of partials. The supports of the vocal partials are used to determine the position of the vocal segments. We have shown that the performance of this approach is close to performance of classical audio classification approaches (model of vocal and non-vocal classes using audio features extracted on the amplitude spectrum of stationary portions of the signal). The method has a very high precision, which indicates that the vibrato and portamento are efficient clues to discriminate the singing voice from other musical instruments. However, these criteria are not sufficient to detect all partials produced by the singing voice (low Recall). Since the analysis is based on the sinusoidal model, non-voiced vocal segments cannot be detected as vocal. But the detailed analysis of the results shows that some vocal partials with a short duration or low frequency are also not detected.

To overcome the limitations of the first approach, in the second part we proposed (Sec.4) a method to retrieve all partials produced by the singing voice more accurately. The proposed method uses some common CASA cues (mainly based on the synchronous variations of the frequency) to group partials harmonically related. ASA cues are used to define a distance between partials, which is by definition equal to zero for two partials related to the same fundamental frequency. Then harmonically related partials are grouped automatically using traditional clustering methods based on the distance (similarity) between partials. The clustering method was first evaluated by measuring the coherence of the clusters obtained on a set of multi-tracks recordings. This evaluation shows that the proposed approach is efficient to group partials emitted simultaneously by the same instrument. However, the case analyzed was not realistic since the partial tracking was performed on each track separately, which encounters the major problem of tracking partials on frequency regions where concurrent partials overlap.

The partial grouping method was then applied on two applications where partials are extracted directly on the mixture: the localization of vocal segments and the transcription of the vocal melody. We have proposed to locate the vocal segments by identifying vocal clusters. A vocal cluster is a group of harmonically related partials where a certain number of partials have been detected as vocal partials. This was done using the method developed in Sec.3. By comparing the results of the previous approach with the method based on clusters, we found that clusters can considerably improve the

recall of vocal segments. This evaluation shows that partial clustering improves the localization of vocal segments.

Finally, we have proposed a simple method to estimate the fundamental frequency associated with a cluster of partials. On a set of accompanied vocals where the partials were tracked directly in the mixture, we applied the partial clustering method. For each cluster the fundamental frequency was estimated and the vocal melody was obtained by considering the fundamental frequency of the vocal clusters only. This method performs relatively well.

These different evaluations (cluster purity, localization of vocal segments based on vocal clusters identification and the vocal melody transcription) show that the method proposed to identify and group vocal partials can be used as the foundation for numerous applications.

Chapter 5

Singer Identification

The lead vocal of a song attracts the attention of most listeners because the vocal melody is the most memorable element of a song. In the same way, the voice, which carries this melody, is an element that helps listeners to categorize songs. It is common that listeners identify their favorite bands by recognizing the lead singer. The voice style and the vocal technique are clues used by listeners to classify songs into music genres. For all these reasons, the automatic recognition of singers has attracted a lot of attention this past decade.

Singer identification is generally treated as an audio classification task. The underlying problem of this task is to find in the signals of songs a "voiceprint" characterizing univocally a singer's voice. In other words, a part of the solution of the problem lies in the choice of appropriate features to describe the singer's voice. The other part of the solution lies in the choice of an appropriate statistical model to build reliable singer models. In this chapter we focus on the study of appropriate features to describe signals of singing.

In the previous chapter we developed a method to recognize vocal partials using criterion based on the intonation of singing voice (periodic and monotonic variations of the time-varying frequency and amplitude capturing information related to the vocal vibrato/tremolo and the portamento/legato). The parameters extracted from partials are referred to as "intonative" features. The main idea of the present chapter is to evaluate if intonative features, that have been proven to be efficient to distinguish the voice from among other instruments, can be used as features for singer identification. We study the performance of these features and the performance of classical features computed on the spectral envelope of sounds. These features are referred to as "timbral" features. In this chapter we also propose a method to combine these features in order to improve the overall performance of singer identification systems. The underlying idea is that timbral features convey information related to the vocal tract of the singer, which is linked to a physical aspect of the singer, while intonative features are related to the style and the technique of the singer. These elements can be used intentionally to add expression but they can also reflect the abilities and limitations of the performer due to its technique. We believe that more complete singer models can be obtained by combining these complementary information. The task of singer identification is known to be challenging because the singing voice is a highly variable instrument (large tessitura, full range of phoneme, various expressive attributes) and in addition the voice is usually accompanied by other musical instruments. For these reasons we propose to evaluate how intonative and timbral features vary with changes in the voice (change pitch

or loudness) and how they vary when they are extracted on accompanied vocals.

This chapter is organized as follows. First, in Sec.1 we formalize the problem of singer identification. In Sec.2 we review some methods proposed to identify singer and artist. We also present in this section some conclusions that have been obtained from perceptual experiments on singer identification. In Sec.3 we describe the approach we developed to perform singer identification based on the combination of timbral and intonative features. In Sec.4.2 we present a set of experiments conducted to evaluate the robustness of intonative and timbral features against changes in voice and the presence of instrumental accompaniment. In Sec.5, the main points of this chapter are summarized and conclusions are drawn.

1 Problem statement

The problem of singer identification can be formally stated in a way similar to the speaker identification problem. We present the problem in the case of a closed-set identification. In this case the class set is composed with a finite of singers. For each singer a singer model is built during the training phase. The model of singer i is denoted by ω_i and the set of all models is denoted by Ω.

The goal of singer identification is to find out which singer has produced a given sample x represented by the feature vector X. In the case of closed-set identification it is assumed that x has been produced by one and only one singer of Ω. The probability that singer c (whose model is given by ω_c) has emitted X is given by :

$$P(\omega_c|X) = \frac{p(X|\omega_c)P(\omega_c)}{P(X)} \tag{5.1}$$

where $p(X|\omega_c)$ designates the conditional probability of X given the singer model ω_c. Assuming that all ω_i are mutually exclusive ($\omega_i \cap \omega_j = \emptyset, \forall i \neq j$) and collectively exhaustive ($\omega_1 \cup \omega_2 \cup \ldots \omega_n = \Omega$) the probability given by Eq.(5.1) can be rewritten as:

$$P(\omega_c|X) = \frac{p(X|\omega_c)P(\omega_c)}{\sum_{\omega_n \in \Omega} p(X|\omega_n)P(\omega_n)} \tag{5.2}$$

The identification process is a straightforward maximum likelihood classifier. The objective is to find, given the set of singer models $\Omega = \{\omega_1, \ldots \omega_N\}$, the model which has the maximum posterior probability for the input vector X. Using the Bayes' rule, the minimum error is given by:

$$\hat{n} = \arg \max_{1 \leq n \leq N} P(\omega_n|X) \tag{5.3}$$

Using Eq.(5.1) and assuming that all singers have the same prior probability ($\forall n$, $P(\omega_n) = \frac{1}{N}$) the denominator and the term $P(\omega_n)$ can be ignored in Eq.(5.3). Finally, Eq.(5.3) can be rewritten as Eq.(5.4) or as Eq.(5.5) in the log domain.

$$\hat{n} = \arg \max_{1 \leq n \leq N} p(X|\omega_n) \tag{5.4}$$

$$\hat{n} = \arg \max_{1 \leq n \leq N} logp(X|\omega_n) \tag{5.5}$$

An identification system is generally evaluated with the measure of classification performance. As explained in Chap.2 Sec.1.3.1, if the classes of the testing set are well balanced the performance can be measured with accuracy (i.e. the percentage of songs assigned to their correct singer). If the data set is not well balanced, the mean of Recall for each class is the best indicator of performance of the system.

2 Related works

The goal of singer identification (SID) is to retrieve the name of the singer performing a given song. It differs from the artist identification. An artist is defined as a music band that performs under an identifiable name. An artist can have more than one lead singer and a given singer can sing with different bands. Ideally, a singer identification system is a system able to detect when the same singer is performing in different bands. Such a system should be based on the vocal characteristics of the lead singer of a song. Most studies conducted on the identification of singers report that the main difficulty of the task lies in the presence of the instrumental accompaniment.

A large number of studies have been conducted on the task of artist and singer identification because the information obtained with such systems is essential to classify music automatically. In this section first we present methods that have been proposed to perform artist identification,then we present the ideas proposed to transform artist identification systems into singer identification systems. Finally, we present some results of perceptual studies on singer recognition.

2.1 Artist identification

When performing artist identification special care has to be taken to avoid biasing the system with the "album effect" or "producer effect". Songs coming from the same album often share similar overall spectral characteristics because of similar production (equipment, orchestration, audio effects, etc.) and post-production (additional equalization, compression, expansion, etc.) techniques. Unless care is taken, an artist identification system will identify an artists' album rather than the artist themselves as shown by Kim *and al.* in [KWP06]. Obviously this problem occurs especially when the recognition system is based on features extracted from the spectrum. To avoid this effect it is necessary to be careful when splitting the data into training and testing sets. Ideally, songs coming from the same album should be placed into the same set. It supposes that the data set is formed with songs coming from different albums of each artist.

The first study on artist identification has been proposed by Whitman [WFL01]. Mel Frequency Cepstral Coefficients (MFCC) are used to train a series of SVM classifiers. An artificial neural network (ANN) is applied on the outputs of the SVM as a "meta-learner" to assign each song to one artist. This method achieves up to 50% classification accuracy on the Minnowmatch data set composed of 21 singers. The near perfect classification accuracy obtained on the training-set suggests that this artist identification system generalizes very poorly on new data.

Using the same data set, Berenzweig *and al.* [BEL02] propose to increase the artist identification accuracy by performing the recognition on vocal segments only. They report an accuracy of 65% for a neural network (NN) classifier trained on MFCC extracted on vocal segments.

Still working with the same data set, Kim and Whitman [KW02] report no improvement by working with vocal segments instead of the whole song. Using wrapped LPC as inputs of GMM and SVM classifiers they obtain an accuracy of 45% for a subset of 17 (over 21) singers.

Cai *and al.*[CLG11] also proposed to perform the identification using vocal segments only. For each singer they train a GMM using "auditory" features (LPMCC and GTCC) extracted on vocal segments. LPMCC are a variant of LPCC computed using a Mel scale and GTCC are a variant of MFCC where the spectrum is filtered by the gamma-tone filter bank instead of the Mel-frequency filter bank. The study obtains 92.5% accuracy for a set of 10 Chinese singers. However, the data set is composed of songs coming from one album per singer. It is likely that the good performance is due to the album effect rather than the choice of the features.

Even if these previously mentioned studies work on vocal segments, they still perform artist identification since they use features extracted on the spectrum of the mixture.

An alternative approach, based on song similarity, has been proposed to retrieve the artist's name of a given song. In [ME05], Mandel and Ellis propose to represent an artist with a set of song-models obtained for each song of the artist. A distance computed between their models gives the similarity between songs. Finally, a query song is labeled with the artist's name of the song that is the most similar. For a set of 18 artists whose songs are modeled by GMM trained with MFCC, they report a maximum accuracy of 84%. In this study the similarity between two songs is computed with the Kullback-Leibler divergence approximated with Monte-Carlo method. Similar approaches for song similarity have been presented by Logan and Salomon [Log], [LS01] and West and Lamere [WL07].

An artist identification system is, by definition, not able to detect when the same singer is performing in different bands. Because of this limitation different approaches have been proposed to model the singer identity directly from the mixture independently of the instrumentals accompaniment. We present in the next paragraph these alternatives.

2.2 Singer identification

In separate studies Maddage *and al.* [MXW04] and Tsai and Wang [TW06] propose to model on one hand the instrumental background and on the other hand the accompanied vocals using the purely instrumental and the vocal portions of the song. Then, a solo-singer model is estimated by subtracting the instrumental model from the accompanied-vocals model. Both studies model singers with GMM trained on cepstral LPC. For a set of 8 singers, Tsai and Wang [TW06] report an accuracy of 84% using solo-singer models. In [MXW04] the authors compare the accuracy of SID when using solo-singer models and purely instrumental models independently or when combining both. Finally, the best accuracy (87% for 8 singers) is obtained when the decisions for instrumental and singer models are combined. These methods suppose that the stochastic characteristics of the instrumental accompaniment remain similar on purely instrumental and vocal segments. They also requires that there exist non-vocal portions in songs and that these portions are estimated accurately. Tsai and Lin propose another approach to learn solo-singer model from samples of a cappella and accompanied vocals. In [TL11] they propose a method to learn the transformation of cepstral features (MFCC) between solo and accompanied vocals. This step requires a large amount of manually mixed vocals. The transformation is estimated with techniques used previously in voice conversion studies. They evaluate the performance of their method on various sets of data (manually mixed data obtained during karaoke

sessions and real recordings). For all experiments their approach increases significantly the performance of singer identification. For the entire data set tested, whose number of singers vary from 10 to 30, they obtain an accuracy above 90% with the proposed approach. However, the performances of singer identification performed on the same data sets, without applying any instrumental background removal technique, obtain an accuracy close to 80%.

Most other approaches developed to perform singer identification work on isolated vocals. Methods to separate the voice from the instrumental accompaniment using source separation algorithm or vocal melody transcription have been presented in Sec.2.3 and Sec.2.4 of the previous chapter.

In [FKG+05], Fujihara *and al* synthesize the harmonic components related to the dominant melody estimated with the method proposed by Goto [Got04]. Next, for each frame of the synthesized melody, they extract spectral and cepstral features and compare them to a vocal and an non-vocal GMM to determine if the frame under analysis is dominated by the voice or not. The singer identification is performed using only the frames dominated by the voice. On a test-set of 10 singers they report an accuracy of 53% when the identification is performed on the mixtures and an accuracy of 95% when the identification is performed on selected frames of isolated vocals. This system for singer identification, using isolated vocals and reliable frame selection, has then been re-used by the same authors [FGKO10] to develop a system for music recommendation based on singer voice timbre.

