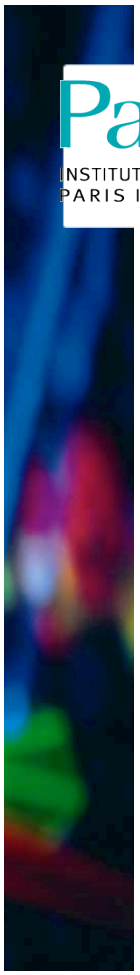
Models of music signals informed by physics. Application to piano music analysis by non-negative matrix factorization.

François Rigaud



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0073

EDITE ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité “ Signal et Images ”

présentée et soutenue publiquement par

François RIGAUD

le 2 décembre 2013

Modèles de signaux musicaux informés par la physique des instruments. Application à l'analyse de musique pour piano par factorisation en matrices non-négatives.

Models of music signals informed by physics. Application to piano music analysis by non-negative matrix factorization.

Directeur de thèse : **Bertrand DAVID**
Co-encadrement de la thèse : **Laurent DAUDET**

Jury

M. Vesa VÄLIMÄKI, Professeur, Aalto University	Rapporteur
M. Emmanuel VINCENT, Chargé de Recherche, INRIA Nancy	Rapporteur
M. Philippe DEPALLE, Professeur, McGill University	Président du jury
M. Simon DIXON, Professeur, Queen Mary University of London	Examinateur
M. Cédric FÉVOTTE, Chargé de Recherche, CNRS, Université de Nice Sophia Antipolis	Examinateur
M. Bertrand DAVID, Maître de Conférence, Télécom ParisTech	Directeur de thèse
M. Laurent DAUDET, Professeur, Université Paris Diderot - Paris 7	Directeur de thèse

T H È S E

Télécom ParisTech
école de l’Institut Mines Télécom – membre de ParisTech
46, rue Barrault – 75634 Paris Cedex 13 – Tél. + 33 (0)1 45 81 77 77 – www.telecom-paristech.fr

Abstract

This thesis builds new models of music signals informed by the physics of the instruments. While instrumental acoustics and audio signal processing target the modeling of musical tones from different perspectives (modeling of the production mechanism of the sound versus modeling of the generic "morphological" features of the sound), this thesis aims at mixing both approaches by constraining generic signal models with acoustics-based information. Thus, it is here intended to design instrument-specific models for applications both to acoustics (learning of parameters related to the design and the tuning) and signal processing (transcription).

In particular, we focus on piano music analysis for which the tones have the well-known property of inharmonicity, *i.e.* the partial frequencies are slightly higher than those of a harmonic comb, the deviation increasing with the rank of the partial. The inclusion of such a property in signal models however makes the optimization harder, and may even damage the performance in tasks such as music transcription when compared to a simpler harmonic model. The main goal of this thesis is thus to have a better understanding about the issues arising from the explicit inclusion of the inharmonicity in signal models, and to investigate whether it is really valuable when targeting tasks such as polyphonic music transcription.

To this end, we introduce different models in which the inharmonicity coefficient (B) and the fundamental frequency (F_0) of piano tones are included as parameters: two NMF-based models and a generative probabilistic model for the frequencies having significant energy in spectrograms. Corresponding estimation algorithms are then derived, with a special care in the initialization and the optimization scheme in order to avoid the convergence of the algorithms toward local optima. These algorithms are applied to the precise estimation of (B, F_0) from monophonic and polyphonic recordings in both supervised (played notes are known) and unsupervised conditions.

We then introduce a joint model for the inharmonicity and tuning along the whole compass of pianos. Based on invariants in design and tuning rules, the model is able to explain the main variations of piano tuning along the compass with only a few parameters. Beyond the initialization of the analysis algorithms, the usefulness of this model is also demonstrated for analyzing the tuning of well-tuned pianos, to provide tuning curves for out-of-tune pianos or physically-based synthesizers, and finally to interpolate the inharmonicity and tuning of pianos along the whole compass from the analysis of a polyphonic recording containing only a few notes.

Finally the efficiency of an inharmonic model for NMF-based transcription is investigated by comparing the two proposed inharmonic NMF models with a simpler harmonic model. Results show that it is worth considering inharmonicity of piano tones for a transcription task provided that the physical parameters underlying (B, F_0) are sufficiently well estimated. In particular, a significant increase in performance is obtained when using an appropriate initialization of these parameters.

Table of contents

Notation

This section gathers acronyms and variables repeatedly used along the manuscript. Those which are not listed here are locally used and properly defined when mentioned in a section.

List of acronyms

EM	Expectation-Maximization (algorithm)
ET	Equal Temperament
EUC	Euclidian (distance)
FN	False Negative
FP	False Positive
IS	Itakura-Saito (divergence)
KL	Kullback-Leibler (divergence)
MAPS	MIDI Aligned Piano Sounds (music database)
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
NMF	Non-negative Matrix Factorization (model)
Ha-NMF	Harmonic (strict) NMF
Inh-NMF	Inharmonic (strict) NMF
InhR-NMF	Inharmonic (Relaxed) NMF
PFD	Partial Frequency Deviation (algorithm)
PLS	Probabilistic Line Spectrum (model)
RWC	Real World Computing (music database)
TN	True Negative
TP	True Positive
WC	Whole Compass

List of variables

Time-frequency representation:

$t \in [1, T]$	time-frame index
$k \in [1, K]$	frequency-bin index
f_k	frequency of bin with index k
τ	analysis window's length
$g_\tau(f_k)$	magnitude of the Fourier Transform of the analysis window
V	magnitude (or power) spectrogram (dimensions $K \times T$)

Standard NMF with β-divergence applied to audio spectrograms:

W	dictionary of spectra (dimensions $K \times R$)
H	activation matrix (dimensions $R \times T$)
$\hat{V}$	matrix product of W and H approximating V
$C(W, H)$	NMF cost-function
β	parameter of the β-divergence
d_β	β-divergence

NMF inharmonic models:

$r \in [1, R]$	note index
$n \in [1, N_r]$	partial rank of note with index r
$\theta^{Ha/Inh/InhR}$	set of parameters, respectively, for *Ha/Inh/InhR-NMF* models
a_{nr}	amplitude of partial with rank n of note with index r
f_{nr}	frequency of partial with rank n of note with index r
F_{0r}	fundamental frequency of note with index r
B_r	inharmonicity coefficient of note with index r
T_{on}	onset detection threshold of H matrix for transcription application

Whole compass model of piano tuning:

$m \in [21, 108]$	index of note in MIDI norm from A0 to C8
$\xi = \{s_b, y_b\}$	set of parameters for the WC model of inharmonicity
ρ	octave type
$\phi = \{\kappa, m_0, \alpha\}$	set of parameters for the WC model of octave type
d_g	global tuning deviation

Chapter 1
Introduction

1.1 Musical sound modeling and representation

Acoustics-based vs. signal-based approaches

The issue of the representation and the modeling of musical sounds is a very active research topic that is often addressed from two main different perspectives, namely those of acoustics and signal processing.

Historically related to the branch of wave physics, and notably led by Helmholtz's advances in physical and physiological acoustics [Helmholtz, 1863], the community of instrumental acoustics has formed around this problem by studying the mechanisms that produce the sound, *i.e.* the musical instruments. Such an approach usually relies on a precise physical modeling of the vibrations and couplings taking place in musical instruments when subject to a player excitation [Rossing and Fletcher, 1995; Fletcher and Rossing, 1998; Chaigne and Kergomard, 2008]. For each specific instrument, the characteristics of the tones are related to meaningful low-level parameters such as, the physical properties of the materials, the geometry of each component (*e.g.* strings, plates, membranes, pipes), the initial conditions related to playing techniques, *etc.* For different designs of the same instrument, these modeling parameters provide a representation useful in various applications. For instance, in synthesis, the control of the characteristics of the tones can be done intuitively by modifying the physical properties of the instrument. Such representations are also useful for psycho-acoustics research, when studying the effects of the instrumental design and playing techniques on the sound and its perception.

In contrast to the physics-based approach, the community of music signal processing aims at modeling musical sounds according to their "morphological" attributes, without necessarily paying due regard to the particular instruments that are played. Indeed, most music signals are generally composed of tonal components (quasi-sinusoidal waves, also called partials, having amplitude and frequency parameters that may vary over time), but also percussive components corresponding to attack transients, and noise. However, such kind of signal-based characteristics are hardly identifiable when having a look at a waveform representation.

Thus, a substantial part of the research in the 60's-70's had focused on the design of transforms/representations that allow for an intuitive visualization and manipulation of these signal-based features in the time-frequency domain (as for instance the well-known

spectrogram based on the Short-Time Fourier Transform). Also, the inversion property of such transforms led the way for analysis/synthesis applications to audio coding [Flanagan and Golden, 1966] and sound transformations (*e.g.* transposition, time-stretching, cross synthesis) [Allen, 1977]. Based on these representations a number of parametric models of speech/music signals have been developed, such as the sinusoidal additive model [McAulay and Quatieri, 1986; Serra and Smith, 1990] or the source-filter model (first introduced as a signal-model mimicking the mechanism of speech production [Fant, 1960], but further widely applied to various types of instrumental tones). In the following decade, a number of enhanced transforms have been introduced (*e.g.* spectral reassignment, Modified Cosine Transform, Wavelet Transform, Constant Q Transform) in order to face the time-frequency resolution trade-off or improve the coding performance by accounting for auditory perception [Mallat, 2008; Roads, 2007].

Since then, a number of theoretical frameworks, often taken from other domains such as image processing, statistics and machine learning, have been adopted in order to replace ad-hoc modelings and heuristic estimation techniques. To cite only a few, sparse coding [Mallat and Zhang, 1993; Chen et al., 1998], Bayesian modeling [Gelman et al., 2003], high-resolution methods [Badeau et al., 2006] or approaches based on rank reduction [Cichocki et al., 2009] have been widely spread across the audio community since the 90's. In contrast with acoustics-based modelings, these frameworks are generic enough to process complex mixture of sounds that can include various types of instruments. In parallel, a new field of applications, referred to as Music Information Retrieval (MIR), has emerged with a need of indexing and classifying the ever-growing amount of multimedia data available on the internet [Downie, 2006; Casey et al., 2008]. In this context, new kind of mid-level representations (*e.g.* chromagrams, onset detection functions, similarity matrices) have been found useful for extracting content-based information from music pieces (*e.g.* chords, melody, genre).

Mixing both approaches

Although acoustics and signal processing model musical tones according to different perspectives, both communities tend to borrow tools from each other.

For instance, synthesis applications based on physical modeling require an important number of parameters that cannot be easily estimated from instrument design only. Indeed, such models based on coupled differential equation systems are usually non-invertible and a joint estimation of all parameters from a target sound is hardly practicable. Thus, the learning of the synthesis parameters often requires physical measurements on real instruments and/or estimation techniques derived from signal-based modeling (*e.g.* High Resolution methods applied to modal analysis [Ege et al., 2009; Elie et al., 2013]). Also, signal processing tools such as numerical methods for differential equation solving (*e.g.* finite element and finite difference methods), and digital filtering are commonly used to perform the synthesis [Rauhala et al., 2007b; Bank et al., 2010].

Conversely, a general trend in signal processing this last decade (as shown in Chapter 2) tends to include prior information about the structure of the tones in order to tailor more efficient models and estimation algorithms. For instance, accounting for the harmonic structure of sustained tones, the smoothness of their spectral and temporal envelopes, can outperform generic models in various applications such as transcription or source separation.

1.2 Approach and issues of the thesis

General framework of the thesis

The goal of this thesis is to go a step further in the mixing of both approaches. From generic signal-based frameworks we aim at including information about the timbre of the tones in order to perform analysis tasks that are specific to a given instrument. Issues related to both acoustics and signal processing applications are then investigated:

- Can such classes of models be used to efficiently learn physics-related parameters (*e.g.* information about the design and tuning) of a specific instrument?
- Does refining/complexifying generic signal models actually improve the performance of analysis tasks targeted by the MIR community?

It however remains unrealistic to try to combine full synthesis models, built from acoustics only, into complex signal-based analysis methods, and to optimize both types of parameters. Instead, we here focus on a selection of a few timbre features that should be most relevant. This relates to one of the core concepts of musical acoustics – but also one of the hardest to apprehend: timbre. The simplest, and standard definition of timbre[1], is what differentiates two tones with the same pitch, loudness, and duration. However, timbre is much more complex than this for instrument makers, as one instrument must be built and tuned in such a way that its timbre keeps a definite consistency from note to note (while not being exactly the same) across the whole tessitura, loudness and playing techniques.

Focus on piano music

Such methodology is applied in this thesis to the case of the piano. Its study is particularly relevant as it has been central to Western music in the last two centuries, with an extremely wide solo and orchestral repertoire. Moreover, the analysis of piano music represents a challenging problem because of the versatility of the tones. Indeed, the piano register spans more than seven octaves, with fundamental frequencies ranging approximately from 27 Hz to 4200 Hz, and the action mechanism allows for a wide range of dynamics. In addition, the tones are well known for their property of inharmonicity (the partial frequencies are slightly higher than those of a harmonic comb) that has an important influence in the perception of the instrument's timbre as well as on its tuning [Martin and Ward, 1961]. Thus, in this thesis particular attention is paid to the inclusion of the inharmonicity in signal-based models. Variability of the inharmonicity along the whole compass of the piano and its influence on the tuning are also investigated.

Signal-based frameworks

This thesis focuses on Non-negative Matrix Factorization (NMF) models [Lee and Seung, 1999]. Such models essentially target the decomposition of time-frequency representations of music signals into two non-negative matrices: one dictionary containing the spectra/atoms of the notes/instruments, and one activation matrix containing their temporal

[1] "attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar", American Standards Association definition 12.9

activations. Thus a substantial part of this thesis consists of finding a way of enforcing the inharmonic structure of the spectra of the dictionary for applications related to both acoustics (estimation of the inharmonicity parameters) and signal processing (transcription). An alternative model based on a Bayesian framework is also proposed in this thesis.

1.3 Overview of the thesis and contributions

Chapter 2 - State of the art: This chapter introduces the theoretical background of this thesis. First, a state-of-the-art on the Non-Negative Matrix (NMF) factorization framework is presented. Both standard and constrained approaches are detailed, and practical considerations about the problem formulation and the optimization are discussed. Second, basic notions of piano acoustics are presented in order to highlight some properties of the tones that should be relevant to include in our models. Then, a state of the art on methods taking into account inharmonicity for piano music analysis is presented. Finally, issues resulting from the inharmonicity inclusion in signal-based models are discussed.

Chapter 3 - Estimating the inharmonicity coefficient and the F_0 of piano tones: This chapter deals with the estimation of the inharmonicity coefficient and the F_0 of piano tones along the whole compass, in both monophonic and polyphonic contexts.

> **Contributions:** Two new frameworks in which inharmonicity is included as a parameter of signal-based models are presented. All presented methods exhibit performances that compare favorably to a state of the art algorithm when applied to the supervised estimation of the inharmonicity coefficient and the F_0 of isolated piano tones.
>
> First in Section 3.1, two NMF-based parametric models of piano tone spectra for which different types of inclusion of the inharmonicity is proposed (strict and relaxed constraints) are described. Estimation algorithms are derived for both models and practical considerations about the initialization and the optimization scheme are discussed. These are finally applied on isolated note and chord recordings in a supervised context (*i.e.* with the knowledge of the notes that are processed).
>
> Second in Section 3.2, a generative probabilistic model for the frequencies of peaks with significant energy in time-frequency representations is introduced. The parameter estimation is formulated as a maximum *a posteriori* problem and solved by means of an Expectation-Maximization algorithm. The precision of the estimation of the inharmonicity coefficient and the F_0 of piano tones is evaluated first on isolated note recordings in both supervised and unsupervised cases. For the unsupervised case, the algorithm returns estimates of the inharmonicity coefficients and the F_0s as well as a probability for each note to have generated the observations in each time-frame. The algorithm is finally applied on a polyphonic piece of music in an unsupervised way.

Chapter 4 - A parametric model for the inharmonicity and tuning along the whole compass: This chapter presents a study on piano tuning based on the consideration of inharmonicity and tuner's influences.

Contributions: A joint model for the inharmonicity and the tuning of pianos along the whole compass is introduced. While using a small number of parameters, these models are able to reflect both the specificities of instrument design and tuner practice. Several applications are then proposed. These are first used to extract parameters highlighting some tuners' choices on different piano types and to propose tuning curves for out-of-tune pianos or piano synthesizers. Also, from the study on several pianos, an average model useful for the initialization of the inharmonicity coefficient and the F_0 of analysis algorithms is obtained. Finally, these models are applied to the interpolation of inharmonicity and tuning along the whole compass of a piano, from the unsupervised analysis of a polyphonic musical piece.

Chapter 5 - Application to the transcription of polyphonic piano music: This chapter presents an application to a transcription task of the two NMF-based algorithms introduced in Section 3.1.

Contributions: In order to quantify the benefits that may result from the inharmonicity inclusion in NMF-based models, both algorithms are applied to a transcription task and compared to a simpler harmonic model. We study on a 7-piano database the influence of the model design (harmonicity vs. inharmonicity and number of partials considered), of the initialization (naive initialization vs. mean model of inharmonicity and tuning along the whole compass) and of the optimization process. Results suggest that it is worth including inharmonicity in NMF-based transcription model, provided that the inharmonicity parameters are sufficiently well initialized.