This idea of re-synthesizing the voice has also been used by Mesaros *and al.* [MVK07]. Evaluations are conducted on a set of a cappella recordings where the same instrumental accompaniment is added with various singing-to-accompaniment ratio (SAR) for the purpose of the study. The goal of the study is to evaluate the reliability of singer models obtained on accompanied vocals. They conduct various experiments using MFCC to feed GMM, linear and quadratic classifiers and come to the conclusion that singer identification performances are greatly affected by the presence of instrumentals, especially for low SAR. For all experiments the performance is considerably improved when working with isolated vocals (obtained by synthesizing the harmonic contents related to the vocal melody estimated with the method proposed in [RK06]). For an SAR equal to zero, on a set of 13 singers, the system reaches 51% accuracy on accompanied segments, and 75% accuracy on isolated vocals.

A large set of experiments conducted on "artificial" and "real" accompanied-vocals are presented in [Bar04]. In most experiments singers are modeled using a non-parametric estimate of the spectral envelope computed from the instantaneous amplitude and frequency of signal's partials (CTF: Composite Transfer Function) or power spectral density (PSD) modeled with Gaussian distribution. In this study, the performance obtained on isolated vocals with CTF is not better than the performance obtained on accompanied vocals using PSD. Bartsch suggests that this result may be caused by the source separation methods that introduces artifacts on the solo voice. Working with a cappella recordings from 12 female singers he examines how singers' identities (modeled with the CTF) are retrieved across a set of different vowels and different pitches [BW04]. When singer models are trained on all vowels, the system yields an accuracy of 95%. The performance is reduced to 80% when the singer models are learned and tested using separated sets of vowels.

All studies mentioned in this section propose systems for an automatic singer. The problem of identifying singers has also been addressed using perceptual experiments. In the next section, we present some results found on the ability of humans to recognize singers.

2.3 Perceptual recognition of singers

Using the same recordings collection as [BW04] (12 singers, all vowels on different pitches), Mellody [Mel01] performs a series of perceptual evaluations of singer identification. The study shows that the way the singers alter their vowels as a function of pitch is singer specific and the transformations can be learned and recognized by auditors after a long training phase. After 12 hours of training, listeners are able to identify the 12 singers with an accuracy of 82% across all variations studied.

The study done by Wakefiled and Bartsh [WB03], on the same data set, shows that the long training phase is necessary. When the pitch of the query is different from the pitch used to define the singer, the listeners identify the performer of the sample with an accuracy close to chance. They conclude that listeners are not able to generalize stimulus beyond a 1/2 octave.

In [MHW01], Mellody *and al.* show that listeners are able to perceive small variations of vibrato parameters (frequency and amplitude modulations) and are also able to classify vibrato into different categories depending on the values of the sinusoidal parameters. In previous studies, Sundberg and Rossing [SR90] show that the vibrato characteristics are constants for a singer and that singers are usually not able to alter their vibrato. The two last studies mentioned do not attempt to perform singer recognition using vocal vibrato. However, they prove that vibrato is a singer-specific characteristic.

Another set of perceptual evaluations has been conducted by Erickson and Perry [EP03] to evaluate if listeners are able to recognize singers across variations in pitches. They show that for all auditors two sung tones performed by different singers with close pitches are perceived as more similar than two notes with distant pitches performed by the same singer. Using a 3-oddball and 6-oddball experiment, they also show that the ability of listeners to recognize a singer highly depends on the number and the variety of samples used to train the listeners.

The ability of listeners to identify singers across the voice's variation has not yet been compared with results obtained with machine learning methods. It is legitimate, therefore, to wonder if machines are better than human at identifying singers across the voice's variation, or if the same requirements (very large training sets that cover all possible variations) are also necessary for machines. This problem will be addressed in Sec.4. First, we evaluate the performance of timbral and intonative features (and their combination) on ideal cases of a cappella recordings when singer models are learned on samples covering the maximum variation of the voice.

3 Proposed approach for singer identification

In this section we present a method to identify singers. All studies conducted on this task report that the main difficulty lies in the presence of the instrumental accompaniment. To circumvent this problem numerous methods suggest performing singer identification on isolated vocals. These studies generally build singer models using features extracted on the amplitude spectrum, and more specifically on the spectral envelope. The use of spectral or cepstral features derived from the envelope, for tasks of speech and singing recognition, is fully justified by the interpretation of the source-filter model (as long as the signal contains voice only). In our method for singer identification we propose building singer models using these type of features combined with features computed on partials ob-

tained using the harmonic sinusoidal model. [1] In the previous chapter, we showed that features related to intonation were capable to discriminate singing from other musical instruments. According to the results of the studies presented in Sec.2.3 it is likely that these features (parameters of vibrato, relation between the rate of vibrato and tremolo and parameters of portamento computed on note transition) convey information on the singer identity. We suppose that this information is orthogonal and complementary to the information conveyed by timbral features traditionally used for audio classification tasks.

The main idea of the proposed method is to increase the performance of singer identification systems by combining timbral and intonative features.

In Sec.3.1, after a discussion on the complementary aspect of the two types of features, we give the details of the proposed approach for singer identification. This method is then evaluated on two independent sets of a cappella recordings in Sec.3.2.

3.1 Description of the proposed approach based on the combination of timbral and intonative features

The main point of the proposed method is to exploit the synergy offered by the two majors representations of audio signals and the features we can derive from them. First, in Sec.3.1.1, we discuss the complementary aspect of the timbral and intonative features obtained with the source-filter and the sinusoidal model respectively. We also detail the reasons that make the direct combination of these features impossible. To combine information offered by these different features it is necessary to perform independent classifications with each feature type separately and to combine the decisions obtained. The details of the combination method are given in Sec.3.1.2.

3.1.1 Sound descriptions complementarity

As shown by the common representation of sound, the spectrogram, a sound is a pattern varying along two dimensions: the time and the frequency axis. So, to describe a sound, one can chose to:

- describe the relative amplitudes of frequencies at a given time or to
- describe the temporal variations of one frequency (or one band of frequencies) during a given interval of time.

These two descriptions literally adopt orthogonal points of view on the sound to be described. The first description clearly corresponds to the idea developed in the source filter model while the second description corresponds to the idea that we have developed in the intonative model applied on partials obtained with the sinusoidal model. Features computed on the spectral envelope describe a characteristic of sounds that varies all along its duration. Thus, to proceed to the complete description of a sound, these features are extracted onto frames and repeated all over the duration a sound. Conversely, to be relevant, intonative features have to be computed on long portions of a signal. Vibrato parameters, obtained on time-varying frequency of partials, have to be computed on a segment covering several cycles of the modulation. The problem is similar for tremolo parameters computed on the time-varying amplitude of partials. Concerning the parameters of portamento, they have to be

[1] Details on the source-filter and the sinusoidal models are given in Chap.3

computed on a segment that covers a note transition. In the following, we compute these parameters on partials segmented using the BIC criterion as explained in Chap.3 Sec.3.2.

By analogy, with the image signal processing area, features computed on the spectral envelope of a given frame are refereed to as *local features*, while features computed on the whole duration of a note (or note transition) are refereed to as *global features*.

It is rather complex to prove formally the complementarities of these features. However, the interpretation of these features given above clearly shows that they convey non-overlapping information and that they come from orthogonal descriptions of audio signal. We note that both types of features can only be extracted on voiced portions of a signal.

In this study we use LPC, MFCC, and TECC as timbral features. The details of the computation of these features can be found in Chap.3, Sec.3.1. Experimentally we have chosen to use 15 LPC, 20 MFCC and 25 TECC. The aim of the proposed method is to increase the singer identification performance by combining complementary information on the signal to be classified. There is no point of combining various timbral features since they all convey the same information. The idea is then to combine timbral features with intonative features.

As explained in Chap.2 Sec.2 there exist different levels of combination. In this case the features cannot be combined to form a unique description of the sound for several reasons. First of all, they have different dimensions. For a given sample, decomposed into F overlapping frames, on which P partials are extracted, the timbral features lead to a matrix of features of size NxF. In contrast, the intonative features lead to a matrix of features of size PxL where N and L indicates the number of timbral and intonative coefficients respectively.

Even if it is possible to transform these two features' matrices (either by repeating or deleting information) to obtain two matrices of the same dimension that can be then compacted into a single matrix, the information conveyed by these matrices have different meanings. As explained previously, in this case it is more appropriate to combine the decisions of classifiers trained on each feature independently.

3.1.2 Combining decisions obtained with each sound description

As explained in Chap.2 Sec.2 there are two main schemes to combine classification decision: sequential and parallel. In the present case, we have only two types of information to be combined. From preliminary experiments, we assume that timbral features have a better performance than intonative features for this task. In this situation, parallel combination rules cannot be applied directly. It would be necessary to learn the combination rule using the outputs of each classification as new features. This solution, however, requires a very large amount of data to avoid over-fitting problems, as explained in Sec.2.3. The remaining solution is to use a sequential scheme of combination. The drawback of all sequential combination schemes is that there is no backward analysis of the decisions taken at each step. Therefore, if a wrong decision is made at one stage there is no chance to obtain a correct classification at the end.

Considering these last points, we propose a combination method that combines the advantages of parallel and sequential combination schemes. The underlying idea of this method is the following: the performance of any system of classification increases when the number of possible classes decreases. This makes it so that the decision of a system of classification can be combined efficiently with the

decisions of more accurate systems if the problem given to the weaker system is simplified. By "simplified problem" we mean a problem with a smaller number of classes.

In the present case, a query sample is given as input of the system of classification based on timbral features. The outputs of this system are then analyzed to retain a small number of possible classes. Then the same query sample is given as input of the system of classification based on intonative features, together with the subset of selected classes. The system returns its decisions for the subset of classes. Finally, the consensual decision is taken using classical parallel decision rules applied on the decisions of the two systems for the reduced set of classes. Here, the decision is made after two iterations because we only have two descriptions of the data. Theoretically, the process can be iterated as long as the last classifier does not return a single class or another system (a new description or a new classifier) is still available. If the method is iterated until only one class remains possible then the method works as a decision tree. Reciprocally, if the class-set is not reduced at each step, then the method is equivalent to a classical parallel scheme of combination.

We present next the general framework of the method that is specially adapted to:

– Combine classifications with different levels of performance.

– Solve problems involving a large number of classes with no hierarchical organization of the data.

– Combine a low number of representations (when a cascade classification can not be processed until only a single class remain possible).

Using the notation introduced in Sec.2 of Chap.2 we have:

- Each pattern z:
  - belongs to one of the N possible **classes** of $\Omega = \{\omega_1 \dots \omega_N\}$.
  - can be described with different **sets of features**
  - is classified with on of the available **classifiers**
- If $\mathscr{D} = \{D_1, \dots, D_L\}$ denotes the set of all available descriptions (features) of z and $\mathscr{C} = \{C_1, \dots, C_M\}$ denotes the set of available classifiers, a system of classification is given by $S^k = \left(D^k, C^k\right)$ where $D^k \in \mathscr{D}$ and $C^k \in \mathscr{C}$.
- During the sequential phase, the number of possible classes is reduced at each step k. Thus, we have: $\Omega^{k+1} \subset \Omega^k \subset \Omega^0 = \Omega$.
- For a given classification task, all systems of classification have the same original set of training samples. The training data set associated with the system S^k (i.e which is described using D^k) is denoted by T^k and the training set reduced to samples belonging to Ω^k is denoted by $T^k_{\downarrow}$.
- We assume that all classifiers outputs can be converted into membership measurement for each class : $M^k(z) = \left[m_1^k(z), \dots, m_N^k(z)\right]$. Then, for a combination involving K systems of classification and a set of classes reduced to N classes at step K, the final decision is taken by analyzing the $K \times N$ membership measurement stored in a decision profile matrix denoted by M.

Alg. 1 HybrideClassif(z,$\mathscr{S}$, T, Ω, r) Iterative algorithm to predict the class of pattern z given a set of K classification systems $\mathscr{S} = \{S^1 \dots S^k\}$, a training-set T and r a combination rule

Inputs: z, the training data set T, the set of possible classes Ω and K classifiers

Output: the predicted class for z: $\hat{w}$

begin
  $\Omega^1 = \Omega^0$
  for k from 1 **to** K **do**
    if $|\Omega^k| > 1$
      Define $T^k_{\downarrow}$ the training set reduced to pattern from class in Ω^k
      Train S^k using classifier C^k and $T^{(k)}_{\downarrow}$
      Compute M^k the membership measurements vector returned by S^k
      Deduce Ω^{k+1} the subset of the N^k most probable classes.
    else
      return $\hat{w} = \Omega^{(k)}$
  endo `// at this stage, N^K classes remain possible
  for k from 1 **to** K **do**
  `// the highest N^K values of each M^k are conserved in a matrix M.
    $M(1 : N^K, k) = M^k(1 : N^K)$
  endo
  $\hat{w} \leftarrow r(M)$ `// make the final decision using the rule r on M
  return $\hat{w}$
end

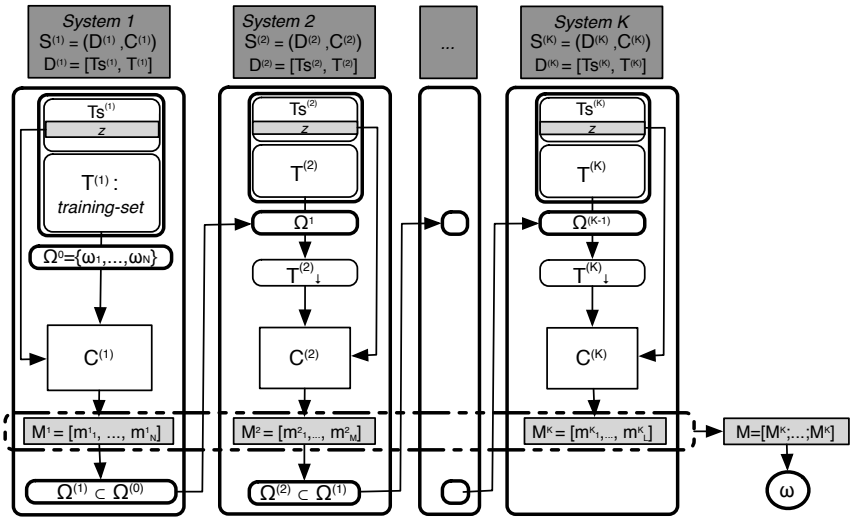


Figure 5.1: Scheme of the proposed method to combine K systems

Based on these notations, the algorithm of the method is given in Alg.1. An illustration of the method is given in Fig.5.1, where each column block represents one system of classification. The output of a classification system is given to the next classifier until no system remains available. The final decision is made by analyzing the M^k of the K systems.

We now discuss the choice of the different parameters of the method in a general case and then present the set of parameters adapted to our experiments.

Choice of parameters

- **Selection of the remaining classes:** The set of classes selected at each step can be chosen using a pre-determined relation: $N^k = \frac{N^{k-1}}{c^k}$ where c^k are defined beforehand. The classes can also be selected using a rule of thumb on the values of M^k.

- **Choice of feature space and classifier:** There is no restriction on the choice of the descriptions D^k and the classifiers C^k. Thus for $i \neq j$, the combination system can be set up with $D^i = D^j$ or $C^i = C^j$. From our experiments, the proposed method is still accurate if $S^i = S^j$.

- **Sequential organization of the S^k:** If knowledge on the relative performances of the K systems is available, the most performant systems should be placed at the top of the iterative process. If all systems have equivalent performances, or if the relative performances cannot be estimated, systems that require the lowest number of computation can be placed at the end of the process to minimize the cost.

- **Parallel combination rule:** At the end of the sequential stage N^K classes remain possible. The K vectors of measurements M^k are reduced to values of classes in $\Omega^{(K)}$ before being combined into a decision profile matrix M. We suggest to first normalize each column of M on the $N^{(K)}$ remaining classes, so that the sum of the $N^{(K)}$ membership values for each classifier is equal to one. Then the *sum-rule* is applied for the reasons explained in [KHDM02].

In the following experiments we deal with the combination of decisions obtained with two independent descriptions. The first description is given by timbral features ($D^{(1)} = TECC$ or $D^{(1)} = MFCC$ or $D^{(1)} = LPC$) and the second description is given by the intonative features ($D^{(2)} = INTO$). The combination method is tested with 3 classifiers (SVM, GMM, kNN) chosen for their variety as explained in Chap.2 Sec.1.2. The number of classes remaining at the end of the first stage is chosen dynamically. The membership values are normalized, such that their sum is equal to one. The classes that explain 80% of the cumulative posterior probabilities are retained to form the new subset of classes of size $N^{(1)}$. Once the membership measurements for the remaining classes given by the two classification systems have been normalized and concatenated to form a decision profile matrix, we apply a "sum-rule" to take the final decision.

3.2 Evaluation of the combination method

We now evaluate the proposed method. The evaluations are conducted on short samples covering either a sustained tone or a note transition. The task here is to retrieve the singer who performs a given sample. In [VM07] it is proven that listeners are able to recognize the singer of a song on a very short excerpt of the song (500 ms). As mentioned before, singer identification systems can be used to classify music automatically. Moreover, such systems can be used by music producers to detect when an excerpt of a song, to which they own the rights, has been used in a remix. If a system can retrieve the singer identity on a short sample, its performance will be even better on a longer excerpt of the song.

In this experiment we evaluate the performance of each type of feature when used independently and the performance of the global system when the features are combined. Each experiment is conducted using a 3 folds cross-validation. All experiments in this section are conducted on a cappella samples to evaluate the performance of features in an ideal case. These experiments can be considered as preliminary experiments for the identification of a singer on isolated vocals.

3.2.1 Data sets

The different evaluations are conducted on two distinct data sets, both composed of isolated sung notes. These datasets are chosen for their complementariness: a set of recordings performed in laboratory conditions by lyrical singers and a set of notes extracted from the lead vocal track of pop-rock songs. The two datasets are described below. For each set we give the detail of the construction of folds made for the experiments.

- **Lyrics Singers** The first dataset, named **LYR**, is composed of isolated notes performed by 17 female lyrical singers. The singers are classified into three levels of practice (5 advanced singers, 6 intermediate-level students and 6 beginners). All samples have been recorded in laboratory conditions with the same material (same room, microphone, position of the microphone et.c) for a study made at the University of Utah. The detailed description of the data acquisition method is given in [DRH09]. All samples composing LYR are recorded on the vowel /a/ performed in 9 different conditions:

[1] The selection is done by sorting the pseudo posterior probability in decreasing order and computing their cumulative sum

- 3 pitches : G4 (392 Hz), D5 (587 Hz) and A5 (880 Hz) each sung with
- 3 distinct levels of loudness referred to as: *p, mf, f*

The three levels of loudness are not linked to any decibel target and thus they are specific for each singer. For each pitch the singers were asked to sing at a comfortable level and then at soft and loud levels. Each note (a given pitch and a given loudness) is recorded three times. Finally, we obtained 27 samples per singer and a total of 459 samples for the whole set. The data available for one singer can be summarized as shown in Tab.5.1.

LYR	*p*			*mf*			*f*		
A5	1	2	3	1	2	3	1	2	3
D5	1	2	3	1	2	3	1	2	3
G4	1	2	3	1	2	3	1	2	3

Table 5.1: Data-set for one singer of LYR, repartition into 3 folds

For the experiments, the data set is divided into three folds. Singer models are learned from the data of two folds and the validation is conducted on the data of the remaining fold. To cover the maximum variability of one singer and obtain the most general singer model, we put into one fold samples with all available pitches and loudnesses. However, to avoid having too similar samples in the training and testing sets, all repetitions of the same note (same pitch and same loudness) are put into the same fold. In Tab.5.1 each fold is represented by one color.

On this data set, the identification task can be referred to as "closed-set, note dependent identification" by analogy with the "closed set, text dependent" classification task in the speaker recognition area.

- **Pop-Rock Singers: POP**

  The second set of isolated notes, named **POP**, has been created by segmenting the lead vocal track of "pop-rock" songs from 18 singers (8 males and 10 females). The songs, under creative common license, have been downloaded from the *ccmixter* internet site [2]. For each singer, we have randomly selected 3 songs. For each song we have manually extracted an average of 50 notes. The dataset is composed of 54 songs and a total of 2,592 notes. We do not have any information on the system of recording used for each song. By listening the different songs we can only say that some singers use the same system of recording for all songs and some others have completely different systems of recording for each song.

  On this data set evaluations are also conducted with a 3 folds cross-validation. To ensure that the identification is performed on the singer identity and not the song (album effect) we chose to put into one fold all notes extracted from one song. Thus, for each fold evaluation, the singer identity is learned using the notes of two songs, and the validation is performed on the notes of the remaining song.

  On this dataset, the task can be referred to as "closed-set, note independent classification".

[2] http://ccmixter.org/

3.2.2 Results

We report in Tab.5.2 the results obtained with LYR and in Tab.5.3 the results obtained with POP.

For both tables, the configurations of the first system of classification (S^1) are presented in the first row and the configurations of the second system of classification (S^2) are presented in the first column. The number in brackets placed beside the name of each classifier indicates the accuracy of the system when a single classification is applied. Finally, the performances obtained with the combination are reported at the intersection of the two systems used. Each performance reported corresponds the mean accuracy obtained trough the 3 folds evaluated. The task evaluated here is challenging since only a short segment (a note of few seconds length) is used to recognize the singer. We comment first on results obtained with a single classification and then we comment on the results obtained with the combination of classification.

	Feature I		**TECC**	
Feature II		SVM (84.97)	kNN (83.22)	GMM (77.56)
Into	SVM (42.48)	89.11	87.8	84.97
	GMM (42.70)	86.93	86.27	84.97
	kNN (39.22)	87.58	87.8	81.7
	Feature I		**MFCC**	
Feature II		SVM (73.42)	kNN (72.55)	GMM (65.36)
Into	SVM (42.48)	79.3	78.21	76.69
	GMM (42.70)	79.52	80.39	76.69
	kNN (39.22)	78.21	77.12	73.86
	Feature I		**LPC**	
Feature II		SVM (80.61)	kNN (77.78)	GMM (69.06)
Into	SVM (42.48)	86.93	84.97	79.3
	GMM (42.70)	86.93	83.88	79.74
	kNN (39.22)	84.97	83.88	74.29

Table 5.2: Results of combination method for lyrical singer identification (**LYR**)

	Feature I		**TECC**	
Feature II		SVM (73.57)	kNN (69.02)	GMM (69.60)
Into	SVM (50.12)	78.97	72.92	74.07
	GMM (46.91)	70.72	68.29	72.80
	kNN (43.25)	73.92	69.80	71.99
	Feature I		**MFCC**	
Feature II		SVM (64.66)	kNN (59.26)	GMM (56.21)
Into	SVM (50.12)	71.37	67.94	56.05
	GMM (46.91)	64.97	61.00	60.03
	kNN (43.25)	67.01	63.70	61.81
	Feature I		**LPC**	
Feature II		SVM (69.10)	kNN (63.81)	GMM (56.17)
Into	SVM (50.12)	74.85	69.60	66.05
	GMM (46.91)	69.92	66.74	61.23
	kNN (43.25)	70.41	65.97	64.37

Table 5.3: Results of combination method for singer pop-rock singer identification (**POP**)

Results with single classification

- **Single classification with timbral features**
  The results obtained with TECC, MFCC and LPC (first row of each table) clearly show that timbre-based features are performant to describe a singer's voice on a cappella recordings. In all cases, SVM trained with TECC yield the best accuracy. For all features, the best performance is obtained using the SVM classifier.

  We note that the results obtained on LYR are better than results obtained on POP. We can suppose that the difference is due to the "producer effect". In LYR, all samples have been recorded and produced with the same equipment. In addition, these samples have been recorded in ideal conditions (no reverberation). Thus, we can suppose that the spectral differences between these recordings are only related to differences between the voices. Conversely, the samples used to train and test singers models in the POP data set come from different songs. In many cases, these songs have different spectral characteristics created by different recording and producing equipment. We have also evaluated the performance of classification on POP when the singer models are learned on two thirds of each song and the validation is done on the remaining data. With a 3 folds cross-validation the average accuracy obtained is equal to 96% for a SVM classifier trained on TECC.

  We can conclude that timbral features (especially TECC) capture an important part of a singer's signature. However, they are not sufficient to obtain a perfect classification accuracy.

- **Single classification with intonative features**
  As far as we know, intonative features have never been used to perform singer identification. These features can somehow find an equivalent in the prosodic features used in speaker identification [CPLTB96].

  On both data sets, classifications obtained with INTO features show a relatively good performance. We remind that a random classification would reach about 5% accuracy for a problem with 17 or 18 singers. Most people think that vocal vibrato is an attribute of classical singing voice. The good performances obtained on POP clearly show that pop-rock singers also have a characteristic vibrato. The analysis of the waveforms of partials from lyrical or pop-rock singers shows that the vibratos of lyrical singers are more regular and have larger amplitude than the vibrato of pop-rock singers. The vibratos of pop-rock singer have a greater variability that can explain the better performance of these features on POP. Conversely, all singers in LYR have a fairly similar vocal technique, and their vibratos sound and look somewhat similar.

  For both data sets we found that the rate of vibrato remains rather constant within the samples of a given singer, even when these samples come from songs with different tempi. Unfortunately, the rates of vibrato across the singers are not spaced apart enough to make a clear distinction.

  Each coefficient of the intonative features has a clear meaning. We propose to evaluate the relevance of each coefficient using the Sequential Forward Feature Selection (SFFS) algorithm. This algorithm starts with a null feature set and, for each step, the best features that maximize the classification accuracy is included with the current feature set (i.e. one step of the sequential forward selection is performed). The algorithm also verifies the possibility of improvement

if some features are excluded. At the end of the search, the (sub) optimal set of features is found. The details of this algorithm can be found in [PNK94]. For both data sets, the most discriminative feature is the vibrato rate. The performance obtained with this coefficient is improved by adding the vibrato extent and the tremolo rate. In the LYR data set, the coefficients of the polynomial are not important because the recordings do not cover notes' transition. In the POP data set, which contains samples covering note transitions, the 2 first coefficients of the polynomial characterizing the portamento increase the classification performance when they are added to the features selected previously.

Results with combination In most cases, the combination of timbral and intonative features increases the classification accuracy. In average (over all experiments per data set), a gain of 6.23% and 4.48% is obtained on LYR and POP respectively. The improvement given by the intonative feature is higher when the timbre-based classification has a lower performance. The gaps between the different systems performance is greatly reduced with the combination of classification. For instance, the variance of the results obtained with LPC only for all classifiers is equal to 15. When the classifications based on LPC are combined with INTO features, the variance is reduced to 3.44.