Chapter 6 - Conclusion and prospects: This last chapter summarizes the contributions of the present work and proposes some prospects for the design of a competitive piano transcription system and the application of such approaches for other string instruments such as the guitar.

1.4 Related publications

— Peer-reviewed journal article —

Rigaud, F., David, B., and Daudet, L. (2013a). A parametric model and estimation techniques for the inharmonicity and tuning of the piano. *Journal of the Acoustical Society of America*, 133(5):3107–3118.

— Peer-reviewed conference articles —

Rigaud, F., David, B., and Daudet, L. (2011). A parametric model of piano tuning. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*, pages 393–399.

Rigaud, F., David, B., and Daudet, L. (2012). Piano sound analysis using non-negative matrix factorization with inharmonicity constraint. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 2462–2466.

Rigaud, F., Drémeau, A., David, B., and Daudet, L. (2013b). A probabilistic line spectrum model for musical instrument sounds and its application to piano tuning estimation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Rigaud, F., Falaize, A., David, B., and Daudet, L. (2013c). Does inharmonicity improve an NMF-based piano transcription model? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15.

Chapter 2
State of the art

This thesis mainly focuses on the inclusion of specific properties of the timbre of piano tones in Non-negative Matrix Factorization (NMF) models. Thus, this chapter first presents in Section 2.1 a state-of-the-art on the modeling of audio spectrograms based on the NMF framework. Both standard and constrained formulations accounting for specific audio properties are detailed. Then in Section 2.2, some basic notions of instrumental acoustics are introduced in order to highlight some characteristics of piano tones that should be relevant to blend into the NMF model. Finally, a state-of-art of methods including the inharmonic nature of the tones in piano music analysis is presented and issues resulting from its inclusion in signal models are discussed in Section 2.3.

2.1 Modeling time-frequency representations of audio signals based-on the factorization of redundancies

Most musical signals are highly structured, both in terms of temporal and pitch arrangements. Indeed, when analyzing the score of a piece of western tonal music (*cf.* for instance Figure 2.1) one can notice that often a restricted set of notes whose relationship are given by the tonality are used and located at particular beats corresponding to subdivisions of the measure. All these constraints arising from musical composition rules (themselves based on musical acoustics and psycho-acoustics considerations) lead to a number of redundancies: notes, but also patterns of notes, are often repeated along the time line.

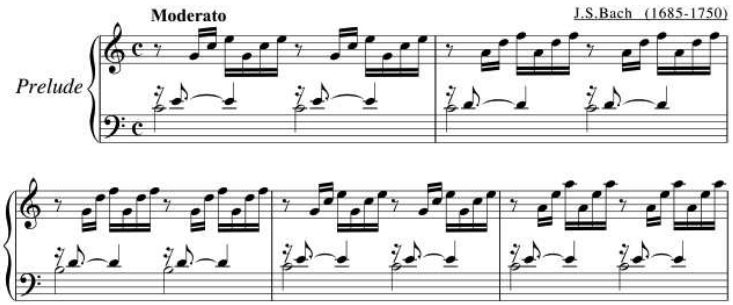


Figure 2.1: Score excerpt of the Prelude I from *The Well-Tempered Clavier Book I* - J.S. Bach. Source: *www.virtualsheetmusic.com/*

Taking these redundancies into account to perform coding, analysis and source separa-

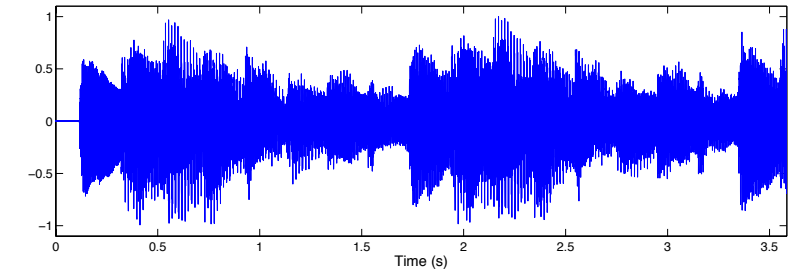
tion of music signals has become a growing field of research in the last twenty years. Originally focused on speech and image coding problems, the sparse approximation framework aims at representing signals in the time domain as a linear combination of a few patterns (*a.k.a.* atoms) spanning different time-frequency regions [Mallat and Zhang, 1993; Chen et al., 1998]. The dictionary of atoms is usually highly redundant, it may be given *a priori* or learned from a separate dataset. It may also be structured in order to better fit the specific properties of the data (*e.g.* harmonic atoms for applications to music [Gribonval and Bacry, 2003; Leveau et al., 2008]).

In the last decade, the idea of factorizing the redundancies inherent to the structure of the music has been applied to signals in the time-frequency domain. Unlike the signal waveform, time-frequency representations such as the spectrogram allow for an intuitive interpretation of the pitch and rhythmic content by exhibiting particular characteristics of the musical tones (*e.g.* a comb structure evolving over time for pitched sounds, *cf.* Figure 2.2). Beyond the recurrence of the notes that are played (*cf.* note E5 whose spectro-temporal pattern is highlighted in black in Figure 2.2(b)), adjacent time-frames are also highly redundant. Thus, methods based on redundancy factorization such as the Non-negative Matrix Factorization (NMF) [Lee and Seung, 1999], or it probabilistic counterpart the Probabilistic Latent Component Analysis (PLCA) [Smaragdis et al., 2008] have been applied to musical signals with the goal to decompose the observations into a few meaningful objects: the spectra of the notes that are played and their activations over time. Another seducing property of these approaches is that the dictionary of spectra/atoms may be directly learned from the data, jointly with the estimation of the activations. However, this great flexibility of factorization approaches often makes difficult the interpretation of activation patterns as individual note activation. In order to be successfully applied to complex tasks such as polyphonic music transcription or source separation, the structure of the decomposition has to be enforced by including prior information on the specific properties of audio data in the model (for instance harmonic structure of sustained note spectra or smoothness of temporal activations).

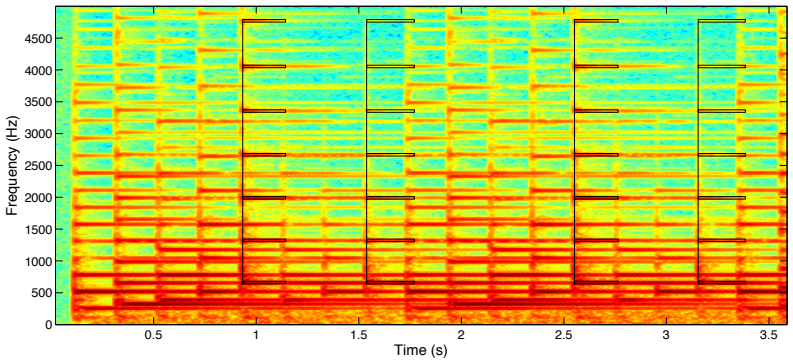
2.1.1 Non-negative Matrix Factorization framework

The NMF is a decomposition method for multivariate analysis and dimensionality/rank reduction of non-negative data (*i.e.* composed of positive or null elements) based on the factorization of the redundancies naturally present. Unlike other techniques such as Independent Component Analysis [Comon, 1994] or Principal Component Analysis [Hotelling, 1933] for which the factorized elements may be composed of positive and negative elements, the key point of NMF relies on the explicit inclusion of a non-negativity constraint that enforces the elements of the decomposition to lie in the same space as the observations. Thus, when analyzing data composed of non-negative elements, such as the pixel intensity of images, the decomposition provided by the NMF exhibits a few meaningful elements that can be directly interpreted as distinctive parts of the initial images (for instance the eyes or the mouth of face images [Lee and Seung, 1999]).

Although the first NMF problem seems to be addressed in [Paatero and Tapper, 1994] under the name Positive Matrix Factorization, is has been widely popularized by the studies of *Lee and Seung* [Lee and Seung, 1999, 2000] which highlighted the ability of the method for learning meaningful parts of objects (images and text) and proposed efficient algorithms. Since then, an important number of studies have been dealing with extend-



(a)



(b)

Figure 2.2: (a) waveform and (b) spectrogram of the first bar of the score given in Figure 2.1. The spectral pattern of note E5 is highlighted in black.

ing NMF models and algorithms to applications in data analysis and source separation [Cichocki et al., 2009] in various domains such as, to name only a few, image [Lee and Seung, 1999; Hoyer, 2004], text [Lee and Seung, 1999; Pauca et al., 2004], biological data [Liu and Yuan, 2008] or audio processing [Smaragdis and Brown, 2003].

2.1.1.1 Model formulation for standard NMF

Given the observation of a non-negative matrix V of dimension $K \times T$, the NMF aims at finding an approximate factorization:

$$V \approx WH = \hat{V}, \tag{2.1}$$

where W and H are non-negative matrices of dimensions $K \times R$ and $R \times T$, respectively. Thus, the observation matrix V is approximated by a positive linear combination of R atoms, contained in the dictionary W whose weightings are given by the coefficients of H, also called the activation matrix. For each element V_{kt}, Equation (2.1) leads to:

$$V_{kt} \approx \sum_{r=1}^{R} W_{kr} H_{rt} = \widehat{V}_{kt} \ . \tag{2.2}$$

The model is illustrated in Figure 2.3 on a toy example, where the exact factorization of a rank-3 matrix is obtained. In real-life applications, V is usually full-rank and only an approximate factorization can be obtained. In order to obtain a few meaningful atoms in the dictionary W, the order of the model R is usually chosen so that $KR + RT \ll KT$, which corresponds to a reduction of the dimensionality of the data. The optimal tuning of

R is not straightforward. It is usually arbitrarily chosen, depending on which application is targeted. For the interested reader, more theoretical considerations and results about the optimal choice of R and the uniqueness of NMF may be found in [Cohen and Rothblum, 1993; Donoho and Stodden, 2004; Laurberg et al., 2008; Klingenberg et al., 2009; Essid, 2012].

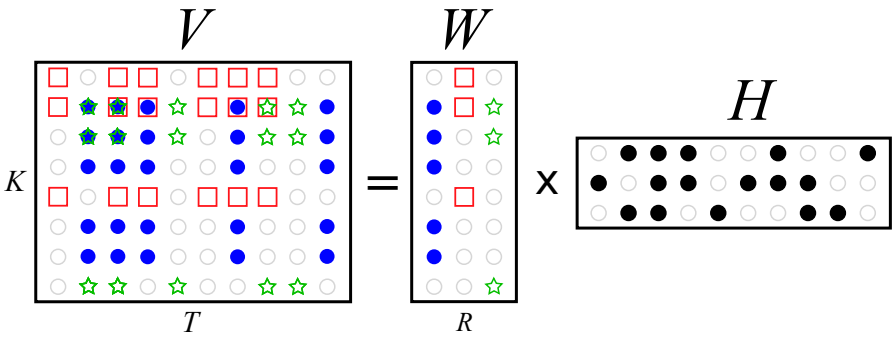


Figure 2.3: NMF illustration for an observation matrix V of rank 3. The superposition of the elements (squares, disks and stars) in V corresponds to a summation. In this particular example, the activation matrix H is composed of binary coefficients.

2.1.1.2 Applications to music spectrograms

In the case of applications to audio signals, V usually corresponds to the magnitude (or power) spectrogram of an audio excerpt, k being the frequency bin index and t being the frame index. Thus, W represents a dictionary containing the spectra of the R sources, and H their time-frame activations. Figure 2.4 illustrates the NMF of a music spectrogram in which one can see that two notes are first played separately, and then jointly. For this particular example, when choosing $R = 2$ the decomposition accurately learns a mean stationary spectrum for each note in W and returns their time-envelopes in H, even when the two notes are played jointly.

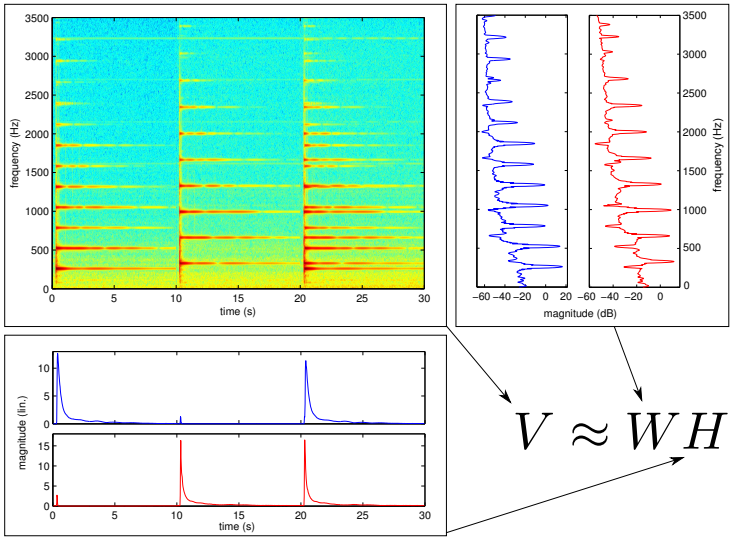


Figure 2.4: NMF applied to the decomposition of an audio spectrogram.

This convenient and meaningful decomposition of audio spectrograms led the way for applications such as automatic music transcription (the transcription task is presented in more detail in Chapter 5) [Smaragdis and Brown, 2003; Paulus and Virtanen, 2005; Bertin, 2009], audio source separation [Virtanen, 2007; Ozerov and Févotte, 2010], audio-to-score alignment [Cont, 2006], or audio restoration [Le Roux et al., 2011].

However, it should be noticed that although the additivity of the sources[1] is preserved in the time-frequency domain by the Short-Time Fourier Transform, a mixture spectrogram is not the sum of the component spectrograms because of the non-linearity of the absolute value operator. Thus, the NMF model, as it is given in Equation (2.1), is strictly speaking incorrect and some studies have introduced a modeling of the phase components [Parry and Essa, 2007a,b] in order to refine the model. However, in most cases the validity of the standard NMF model is still justified by the fact that for most music signals the sources are spanning different time-frequency regions and each time-frequency bin's magnitude is often explained by a single predominant note/instrument [Parvaix and Girin, 2011]. When several sources are equally contributing to the energy contained in a time-frequency bin, the exponent of the representation (for instance 1 for the magnitude spectrogram and 2 for the power spectrogram) may have an influence on the quality of the NMF approximation. In the case of independent sources (for instance corresponding to different instruments), it seems that a better approximation is obtained for an exponent equal to 1 (*i.e.* V is a magnitude spectrogram) [Hennequin, 2011].

Besides the factorization of spectrograms, other types of observation matrices may be processed when dealing with musical data. For instance, the NMF has been applied to the decomposition of self-similarity matrices of music pieces for music structure discovery [Kaiser and Sikora, 2010], or matrices containing audio features for musical instrument [Benetos et al., 2006] and music genre [Benetos and Kotropoulos, 2008] classification.

2.1.1.3 Quantification of the approximation

In order to quantify the quality of the approximation of Equation (2.1), a measure of the dissimilarity between the observation and the model is evaluated by means of a distance (or divergence), further denoted by $D(V \mid WH)$. This measure is used to define a reconstruction cost-function

$$C(W,H) = D(V \mid WH), \tag{2.3}$$

from which H and W are estimated according to a minimization problem

$$\hat{W}, \hat{H} = \underset{W \in \mathbb{R}_+^{K \times R}, H \in \mathbb{R}_+^{R \times T}}{\operatorname{argmin}} C(W,H). \tag{2.4}$$

For a separable metric, the measure can be expressed as:

$$D(V \mid WH) = \sum_{k=1}^{K} \sum_{t=1}^{T} d\left(V_{kt} \mid \widehat{V}_{kt}\right), \tag{2.5}$$

where $d(. \mid .)$ denotes a measure of dissimilarity between two scalars, corresponding for instance to each time-frequency bin of the observed spectrogram and the model. A wide

[1] This holds for music signals obtained by a linear mixing of the separated tracks, thus by neglecting the non-linear mixing operations such as for instance the compression applied at the mastering.

range of metrics have been investigated for NMF, often gathered in families such as the Csiszar [Cichocki et al., 2006], β [Févotte and Idier, 2011] or Bregman divergences [Dhillon and Sra, 2006]. It is worth noting that divergences are more general than distances since they do not necessarily respect the properties of symmetry and triangular inequality. However they are still adapted to a minimization problem since they present a single minimum for $V_{kt} = \widehat{V}_{kt}$.

β-divergences family: The family of β-divergences is widely used in a number of NMF applications because it encompasses 3 common metrics: a Euclidean (EUC) distance, the Kullback-Leibler (KL) divergence (introduced in [Kullback and Leibler, 1951] as a measure of the difference between two probability distributions) and the Itakura-Saito (IS) divergence (obtained in [Itakura and Saito, 1968] from a Maximum Likelihood estimation problem for auto-regressive modelings of speech spectra). Its general formulation is given by:

$$d_\beta(x \mid y) = \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}), \quad \beta \in \mathbb{R} \setminus \{0,1\}. \tag{2.6}$$

EUC is obtained by setting $\beta = 2$, while KL and IS are obtained by taking the limit, respectively, for $\beta \to 1$ and $\beta \to 0$.

$$d_{\beta=2}(x \mid y) = d_{\mathrm{EUC}}(x \mid y) = \frac{(x-y)^2}{2}, \tag{2.7}$$

$$d_{\beta\to1}(x \mid y) = d_{\mathrm{KL}}(x \mid y) = x \log \frac{x}{y} + (y-x), \tag{2.8}$$

$$d_{\beta\to0}(x \mid y) = d_{\mathrm{IS}}(x \mid y) = \frac{x}{y} - \log \frac{x}{y} - 1. \tag{2.9}$$

Figure 2.5 depicts these divergences for $x = 1$. A convenient property of β-divergences is that their first and second order partial derivatives with respect to y (the model) are continuous in $\beta \in \mathbb{R}$.

$$\frac{\partial d_\beta(x \mid y)}{\partial y} = y^{\beta-2}(y-x), \tag{2.10}$$

$$\frac{\partial^2 d_\beta(x \mid y)}{\partial^2 y} = y^{\beta-3}((\beta-1)y + (2-\beta)x). \tag{2.11}$$

Thus, when deriving update rules for optimization algorithms based on gradient descent, a general formulation may be obtained for all metrics gathered in the β-divergence family (*cf.* Section 2.1.1.5).