This combination method has been developed because none of the traditional combination methods provided an improvement of the performance already obtained with timbre-based features. In practice, it is rather difficult to improve a high classification accuracy by introducing information leading to a lower accuracy. So, we can conclude that the proposed method solves this challenging problem. The singer identity of a short sample is retrieved with an accuracy of about 90% and 80% for LYR and POP respectively when combining the timbral and intonative informations.

4 Proposed approach to evaluate the features robustness for singer recognition

The ultimate goal of singer identification is to find a "voiceprint", univocally characterizing a singer voice. This voiceprint should remain invariant through all excerpts performed by the same singer. However, voice is a highly variable instrument and this task is far from being straightforward even when the signals analyzed are performed a cappella. In this section we propose a method to evaluate the robustness of timbral and intonative features against variations that appear on the voice itself along a single song, and against variations introduced by the presence of instrumental accompaniment.

The characteristics of a voice can change considerably over a short excerpt of singing. A single musical phrase is formed with multiples notes sung on different phonemes and with various expressions (or musical intention). Each note is characterized by a pitch, a duration and a dynamic. Each phoneme is characterized by a specific set of formants. To built robust singer models it is necessary to find features that remain invariant against the variation of pitch, intensity, phoneme, expression etc. In the following, we refer to these variations as "intra-song" singing voice variations since all of them can occur along the melody of a song and they only depend on the singer voice. These variations are opposed to "inter-song" singing voice variations. The latter refer to the variations created by the changes of accompaniment and changes of post-processing treatments that can be found through different songs performed by the same singer. In the review of works presented in Sec.2 we have seen

that the "inter-song" voice variations created by changes in phonemes have already been studied in [BW04]. The study shows that the singer identification was affected when the phonemes of the query samples were different from the phonemes used to train the singer models. We propose to evaluate, in Sec.4.2.2, the robustness of intonative and timbral against variations of pitch and intensity. In addition, we have seen that some studies attempt to perform singer recognition using features extracted on vocal segments. We have presented such systems as artist identification systems because these methods extract spectral features from accompanied vocals. It is legitimate to wonder, regarding the performance of these studies, if spectral features can capture information on the singer's voice when they are extracted on accompanied vocals. We perform a second experiment, presented in Sec.4.2.3, to evaluate how the performances of singer identification systems, trained with intonative or timbral features, are affected by the presence of instrumental accompaniment.

4.1 Description of the proposed approach

Theoretically, a singer is similar to him/herself when he/she repeats the same note with the same pitch, same intensity, same phoneme, same expression etc. So that, any features used to characterize a voice should remain invariant on different repetitions of the same expert. Thus, the best performance of any feature for a system of recognition is obtained when the testing samples are similar to the samples of the training-set. To evaluate the performance of a feature against a specific variation, we propose to put into the training and testing sets samples that show significant variations. For instance, if the variation studied is the change of vowel, we can learn the singer models on all vowels except for /a/ and perform the identification on samples sung on /a/. If the feature studied is robust against the change of vowel, the performance for this evaluation should remain more or less similar to the performance obtained with models trained on samples covering all possible vowels. To evaluate the robustness of a feature against a specific variation it is necessary to have a baseline for the comparison.

4.2 Evaluation of the features of robustness against intra-song and inter-song variations

4.2.1 Data set

The LYR data set, described in Sec.3.2.1, is appropriate to study the robustness of features against change in pitch and loudness. In addition, the repetitions of the same note (pitch and loudness) can be used to set a baseline for the experiments. Since all samples of LYR have been recorded and processed with the same equipment, we can ensure that the experiments are not biased in regards to the produced effect.

To study the robustness of features against change in the instrumental accompaniment, we create a new data set by mixing these a cappella samples with excerpts of instrumental tracks. We choose two mono-instrumental tracks: a guitar track with a fast tempo and a rich harmonic structure, and a piano track with a slow melody accompanied by chords on the down beats. The accompanied vocals are created by mixing a cappella samples with one instrumental track. Before mixing the instrumental and the a cappella track, the energy of both tracks are normalized to have equal root-mean-square (RMS) power. For each mix, the singing-to-accompaniment ratio (SAR), as defined in [MVK07], is then equal to 0. Like in the previous experiments (Sec.3.2), all evaluations are conducted using 3 folds

cross-validation. The data set is split into three folds, two folds are used to train the singer models and the validation is done on the data of the remaining fold. The 3 folds cross-validation is done by rotating the folds.

4.2.2 Results for intra-song variations on a cappella recordings

To evaluate the performance of features against the pitch variation, we placed all samples with the same pitch into the same fold to create 3 folds. Then we evaluate if singer models learned from two pitches (e.g. D and A) are accurate to identify the singer of a sample sung on the remaining pitch (in this case G). Similarly, to study the robustness against variations of loudness ("Loud"), the three folds are created by placing all samples with the same loudness into the same fold.

The performances for pitch and loudness variations are denoted by "pitch" and "loud" respectively. The baseline, obtained on the repetition of the same note, is denoted by "rep". These performances, for GMM, SVM and kNN classifiers are plotted in Fig.5.2, 5.3 and 5.4. On all figures, the accuracies obtained after voting are plotted with dashed lines.

When working with spectral features, one decision is made for each frame. To take a unique decision for a sample to be classified, the classes of each frame is analyzed and the class that occurs the most often is chosen to label the sample (majority voting). The decision for samples described with intonative features is made in the same way by studying the classes assigned for each partials of the sample under analysis.

In this study, partials are extracted with two strategies denoted by "Harmo" and "Inharmo". In the case of "Harmo", the fundamental frequency of the singing melody is given as input of the partials tracking algorithm. Thus, the algorithm tracks partials harmonically related to this f_0. In case of Inharmo, all partials of the signal are tracked and the vocal partials are detected with the 3^{rd} step of the method described in Chap.4 Sec.3.1.

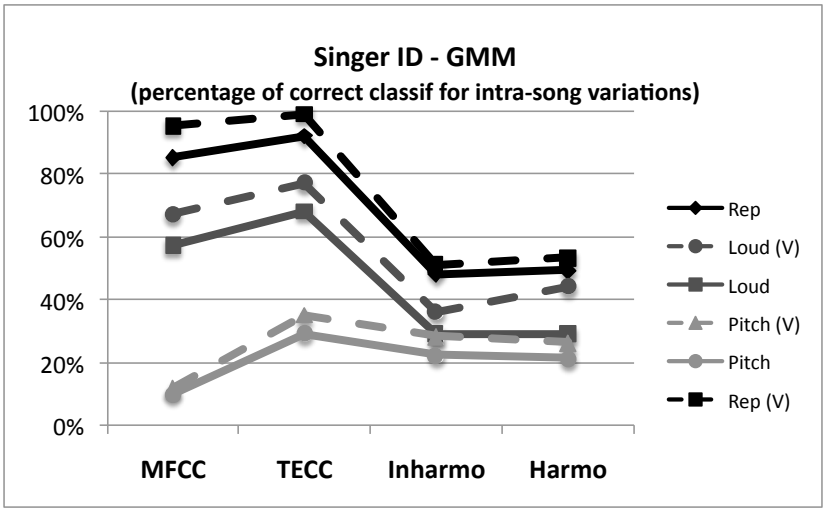


Figure 5.2: Classification accuracy for pitch and loudness variations for GMM. Results obtained after voting (V) are plotted with dashed lines

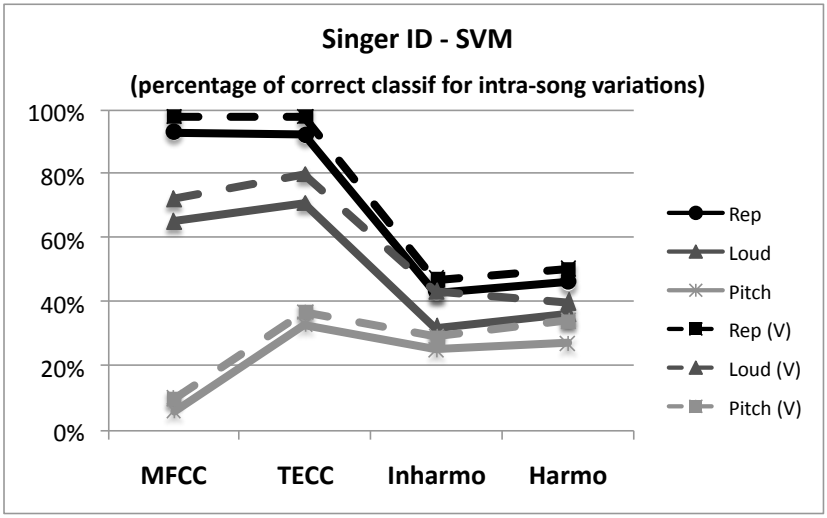


Figure 5.3: Classification accuracy for pitch and loudness variations for SVM. Results obtained after voting (V) are plotted with dashed lines

Timbral features As proven by the results of "Rep", timbral features (MFCC and TECC) are highly capable of modeling singers on a cappella recordings, as long as the query samples do not show major differences with the samples used to train the singer models. Indeed, the TECC, after voting, reach a near perfect classification and the performance of MFCC is higher than 95% accuracy. It should be noted that all singers in the LYR data set have a fairly similar vocal technique since they all study classical opera singing in the same school. The performance reached by timbral features to identify these singers across repetitions of the same notes is certainly much higher than the performance that any listeners would reach for the same task. From the results obtained by varying the loudness ("loud"), we can deduce that the relative amplitudes of frequencies of a sound vary slightly with the dynamic. For both feature sets, the performance is decreased by about 20%. Still, the performance of TECC is better than the performance of MFCC. We could have expected that all frequencies were raised or reduced in the same way when changing loudness. In this case, the performance would not have been affected since MFCC and TECC are both obtained with the DCT coefficients computed on the estimated envelope except for the first coefficients (which indicates the mean values of these coefficients).

From the experiments conducted on the variation of pitches, we can confirm that timbre-based features are strongly correlated to pitch. The overall spectral shape changes considerably when changing pitch and features based on the spectral envelope are not robust at all against the pitch variations. We can suppose that the very low performance obtained with the MFCC is due to the bad estimation of the envelope obtained by the cepstrum on high-pitched harmonic sounds. As mentioned in Chap.3 Sec.3.1 the true envelope is the most appropriate alternative in this case. It is also possible that the performance degradation is caused by the formant tuning.

These results are not very surprising. Similar conclusions have been obtained from perceptual experiments [EP03], and [WB03]. We propose to verify, in the next experiments, if and automatic system for singer identification has the same limitations as listeners shown [EP03], i.e. notes with close pitches are perceived as more similar than notes with distant pitches, even when the notes with

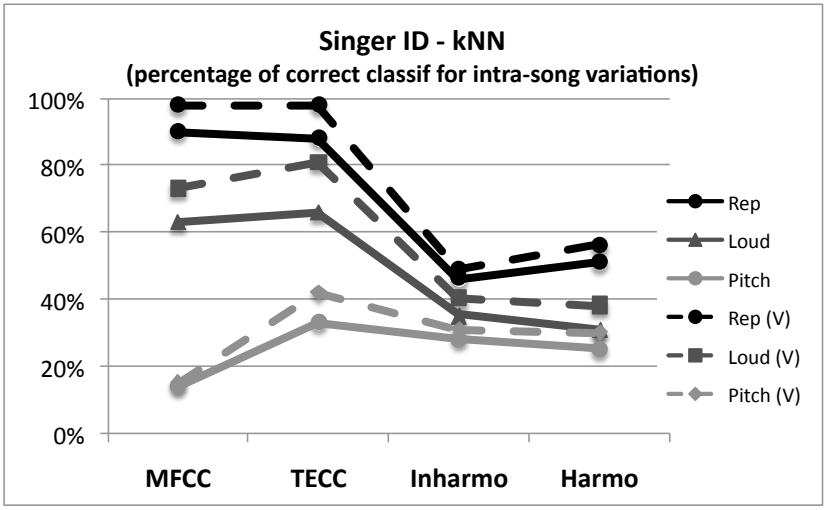


Figure 5.4: Classification accuracy for pitch and loudness variations for kNN. Results obtained after voting (V) are plotted with dashed lines

close pitches are performed by different singers. We propose to measure if the distance between two note-models from the same singer with distant pitches is smaller than the distance between two note-models with equal pitch sung by two different singers. This experiment is performed independently with TECC and MFCC features. For each singer, we build three note-models (GMM). Each note-model is trained with all samples performed on a given pitch. These models can also be interpreted as pitch-based singer model. The distance (or similarity) between two pitch-based singer models is given by the distance between their respective models computed with the Kullback-Leilber (KL) divergence. Working with mixtures of gaussians, the KL divergence is approximated using the Monte-Carlo method and converted into a distance using the method proposed in [ME05]. We report in Fig.5.5 and the distance matrices of three note-models of two singers for MFCC and TECC respectively. In this figure, the pitch-based models of the first singers are denoted by A1, D1 and G1 and the ones of the second singers by A2, D2 and G2.

As shown on these figures, we can observe that the models obtained with MFCC are strongly correlated to the pitch. This is because the note models of two different singers computed on the same pitch are always more similar than the note models of the same singers computed on a different pitch. The models computed with the TECC are clearly more robust to pitch variations. We remark that the distance between the note models is not proportional with the interval between the pitches. Contrary to the conclusions given in [HE01], which is that the timbre remains invariant along an octave, we found that the timbre varies considerably between notes spaced apart by a 5^th. The same experiments conducted with all singers lead to the same conclusions.