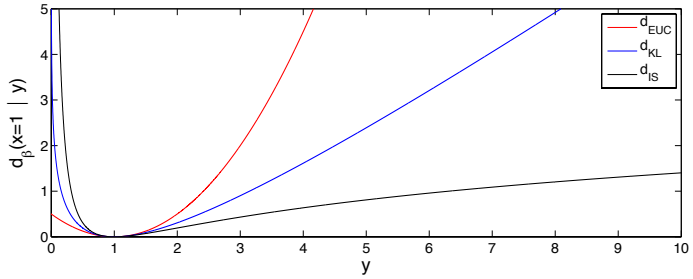


Figure 2.5: EUC, KL and IS divergences for $x = 1$.

Scaling properties: When dealing with a specific application, the choice of the cost-function should be done according to the particular properties of the data. In the case where the NMF is applied to the factorization of audio spectrograms, an important consideration to take into account is that magnitude spectra have a large dynamic range. As illustrated on Figure 2.6(a), audio spectra are usually characterized by a fast decrease of the spectral envelope and only a few partials with low rank have a large magnitude. However, some higher-rank partials with lower magnitudes still have significant loudness for auditory perception, this latter being more closely related to a logarithmic scale (*cf.* Figure 2.6(b)). Likewise, along the time-frame axis, spectrograms contain much more energy at attack transients than in decay parts. Thus, an important feature to consider when choosing a divergence is its scaling property [Févotte et al., 2009]. When dealing with $\beta-$divergences the following expression is obtained from Equation (2.6):

$$d_\beta(\gamma x \mid \gamma y) = \gamma^\beta \cdot d_\beta(x \mid y), \quad \gamma \in \mathbb{R}^+. \tag{2.12}$$

It implies that according to the value of β, different relative weights will be given in the cost-function (Equation (2.5)) for each time-frequency bin depending on its magnitude. For instance, for $\beta > 0$ the reconstruction of time-frequency bins presenting highest magnitudes will be favored since a bad fit on these coefficients will cost more than an equally bad fit on these with lowest magnitude. For $\beta < 0$ a converse effect will be obtained. Only $\beta = 0$, which corresponds to the IS divergence, exhibits a scale invariance property which should be suitable for applications to audio source separation since it should allow for a reliable reconstruction of all coefficient regardless of their magnitude, even low-valued ones that may be audible [Févotte et al., 2009]. For these reasons, most studies related to audio applications usually consider values of $\beta \in [0, 1]$ [Virtanen, 2007; Ozerov and Févotte, 2010; Vincent et al., 2010; Dessein et al., 2010]. EUC is rarely used, and it is sometimes replaced by weighted Euclidean distances using perceptual frequency-dependent weights [Vincent and Plumbey, 2007; Vincent et al., 2008]. However, a fine tuning of β within the range $[0, 1]$ is not straightforward and seems to be dependent on the application and on the value of the spectrogram exponent (*e.g.* 1 for a magnitude or 2 for a power spectrogram). A detailed study on the β parameter has shown that a KL divergence for processing magnitude spectrograms in source separation applications was leading to highest performances (evaluated in terms of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR)) [FitzGerald et al., 2009]. When processing power spectrograms for the same dataset, performances were found optimal for $\beta = 0.5$, but lower than those obtained on magnitude spectrograms. In application to polyphonic transcription [Vincent et al., 2010], optimal pitch estimation performances (F-measure) were obtained for $\beta = 0.5$ when processing magnitude spectrograms.

Convexity: Another important property to ensure that the optimization problem holds a single minimum is the convexity of the cost-function. Equation (2.10) shows that β-divergences present a single minimum for $x = y$ and increase with $\mid x - y \mid$. However, the convexity with respect to y for x fixed is limited to the range $\beta \in [1, 2]$ (*cf.* Equation (2.11)). Thus, IS is non-convex and present an horizontal asymptote for $\beta \to +\infty$. Moreover, even for $\beta \in [1, 2]$, the convexity of β-divergences does not necessarily lead to the convexity of the reconstruction cost-function given in Equation (2.5) since the optimization of the parameters $\{W_{k,r}\}_{k\in[1,K],r\in[1,R]}$ and $\{H_{r,t}\}_{r\in[1,R],t\in[1,T]}$ should be jointly

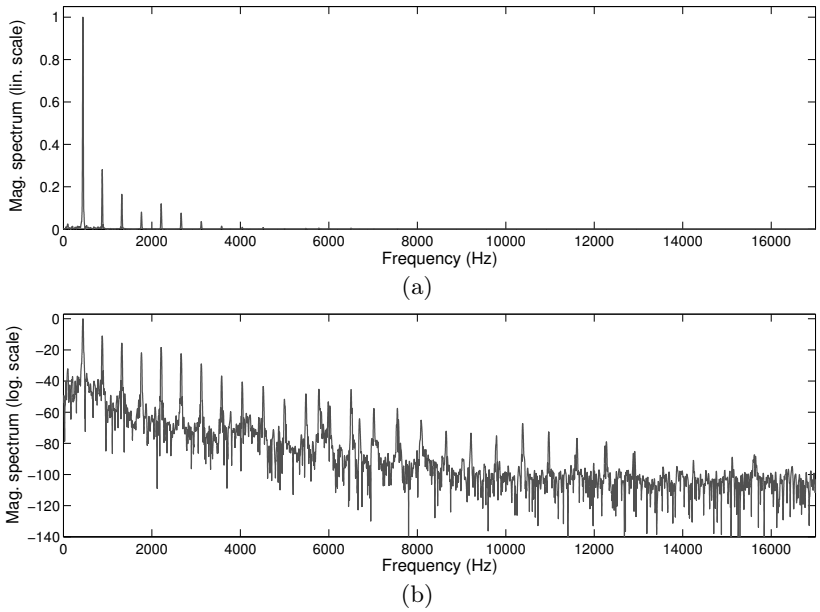


Figure 2.6: Magnitude spectrum of a piano note A3 in (a) linear and (b) dB scale.

performed. Thus, most algorithms perform updates of the parameters in an iterative way, each coefficient being independently updated while others are kept fixed.

In the case where dependencies are considered for the elements of W or H, for instance by setting parametric models (*cf.* Section 2.1.2.2), the convexity of the cost-function with respect to the new highest-level parameters will not necessarily be verified and heuristics may be used to ensure that the algorithm does not stop in a local minimum (as shown in Section 3.1.2.2 where an NMF parametric model for piano sound analysis is introduced).

2.1.1.4 Notes about probabilistic NMF models

As seen so far, solving an NMF problem requires the choice of a cost-function, this latter being usually done according to a trade-off between the specific properties of the observations and the convexity of the problem. An alternative method consists of considering the set of observations as a realization of a generative process composed of random variables. Thus, the choice of the probability distribution that is associated to each variable provides a cost function, for instance by considering a maximum-likelihood formulation. A main benefit of the probabilistic framework relies in the possibility of including information about the data by setting *a priori* distributions on the NMF parameters, and according to Bayesian inference solving a maximum *a posteriori* problem. For instance, in [Yoshioka and Sakaue, 2012] a log-normal distribution is assumed for each element of the dictionary of spectra W_{kt}. This choice leads to a cost-function corresponding to the squared error between the observation and the model spectrograms with logarithmic magnitude. Such a model can account for the high dynamic range of audio signals, as discussed above.

In some cases, both deterministic and probabilistic approaches can be shown to be equivalent [Févotte and Cemgil, 2009]. For instance, taking standard NMF cost-function (Equation (2.5)) with KL divergence is equivalent to considering that the sources, indexed by r, that contribute to the generation of a time-frequency bin of the observed magnitude spectrogram V_{kt}, are independent and distributed according to a Poisson distribution

with parameter $W_{kr}H_{rt}$ [Virtanen et al., 2008]. For IS an equivalence is obtained by assuming that each source independently contributes to the time-frequency bin of the power spectrogram $\mid V_{kt} \mid^2$ according to a complex Gaussian distribution with null mean and variance $W_{kr}H_{rt}$ [Févotte et al., 2009].

2.1.1.5 Optimization techniques

A wide variety of continuous optimization algorithms can be used to solve problem (2.5) [Cichocki et al., 2008]. For instance, standard algorithms based on gradient descent [Lin, 2007; Wang and Zou, 2008] or Newton methods [Zdunek and Cichocki, 2007] may be applied to NMF problems. However, in order to preserve the non-negativity of the decomposition obtained with these methods a projection onto the positive orthant is required after each update. Thus, other methods that implicitly preserve the non-negativity are usually preferred. Among such techniques, the most popular are very likely the multiplicative algorithms whose update rules can be obtained from heuristic decompositions of the cost-function or in a more rigorous way by using Majorization-Minimization (MM) methods [Lee and Seung, 2000; Févotte and Idier, 2011]. In the case of probabilistic NMF models, methods such as the Expectation-Maximization algorithm [Dempster et al., 1977] or variants such as the Space-Alternating Generalized Expectation-Maximization (SAGE) algorithm [Fessler and Hero, 2010; Bertin et al., 2010] are commonly used.

Multiplicative algorithms are the most employed in practice because they guarantee the non-negativity of the decomposition and it is often experimentally verified that they lead to satisfactory solutions with a reasonable computational time. Thus, for the NMF models proposed in this thesis we chose to focus on this method in order to perform the optimizations.

Multiplicative algorithms: The heuristic approach for deriving multiplicative updates consists of decomposing the partial derivatives of a cost function, with respect to a given parameter θ^*, as a difference of two positive terms:

$$\frac{\partial C(\theta^*)}{\partial \theta^*} = P(\theta^*) - Q(\theta^*), \qquad P(\theta^*),\, Q(\theta^*) \geq 0 \tag{2.13}$$

and at iteratively updating the corresponding parameter according to:

$$\theta^* \leftarrow \theta^* \times Q(\theta^*)/P(\theta^*) \tag{2.14}$$

Since $P(\theta^*)$ and $Q(\theta^*)$ are positive, it guarantees that parameters initialized with positive values stay positive during the optimization, and that the update is performed in the descent direction along the parameter axis. Indeed, if the partial derivative of the cost function is positive (respectively negative), then $Q(\theta^*)/P(\theta^*)$ is smaller (resp. bigger) than 1 and the value of the parameter is decreased (resp. increased). At a stationary point, the derivative of the cost function is null so $Q(\theta^*)/P(\theta^*) = 1$.

When applied to the NMF model with β-divergences, the following expressions for the gradient are obtained by using Equations (2.5) and (2.10):

$$\nabla_H C(W,H) = W^T \left((WH)^{\cdot[\beta-2]} \otimes (WH - V) \right), \tag{2.15}$$

$$\nabla_W C(W,H) = \left((WH)^{\cdot[\beta-2]} \otimes (WH - V) \right) H^T, \tag{2.16}$$

where $\otimes$ and $\cdot^{[]}$ respectively denote element-wise multiplication and exponentiation, and ^{T} denotes the transpose operator. Since all matrices are composed of non-negative coefficients an obvious decomposition leads to the following updates:

$$H \leftarrow H \otimes \frac{W^T \left((WH)^{\cdot[\beta-2]} \otimes V\right)}{W^T (WH)^{\cdot[\beta-1]}}, \tag{2.17}$$

$$W \leftarrow W \otimes \frac{\left((WH)^{\cdot[\beta-2]} \otimes V\right) H^T}{(WH)^{\cdot[\beta-1]} H^T}, \tag{2.18}$$

where the fraction bar denotes element-wise division.

In the general case, no proof is given for the convergence of the algorithm or even for the decrease of the cost-function since the decomposition (2.13) is not unique and is arbitrarily chosen. When using MM algorithms, update rules are obtained by building a majorizing function which is minimized in an analytic way in a second step, thus proving the decrease of the criterion. In the case of NMF with β-divergences, identical rules to (2.17) and (2.18) have been obtained for $\beta \in [1, 2]$ with a proof of the decrease of the criterion by using MM algorithm formalism [Kompass, 2007]. However it should be noticed that the property of the decrease of the criterion does not ensure that the algorithm will end in a local and even less in the global minimum. Moreover, it can be observed in some case that an increase of the criterion at some iterations may lead to a much faster convergence toward a local minimum [Badeau et al., 2010; Hennequin, 2011].

2.1.2 Giving structure to the NMF

Besides the generic "blind" approach, prior information is often considered in order to better fit the decomposition to specific properties of the data. Indeed, the standard NMF problem, as it is given by Equation (2.4), does not guarantee that the optimal solution is suitable regarding the targeted application. For instance, in a music transcription task it cannot be ensured that each column of W will contain the spectrum of a unique note that has been played and not a combination of several notes. Moreover, there is no guarantee that an optimal solution (with respect to a given application) is obtained since different local optima may be reached by the algorithm, depending on the initialization of the parameters.

Thus, taking into account prior information in the initialization is a matter of importance since it should help in avoiding the convergence toward minima that may not be optimal for the considered application. For instance, since most instrumental tones share an harmonic structure, initializing W with harmonic combs having different fundamental frequencies is valuable when targeting polyphonic music source separation [Fritsch and Plumbey, 2013]. Likewise, for drum separation applications, the initialization of the activations of H by an onset detection function combined with a dictionary W for which the drum sources span different frequency bands allows the recovery of the different elements (bass-drum, snare, ...) in an efficient way [Liutkus et al., 2011].

Moreover, in order to reduce the search space to meaningful solutions, such information is often also included during the optimization process. For instance, when additional data is available (*e.g.* the score of the piece of music or isolated note recordings of the same instruments with pitch labels), one can directly use the supplementary information to perform the decomposition in a supervised way. When it is not the case, the data properties can be explicitly included in the modeling in order to constrain the decomposition in

a semi-supervised way. In the case of audio, these can be for instance physics-based (properties about the timbre of the tones), signal-based (*e.g.* sparsity of notes simultaneously played, smoothness of activations and temporal envelopes) or derived from musicological considerations (*e.g.* the note activation should be synchronized on the beat, or chord transitions may have different probabilities depending on the tonality). For each method presented in the following, the inclusion of a specific property is highlighted but it should be emphasized that models usually combine different types of information.

2.1.2.1 Supervised NMF

Supervised NMF methods usually consist in initializing some parameters of the decomposition according to some information already available, for instance provided as additional data, or obtained by some preliminary analysis process. These parameters are then fixed during the optimization.

For instance, if the score of a piece of music is known, elements of the activation matrix H are forced to 0 where is it is known that a note is not present and initialized to 1 otherwise. Such approaches may be used in audio source separation [Hennequin et al., 2011b; Ewert and Müller, 2012] where thus the problem is to estimate the remaining parameters (H for non-zero coefficients and W).

In the case where a training set is available, the dictionary of spectra W may be learned in a first step. For instance in a music transcription task, for which isolated note recordings of the same instrument and with the same recording conditions are available, the spectra can be learned independently for each note by applying a rank-one decomposition [Niedermayer, 2008; Dessein et al., 2010]. Then, the supervised NMF problem reduces to the estimation of the activation matrix H for the considered piece of music. Similar approaches have been applied to speech enhancement and speech recognition [Wilson et al., 2008; Raj et al., 2010].

2.1.2.2 Semi-supervised NMF

If no additional data is available, the information about the properties of the data may be explicitly considered in the optimization (*cf.* Paragraph Regularization) or in the modeling (*cf.* Paragraph Parameterization).

Regularization: As commonly done with ill-posed problems, the NMF can be constrained by considering a regularized problem. In practice, it consists of adding to the reconstruction cost-function some penalty terms that emphasize specific properties of the variables. When the constraints are considered independently for W and H, which is usually the case, the optimization problem can be expressed as the minimization of a new cost-function given by:

$$C(W,H) = D(V \mid WH) + \lambda_H \cdot C_H(H) + \lambda_W \cdot C_W(W). \tag{2.19}$$

where $C_H(H)$ and $C_W(W)$ correspond to penalty terms respectively constraining H and W matrices and whose weightings are given by λ_H and λ_W. The choice of the regularization parameters λ_H and λ_W can be done using empirical rules, or through cross-validation. In the probabilistic NMF framework, this type of constraint is equivalent to adding *a priori* distributions on H and W, as $-D(V \mid WH)$ corresponds to the log-likelihood, up to a constant.

A number of penalties have been proposed in the literature, for instance in order to enforce sparsity of H, *i.e.* reducing the set of notes that should be activated [Eggert and Körner, 2004; Hoyer, 2004; Virtanen, 2007; Joder et al., 2013], smoothness of the lines of H in order to avoid well-localized spurious activations [Virtanen, 2007; Bertin et al., 2009], or the decorrelation of the lines of H [Zhang and Fang, 2007].