Timbral features Returning to the original experiment, the performance of the intonative features (Fig.5.2–5.4), based on the characteristics and variations of the fundamental frequency (and its harmonic partials), is clearly below the performance of timbre-based features. However, they present the advantage of being less affected by variations of pitch and loudness than timbral features. We note the performance of Harmo is slightly better than the performance of Inharmo feature sets. On the baseline

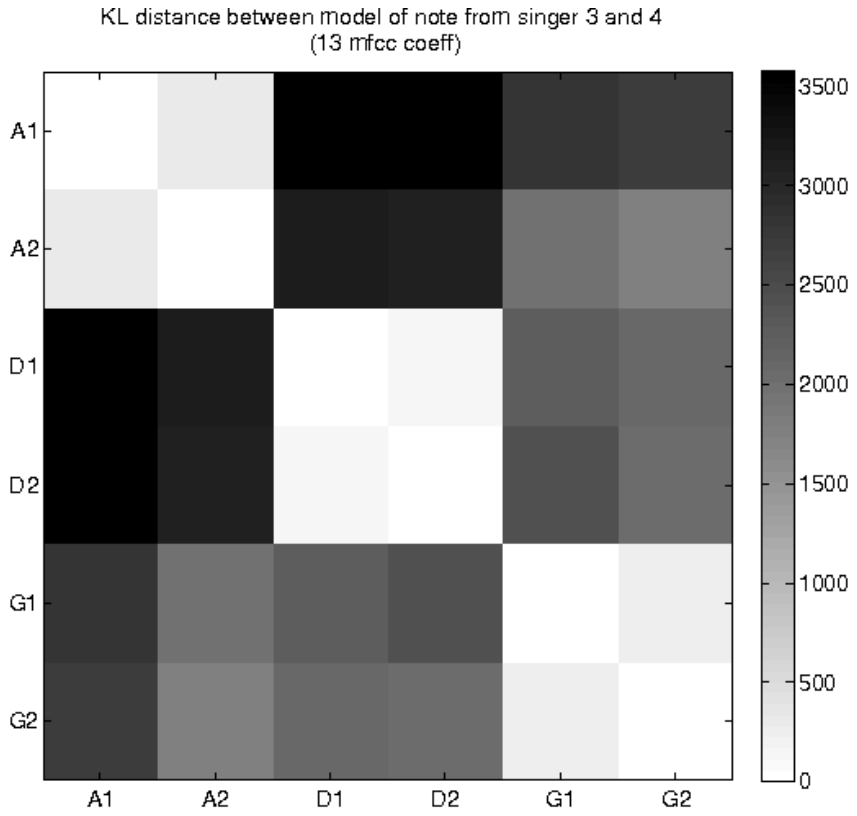


Figure 5.5: Distance between note-model for 2 singers with MFCC

experiment, intonative features reach 55% accuracy, which is about ten times the random accuracy for a problem with 17 classes. The performance drops to 40% for "pitch" and "loud" experiments.

In studies conducted on the control of vibrato, Maher [Mah08],[MB90] suggests that the control of vibrato depends on the level of practice of singers. We propose in the next experiment to evaluate if the intonative features remain more invariant for advanced singers than beginners. The singers of the LYR data set are grouped into three categories: Advanced (5 singers), Intermediate (6 singers) and Beginners (6 singers). We report in Fig.5.7, the results of the same experiments for each category of singer.

The vibrato/tremolo of trained singer seems to be more constant against all variations studied than the vibrato of beginners. For the trained singers, the vibrato is not affected by change in loudness. For all level of practice, the vibrato parameters vary with the pitch.

The results obtained with the "Rep" experiments show that the singer identification is accurate when there is no major voice variation between the samples of the training and testing sets. The only way to obtain an accurate system of classification is probably to train the models with a very large number of samples. Ideally, the training set should contain samples with all possible pitches sung on all possible vowels. We now investigate if the timbral and intonative features are still able to recognize a singer when different experts of instrumental tracks accompany the repetitions of the same note.

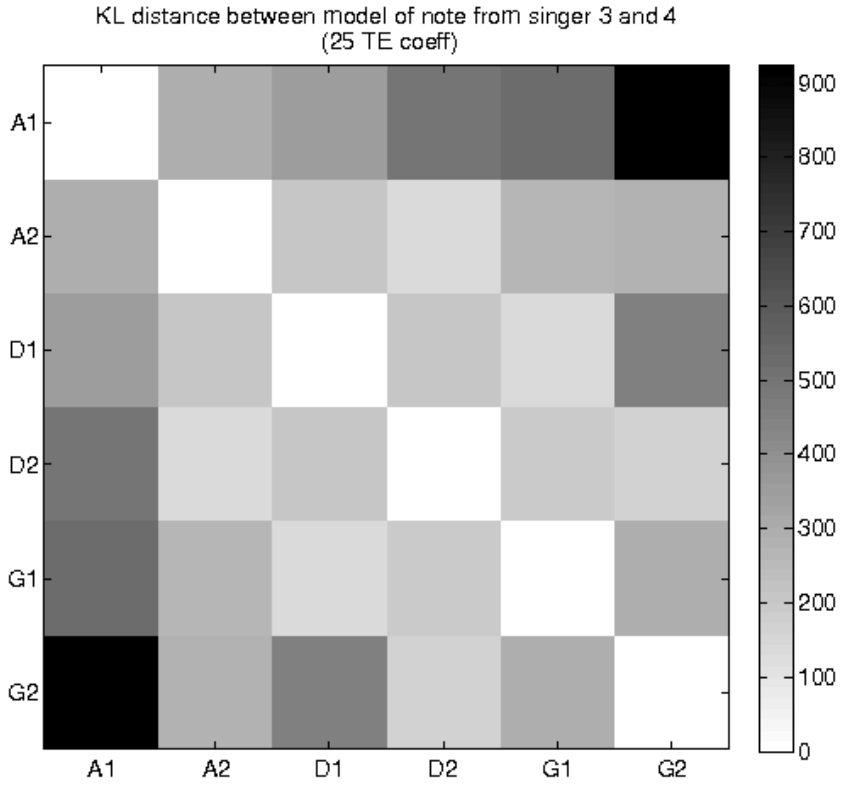


Figure 5.6: Distance between note-model for 2 singers with TECC

4.2.3 Results for inter-song variations

To study the robustness of features against variation in the instrumental accompaniment we propose two experiments. Similar to the previous experiments, evaluations are conduced with a 3 folds cross-validation.

The first experiment is done to evaluate the performance of features to retrieve the singer identity when the same instrument accompanies the voice. We conduct this evaluation twice: once when the a cappella samples are accompanied by a guitar, and again when the samples are accompanied by a piano. The second experiment is conducted to evaluate if the singer is still recognized when different instruments in training and testing sets accompany the voice. For this evaluation, the first repetition of each note is mixed with an excerpt of the piano track, the second repetition is mixed with an excerpt of the guitar track, and the third repetition is left a cappella. This experience is denoted by **P/G/A**. For each sample, the excerpt of the instrumental track is chosen randomly. As in the previous experiments, the baseline is the evaluation conducted on a cappella samples (where singer models are learned using two repetitions of the same note and the evaluation is done on the third repetition).

In the intra-song variation experiments we found no major differences between the performances of the different classifiers. In Fig.5.8 and Fig.5.9 we plot the results (before and after voting) obtained with GMM.

We first comment on the results obtained before voting for timbral feature and then for intonative

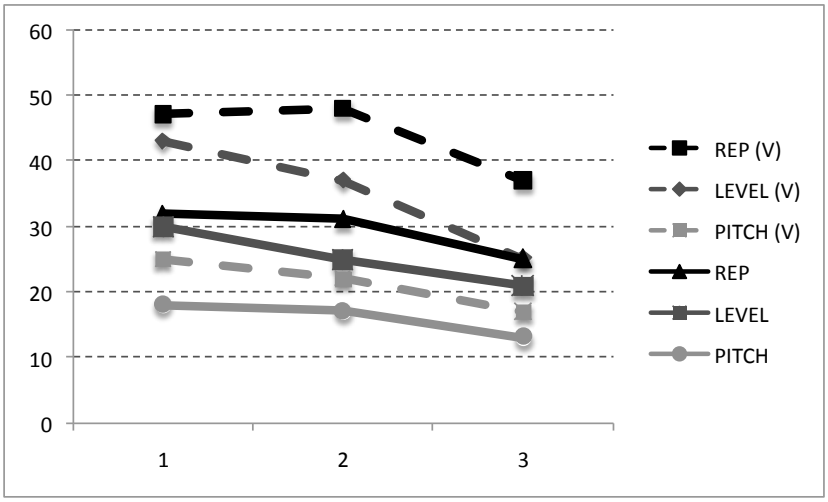


Figure 5.7: Performance of intra-song variation per of level practice : 1 Advanced, 2 Inter, 3 Beginner. Results after voting (V) are plotted with dashed lines

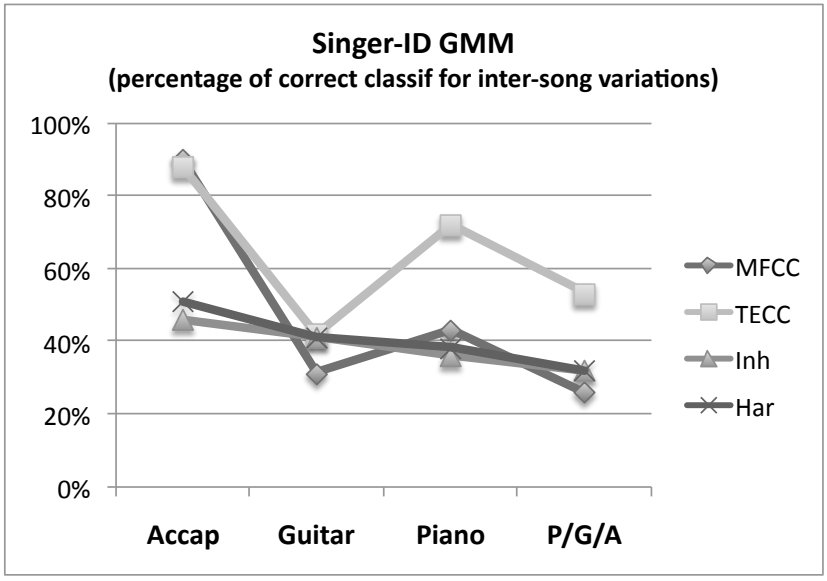


Figure 5.8: Classification accuracy for various accompaniment before majority voting

features.

Timbral features These results show that timbral features are clearly affected by the presence of instrumental accompaniment. We note that the performance degradation is correlated to the "density" of the instrumental track: with the piano accompaniment, which is composed of a slow melody accompanied by few chords, the results are much better than the results obtained on the samples accompanied by guitar (which is very dense harmonically and very fast). In all cases, the TECC clearly outperform the MFCC. The performance of MFCC, which obtain about 90% accuracy on a cappella samples, drops to 25% for the P/G/A experiments. We can assume that these coefficients capture more information on the instrumental background than on the singing voice. Obviously, when the

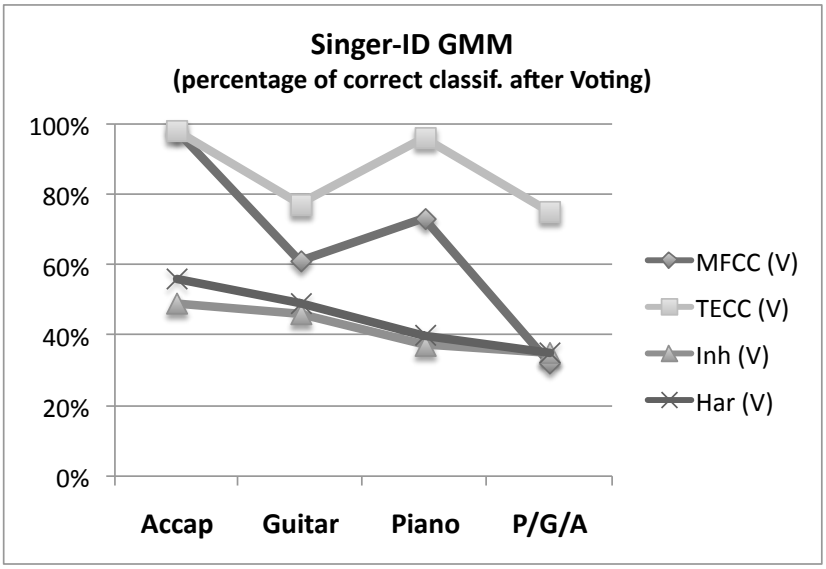


Figure 5.9: Classification accuracy for various accompaniment after majority voting

mixture analyzed is formed by a voice plus another instrument, the spectral envelope completely lost its primary function, which is to describe the vocal tract of the singer. In the related works, we have seen that numerous studies perform the singer identification based on these timbral features and they obtain relatively good performance. It is very likely that the evaluation of these systems is biased by the album effect.

Intonative features We note that the performances of intonative features are much less affected by these variations than timbral features. It appears that these features even obtain better results on accompanied samples than MFCC. Theoretically, if all sung partials were retrieved independently of the instrumental background the results for intonative features should not vary with accompaniment changes. In practice, there is a degradation that can be caused by a bad selection of a sung partial. The Harmo feature-set is composed of partials harmonically related to the f_0 of the sung melody so that it is composed of sung partials only. This f_0 was estimated, with YIN, on the a cappella versions of these samples before the mixing step. But the performance for Harmo is not much better than the performance for Inharmo, so we can deduce that the degradation is not due to the vocal partial selection. The main problem certainly remains in the overlapping frequencies of the different audio sources. When two partials of different audio sources overlap in a frequency, the amplitudes on this region are modified. Thus, the relation between tremolo and vibrato amplitude values, which convey interesting information on the singer vocal tract (see in Chap.3 Sec.2.4), is modified.

The performances of timbral features are considerably improved after voting. We can assume that an important proportion of frames are dominated by the voice. The only consideration taken when mixing the vocal and instrumental track was to have an SAR equal to 0 for the whole expert. On frames dominated by the voice, the recognition should be performed more accurately as shown by the results of [MVK07]. We note also that the voting rule applied on the classes assigned for each partial, in the case of intonative features, do not improve the performance significantly. This is due to the fact

that there are only a few vocal partials per sample and less generalizations can be made.

5 Summary and conclusions

In this chapter we have investigated the problem of singer identification. This problem has been addressed as a problem of audio classification and the emphasis has been put on the study of appropriate features to capture the voice characteristics of singers. Two complementary approaches have been chosen to describe the content of singing signals: the traditional description based on the source-filter model interpretation which consists of describing the spectral envelope of stationary portions of signal, and an additional description based on the description of partials obtained with the sinusoidal model. The latter, which gives an interpretation of the temporal variation of frequency and amplitude of partials had never been used before for the characterization of a singer's voice.

In this chapter we have conducted a series of experiments to evaluate different aspects of these two types of features. All experiments were conducted on samples of voice representing a sustained note or a note transition. To give a more complete evaluation of these features we chosed to perform the experiments on two distinct sets of singers: 17 lyrical singers and 18 pop-rock singers. Using a cappella samples, we have evaluated the performance of timbral and intonative features to recognize the singer of a given sample when the singer models are trained with samples covering the maximum variability of the singers' voice. Then, using accompanied vocals created for the purpose of the study, we have evaluated the capacity of these features to retrieve the singer identity when the voice is accompanied.

In Sec.3 we have shown that intonative features can be used in supervised learning systems for singer identification. On two distinct sets made of a cappella samples of 17 lyrical and 18 pop-rock singers they obtain about 45% accuracy. This performance, however, is much lower than the performance of timbral features such as TECC, MFCC and LPC, which lead to a correct classification of about 75%. We have shown that these two descriptions convey complementary information and then can be combined to improve the overall performance of singer identification systems. We have proposed a method to combine these features astutely. The proposed combination method suggests one uses the timbral features to determine the most likely classes for a sample to be classified, and then to use to intonative feature to refine the decisions on this selected subset of classes. The proposed approach obtains good results. In the best cases, the classification after combination reach 90% and 80% of correct classification for the lyrical and pop-rock singers respectively. These performances are obtained using TECC and INTO features with a SVM classifier.

This evaluation has been conducted on a cappella samples to verify if these features are really capturing information on a singer's voice in ideal cases. In previous studies on this task it has been suggested to perform singer recognition on isolated vocals. If methods to separate sources or to re-synthesize the lead vocal of a song had a good performance this method could be applied. The performance of singer identification systems highly relies on the choice of the data used to train the singer models.

Indeed, the singing voice is a highly variable instrument whose characteristics can change considerably over an excerpt of few seconds. In a set of experiments performed in Sec.4.2.2 we have shown that most features are not invariant against variations of pitch and loudness. Therefore, to

obtain an accurate singer model it is necessary to train the model using samples covering the whole *tessitura* of the singer. Ideally, the set of training samples should cover all possible variations of the singer voice (pitch, loudness and phoneme). When there exists in the training set a sample similar to the query sample, a supervised learning approach will classify this sample accurately. In a last series of experiments, presented in Sec.4.2.3, we have evaluated if the classification remains accurate when the similar samples of singing voice are accompanied by a different excerpt of the instrumental track. This experience was conducted using different excerpts of the same instrumental track and using excerpts of instrumental tracks played by different instruments. In both cases, the performance of timbral features depends on the nature of the accompaniment. When the accompaniment is very dense harmonically and played in a fast tempo the performance drops drastically when considering the frame-by-frame classification obtained with timbral features. By using the information obtained on successive frames the classification accuracy can be improved significantly if a large portion of frames is dominated by the voice. The results obtained by TECC for these experiments, after voting are rather impressive compared to the performance of MFCC. In general, intonative features are much less affected by the presence of instrumental accompaniment.

From all these experiments, we retain that the use of intonative features can increase the performance of singer identification when they are combined with timbral features. These features present the advantage of being much more robust to intra and inter song variations. We also note that the TECC clearly outperform the traditional MFCC for these experiments. We have also proven in this chapter that the singer identification task is more complex than it appears at first sight. Due to the variability of the voice, to obtain accurate singer models, it is necessary to be very careful when choosing the set of data to train the singer models.

Chapter 6

Conclusions

1 Summary

In this dissertation we addressed the problem of characterizing the singing voice in order to differentiate the singing voice from the other musical instruments and to differentiate singers from one another. This research was motivated by the fact that the singing voice is the element that attracts the attention of most auditors and it is common to retrieve a given song using information related to the singing voice: the name of the singer, the melody carried by the singer, the lyrics of the song, etc. Considering these points, it has became necessary to develop systems that are able to extract information related the lead vocal of a song, in order to automatically organize large collections of music. The originality of the work presented in this document is in regards to the descriptions of the singing voice using long-term features related to the temporal variations of the fundamental frequency that appear naturally in singing: the vibrato and the portamento. This idea came from the observations of spectrograms, on which it is fairly easy to follow over time the spectral components corresponding to the singing voice: the sustained notes are characterized by the presence of harmonic partials whose frequency trajectories follow a quasi-periodic modulation and the transitions between notes are characterized by a continuous variation of the fundamental frequency from one pitch to another. In previous studies on the understanding of the vocal production and on perception of singing, it has been shown that the temporal variations of the frequency in singing help the voice to be easily heard, even in a presence of a loud accompaniment. It has also been demonstrated that the vibrato is a natural effect of singing that can hardly be modified. The purpose of this study was to evaluate if the parameters of the vocal vibrato and the portamento can be used in pattern recognition methods to distinguish the singing voice among the other instruments and to model the signature of a singer. More precisely, it was proposed to evaluate if the information conveyed by the features extracted on the temporal variation of the singing voice is complementary to the information conveyed by features which aim to characterize the timbre of the singing voice. These features could thus be combined to form a more complete description of the signature of a singer.

In practice, the parameters of vibrato and portamento are extracted from the time-varying frequency and amplitude of partials obtained with the harmonic sinusoidal model. To obtain an accurate description of the vibrato and portamento, we proposed to model each partial frequency as the sum of a slow and continuous variation to represent the portamento, plus a quasi-periodic modulation which

represents the vibrato. The coefficients obtained with this model form a set of features referred to as intonative features. Contrary to the features computed on the amplitude spectrum, the intonative features aim to describe the elements related to singing voice in a polyphonic mixture independently of the other instruments accompanying the voice.

In a first set of experiments we showed that vocal partials could be distinguished among the partials of other instruments using a series of decisions based on the characteristics of the vocal vibrato and the portamento. Typically, the vocal vibrato corresponds to a quasi-sinusoidal modulation of the frequency with a rate comprised between 5 and 8 Hz and is quasi-constant for a given singer. The vocal vibrato extent is relatively large compared to the other instruments. The vocal vibrato has the particularity of being accompanied by a modulation of amplitude created by the filter that constitutes the vocal tract of a singer. The singing voice also has the particularity of being highly harmonic. To improve the recognition rate of vocal partial, we proposed to automatically group harmonically related partials using a set a comparison functions, which can be compared to CASA cues. Once the vocal partial of a mixture are identified, the vocal segments of a song can be directly localized. On these segments, the voice can be isolated from the instrumental accompaniment and the vocal melody line can be transcribed.

In this document we evaluated if the features extracted from the vocal partials can also be used to characterize the signature of a singer. The results obtained through a series of experiments show that features related to the intonation describe a part of a singer's identity, and these features can be combined with features related to the timbre to improve the overall performance of singer identification. The improvement lies in the complementary information offered by the two descriptions of the singing voice.

The research presented in this document is a step toward a novel approach to describe the musical signals based on the temporal variations of the frequencies and amplitude of the frequencial components of harmonic sounds. The results obtained with the proposed description of sound are encouraging. For the task of vocal segments localization, the proposed features obtained a performance comparable with the classical approaches (i.e. pattern recognition methods based on features extracted from the amplitude of the sort-term spectrum). For the singer identification task, we showed that intonative features convey information complementary to the information conveyed by timbral features, and can thus be combined to increase the recognition rate of singer identification.

2 Future works

The performance of the proposed methods could be further improved using a better estimation of the vibrato parameters and a method to automatically determine the best segment of f_0 on which to perform the analysis. Concerning the method for vocal partial recognition, the set of thresholds associated with each decision functions could be better optimized if they were estimated conjointly. In the method developed to group harmonically-related partials, the weights of the comparison functions were found empirically on a training data set composed of partials extracted from mono-instrumental tracks. To increase the performance on real mixture of instruments, these weights should be learned on partials directly extracted from the mixture, which requires a special training set where groups of harmonically related partials for each instrument are annotated manually. In other words, the perfor-

mance of partial grouping can be increased by learning the difference between the partial tracking results obtained on independent tracks and on mixed tracks. Concerning the identification of singers, we showed that the combination of timbral and intonative features improves the performance of the identification on a cappella recordings. We also showed that intonative features are more robust that timbral features to describe the singing voice in the presence of accompaniment. However, the performance of timbral features on accompanied vocal recordings, especially for the TECC, is considerably improved when considering a single decision per recording obtained by a majority-voting rule. This result is due to the fact that the voice dominates the other instruments on the majority of the frames. To improve the recognition of the singer on accompanied recordings, it could be interesting to determine the frames dominated by the voice by analyzing the relative amplitude of the vocal partials against the amplitude of the partial of the other instruments.

The analysis of the temporal variation of frequency can be pushed ever further. One can imagine analyzing the temporal variations on a thinner scale in order to characterize the jitter and shimmer of a voice, which are considered elements composing the vocal quality. However, to obtain a precise measure of these small and fast variations it is necessary to work with a very precise transcription of the fundamental frequency. The temporal variations of the singing voice could also be analyzed on a larger scale to determine patterns of note sequences that could be characteristic of singers. These patterns should consider both the temporal organization of the successive notes and the specificities of the transition between the notes. Such a model could be given by a Hidden Markov Model based on the notes and note transitions characterization proposed in this document.

In addition, the features developed in this document could also be applied on other instruments, in particular the string instruments with which it is also possible to produce vibrated tones. Contrary to the singing voice, the modulations of frequency of string instruments are not related to the instrument itself and they depend primarily on the technique of the performer. It could be interesting to evaluate if these features can be used to distinguish two performers playing the same musical piece on the same instrument. More generally, it could be interesting to evaluate if these features can be used and or adapted to classify instruments.

The results obtained in this research prove that the vibrato is an important characteristic of the singing voice. The different elements highlighted in this research could also be used and adapted to create more natural and expressive synthetic singing voices. In particular, it could be interesting to develop some vibrato models adapted to different types of vocal tract. Such models could also be used to correct some imperfections on singing recordings. For instance, we can imagine an auto-tune post-processing that could correct the pitch but preserve some temporal characteristics (such as the vibrato/tremolo) of the original recordings to preserve a part of the singer's signature.

Bibliography

[ABLS02] X. Amatriain, J. Bonada, A. Loscos, and X. Serra. Spectral processing. In *DAFX*, pages 373–438. Wiley Online Library, 2002.

[AC04] I. Arroabarren and A. Carlosena. Vibrato in singing voice: the link between source-filter and sinusoidal models. *EURASIP Journal on Applied Signal Processing*, 2004(7):1007–1020, 2004.

[AF82] T. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Communication*, 1(3-4):167–184, 1982.

[AKZ99] R. Althoff, F. Keiler, and U. Zolzer. Extracting sinusoids from harmonic signals. In *Proceedings of the Digital Audio Effects (DAFx) Workshop*, pages 97–100. Citeseer, 1999.

[AS05] M. Abe and J. Smith. Am/fm rate estimation for time-varying sinusoidal modeling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 201–204. 2005.

[Aug95] F. Auger. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5), 1995.

[Bar34] W. Bartholomew. A physical definition of ÂŞgood voice-qualityÂŤ in the male voice. *the Journal of the Acoustical Society of America*, 5:224, 1934.

[Bar04] M. Bartsch. *Automatic singer identification in polyphonic music*. Ph.D. thesis, The University of Michigan, 2004.

[BDR06] R. Badeau, B. David, and G. Richard. High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials. *Signal Processing, IEEE Transactions on*, 54(4):1341–1350, 2006.

[BE01] A. Berenzweig and D. Ellis. Locating singing voice segments within music signals. *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 119–122, 2001.

[BEL02] A. Berenzweig, D. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. *AES 22nd International Conference*, 2002.

[Ben81] G. Bennett. Singing synthesis in electronic music. *Royal Swedish Academy of Music*, 33:34–50, 1981.

[BHM01] J. Brown, O. Houix, and S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *The Journal of the Acoustical Society of America*, 109:1064, 2001.

[BHT63] B. Bogert, M. Healy, and J. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, pages 209–243. New York: John Wiley and Sons, Inc, 1963.