Parameterization: Unlike regularization techniques that introduce the constraint in a "soft" way by favoring some particular solutions during the optimization, another common method to impose structure on the decomposition is to use parametric models for the matrices W and H.

A common example is a parametrization of the dictionary W to fit to the structure of the spectra. Here, the number of parameters reduces from $K \times T$ to a few meaningful parameters corresponding for instance to magnitudes and frequencies of the partials. In [Hennequin et al., 2010; Ewert et al., 2013] the spectra are modeled by harmonic combs, parameterized by F_0 and the partials' magnitude. In [Bertin et al., 2010] each spectrum is modeled as a linear combination of narrow-band sub-spectra, each one being composed of a few partials with fixed magnitude and all sharing a single parameter F_0. This latter model of additive sub-spectra brings smoothness to the spectral envelope, which reduces the activation of spurious harmonically-related notes (octave or fifth relations, for instance, where partials fully overlap) [Klapuri, 2001; Virtanen and Klapuri, 2002].

For the matrix H, the temporal regularity underlying the occurrences of note onsets is modeled in [Ochiai et al., 2012]. Thus, each line of H is composed of smooth patterns located around multiples or fractions of the beat period. Besides avoiding spurious activations, this parametrization also makes post-processing of H easier in order to perform transcription.

Several studies also exploit in NMF the non-stationarity of musical sounds. For instance in [Durrieu et al., 2010, 2011] a source-filter model for main melody extraction is presented. The main melody is here decomposed in two layers, one standard-NMF (the source) where the dictionary of spectra is fixed with harmonic combs, multiplied by a second NMF (the filter) adjusting the time-variations of the spectral envelopes with a combination of narrow-band filters contained in the second dictionary. In [Hennequin et al., 2011a], the temporal variations of the spectral envelopes are handled by means of a time-frequency dependent activation matrix $H_{rt}(f)$ based on an auto-regressive model. The vibrato effect is also modeled in [Hennequin et al., 2010] by considering time-dependent F_0s for the dictionary of spectra. In a more flexible framework, all these temporal variations (*e.g.* changes in the spectral content of notes between attack to decay transitions, spectral envelope variations or vibrato) are taken into account by considering several basis spectra for each note/source of the dictionary and Markov-chain dependencies between their activations [Nakano et al., 2010; Mysore et al., 2010].

Such methods for enforcing the structure of the NMF will inspire the models presented in Chapters 3 and 5 of this thesis, where we focus on specific properties of the piano.

2.2 Considering the specific properties of piano tones

This section introduces a few basic notions from musical acoustics in order to point out some particular features of piano tones that should be relevant for piano music analysis. Thus, it is not intended to give a complete description and modeling of the instrument's elements and their interaction. We rather focus on the core element of the piano, namely the strings, whose properties are mainly responsible for the inharmonic nature of the tones. A more detailed study about the influence of the inharmonicity in the string design and tuning is proposed in Chapter 4.

2.2.1 Presentation of the instrument

From the keyboard, which allows the player to control the pitch and the dynamics of the tones, to the propagation of the sound through the air, the production of a piano note results from the interaction of various excitation and vibrating elements (*cf.* Figure 2.7). When a key is depressed, an action mechanism is activated with a purpose to translate the depression into a rapid motion of a hammer, this latter inducing a free vibration of the strings after the stroke. Because of their small radiating surface, the strings cannot efficiently radiate the sound. It is the coupling at the bridge that allows the transmission of the vibration from the strings to the soundboard, this latter acting as a resonator and producing an effective acoustic radiation of the sound. In order to produce tones having, as much as possible, a homogeneous loudness along the whole compass, the notes in the treble-medium, bass and low-bass register are respectively composed of 3, 2 and 1 strings. Finally, at the release of the key, the vibration of the strings is muted by the fall of a damper.

In addition to the keyboard, pianos usually include a pedal mechanism that allows the instrumentalist to increase the expressiveness of his playing by controlling the duration and the loudness of the notes. The most commonly used is the *sostenudo* pedal, that raises the dampers off the played strings so they keep vibrating after the key has been released. In a similar way, the damper pedal activation leads to a raise of *all* the piano dampers, increasing the sustain of the played notes and allowing for the sympathetic resonances of other strings. In order to control the loudness and the color of the sound, the *una corda* pedal, when activated, provokes a shift of the action mechanism so that the hammer strikes only two strings over the three composing the notes in the medium-treble range.

The realistic synthesis of piano tones by physical modeling requires a precise description of the different elements and their interactions. Thus, a number of works have focused on modeling specific parts and characteristics of the piano such as, the action mechanism [Hayashi et al., 1999; Hirschkorn et al., 2006], the hammer-string interaction [Suzuki, 1986a; Chaigne and Askenfelt, 1993; Stulov, 1995], the pedals [Lehtonen et al., 2007], the phenomenon of sympathetic resonance [Le Carrou et al., 2005], the vibration of the strings [Young, 1952; Fletcher, 1964; Fletcher and Rossing, 1998], their coupling [Weinreich, 1977; Gough, 1981], the soundboard vibrations [Suzuki, 1986b; Mamou-Mani et al., 2008; Ege, 2009]... In order to perform the synthesis, a discretization of the obtained differential equations is usually required [Bensa, 2003; Bensa et al., 2003; Bank et al., 2010; Chabassier, 2012] and the synthesis parameters are often directly obtained from measurements on pianos. Thus, the inversion of such models is usually not straightforward. For the goal of our work, which targets the inclusion of information from physics in signal-based analysis

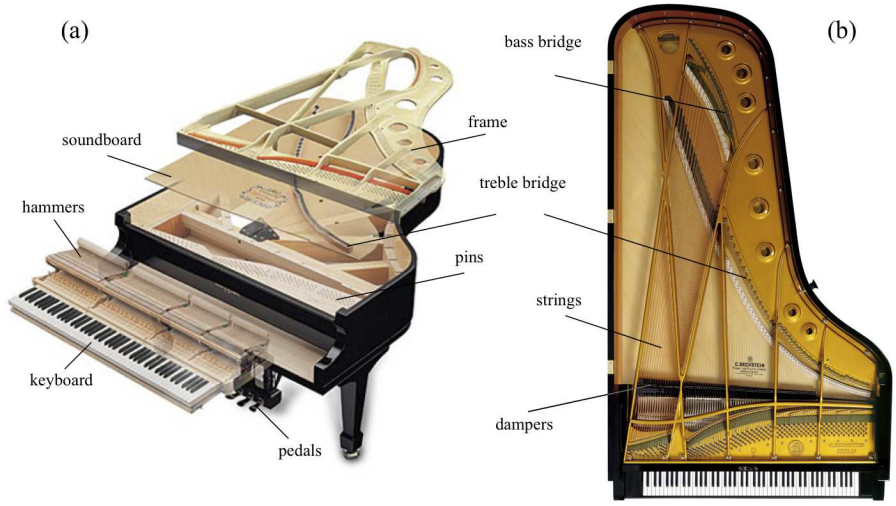


Figure 2.7: Design of grand pianos.
Sources: (a) *www.pianotreasure.com*, (b) *www.bechstein.de*

models, we only focus on the transverse vibrations of the strings. This simple model explains, to a great extent, the spectral content of piano tones and their inharmonic property.

2.2.2 Model for the transverse vibrations of the strings

• **Flexible strings:** First consider a string characterized by its length L and linear mass μ, being subject to a homogeneous tension T and no damping effects. At rest position, the string is aligned according to the axis (Ox). When excited, its transverse displacement over time t, $y(x,t)$ follows, at first order, d'Alembert's differential equation:

$$\frac{\partial^2 y}{\partial^2 x} - \frac{\mu}{T}\frac{\partial^2 y}{\partial^2 t} = 0. \tag{2.20}$$

By considering fixed end-point boundary conditions, the solutions correspond to stationary waves with frequencies related by an harmonic relation $f_n = nF_0, n \in \mathbb{N}^\star$, whose fundamental frequency is given by:

$$F_0 = \frac{1}{2L}\sqrt{\frac{T}{\mu}}. \tag{2.21}$$

• **Bending stiffness consideration:** This latter model for the transverse vibration is not precise enough to accurately describe the actual behavior of piano strings. Because of the important size of the strings and the properties of the piano wire (when compared to other string instruments such as the guitar), the bending stiffness effect has to be considered. Then, by introducing the diameter d of the plain string, having a Young's modulus E and an area moment of inertia $I = \frac{\pi d^4}{64}$, the following differential equation is obtained:

$$\frac{\partial^2 y}{\partial^2 x} - \frac{\mu}{T}\frac{\partial^2 y}{\partial^2 t} - \frac{EI}{T}\frac{\partial^4 y}{\partial^4 x} = 0 \tag{2.22}$$

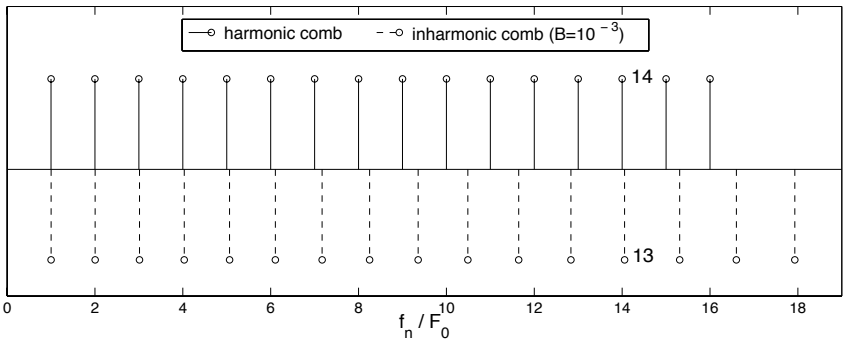


Figure 2.8: Comparison of an harmonic and inharmonic comb spectrum for $B = 10^{-3}$.

The presence of the fourth order differential term reflects the fact that the piano wire is a dispersive medium, *i.e.* the vibrating modes have different propagation speed. By still considering a fixed end-point boundary condition the modal frequencies can be expressed as an inharmonic relation [Morse, 1948; Fletcher and Rossing, 1998]:

$$f_n = nF_0\sqrt{1 + Bn^2}, \qquad n \in \mathbb{N}^\star, \tag{2.23}$$

where

$$B = \frac{\pi^3 E d^4}{64 T L^2} \tag{2.24}$$

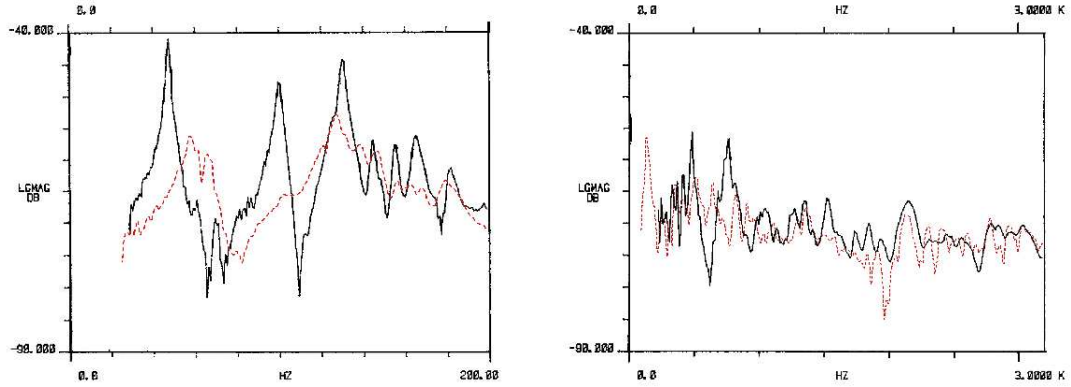
is a dimensionless coefficient called the inharmonicity coefficient. Since the mechanical characteristics of the strings differ from one note to another, obviously F_0 but also B are varying along the compass (typical values for B are in the range 10^{-5}-10^{-2}, from the low-bass to the high-treble register [Young, 1952; Fletcher, 1964]). Perceptual studies have shown that inharmonicity is an important feature of the timbre of piano tones [Fletcher et al., 1962; Blackham, 1965] that should be taken into account in synthesis applications, especially for the bass range [Järveläinen et al., 2001]. In addition, it has a strong influence on the the design [Conklin, Jr., 1996b] and on the tuning [Martin and Ward, 1961; Lattard, 1993] of the instrument. This latter point is presented in more detail in Chapter 4.

In the spectrum of a note, inharmonicity results in a sharp deviation of the partial frequencies when compared to a harmonic spectrum, and the higher the rank of the partial, the sharper the deviation. Because the inharmonicity coefficient values are quite low, the deviation is hardly discernible for low rank partials. However, for high rank partials, the deviation becomes important and the inharmonicity consideration must be taken into account when targeting spectral modeling in piano music analysis tasks. For instance, the 13th partial of an inharmonic spectrum will be sharper than the 14th partial of a harmonic spectrum when considering an inharmonicity coefficient value $B = 10^{-3}$ (*cf.* Figure 2.8).

2.2.3 Notes about the couplings

• **Influence of the soundboard mobility:** The modal frequencies of transverse vibrations given by Equation (2.23) are obtained by assuming a string fixed at end-points. In practice, this consideration is a good approximation for the fixation at the pin, but it does not reflect the behavior of the string-soundboard coupling at the bridge. Indeed, such a boundary condition considers a null mobility at the bridge fixation, while the actual aim of this coupling is to transfer the vibration of the string to the soundboard.

As well as the strings, the soundboard possesses its own vibrating modes [Conklin, Jr., 1996a; Suzuki, 1986b; Giordano, 1998; Mamou-Mani et al., 2008; Ege, 2009] and their coupling may induce changes in the vibratory behavior of the assembly. This can be seen on Figure 2.9, where measurements of a soundboard's transverse mobility (ratio velocity over force as a function of the frequency) are depicted in black and red, respectively, for a configuration "strings and plate removed" and "whole piano assembled and tuned". When considering only the soundboard (black lines), a few well-marked modes are usually present in the low-frequency domain (approximately under 200 Hz). While going up along the frequency axis the modal density increases, and because of the overlap of the modes the resonances become less and less pronounced, until reaching a quasi-flat response for the high-frequency domain (above 3.2 kHz, not depicted here). When assembling the strings and the soundboard, the vibratory behavior of each element is altered (*cf.* red curve for the soundboard). Thus, the modal frequencies of transverse vibrations of the strings given by Equation (2.23) are slightly modified. This can be observed in the spectrum of a note, as depicted in Figure 2.10, where the partial frequencies and ranks corresponding to transverse vibrations of the strings have been picked up and represented in the graphic f_n^2/n^2 as a function of n^2. When comparing the data to the inharmonic relation (2.23) (corresponding to a straight line in this graphic) one can notice slight deviations, mainly for low rank partials having frequencies for which the soundboard should present a strong modal character.



(a) 0-200 Hz (b) 0 - 3.2 kHz

Figure 2.9: Bridge mobility (direction normal to the soundboard) of a grand piano at the endpoint of the E2 strings, after [Conklin, Jr., 1996a]. Solid black and dashed red curves respectively correspond to "strings and plate removed" and "piano assembled and tuned" configurations.

• **Other string deformations and their coupling:** Beyond the one-dimensional transverse deformations considered in Section 2.2.2, the propagation of several waves occur in the string after the strike of the hammer, these being able to contribute to the distinct timbre of the instrument. For instance, the propagation of longitudinal waves has been shown to be significant for the perception, for notes in the bass range, up to A3 [Bank and Lehtonen, 2010]. For slight deformations, produced for instance when the notes are played with *piano* dynamics, the couplings between those different waves (*e.g.* the two polarizations of the transverse waves, the longitudinal and the torsional waves) are neg-

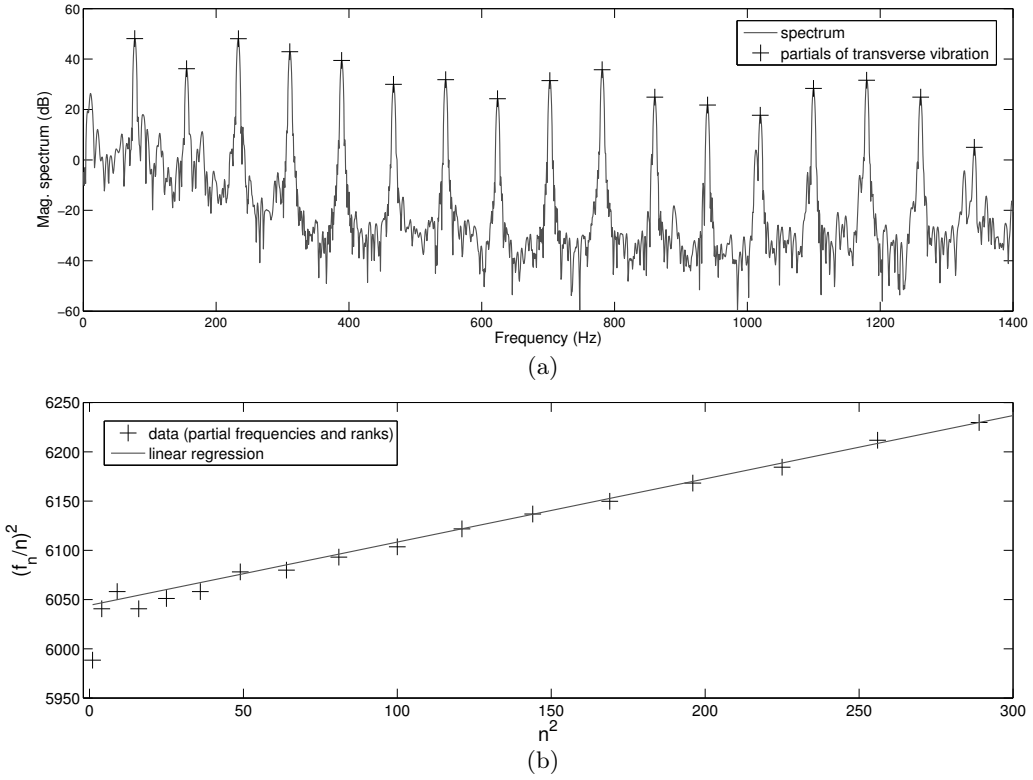


Figure 2.10: Influence of the string-soundboard coupling for a note E2. (a) Magnitude spectrum from which the partials corresponding to transverse vibrations of the strings are emphasized with '+' markers. (b) Partial frequencies depicted in the graphic f_n^2/n^2 as a function of n^2 ('+' markers) and linear regression corresponding to the theoretical relation (2.23).

ligible and each deformation can be considered independently. Thus, when studying the propagation of longitudinal vibration in the strings, a simple harmonic model is obtained for the modal frequencies [Valette and Cuesta, 1993].