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer New York., 2006.

[BP86] G. Bloothooft and R. Plomp. The sound level of the singerÂŠs formant in professional singing. *The Journal of the Acoustical Society of America*, 79:2028, 1986.

[Bra97] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[Bre90] A. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1990.

[BS03] J. Bretos and J. Sundberg. Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos. *Journal of Voice*, 17(3):343–352, 2003.

[Bur98] C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[Bur01] G. Burgess. Vibrato awareness. *Double Reed*, 24(4), 2001.

[BW04] M. Bartsch and G. Wakefield. Singing voice identification using spectral envelope estimation. *IEEE Transactions on Speech and Audio Processing*, 12(2):100–109, 2004.

[BZSJ09] C. Butte, Y. Zhang, H. Song, and J. Jiang. Perturbation and nonlinear dynamic analysis of different singing styles. *Journal of Voice*, 23(6):647–652, 2009. ISSN 0892-1997.

[C+97] J. Campbell et al. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[Cas93] M. Castellengo. Fusion or separation: From vibrato to vocal trill. In *Proceedings of the Stockholm Music Acoustics Conference*. 1993.

[CG01] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 2, 2001.

[CGJV01] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 34(3):267–285, 2001.

[CLG11] W. Cai, Q. Li, and X. Guan. Automatic singer identification based on auditory features. In *Natural Computation (ICNC), 2011 Seventh International Conference on*, volume 3, pages 1624–1628. IEEE, 2011.

[CLLY07] C. Cao, M. Li, J. Liu, and Y. Yan. Singing melody extraction in polyphonic music by harmonic tracking. In *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, pages 373–374. Citeseer, 2007.

[CM96] O. Cappé and E. Moulines. Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3(4):100–102, 1996.

[Coo90] P. Cook. *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing*. Ph.D. thesis, to the Department of Electrical Engineering.Stanford University, 1990.

[CPLTB96] M. Carey, E. Parris, H. Lloyd-Thomas, and S. Bennett. Robust prosodic features for speaker identification. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1800–1803. IEEE, 1996.

[CR08] A. Chanrungutai and C. Ratanamahatana. Singing voice separation for mono-channel music using non-negative matrix factorization. In *Advanced Technologies for Communications, 2008. ATC 2008. International Conference on*, pages 243–246. IEEE, 2008.

[CS92] G. Carlsson and J. Sundberg. Formant frequency tuning in singing. *Journal of Voice*, 6(3):256–260, 1992.

[dCK02] A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002.

[DG06] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM New York, NY, USA, 2006.

[DHAT99] P. Desain, H. Honing, R. Aarts, and R. Timmers. Rhythmic aspects of vibrato. *P. Desain and L. Windsor, Rhythm–Perception and Production*, pages 203–216, 1999.

[DHS95] P. Dejonckere, M. Hirano, and J. Sundberg. Vibrato, ch. 2. *Singular Pub., San Diego*, 1995.

[DHS01] R. Duda, P. Hart, and D. Stork. *Pattern classification*, volume 2. Wiley, 2001.

[Die98] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[DLR+77] A. Dempster, N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[DRD09] J. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 105–108. IEEE, 2009.

[DRDF10] J. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):564–575, 2010.

[DRH09] C. Dromey, L. Reese, and J. Hopkin. Laryngeal-level amplitude modulation in vibrato. *Journal of Voice*, 23(2):156–163, 2009.

[Dri01] J. Drish. Obtaining calibrated probability estimates from support vector machines. In *Final Project for CSE 254: Seminar on Learning Algorithms*. Citeseer, 2001.

[Dud39] H. Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, 11:165, 1939.

[EB03] J. Eggink and G. Brown. Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In *Proc. ISMIR*, pages 125–131. Citeseer, 2003.

[EB04a] J. Eggink and G. Brown. Extracting melody lines from complex audio. In *Proc. ISMIR*, pages 84–91. Citeseer, 2004.

[EB04b] J. Eggink and G. Brown. Instrument recognition in accompanied sonatas and concertos. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 4, pages iv–217. IEEE, 2004.

[EJM91] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *Signal Processing, IEEE Transactions on*, 39(2):411–423, 1991.

[EK00] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II753–II756. IEEE, 2000.

[EP03] M. Erickson and S. Perry. Can listeners hear who is singing? a comparison of three-note and six-note discrimination tasks* 1. *Journal of Voice*, 17(3):353–369, 2003.

[ERD06] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):68–80, 2006.

[Fan81] G. Fant. The source filter concept in voice production. In *IV FASE Symposium on Acoustics and Speech, Venezta*. 1981.

[Faw04] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, 31, 2004.

[FGKO10] H. Fujihara, M. Goto, T. Kitahara, and H. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):638–648, 2010.

[FKG+05] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. *Proc. ISMIR*, pages 329–336, 2005.

[FKG+06] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume 5. 2006.

[FS95] Y. Freund and R. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[GL+88] D. Griffin, J. Lim, et al. Multiband excitation vocoder. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(8):1223–1235, 1988.

[GN02] S. A. Giusti N, Massulli F. Theoretical and experimental analysis of a two-stage system for classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2002.

[Got04] M. Goto. A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[GR91] T. Galas and X. Rodet. Generalized functional approximation for source-filter system modeling. In *Second European Conference on Speech Communication and Technology*. 1991.

[Gre75] J. Grey. An exploration of musical timbre. *Journal of the Acoustical Society of America*, 1975.

[GTA08] M. Gonen, A. Tanugur, and E. Alpaydm. Multiclass posterior probability support vector machines. *Neural Networks, IEEE Transactions on*, 19(1):130–139, 2008.

[Har78] F. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[Har88] W. Hartmann. Pitch perception and the segregation and integration of auditory entities. *Auditory function: Neurobiological bases of hearing*, pages 623–347, 1988.

[HBPD03] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003. ISSN 0929-8215.

[HCL+03] C. Hsu, C. Chang, C. Lin, et al. A practical guide to support vector classification, 2003.

[HdD01] N. Henrich, C. d'Alessandro, and B. Doval. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. In *Seventh European Conference on Speech Communication and Technology*. 2001.

[HE01] S. Handel and M. Erickson. A rule of thumb: The bandwidth for timbre invariance is one octave. *Music Perception*, 19(1):121–126, 2001.

[Hen01] N. Henrich. *Study of the glottal source in speech and singing: Modeling and estimation, acoustic and electroglottographic measurements, perception*. Ph.D. thesis, Univ. Paris VI, 2001.

[HH88] Y. Horii and K. Hata. A note on phase relationships between frequency and amplitude modulations in vocal vibrato. *Folia phoniatrica*, 40(6):303, 1988.

[HHS02a] T. Ho, J. Hull, and S. Srihari. Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(1):66–75, 2002. ISSN 0162-8828.

[HHS02b] T. Ho, J. Hull, and S. Srihari. On multiple classifier systems for pattern recognition. In *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, pages 84–87. IEEE, 2002. ISBN 0818629150.

[HJ10a] C. Hsu and J. Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2):310–319, 2010.

[HJ10b] C. Hsu and J. Jang. Singing pitch extraction at mirex 2010. *Music Information Retrieval Evaluation eXchange Audio Melody Extraction Contest Abstracts*, 2010.

[HKV09] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. *10th ISMIR*, pages 327–332, 2009.

[Hor89] Y. Horii. Frequency modulation characteristics of sustained//sung in vocal vibrato. *Journal of Speech, Language and Hearing Research*, 32(4):829–836, 1989.

[HS93] Y. Huang and C. Suen. The behavior-knowledge space method for combination of multiple classifiers. In *IEEE Computer Society connferenceon computer vision and pattern recognition*, pages 347–347. IEEE, 1993. ISSN 1063-6919.

[IA79] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Electron. and Commun. in Japan*, pages 10–17, 1979.

[Jin04] P. Jincahitra. Polyphonic instrument identification using independent subspace analysis. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 2, pages 1211–1214. IEEE, 2004.

[Jol02] I. Jolliffe. Principal component analysis. In *Encyclopedia of Statistics in Behavioral Science*. Wiley Online Library, 2002.

[JZ02] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 2002. ISSN 0162-8828.

[Kay88] S. Kay. *Modern spectral estimation: theory and application*. Prentice Hall, 1988.

[KGV78] K. Kodera, R. Gendrin, and C. Villedary. Analysis of time-varying signals with small bt values. *IEEE Transactions on Acoustics Speech and Signal Processing*, 26(1):64–76, 1978.

[KHDM02] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239, 2002. ISSN 0162-8828.

[Kla03] A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *Speech and Audio Processing, IEEE Transactions on*, 11(6):804–816, 2003.

[KM99] K. Kashino and H. Murase. A sound source identification system for ensemble music based on template adaptation and music stream extraction1. *Speech Communication*, 27(3-4):337–349, 1999.

[KST99] T. Kinoshita, S. Sakai, and H. Tanaka. Musical sound source identification based on frequency component adaptation. In *Proc. IJCAI Worshop on CASA*, pages 18–24. 1999.

[Kun04] L. Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004. ISBN 0471210781.

[Kur88] M. Kurzynski. On the multistage bayes classifier. *Pattern Recognition*, 21(4):355–365, 1988. ISSN 0031-3203.

[KW02] Y. Kim and B. Whitman. Singer identification in popular music recordings using voice coding features. In *ISMIR*, pages 164–169. 2002.

[KWP06] Y. Kim, D. Williamson, and S. Pilli. Towards quantifying the album effect in artist identification. In *ISMIR*, pages 393–394. Citeseer, 2006.

[Lar89] J. Laroche. *Etude dun systeme d' analyse et de synthese utilisant la méthode de Prony. Application aux instruments de musique de type percussif.* Ph.D. thesis, Ecole Normale Superieur des Telecommunications, 1989.

[Leb99] R. Lebon. *The professional vocalist: a handbook for commercial singers and teachers*. Scarecrow Pr, 1999.

[LGD07] H. Lukashevich, M. Gruhne, and C. Dittmar. Effective singing voice detection in popular music using arma filtering. *Workshop on Digital Audio Effects (DAFx'07)*, 2007.

[Liu05] C. Liu. Classifier combination based on confidence transformation. *Pattern Recognition*, 38(1):11–28, 2005. ISSN 0031-3203.

[LL09] R. Liu and S. Li. A review on music source separation. In *Information, Computing and Telecommunication, 2009. YC-ICT'09. IEEE Youth Conference on*, pages 343–346. IEEE, 2009.

[LMMT08] M. Lagrange, L. Martins, J. Murdoch, and G. Tzanetakis. Normalized cuts for predominant melodic source separation. *IEEE Transactions on Audio Speech and Language Processing*, 16(2):278, 2008.

[Log] B. Logan. Nearest-neighbor artist identification. *Proceeding of Music Information Retrieval Evaluation eXchange*, pages 192–194.

[LS01] B. Logan and A. Salomon. A content-based music similarity function. *Cambrige Res. Lab*, 2001.

[Lu02] H. Lu. *Toward a high-quality singing synthesizer with vocal texture control*. Ph.D. thesis, Stanford University, 2002.

[LW05] Y. Li and D. Wang. Detecting pitch of singing voice in polyphonic audio. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, volume 3. 2005.

[LW07] Y. Li and D. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[Mah08] R. Maher. Control of synthesized vibrato during portamento musical pitch transitions. *JOURNAL-AUDIO ENGINEERING SOCIETY*, 56(1/2):18, 2008.

[Mak75] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[MB90] R. Maher and J. Beauchamp. An investigation of vocal vibrato for synthesis. *Applied Acoustics*, 30(2-3):219–45, 1990.

[MBN02] L. Molina, L. Belanche, and À. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on*, pages 306–313. IEEE, 2002.

[MBTL07] L. Martins, J. Burred, G. Tzanetakis, and M. Lagrange. Polyphonic instrument recognition using spectral clustering. In *Proc. Int. Conf. on Music Inf. Retrieval*. 2007.

[McA84] S. McAdams. *Spectral fusion, spectral parsing and the formation of auditory images*. 22. Stanford University, 1984.

[MD92] C. McIntyre and D. Dermott. A new fine-frequency estimation algorithm based on parabolic regression. In *icassp*, pages 541–544. IEEE, 1992.

[ME05] M. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In *Proc. ISMIR*, volume 5. Citeseer, 2005.

[Mel92] D. Mellinger. *Event formation and separation in musical sound*. Ph.D. thesis, Stanford University Stanford, CA, USA, 1992.

[Mel01] M. Mellody. *Signal analysis of the female singing voice: features for perceptual singer identity*. Ph.D. thesis, The University of Michigan, 2001.

[Met32] M. Metfessel. *The vibrato in artistic voices*, volume The vibrato: Studies in the psychology of music. In C. E. Seashore (Ed.), 1932.

[MG76] J. Markel and J. Gray. *Linear Prediction of Speech Signals*. Springer-Verlag, New York, 1976.

[MHW01] M. Mellody, F. Herseth, and G. Wakefield. Modal distribution analysis, synthesis, and perception of a soprano's sung vowels* 1. *Journal of Voice*, 15(4):469–482, 2001.

[Mil86] R. Miller. *The structure of singing: system and art in vocal technique*. Schirmer, 1986.

[MK98] K. Martin and Y. Kim. Musical instrument identification: A pattern-recognition approach. *The Journal of the Acoustical Society of America*, 104:1768, 1998.

[MM91] C. Marin and S. McAdams. Segregation of concurrent sounds. ii: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. *The Journal of the Acoustical Society of America*, 89:341, 1991.