In practice, for an accurate modeling of the vibratory behavior of the strings for all dynamics, all these deformations should be jointly considered, this leading to a non-linear differential equation system and explaining the apparition of the so-called "phantom partial" in the spectrum of piano tones [Conklin, Jr., 1999; Bank and Sujbert, 2005]. Moreover, for notes in the medium and treble register, one should take into account the coupling of doublets or triplets of strings (usually slightly detuned in order to increase the sustain of the sound) through the bridge [Weinreich, 1977; Gough, 1981] in order to model the presence of multiple partials (*cf.* Figure 2.11) and the double decays and beats in the temporal evolution of partials [Aramaki et al., 2001] (*cf.* Figure 2.12).

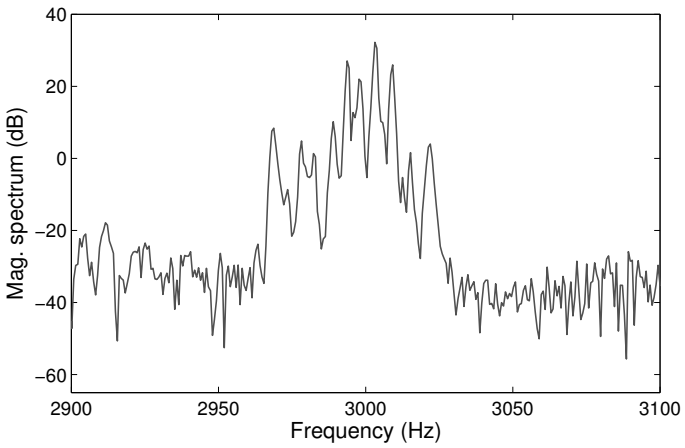
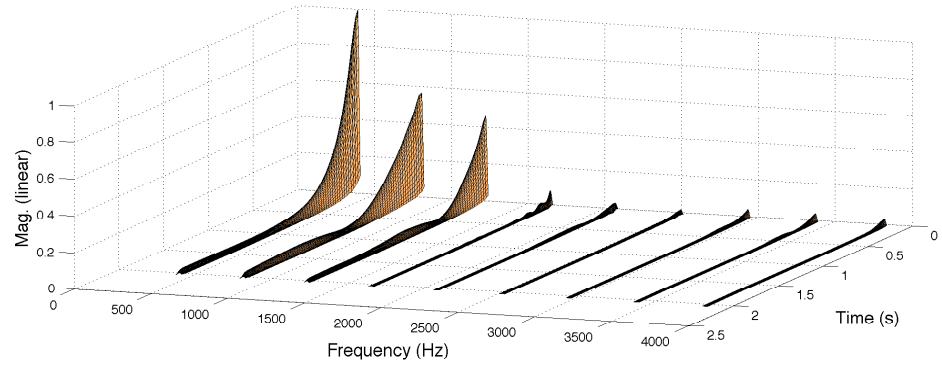
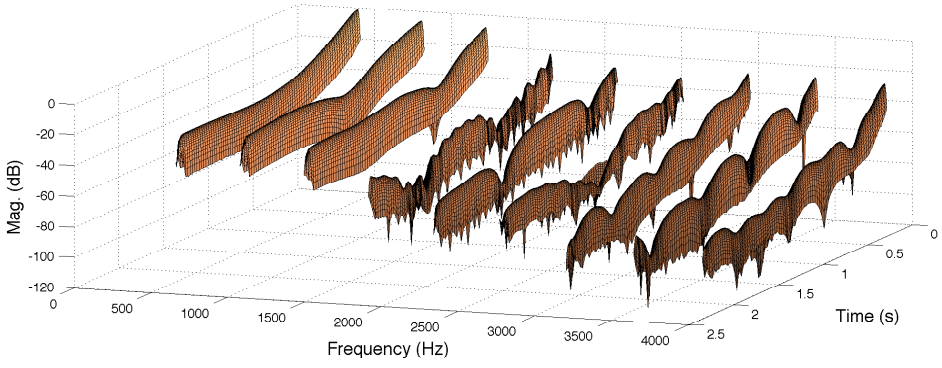


Figure 2.11: Zoom along the frequency axis of the spectrum of a note G♭6 composed of a triplet of strings. Multiple peaks are present while the simple model of transverse vibration for a single string assumes a single partial.



(a)



(b)

Figure 2.12: Double decay and beating of notes composed of multiple strings. Temporal evolution (a) in linear and (b) in dB, of the first 9 partials of transverse vibration for a note G♯4 played with *mezzo-forte* dynamics.

2.3 Analysis of piano music with inharmonicity consideration

Beyond applications related to music synthesis, the estimation of the inharmonicity coefficient of piano tones is an issue of importance in a number of analysis tasks such as F_0 and multiple F_0 estimation [Emiya et al., 2007b,a; Rauhala and Välimäki, 2007; Blanco-Martín et al., 2008], polyphonic music transcription [Emiya et al., 2010a; Benetos and Dixon, 2011], temperament and tuning estimation [Lattard, 1993; Dixon et al., 2012] or piano chord recognition [Ortiz-Berenguer et al., 2004; Ortiz-Berenguer and Casajús-Quirós, 2002].

The methods proposed in the literature can be classified according to two main types of approaches, the first one considers the estimation of (B, F_0) from the search of the partials related to transverse vibrations of the strings in spectra, while the second one considers them as particular parameters of a signal model.

2.3.1 Iterative peak-picking and refinement of the inharmonicity coefficient

This first type of method is usually based on a two-step iterative scheme that requires the knowledge of the notes that are being played. From rough initial values of (B, F_0) (F_0 being usually initialized according to Equal Temperament - *cf.* Appendix A - and B being set to a low value, typically in the range $10^{-5} - 10^{-4}$) a few partials located around the theoretical frequencies (given by the Equation (2.23)) are selected in the magnitude spectrum in a peak-picking step. Having the frequencies and ranks of these partials, (B, F_0) are refined according to the inharmonicity relation or alternative forms of it, and the search for partials with higher ranks is iterated. Then, the procedure is usually run until no peak is found above a threshold on the magnitude of the spectrum. Note that in order to increase the precision of the estimation, most methods introduce a step of refinement of the partial frequencies after the peak-picking, this latter being often based on a local interpolation of the peak's magnitude. Finally, the core difference for all these methods relies in the choice of the estimator that is used to compute (B, F_0) from the knowledge of the partial frequencies.

For instance, in [Rauhala et al., 2007a; Rauhala and Välimäki, 2007], the *PFD* (Partial Frequency Deviation) algorithm estimates (B, F_0) by minimizing the deviation between the theoretical partial frequencies of the model and the frequencies of the peaks selected in the magnitude spectra.

In [Emiya, 2008] an alternative form of the inharmonicity relation is proposed

$$\frac{f_n^2}{n^2} = F_0^2 + F_0^2 B \cdot n^2. \tag{2.25}$$

and (B, F_0) are estimated by means of a least-square linear regression in the plane f_n^2/n^2 as a function of n^2.

Another alternative approach estimates B by considering pairs of partials. Indeed, in order to estimate B from the frequencies of a couple of partials (f_j, f_k) the following expression can be obtained when inverting Equation (2.23):

$$B_{jk} = \frac{j^2 f_k^2 - k^2 f_j^2}{k^4 f_j^2 - j^4 f_k^2}. \tag{2.26}$$

In practice, B cannot be robustly estimated by the simple consideration of two partials since this estimator is highly sensitive to the frequency deviations produced by the bridge coupling [Ortiz-Berenguer et al., 2004] . Thus, it has been proposed in [Hodgkinson et al., 2009] to estimate B by considering the median over the set of B_{jk} values computed for each possible pair of partials. Such an estimator has been applied to the unsupervised estimation of inharmonicity and temperament, together with a transcription task in a polyphonic recording context [Dixon et al., 2012].

2.3.2 Inharmonicity inclusion in signal models

The second approach consists in explicitly including (B, F_0) as parameters of a signal model and is often used when targeting tasks such as multiple-F_0 estimation or polyphonic transcription.

In [Godsill and Davy, 2005], a Bayesian framework for modeling inharmonic tones in the time domain is introduced. (B, F_0), as well as the partial amplitudes, are considered as random variables and estimated by solving a maximum *a posteriori* problem. In [Davy et al., 2006], the model is extended to inharmonic tones that do not necessarily follow the inharmonicity relation given in Equation (2.23). In the time-frequency domain, a model based on the NMF framework has been proposed in [Vincent et al., 2008] to deal with a transcription task. (B, F_0) are here included as parameters of the dictionary of spectra and optimized by minimizing a reconstruction cost-function based on a weighted Euclidean norm. Surprisingly, the transcription results were found slightly below those obtained by a simpler harmonic model. The same conclusion was found for a similar parametric NMF model [Hennequin, 2011], where the difficulty of updating B was encountered. Such NMF-based models will form the basis of the work presented in Chapter 3, where we investigate how we can go beyond these limitations.

In order to avoid these difficulties in the optimization, other methods split the estimation problem into two sub-problems. First, they estimate (B, F_0) parameters of the signal model by optimizing a detection function designed so that the influence of the inharmonicity parameters is put forward. Then, they estimate the remaining parameters of the model according to the original problem formulation. As proposed by the following studies, the optimization of the detection function is usually performed by means of a grid search over (B, F_0) parameters for each possible note. Such an optimization technique may be found in [Galembo and Askenfelt, 1999], where they estimate the inharmonicity of piano tones by performing an inharmonic comb filtering of the magnitude spectrum. In the case of a transcription tasks, a similar approach based on an inharmonic spectral product has been introduced in [Emiya et al., 2010a]. This latter is applied in a first step in order to obtain a selection of note candidates, jointly with the estimation of their parameters (B, F_0), before the estimation of the amplitude and noise parameters of the model according to a maximum likelihood problem. A similar detection function (referred as pitch salience function) for log-frequency spectrum has been proposed in [Benetos and Dixon, 2011].

2.3.3 Issues for the thesis

As seen in this section, considering the inharmonicity of piano tones in signal models does not seem straightforward, particularly because of optimization issues. Some works bypass these difficulties by separating the estimation problems of (B, F_0) and amplitude/noise

parameters, but in most studies a simpler harmonic model is considered [Monti and Sandler, 2002; Marolt, 2004; Kobzantsev et al., 2005; Bello et al., 2008; Vincent et al., 2010].

The main goal of this thesis is thus to have a better understanding about the issues arising from the inharmonicity inclusion in signal models and to investigate whether it is really valuable when targeting tasks such as polyphonic music transcription. For this, different models in which (B, F_0) are included as parameters (two NMF-based models and a generative probabilistic model for the frequencies having significant energy in spectrograms) and their optimization algorithms are introduced in Chapter 3. These are applied to the precise estimation of (B, F_0) from monophonic and polyphonic recording in both supervised and unsupervised conditions. A special care is taken in the initialization of these parameters, by introducing in Chapter 4 a model for the inharmonicity and tuning along the whole compass of pianos. Based on invariants in design and tuning rules, the model is able to explain the variations of piano tuning along the compass with only a few parameters. Beyond the initialization of the analysis algorithms, it is applied to model the tuning of well-tuned pianos, to provide tuning curves for out-of-tune pianos or physically-based synthesizers and finally to interpolate the inharmonicity and tuning of pianos along the whole compass from the analysis of a polyphonic recording containing only a few notes. Finally the efficiency of an inharmonic model for NMF-based transcription is investigated in Chapter 5.

Chapter 3

Estimating the inharmonicity coefficient and the F_0 of piano tones

This chapter presents two new frameworks for the estimation of (B, F_0). In Section 3.1, two different NMF-based models in which (B, F_0) are included as parameters are presented. For each, update rules are derived and practical solutions concerning the optimization are proposed in order to avoid the convergence of the algorithms toward local minima. Both models are applied to the supervised estimation (*i.e.* the notes are known) of (B, F_0) from isolated note and chord recordings, and the performances are compared to the *PFD* algorithm (described in Section 2.3). Portions of this work have been published in [Rigaud et al., 2012, 2013a]. In Section 3.2, a generative probabilistic model for the frequencies of significant energy in the time-frequency domain is introduced. From a prior peak-picking in a magnitude spectrum, the estimation of (B, F_0) is performed jointly with a classification of each observed frequency into noise or partial components for each note of the model. The algorithm is then applied to the unsupervised estimation of (B, F_0) from isolated note and polyphonic recordings. This latter work has been published in [Rigaud et al., 2013b].

3.1 NMF-based modelings

The purpose of this section is to introduce the information of the inharmonicity of piano tones explicitly into the dictionary of spectra W of NMF-based modelings. The idea is to take into account the parameters (B, F_0) as constraints on the partial frequencies of each note, so as to perform a joint estimation. In order to limit the number of parameters that we need to retrieve, besides the amplitude and frequency of each partial, we make the assumption that for every recording we know which notes are being played, and the corresponding time activations. We refer to this case as supervised estimation. Then, short-time spectra are extracted from the recordings and concatenated to build the observation matrix V (it is therefore not strictly speaking a spectrogram). Because for each column of V the played notes are known, the elements of H are fixed to one whenever a note is played, and zero when it is not. Thereby, only the dictionary W is optimized on the data. In that case, we should notice that the proposed model is not a proper factorization, because there is no explicit modeling along the time axis. However, since the model is developed in the NMF framework, further inclusion in transcription (*cf.* Chapter 5) or source separation algorithms may still be considered, where the activations are not known

and must be jointly optimized (unsupervised case).

3.1.1 Modeling piano sounds in W

The model for the spectra/atoms of the dictionary W is based on an additive model: the spectrum of a note is composed of a sum of partials, in which the frequencies are constrained by the inharmonicity relation introduced in Equation (2.23). Two different ways of enforcing the constraint are proposed. The first model (later called *Inh-NMF*) forces the partial frequencies to strictly comply with the theoretical inharmonicity relation, while the second model (later called *InhR-NMF*) relaxes this constraint, and enhances inharmonicity through a weighted penalty term.

3.1.1.1 General additive model for the spectrum of a note

The general parametric atom used in this work is based on the additive model proposed in [Hennequin et al., 2010]. Each spectrum of a note, indexed by $r \in [1, R]$, is composed of the sum of N_r partials. The partial rank is denoted by $n \in [1, N_r]$. Each partial is parametrized by its amplitude a_{nr} and its frequency f_{nr}. Thus, the set of parameters for a single atom is denoted by $\theta_r = \{a_{nr}, f_{nr} \mid n \in [1, N_r]\}$ and the set of parameters for the dictionary is denoted by $\theta = \{\theta_r \mid r \in [1, R]\}$. Finally, the expression of a parametric atom is given by:

$$W_{kr}^{\theta_r} = \sum_{n=1}^{N_r} a_{nr} \cdot g_\tau(f_k - f_{nr}), \tag{3.1}$$

where f_k is the frequency of the bin with index k, and $g_\tau(f_k)$ is the magnitude of the Fourier Transform of the τ-length analysis window. In order to limit the computational time and to obtain simple optimization rules, the spectral support of g_τ is restricted to its main lobe. For the experiments proposed in this thesis, a Hann window is used to compute the spectra. The magnitude spectrum of its main lobe (normalized to a maximal magnitude of 1) is given by:

$$g_\tau(f_k) = \frac{1}{\pi\tau} . \frac{\sin(\pi f_k \tau)}{f_k - \tau^2 f_k^3} , \quad f_k \in [-2/\tau, 2/\tau]. \tag{3.2}$$

In order to estimate the parameters, the reconstruction cost-function to minimize is chosen as the β-divergence between the observed spectra V and the model $\hat{V} = W^\theta H$:

$$C_0(\theta, H) = \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} d_\beta \left(V_{kt} \mid \sum_{r=1}^{R} W_{kr}^{\theta_r} \cdot H_{rt} \right). \tag{3.3}$$

It is worth noting that since the partials of the model are defined on a limited set of frequency-bins $f_k \in f_{nr} + [-2/\tau, 2/\tau]$, the sum over k of Equation (3.3) is applied on the set $\mathcal{K} = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau], \forall n \in [1, N_r], \forall r \in [1, R]\}$.