[MM99] J. Marques and P. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. *Cambridge Research Laboratory Technical Report Series CRL*, 4, 1999.

[MQ86] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 34(4):744–754, 1986.

[MS88] E. Mandler and J. Schurmann. Combining the classification results of independent classifiers based on the dempster/shafer theory of evidence. In *Pattern recognition and artificial intelligence: towards an integration: proceedings of an international workshop held in Amsterdam, May 18-20, 1988*, page 381. North-Holland, 1988. ISBN 0444871373.

[MS90] D. Miller and H. Schutte. Formant tuning in a professional baritone*. *Journal of Voice*, 4(3):231–237, 1990.

[MVK07] A. Mesaros, T. Virtanen, and A. Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *ISMIR*, pages 375–378. 2007.

[MWXW04] N. C. Maddage, K. Wan, C. Xu, and Y. Wang. Singing voice detection using twice-iterated composite fourier transform. In *ICME*, pages 1347–1350. IEEE, 2004.

[MXW04] N. Maddage, C. Xu, and Y. Wang. Singer identification based on vocal and instrumental models. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2, 2004.

[NH91] S. Nowlan and G. Hinton. Evaluation of adaptive mixtures of competing experts. *Advances in neural information processing systems*, 3:774–780, 1991.

[NSW04] T. L. Nwe, A. Shenoy, and Y. Wang. Singing voice detection in popular music. In H. Schulzrinne, N. Dimitrova, A. Sasse, S. B. Moon, and R. Lienhart, editors, *ACM Multimedia*, pages 324–327. ACM, 2004. ISBN 1-58113-893-8.

[Nut81] A. Nuttall. Some windows with very good sidelobe behavior. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(1):84–91, 1981.

[OPGB05] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 90–93. IEEE, 2005.

[OSB+89] A. Oppenheim, R. Schafer, J. Buck, et al. *Discrete-time signal processing*, volume 1999. Prentice hall Englewood Cliffs, NJ:, 1989.

[Osh87] D. Oshaughnessy. *Speech communications: human and machine*. Universities Press, 1987.

[OSON05] Y. Okada, T. Sahara, S. Ohgiya, and T. Nagashima. Detection of cluster boundary in microarray data by reference to mips functional catalogue database. *GIW2005*, 2005.

[Pee01] G. Peeters. *Modèle et modélisation du signal sonore adapté à ses caractéristiques locales*. Ph.D. thesis, Univ. Paris VI, 2001.

[Pee03] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. *115th AES convention, New York, USA, October*, 2003.

[Pee07a] G. Peeters. A generic system for audio indexing: application to speech/ music segmentation and music genre. *Proc. of DAFX*, 2007.

[PEE+07b] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong. Melody transcription from music-audio: Approaches and evaluation. *IEEE Transactions on Audio Speech and Language Processing*, 15(4):1247, 2007.

[Pis73] V. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal of the Royal Astronomical Society*, 33(3):347–366, 1973.

[PNK94] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994. ISSN 0167-8655.

[Pol02] E. Pollastri. Some considerations about processing singing voice for music retrieval. In *Proceedings of 3rd International Conference on Music Information Retrieval, ISMIR*, volume 2. Citeseer, 2002.

[Pot06] J. Potter. Beggar at the door: the rise and fall of portamento in singing. *Music and Letters*, 87(4):523, 2006.

[Pra94] E. Prame. Measurements of the vibrato rate of ten singers. *The Journal of the Acoustical Society of America*, 96:1979, 1994.

[Pra97] E. Prame. Vibrato extent and intonation in professional western lyric singing. *The Journal of the Acoustical Society of America*, 102:616, 1997.

[Pro95] R. Prony. Essai experimental et analytique. *J. Ec. Polytech.(Paris)*, 2:24–76, 1795.

[Ras78] R. Rasch. The perception of simultaneous notes such as in polyphonic music. *Acustica*, 40(1):21–33, 1978.

[Ras79] R. Rasch. Synchronization in performed ensemble music. *Acustica*, 43(2):121–131, 1979.

[RH07] M. Rocamora and P. Herrera. Comparing audio descriptors for singing voice detection in music audio files. *SBCM - Brasilian Sypmosium on Computer Music 07*, 2007.

[RK06] M. Ryynanen and A. Klapuri. Transcription of the singing melody in polyphonic music. In *Proc. 7th International Conference on Music Information Retrieval*. 2006.

[RMW90] T. Rossing, F. Moore, and P. Wheeler. *The science of sound*, volume 2. Addison-Wesley, 1990.

[Röb05] A. Röbel. Efficients spectral envelope estimation and its application to pitch shifting and envelope preservation. In *DAF'x*. 2005.

[RP09] L. Regnier and G. Peeters. Singing voice detection in music track using direct vibrato detection. In *ICASSP*, pages 441–444. 2009.

[RPB84] X. Rodet, Y. Potard, and J. Barriere. The chant project: from the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, 8(3):15–31, 1984.

[RPK86] R. Roy, A. Paulraj, and T. Kailath. Esprit: a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1340–1342, 1986.

[RR08] M. Ramona and B. Richard, G. David. Vocal detection in music with support vector machine. *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'08). IEEE International Conference on*, pages 1885 – 1888, March 2008.

[RS78] L. Rabiner and R. Schafer. *Digital processing of speech signals*, volume 100. Prentice-hall Englewood Cliffs, NJ, 1978.

[RSSS07] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh. Separating a foreground singer from background music. In *International Symposium on Frontiers of Research on Speech and Music, Mysore, India*. Citeseer, 2007.

[Sal08] J. Salamon. *Chroma-based predominant melody and bass line extraction from music audio signals*. Ph.D. thesis, Universitat Pompeu Fabra Barcelona, Spain, 2008.

[SBC+02] R. Schettini, C. Brambilla, G. Ciocca, A. Valsasna, and M. De Ponti. A hierarchical classification strategy for digital documents. *Pattern Recognition*, 35(8):1759–1769, 2002. ISSN 0031-3203.

[SCB81] H. Schultz-Coulon and R. Battmer. Die quantitative bewertung des s "angervibratos. *Fol Phoniatr*, 33:1–14, 1981.

[Sch78] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[Sch86] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[Sea31] C. Seashore. The natural history of the vibrato. *Proceedings of the National Academy of Sciences*, 17(12):623–626, 1931.

[Sea36] C. Seashore. Psychology of the vibrato in voice and instrument. iowa, 1936.

[Sea38] C. Seashore. Psychology ofmusic, 1938.

[SG96] M. Schmidt and H. Gish. Speaker identification via support vector classifiers. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 105–108. IEEE, 1996.

[SG11] J. Salamon and E. Gómez. Melody extraction from polyphonic music: Mirex 2011. *7th Music Information Retrieval Evaluation exchange (MIREX)*, 2011.

[SIL07] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507, 2007.

[SLS80] T. Shipp, R. Leanderson, and J. Sundberg. Some acoustic characteristics of vocal vibrato. *J Res Singing*, 4(1):18–25, 1980.

[SM91] H. Schutte and D. Miller. Acoustic details of vibrato cycle in tenor high notes*. *Journal of Voice*, 5(3):217–223, 1991.

[SMS95] H. Schutte, D. Miller, and J. Svec. Measurement of formant frequencies and bandwidths in singing*. *Journal of Voice*, 9(3):290–296, 1995.

[SPN+99] P. Somol, P. Pudil, J. Novovicová, et al. Adaptive floating search methods in feature selection. *Pattern recognition letters*, 20(11-13):1157–1163, 1999. ISSN 0167-8655.

[SR90] J. Sundberg and T. Rossing. The science of singing voice. *The Journal of the Acoustical Society of America*, 87:462, 1990.

[SS87] J. Smith and X. Serra. Parshl: A program for the analysis/synthesis of inharmonic sounds based on a sinusoidal representation. In *Int Computer Music Conf*. 1987.

[SS90] X. Serra and J. Smith. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. *5. European Signal Processing Conference.*, 2:1347–1350, 1990.

[SSH84] T. Shipp, J. Sundberg, and S. Haglund. A model of frequency vibrato. In *Transcripts of the Thirteenth Symposium: Care of the Professional Voice*, page 116Á. 1984.

[SSWR83] W. Seidner, H. Schutte, J. Wendler, and A. Rauhut. Dependence of the high singing formant on pitch and vowel in different voice types. In *SMAC 83, Proceedings of the Stockholm Music Acoustics Conference*, pages 261–268. 1983.

[Sun74] J. Sundberg. Articulatory interpretation of the ÂŚsinging formantÂŠ. *J Acoust Soc Am*, 55(4):838–844, 1974.

[Sun94] J. Sundberg. Perceptual aspects of singing*. *Journal of voice*, 8(2):106–122, 1994.

[Sun95] J. Sundberg. Acoustic and psychoacoustic aspects of vocal vibrato. *Vibrato*, pages 35–62, 1995.

[SWW05] A. Shenoy, Y. Wu, and Y. Wang. Singing voice detection for karaoke application. *Proceedings of SPIE*, 5960:596028, 2005.

[TD00] R. Timmers and P. Desain. Vibrato: Questions and answers from musicians and science. *Proc. ICMPC*, 2000.

[TL11] W. Tsai and H. Lin. Background music removal based on cepstrum transformation for popular singer identification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:1196 – 1205, 2011. ISSN 1558-7916.

[TM98] I. Titze and D. Martin. Principles of voice production. *Acoustical Society of America Journal*, 104:1148, 1998. ISSN 0001-4966.

[Tof96] N. Toff. *The flute book: a complete guide for students and performers*. Oxford University Press, USA, 1996.

[TRW04] W. Tsai, D. Rodgers, and H. Wang. Blind clustering of popular music recordings based on singer voice characteristics. *Computer Music Journal*, 28(3):68–78, 2004.

[TSSL02] I. Titze, B. Story, M. Smith, and R. Long. A reflex resonance model of vocal vibrato. *The Journal of the Acoustical Society of America*, 111:2272, 2002.

[TW06] W. Tsai and H. Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(1):330–341, 2006.

[VB05] S. Vembu and S. Baumann. Separation of vocals from polyphonic audio recordings. In *Proc. ISMIR*, pages 337–344. Citeseer, 2005.

[Ven67] W. Vennard. *Singing: the mechanism and the technic*. Carl Fischer Music Dist, 1967.

[VES00] M. Van Erp and L. Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. In *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 443–452. 2000.

[VGD05] V. Verfaille, C. Guastavino, and P. Depalle. Perceptual evaluation of vibrato models. *Proceedings of Conference on Interdisciplinary Musicology*, 2005.

[vLMPB06] D. van Leeuwen, A. Martin, M. Przybocki, and J. Bouten. Nist and nfi-tno evaluations of automatic speaker recognition. *Computer Speech & Language*, 20(2-3):128–158, 2006. ISSN 0885-2308.

[VM07] F. Vallet and M. McKinney. Perceptual constraints for automatic vocal detection in music recordings. In *In Proc. Conference on Interdisciplinary Musicology (CIM 2007)*. 2007.

[VMR08] T. Virtanen, A. Mesaros, and M. Ryyn
"anen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition SAPA*. Citeseer, 2008.

[VR04] E. Vincent and X. Rodet. Instrument identification in solo and ensemble music using independent subspace analysis. In *Proc. ISMIR*, pages 576–581. Citeseer, 2004.

[VT68] H. Van Trees. *Detection, estimation, and modulation theory: Detection, estimation, and linear modulation theory*. Wiley, 1968.

[Wan94] A. Wang. *Instantaneous and frequency-warped signal processing techniques for auditory source separation*. Ph.D. thesis, Standford University, 1994.

[WB03] G. Wakefield and M. Bartsch. Where's caruso? singer identification by listener and machine. In *Cambridge University Music Processing Colloquium*. 2003.

[WB06] D. Wang and G. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press, 2006.

[WBM+01] R. Weiss, W. Brown, J. Moris, et al. Singer's formant in sopranos: Fact or fiction? *Journal of Voice*, 15(4):457–468, 2001.

[WFL01] B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with minnowmatch. *Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pages 559–568, 2001.

[Win53] F. Winckel. Physikalische kriterien f
"ur objektive stimmbeurteilung. *Folia phoniat*, 5:232–252, 1953.

[Win74] F. Winckel. Acoustical cues in the voice for detecting laryngeal diseases and individual behavior. In *Ventilatory and phonatory control Systems. An international Symposium, Oxford Univ. Press, London*, pages 248–264. 1974.

[WL07] K. West and P. Lamere. A model-based approach to constructing music similarity functions. *EURASIP Journal on Applied Signal Processing*, 2007(1):149–149, 2007. ISSN 1110-8657.

[WLW04] T. Wu, C. Lin, and R. Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004. ISSN 1532-4435.

[WP05] B. Wang and M. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Proc. DMRN Summer Conf*, pages 23–24. 2005.

[WWB03] M. Wu, D. Wang, and G. Brown. A multipitch tracking algorithm for noisy speech. *Speech and Audio Processing, IEEE Transactions on*, 11(3):229–241, 2003.

[XKS02] L. Xu, A. Krzyzak, and C. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(3):418–435, 2002. ISSN 0018-9472.

[ZBS07] P. Zhang, T. Bui, and C. Suen. A novel cascade ensemble classifier system with a high recognition performance on handwritten digits. *Pattern Recognition*, 40(12):3415–3429, 2007. ISSN 0031-3203.

[ZE01] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Machine learning international workshop then conference*, pages 609–616. Citeseer, 2001.

[Zha03] T. Zhang. Automatic singer identification. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 1, 2003.