3.1.1.2 Inharmonic constraints on partial frequencies

- *Strictly inharmonic / Inh-NMF:*

The strict inharmonic constraint consists in fixing:

$$f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}, \quad n \in \mathbb{N}^*, \tag{3.4}$$

directly in the parametric model (Equation (3.1)) so that the partials are forced to strictly follow the inharmonicity relation. Then the set of parameters for a single atom reduces to $\theta_r^{Inh} = \{a_{nr}, F_{0r}, B_r \mid n \in [1, N_r]\}$ and the reconstruction cost-function can be rewritten as

$$C^{Inh}(\theta^{Inh}, H) = \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} d_\beta \left(V_{kt} \mid \sum_{r=1}^{R} W_{kr}^{\theta_r^{Inh}} \cdot H_{rt} \right). \tag{3.5}$$

• *Inharmonic relaxed / InhR-NMF:*

An alternative way of enforcing inharmonicity is through an extra penalty term added to the reconstruction cost-function C_0 (Equation (3.3)). Thus, the global cost-function can be expressed as:

$$C^{InhR}(\theta, \gamma, H) = C_0(\theta, H) + \lambda \cdot C_1(f_{nr}, \gamma), \tag{3.6}$$

where the set of parameters of the constraint is denoted by $\gamma = \{F_{0r}, B_r \mid r \in [1, R]\}$. λ is a parameter, empirically tuned, that sets the weight between the reconstruction cost error and the inharmonicity constraint. The constraint cost-function C_1 is chosen as the sum on each note of the mean square error between the estimated partial frequencies f_{nr} and those given by the inharmonicity relation:

$$C_1(f_{nr}, \gamma_r) = K_\tau T \cdot \sum_{r=1}^{R} \sum_{n=1}^{N_r} \left(f_{nr} - nF_{0r}\sqrt{1 + B_r n^2} \right)^2, \tag{3.7}$$

where $K_\tau = \text{Card}\{f_k \in [-2/\tau, 2/\tau]\}$ is the number of frequency-bins for which the partials of the model are defined and T is the number of time-frames. This normalization factor allows a tuning of λ that is independent of these two values. A potential benefit of this relaxed formulation is to allow a slight deviation of the partial frequencies around the theoretical inharmonicity relation, that can be observed for instance in the low frequency range due to the coupling between the strings and the soundboard (*cf.* Section 2.2.3).

3.1.2 Optimization algorithm

3.1.2.1 Update of the parameters

As commonly proposed in NMF modeling, the optimization is performed iteratively, using multiplicative update rules for each parameter. For each modeling, the update rules are obtained from the decomposition of the partial derivatives of the cost-function, in a similar way to [Hennequin et al., 2010] (the detail of the derivation is detailed in Appendix C). In the following, $P(\theta^*)$ and $Q(\theta^*)$ refer to positive quantities obtained by decomposing the partial derivative of a cost function $C(\theta)$ with relation to a particular parameter θ^* so that $\frac{\partial C(\theta)}{\partial \theta^*} = P(\theta^*) - Q(\theta^*)$. The parameter is then updated as $\theta^* \leftarrow \theta^* \cdot Q(\theta^*)/P(\theta^*)$.

The update for a_{nr} are identical for both models since these parameters only appear in the reconstruction cost function C_0 and can be expressed as:

$$a_{nr} \leftarrow a_{nr} \cdot \frac{Q_0(a_{nr})}{P_0(a_{nr})}, \tag{3.8}$$

where

$$P_0(a_{nr}) = \sum_{k \in \mathcal{K}_{nr}} \sum_{t=1}^{T} \left[(g_\tau(f_k - f_{nr}).H_{rt}) .\widehat{V}_{kt}^{\beta-1} \right], \tag{3.9}$$

$$Q_0(a_{nr}) = \sum_{k \in \mathcal{K}_{nr}} \sum_{t=1}^{T} \left[(g_\tau(f_k - f_{nr}).H_{rt}) .\widehat{V}_{kt}^{\beta-2}.V_{kt} \right], \tag{3.10}$$

and $\mathcal{K}_{nr} = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau]\}$. $\widehat{V} = W^\theta H$ denotes the model. Note that for a transcription task, update rules for H will be given in Chapter 5.

The update rules of the remaining parameters are specific for each of the different NMF models. In the following, $g'_\tau(f_k)$ represents the derivative of $g_\tau(f_k)$ with respect to f_k on the spectral support of the main lobe. For a Hann window (normalized to a maximal magnitude of 1, *cf.* Equation (3.2)) and $f_k \in [-2/\tau, 2/\tau]$ its expression is given by

$$g'_\tau(f_k) = \frac{1}{\pi\tau} \frac{(3\tau^2 f_k^2 - 1)\sin(\pi\tau f_k) + \pi\tau(f_k - \tau^2 f_k^3)\cos(\pi\tau f_k)}{(f_k - \tau^2 f_k^3)^2}. \tag{3.11}$$

• *Strictly inharmonic / Inh-NMF:*

$$B_r \overset{Inh}{\leftarrow} B_r \cdot \left(\frac{Q_0^{Inh}(B_r)}{P_0^{Inh}(B_r)} \right)^\gamma, \tag{3.12}$$

$$F_{0r} \overset{Inh}{\leftarrow} F_{0r} \cdot \frac{Q_0^{Inh}(F_{0r})}{P_0^{Inh}(F_{0r})}, \tag{3.13}$$

where

$$\begin{aligned} P_0^{Inh}(B_r) = \sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} \Bigg[& \left(\sum_{n=1}^{N_r} a_{nr} \frac{-C.f_k.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right) .\widehat{V}_{kt}^{\beta-1} \\ & + \left(\sum_{n=1}^{N_r} a_{nr} \frac{-C.f_{nr}.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right) .\widehat{V}_{kt}^{\beta-2}.V_{kt} \Bigg], \end{aligned} \tag{3.14}$$

$$\begin{aligned} Q_0^{Inh}(B_r) = \sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} \Bigg[& \left(\sum_{n=1}^{N_r} a_{nr} \frac{-C.f_k.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right) .\widehat{V}_{kt}^{\beta-2}.V_{kt} \\ & + \left(\sum_{n=1}^{N_r} a_{nr} \frac{-C.f_{nr}.g'_\tau(f - f_{nr})}{f_k - f_{nr}}.H_{rt} \right) .\widehat{V}_{kt}^{\beta-1} \Bigg], \end{aligned} \tag{3.15}$$

$$\begin{aligned} P_0^{Inh}(F_{0r}) = \sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} \Bigg[& \left(\sum_{n=1}^{N_r} a_{nr} \frac{-D.f_k.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right) .\widehat{V}_{kt}^{\beta-1} \\ & + \left(\sum_{n=1}^{N_r} a_{nr} \frac{-D.f_{nr}.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right) .\widehat{V}_{kt}^{\beta-2}.V_{kt} \Bigg], \end{aligned} \tag{3.16}$$

$$Q_0^{Inh}(F_{0r}) = \sum_{k\in\mathcal{K}_r}\sum_{t=1}^{T}\left[\left(\sum_{n=1}^{N_r} a_{nr}\frac{-D.f_k.g'_\tau(f_k-f_{nr})}{f_k-f_{nr}}.H_{rt}\right).\widehat{V}_{kt}^{\beta-2}.V_{kt} + \left(\sum_{n=1}^{N_r} a_{nr}\frac{-D.f_{nr}.g'_\tau(f-f_{nr})}{f_k-f_{nr}}.H_{rt}\right).\widehat{V}_{kt}^{\beta-1}\right], \tag{3.17}$$

with $f_{nr} = nF_{0r}\sqrt{1+B_rn^2}$, $C = \partial f_{nr}/\partial B_r = n^3F_{0r}/2\sqrt{1+B_rn^2}$ and $D = \partial f_{nr}/\partial F_{0r} = n\sqrt{1+B_rn^2}$. The set of frequency-bins for which the model spectrum $W_{kr}^{\theta_r^{Inh}}$ is defined is here denoted by $\mathcal{K}_r = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau], n \in [1, N_r]\}$.
One can note that an exponent $\gamma > 0$ has been included in the update of B_r, in Equation (3.12). This parameter, whose role is similar to the step size in usual gradient descents, allows for the control of the convergence rate of the parameter B_r [Bertin et al., 2009]. We empirically found that setting $\gamma = 10$ was leading to accelerated updates while preserving the decrease of the cost-function.

• *Inharmonic relaxed / InhR-NMF:*

For the inharmonic relaxed model, the following update rules are applied. Note that for F_{0r}, an exact analytic solution is obtained when canceling the partial derivative of the cost-function C_1 (Equation (3.7)).

$$f_{nr} \overset{InhR}{\leftarrow} f_{nr}\cdot\frac{Q_0^{InhR}(f_{nr})+\lambda\cdot Q_1^{InhR}(f_{nr})}{P_0^{InhR}(f_{nr})+\lambda\cdot P_1^{InhR}(f_{nr})}, \tag{3.18}$$

$$B_r \overset{InhR}{\leftarrow} B_r\cdot\frac{Q_1^{InhR}(B_r)}{P_1^{InhR}(B_r)}, \tag{3.19}$$

$$F_{0r} \overset{InhR}{=} \frac{\sum_{n=1}^{N_r} f_{nr}n\sqrt{1+B_rn^2}}{\sum_{n=1}^{N_r} n^2(1+B_rn^2)}, \tag{3.20}$$

where

$$P_0^{InhR}(f_{nr}) = \sum_{k\in\mathcal{K}_{nr}}\sum_{t=1}^{T}\left[\left(a_{nr}\frac{-f_k.g'_\tau(f_k-f_{nr})}{f_k-f_{nr}}.H_{rt}\right).\widehat{V}_{kt}^{\beta-1} + \left(a_{nr}\frac{-f_{nr}.g'_\tau(f_k-f_{nr})}{f_k-f_{nr}}.H_{rt}\right).\widehat{V}_{kt}^{\beta-2}.V_{kt}\right], \tag{3.21}$$

$$Q_0^{InhR}(f_{nr}) = \sum_{k\in\mathcal{K}_{nr}}\sum_{t=1}^{T}\left[\left(a_{nr}\frac{-f_k.g'_\tau(f_k-f_{nr})}{f_k-f_{nr}}.H_{rt}\right).\widehat{V}_{kt}^{\beta-2}.V_{kt} + \left(a_{nr}\frac{-f_{nr}.g'_\tau(f-f_{nr})}{f_k-f_{nr}}.H_{rt}\right).\widehat{V}_{kt}^{\beta-1}\right], \tag{3.22}$$

$$P_1^{InhR}(f_{nr}) = K_\tau T\cdot 2f_{nr}, \tag{3.23}$$

$$Q_1^{InhR}(f_{nr}) = K_\tau T\cdot 2nF_{0r}\sqrt{1+B_rn^2}, \tag{3.24}$$

$$P_1^{InhR}(B_r) = F_{0r} \sum_{n=1}^{N_r} n^4, \tag{3.25}$$

$$Q_1^{InhR}(B_r) = \sum_{n=1}^{N_r} \frac{n^3 f_{nr}}{\sqrt{1 + B_r n^2}}. \tag{3.26}$$

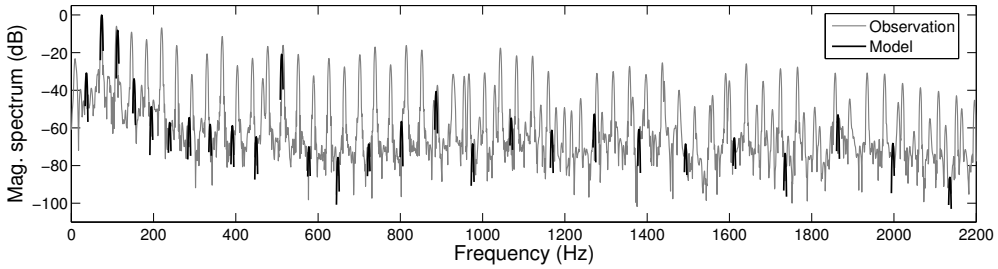
3.1.2.2 Practical considerations

Choice of the reconstruction cost-function: As seen in Section 2.2, high rank partials of piano tones should be taken into account when targeting the precise estimation of (B, F_0). Besides the tuning of N_r, the number of partials for a note of the model, the choice of the reconstruction cost-function may have an influence on the precision of the estimation. For the experiments presented in this thesis a Kullback-Leibler divergence ($\beta = 1$) has been considered. This choice corresponds to a trade-off between the Euclidian distance ($\beta = 2$), for which the estimation would mainly rely on low rank partials - these having usually the greatest amplitudes in the spectrum of a note and whose frequencies may be affected by the bridge coupling - and the Itakura-Saito divergence ($\beta = 0$) that would consider every partial regardless of its amplitude, even high rank partials that may be drowned out in noise.

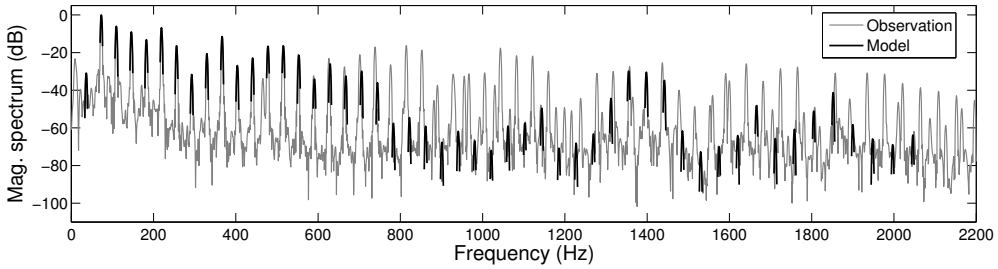
Initialization of the parameters: A good initialization of the optimization algorithm corresponds to the overlap of an important number of partials between the model and the data. As shown in Figure 3.1(a), a naive initialization (F_0 set to Equal Temperament and B to a common value for every note) may lead to the overlap of only a few first partials and cause the algorithm to stop in a local minimum.

In order to avoid these situations, special care is taken to initialize (B_r, F_{0r}). This latter is performed using a model for the inharmonicity and the tuning of pianos along the whole compass, with typical values of the parameters. This model, presented in Chapter 4, is based on generic rules in design and tuning. One can see on Figure 3.1(b) that such initialization leads to the overlap of a greater number of partials (20 for this example). However, it can be noticed around 1400 Hz that partials of the model are overlapping with some partials of the data with different ranks. This situation may correspond to a local minimum of the reconstruction cost-function.

Thus, it is chosen for the algorithm to initialize the spectra of the model with a few partials (it is empirically chosen that N_r^{ini} decreases linearly from 10 for lowest notes to 2 for highest, as displayed with gray '+' markers on Figure 3.2). The number of partials is then iteratively increased after each update of (B_r, F_{0r}) until the maximum number of partials N_r^{fin} is reached (*cf.* black '+' markers on Figure 3.2, where for each note of the model the maximum number of partials below the Nyquist frequency limit is set to 50). This heuristic should broaden the zone where the cost-function is convex in order to iteratively converge toward optimal values of the parameters (*cf.* Figure 3.3). It has been experimentally chosen for the following experiments to add one partial for every note of the model every 3 iterations of the optimization loop.



(a)



(b)

Figure 3.1: Initialization of the NMF models for the analysis of a note $D1$. (a) Rough initialization with F_0 set to Equal Temperament and $B = 5.10^{-3}$. (b) Initialization using the model of inharmonicity and tuning along the whole compass of pianos.

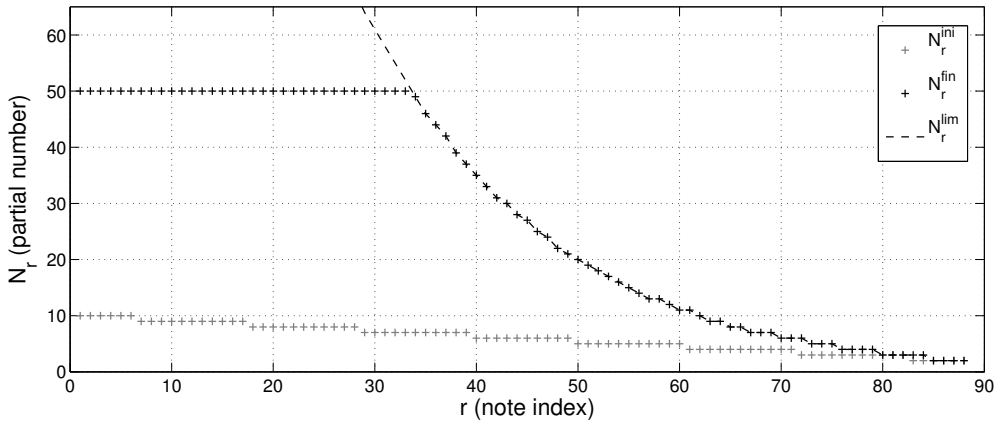


Figure 3.2: Number of partials N_r for the 88 notes of the models ($r \in [1, 88]$, from A0 to C8) at the initialization (gray markers labeled as N_r^{ini}) and at the end of the algorithm (black markers labeled as N_r^{fin}, here depicted for a choice of 50 partials maximum). The black dashed curve labeled as N_r^{lim} corresponds to the maximum number of partials that should be present below the Nyquist frequency (with a sampling frequency of 22050 Hz). For notes in the bass range this limits, not displayed here, goes up to around 180 partials.

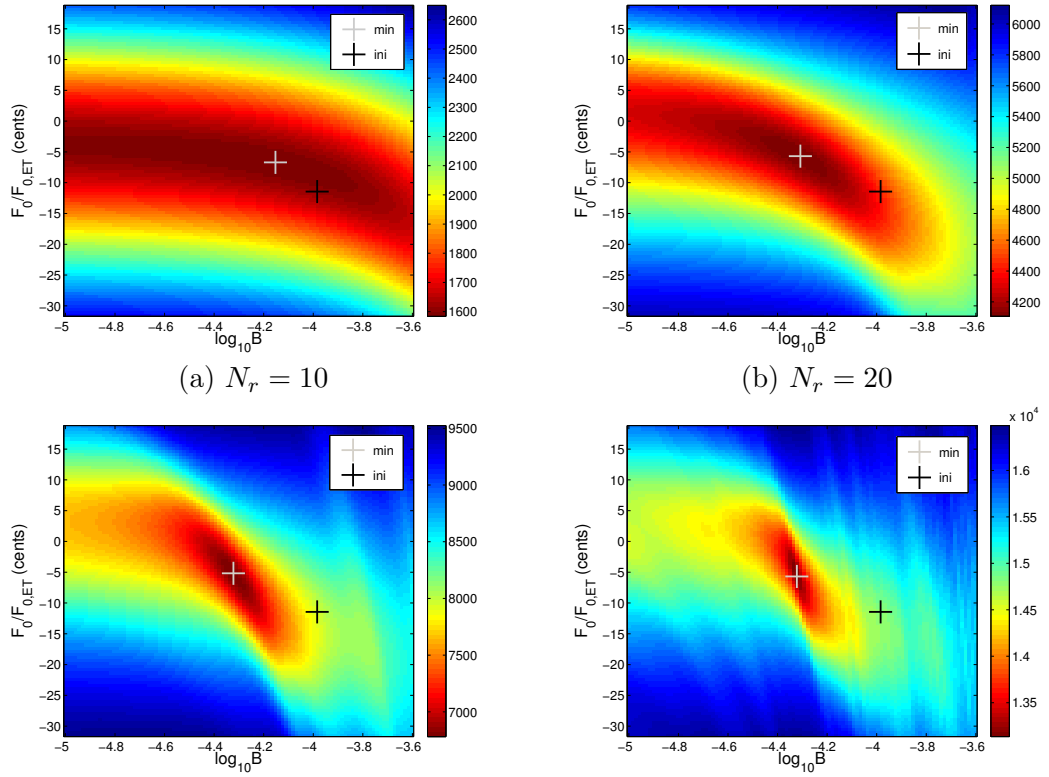
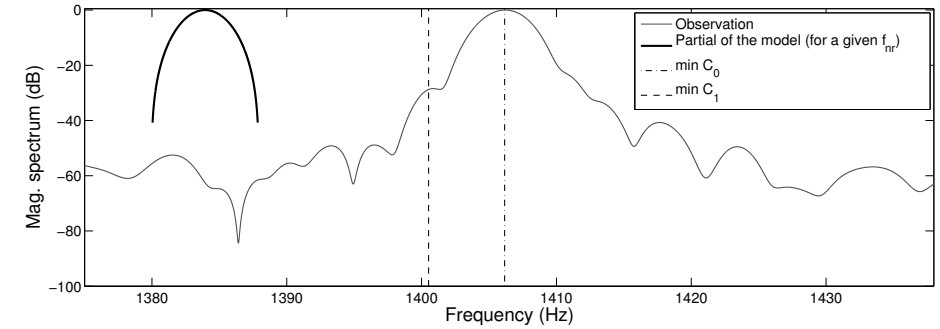
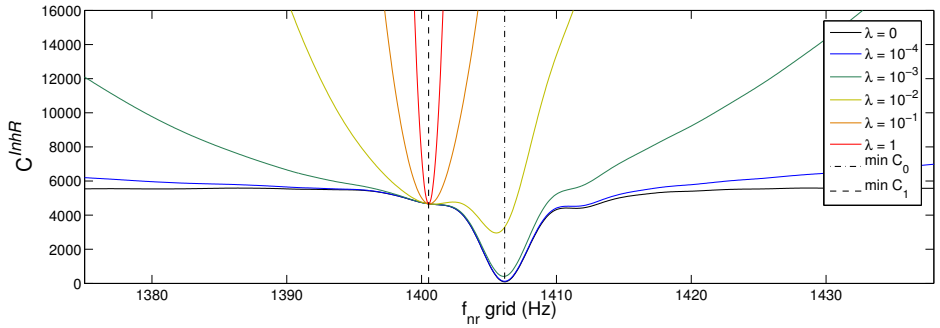


Figure 3.3: Cost-function for *Inh-NMF* (Eq. (3.5) with $\beta = 1$) computed on a grid of (B_r, F_{0r}), and different numbers of partials N_r, having amplitudes a_{nr} fixed to 1, for the analysis of the spectrum of a note $D1$. Black and gray markers respectively correspond to the initialization and the minimum of the cost-function.

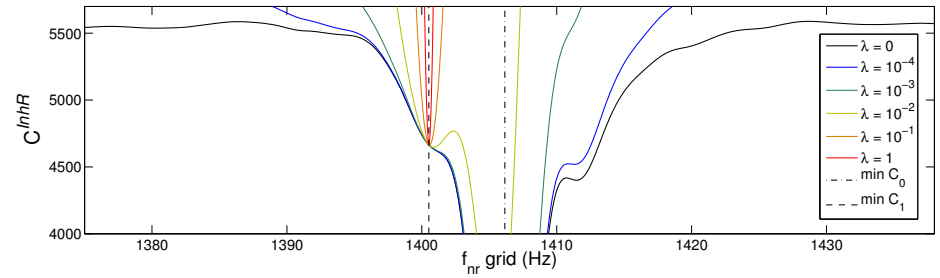
Discussion about the tuning of the regularization parameter for *InhR-NMF*: As mentioned in Section 3.1.1.2, the tuning of the regularization parameter λ influences the weighting between the reconstruction cost-function C_0 and the inharmonicity constraint C_1. A too high value of λ will favor a good fit of the partial frequencies with the inharmonicity relation without considering the slight deviations, *e.g.* due to the soundboard-string couplings, while a too small value will lead to an optimal reconstruction of the spectrum, even if the partials of the model do not match with transverse vibration partials. The influence of the tuning of λ for a given partial is illustrated on Figure 3.4. For this particular example, the optimal values of f_{nr} are different for C_0 and C_1 cost-functions (these are depicted, respectively as dash-dotted and dashed vertical lines). While increasing λ from 0 to 1, one can see on Figure 3.4(b) that the global minimum of C^{InhR} is changing from the optimal value of C_0 to the one of C_1. In this example, the difference between the values of these two optima is coming from the fact that the partial frequency actually deviates from the inharmonicity relation. In this case, a value of $\lambda \in [10^{-4}, 10^{-3}]$ should allow a good match of the partial of transverse vibrations with the observation (*i.e* minimize C_0), while avoiding the local minima of C_0 (as can be seen on Figure 3.4(c) around 1380 or 1412 Hz for $\lambda = 0$).



(a)



(b)



(c)

Figure 3.4: Influence of the tuning of λ on the cost-function of *InhR-NMF*. (a) Partial of transverse vibration of a spectrum (gray) and partial of the model (black) for a given f_{nr} and $a_{nr} = 1$. (b) Corresponding cost-function C^{InhR} (Eq. (3.6)) computed on a grid of f_{nr} for a_{nr} fixed to 1 and different values of λ. (c) Zoom along the y-axis of sub-figure (b). Vertical dashed-dot and dashed line bars correspond to the value of f_{nr}, respectively, minimizing C_0 and C_1.

Since we use a Kullback-Leibler divergence for the reconstruction cost-function C_0, one should note that the tuning of λ may be influenced by the amplitude of the partials of the observations. Indeed, the Kullback-Leibler divergence is not scale invariant and $d_{\mathrm{KL}}(\gamma x \mid \gamma y) = \gamma\ d_{\mathrm{KL}}(x \mid y)$, $\gamma \in \mathbb{R}^+$. In order to avoid a fine tuning of λ that depends on the recording conditions (piano type, nuances, ...) each spectrum composing the observation matrix is normalized to a maximal magnitude of 1. Thus, the illustrations of Figure 3.4 correspond to the tuning of λ for a partial with maximal amplitude. When considering a partial with amplitude 0.1 (-20 dB), a similar network of curves of Figure 3.4(b) will be obtained by multiplying every value of λ by 10. For the experiments that are presented in Section 3.1.3, λ has been set empirically $\lambda = 5 \cdot 10^{-4}$.

Dealing with noise: In theory, when analyzing piano tones, one should consider partials in the whole frequency range. Thus, the number of partials having frequencies below the Nyquist frequency $F_s/2$ (F_s being the sampling frequency) should be given, for each note, by:

$$N_r^{\text{lim}} = \left\lfloor \frac{F_s}{2F_{0r}} \sqrt{\frac{2}{1+\sqrt{1+B_r \frac{F_s^2}{F_{0r}^2}}}} \right\rfloor, \tag{3.27}$$

where $\lfloor . \rfloor$ denotes the integer rounding towards $-\infty$ (as depicted in black line on Figure 3.2). However in practice, some partials, mainly with high ranks, may be missing or drowned out in noise because of their strong damping. When targeting the precise estimation of (B, F_0), taking into account these partials in the model may lead to bad estimates. Thus, for each iteration of the optimization algorithm, we cancel their influence in the estimation of (B_r, F_{0r}) by removing them from the corresponding cost-functions (Eq. (3.5) for *Inh-NMF* and Eq. (3.7) for *InhR-NMF*). For the proposed application, we compute, during a pre-processing step, the noise level $\text{NL}(f_k)$ on each magnitude spectrum composing the matrix V (see Appendix B), and at each iteration we look for the estimated partials that have a magnitude greater than the noise. Thus, we define the set of reliable partials of each note, being above the noise level, by $\Delta_r = \{n \mid a_{nr} > \text{NL}(f_{nr}),\ n \in [1, N_r]\}$. This information is taken into account by replacing the sums over the entire set of partials $\sum_{n=1}^{N_r}$ by sums over the reliable set of partials $\sum_{n \in \Delta_r}$ in the update rules of B_r (Eq. (3.14)-(3.15) for *Inh-NMF*, and Eq. (3.25)-(3.26)) for *InhR-NMF*) and F_{0r} (Eq. (3.16)-(3.17) for *Inh-NMF* and Eq. (3.20) for *InhR-NMF*).

3.1.2.3 Algorithms

Finally, the steps of *Inh-NMF* and *InhR-NMF* algorithms are summarized, respectively in tables Algorithm 1 and Algorithm 2. Note that the number of iterations for each parameter has been determined empirically.

Algorithm 1 *Inh-NMF* for the estimation of (B, F_0)

Input:
V set of magnitude spectra (each normalized to a max. of 1)
H filled with 0 and 1
$\beta = 1$

Pre-processing:
for each column of V compute NL(f_k) the noise level (*cf.* App. B)

Initialization:
(B_r, F_{0r}), $\forall\ r \in [1, R]$ according to the model of Sec. 4.4.3
N_r^{ini} and N_r^{fin}, $\forall\ r \in [1, R]$ as shown in Fig. 3.2
$f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}$, $a_{nr} = 1$, $\forall\ r \in [1, R]$, $n \in [1, N_r^{\text{ini}}]$
W^θ computation (*cf.* Eq. ((3.1))

Optimization:
for $it = 1$ to It **do**
  if mod$(it, 3) = 0$ **then**
    $N_r \leftarrow N_r + 1$, $\forall r \in [1, R]$ provided $N_r < N_r^{\text{fin}}$
  end if
  • a_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.8))
    W^θ update (Eq. (3.1))
    deduce Δ_r by comparing a_{nr} with NL(f_{nr})
  for $u = 1$ to 10 **do**
    • F_{0r} update $\forall r$, $n \in \Delta_r$ (*cf.* Eq. (3.13))
      W^θ update (*cf.* Eq. ((3.1))
    • B_r update $\forall r$, $n \in \Delta_r$ (*cf.* Eq. (3.12))
      W^θ update (*cf.* Eq. ((3.1))
  end for
end for

Output: B_r, F_{0r}, a_{nr}

Algorithm 2 *InhR-NMF* for the estimation of (B, F_0)

Input:
V set of magnitude spectra (each normalized to a max. of 1)
H filled with 0 and 1
$\beta = 1$ / $\lambda = 5 \cdot 10^{-4}$

Pre-processing:
for each column of V compute NL(f_k) the noise level (*cf.* App. B)

Initialization:
(B_r, F_{0r}), $\forall\ r \in [1, R]$ according to the model of Sec. 4.4.3
N_r^{ini} and N_r^{fin}, $\forall\ r \in [1, R]$ as shown in Fig. 3.2
$f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}$, $a_{nr} = 1$, $\forall\ r \in [1, R]$, $n \in [1, N_r^{\text{ini}}]$
W^θ computation (*cf.* Eq. ((3.1))

Optimization:
for $it = 1$ to It **do**
  if mod$(it, 3) = 0$ **then**
    $N_r \leftarrow N_r + 1$, $\forall r \in [1, R]$ provided $N_r < N_r^{\text{fin}}$
  end if
  • a_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.8))
    W^θ update (Eq. (3.1))
    deduce Δ_r by comparing a_{nr} with NL(f_{nr})
  • f_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.18))
    W^θ update (Eq. (3.1))
  for $v = 1$ to 30 **do**
    $\forall r$, $n \in \Delta_r$
    F_{0r} update (*cf.* Eq. (3.20))
    B_r update (20 times) (*cf.* Eq. (3.19))
  end for
end for

Output: B_r, F_{0r}, a_{nr}, f_{nr}

3.1.3 Results

The ability of both NMF models/algorithms to provide correct estimates of (B, F_0) on the whole piano compass is investigated here in a supervised context (*i.e.* the played notes and their time-activations are known). The estimation from isolated note and chord recordings are respectively presented in Sections 3.1.3.2 and 3.1.3.4.

3.1.3.1 Database presentation

The results presented in this section are obtained from 3 separate databases (Iowa [University of Iowa, 1997], RWC [Goto et al., 2003] and MAPS [Emiya et al., 2010b]) covering a total of 11 different pianos with different recording conditions (microphones close to the strings or in room ambient condition) and dynamics (*piano*, *mezzo-forte*, *forte*). The details for all pianos are given in Table 3.1. Note, that all piano synthesizers from MAPS database are using high-quality samples. For all pianos, isolated note recordings along the whole compass, from A0 (21 in MIDI index)[1] to C8 (108), are available. For MAPS pianos, chord recordings are also given. An additional dataset [Rauhala et al., 2007a] composed of synthesized isolated tones from A0 (21) to G3 (55) is used in Section 3.1.3.3 for the evaluation of the precision of B estimates.

	ref. name	piano type	rec. conditions
Iowa	Iowa	grand	close
RWC	RWC1	grand	close
	RWC2	grand	close
	RWC3	grand	close
MAPS	AkPnBcht	grand (synth.)	Software preset
	AkPnBsdf	grand (synth.)	Software preset
	AkPnCGdD	grand (synth.)	Software preset
	AkPnStgb	upright (synth.)	Software preset
	ENSTDkAm ENSTDkCl	upright	ambiant close
	SptkBGAm SptkBGCl	grand (synth.)	ambiant close
	StbgTGd2	grand (synth.)	Software preset

Table 3.1: Details of the databases.

3.1.3.2 Isolated note analysis

The dataset for each piano consists of isolated note recordings of the 88 notes composing the compass. Each recording is here first down-sampled to $F_s = 22050$ Hz. In order to obtain a sufficient spectral resolution for notes in the bass range, the observation spectra are extracted from 300 ms Hann windows, applied to the decay part of the sounds. Then, the matrix V is built by concatenating the 88 spectra (each column corresponding to the magnitude spectrum of a note, from A0 (21) to C8 (108)) and H is fixed to the identity matrix (*cf.* Figure 3.5). For each note, the number of partials N_r is set to $\arg\min_{N_r}(50, f_{N_r,r} < F_s/2)$, as depicted in black '+' markers on Figure 3.2.

[1] In the following, each note is given with its MIDI note number in brackets. More details on the MIDI norm are given in Appendix A.

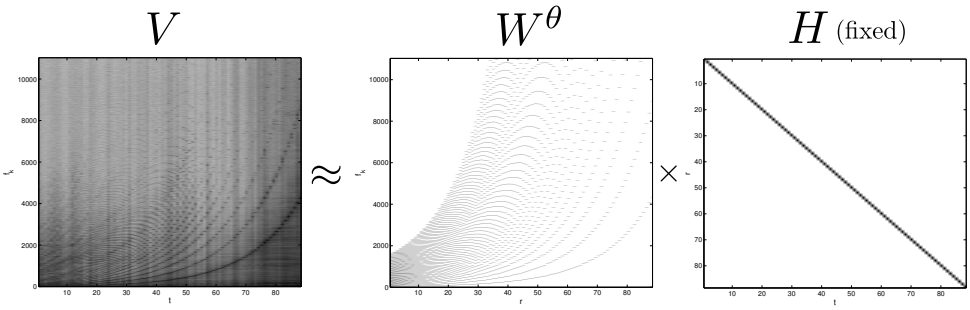


Figure 3.5: Scheme for the NMF-based estimation of (B, F_0) along the whole compass of pianos from isolated note recordings.

It is worth noting here that the experiments have been performed for three sets of dynamics, namely *forte*, *mezzo-forte* and *piano*. Only the results for the *mezzo-forte* dynamics are presented in the following since the performance did not appear significantly dependent. Also, MAPS datasets are composed of tones played with different configurations for the *forte* pedal: active or inactive (randomly distributed).

Two selected examples are presented in Figures 3.6 and 3.7 in order to exhibit characteristics of both algorithms and analyze particular cases that lead to a failure of the estimation. The results for the 11 pianos are given in Appendix D, page 125. Sub-figures (a) correspond to the inharmonicity curves along the compass. As discussed in Chapter 4, all piano models have a similar B behavior for the highest notes. Sub-figure (b) represent the curves of F_0 as the deviation from Equal Temperament (ET), in cents. The initialization of (B, F_0) is depicted as black dashed lines and the blue and red curves respectively correspond to the estimates obtained by *Inh-NMF* and *InhR-NMF* algorithms.

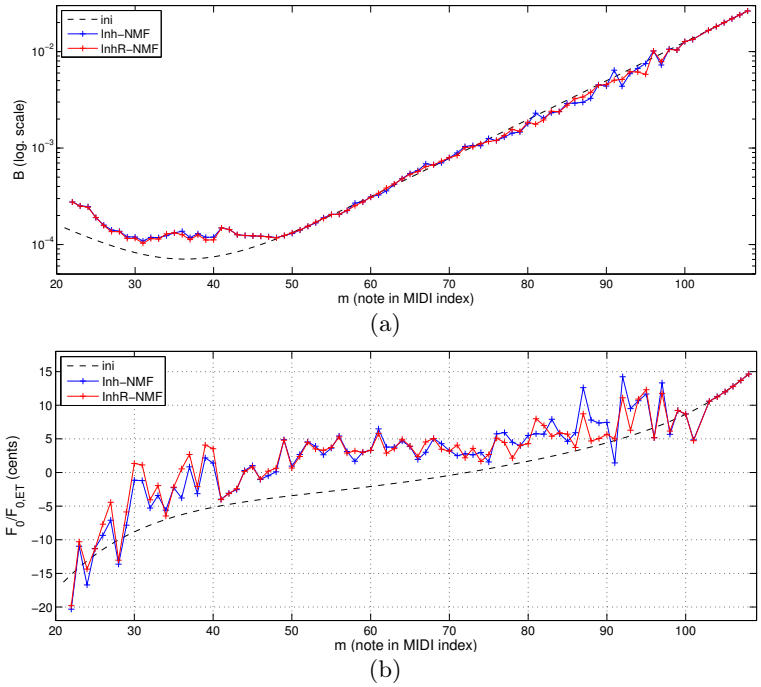


Figure 3.6: Iowa (a) B and (b) F_0 as dev. from ET along the whole compass.

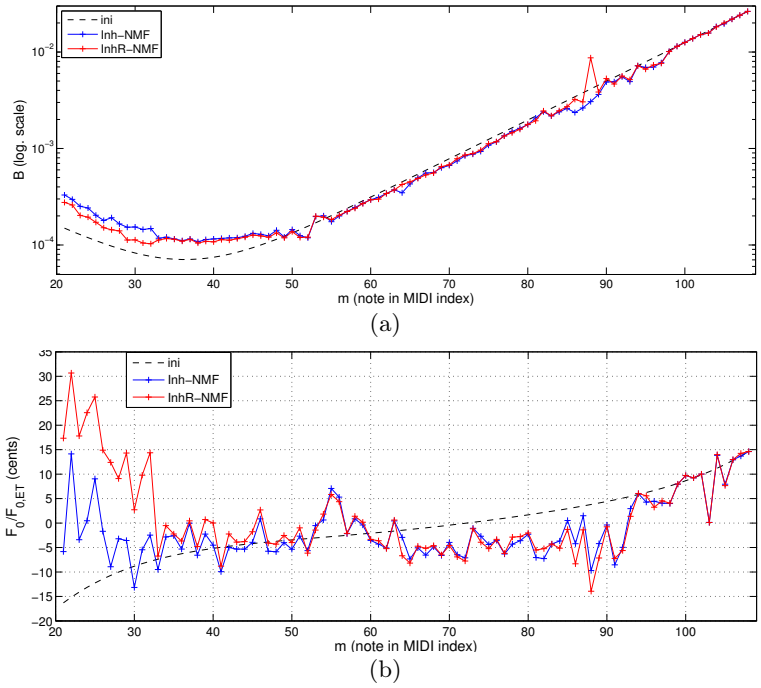
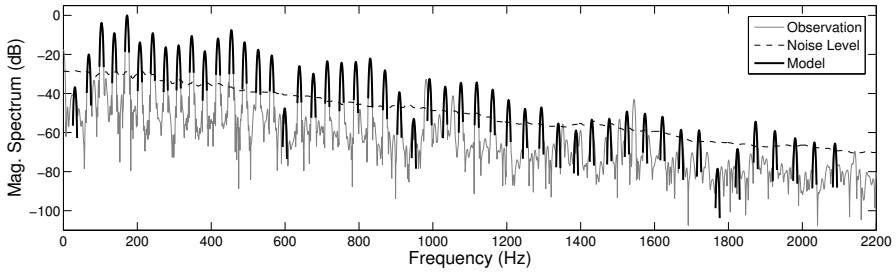


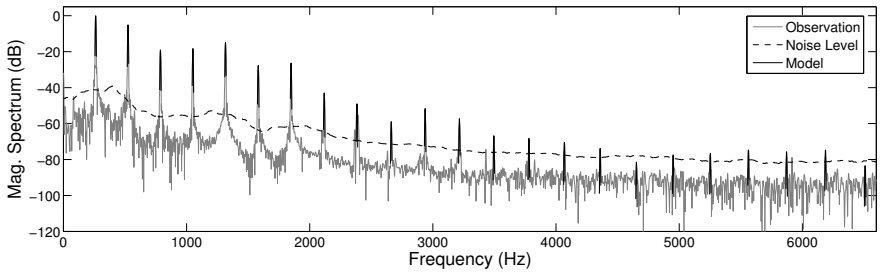
Figure 3.7: ENSTDkAm (a) B and (b) F_0 as dev. from ET along the whole compass.

When comparing the blue and red curves one can see that for most notes, particularly in the medium range, both estimates are in good agreement. In order to qualitatively evaluate the results one can compare the observed spectra and the model. For instance Figure 3.8 depicts the result of *InhR-NMF* algorithm for the analysis of notes C♯1 (25), C4 (60) and G♯5 (80) of Iowa grand piano. *Inh-NMF* exhibits similar results (not shown here).

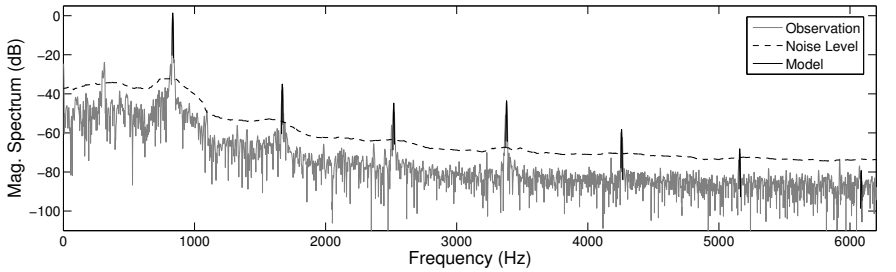
However, for some particular examples a significant difference between *Inh-NMF* and *InhR-NMF* estimates is visible, for instance in the bass range of ENSTDkAM piano between A0 (21) and A1 (33) on Figure 3.7. When comparing the observed spectrum and the models for the note F♯1 (*cf.* Figure 3.9), one can see that *Inh-NMF* failed to fit the partials of transverse vibrations from rank 27 (Figure 3.9(a)), while *InhR-NMF* correctly fitted the 50 partials (Figure 3.9(b)). Figure 3.9(c) shows that the failure of *Inh-NMF* is due to the fact that the partial frequencies significantly differ from the theoretical inharmonicity relation (2.23). The "+" markers correspond to a plot of the partial frequencies obtained by *InhR-NMF* in the graphic $(f_{nr}/n)^2$ as a function of n^2. If the partial frequencies were strictly following the inharmonicity relation they should be distributed according to a straight line of equation $F_{0r}^2(1 + B_r n^2)$. Thus, *InhR-NMF* here succeeds to adapt the parameters (B_r, F_{0r}) (red straight line) during the optimization while permitting the partials with low ranks to deviate from the theoretical law to fit the partials of the observation. On the contrary, the strict constraint of *Inh-NMF* led to the estimation of an inharmonicity relation fitting approximately the first 27 partials (blue straight line). It is worth noting for this very particular example that the distribution of the partial frequencies differs from the usual inharmonicity relation with deviations caused by the bridge coupling (as shown in Figure 2.10, page 31) and may be resulting from other effects such as for instance the wrapping of the strings in the bass range (*cf.* Section 4.2.2.1). In this case, the validity of the estimation of (B, F_0) is hardly quantifiable because the model of



(a)



(b)



(c)

Figure 3.8: Results of *InhR-NMF* for the analysis of notes (a) C♯1, (b) C4 and (c) G♯5 of Iowa grand piano.

transverse vibration is not sufficient to explain the distribution of the partial frequencies. However, a benefit of *InhR-NMF* is to handle these discrepancies so that a recovery of the complete set of partials is still possible.

Other differences between the results provided by both algorithms can be seen in the treble range. For instance, Figure 3.10 depicts the result of *InhR-NMF* for the note B6 (95) of ENSTDkAM piano. In this case, the validity of the estimation cannot be assessed so easily for some partials, mainly because these partials aggregate multiple peaks while the model assumes only a single component. This issue typically happens in the treble range, where the notes are associated with triplets of slightly detuned strings. Then, the algorithm selects one peak per group, that will depend on the initialization and on a balance between peak strength and model fitting. Thus, each algorithm might return slightly different biased estimates for (B, F_0), depending on which partials are selected.

Finally, in some cases corresponding mainly to notes in the high treble range, both algorithms may fail because of initialization issues. This can be seen of Figure 3.6 where the estimates of (B, F_0) for notes above G7 (103) remain fixed to their initial values because no partials of the models were overlapping partials of the data at the initialization. For instance, a difference of 10 cents between the initialization and the actual value of F_0 of

a note G7 leads approximately to a deviation of 18 Hz between the first partial frequency of the observation and the model. In order that the main lobes of the first partial of the observation and the model overlap at the initialization, the limit deviation between the partial frequencies is equal to $4/\tau$ for a Hann window. For the presented results, a 300 ms Hann window was used so this limit is approximately equal to 13 Hz. Under these conditions, it is very likely that no partial of the model overlaps partials of the observed spectrum at the initialization, thus leading to a failure of both algorithms. Similar results may be observed for AkPnBcht and AkPnBsdf in Appendix D. A possible solution to overcome these limitations is to perform a rough estimation by means of a grid search before running the optimization. Also, using a shorter analysis window may help in increasing the overlapping between the observed and modeled partials (for instance a 50 ms Hann window will lead to a limit of 80 Hz).

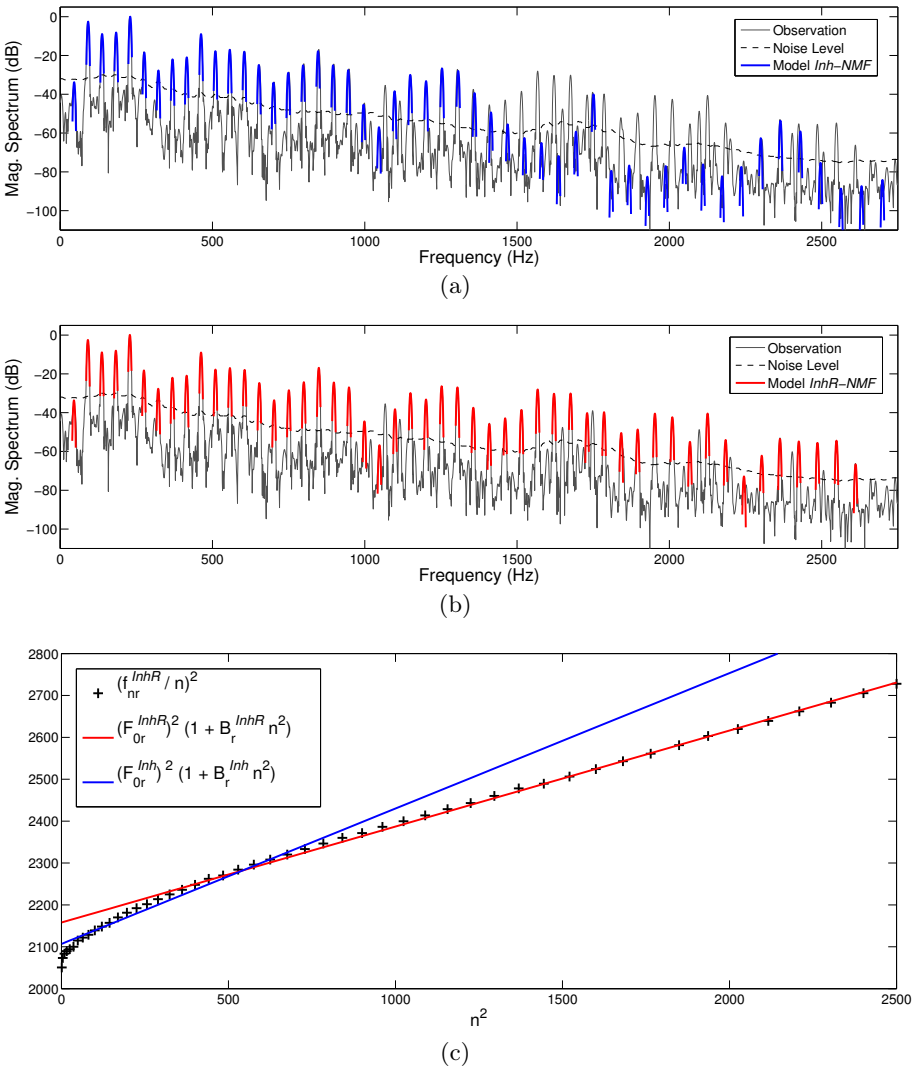


Figure 3.9: Estimation of (B, F_0) for the note F♯1 (30) of MAPS ENSTDkAM piano. Observed spectrum and estimated spectrum by (a) *Inh-NMF* and (b) *InhR-NMF*. (c) Plot $(f_{nr}/n)^2$ as a function of n^2, with the inharmonicity relation estimated by *Inh-NMF* and *InhR-NMF*, respectively depicted as blue and red lines.

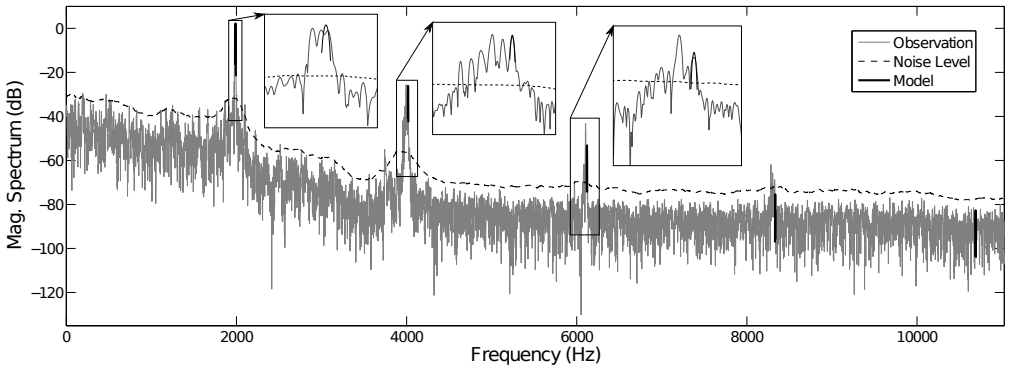


Figure 3.10: Results of *InhR-NMF* for the note B6 (95) of ENSTDkAM piano.

3.1.3.3 Performance evaluation

This section presents the quantitative evaluation of the precision of the estimation of (B, F_0) provided by both algorithms and a comparison with the state-of-the-art *PFD* (Partial Frequencies Deviation) algorithm [Rauhala et al., 2007a; Rauhala and Välimäki, 2007].

Reference extraction: The evaluation is performed on both synthetic and Iowa real piano tones on the limited range A0 (21) - G3 (55) (as done in [Rauhala et al., 2007a]) because of the lack of ground truth. Indeed, the string dimensions and properties are not known for real pianos so the reference values of B need to be manually retrieved from the note spectra. This requires a supervised extraction of the partial frequencies corresponding to transverse vibrations of the strings, which can be quite difficult because of the presence of many partials due to the string couplings, particularly for notes composed of multiple strings. In our experiments, the partials corresponding to transverse vibrations of the strings have been picked, up to a rank of 30, from the spectra of notes played with *piano* dynamics so that the influence of the couplings should be limited. In order to have a precise spectral resolution, the spectra have been computed from 2 seconds of decaying sound on 2^{17} frequency bins. In case multiple partials are produced by the coupling of doublet/triplet of strings, several peaks have been considered (as illustrated on Figure 3.11(a)). In the following equation, the frequencies of these peaks are denoted by $f^{\star}_{r,n,p}$, where r corresponds to the index of the note, $p \in [1, P_{r,n}]$ to the index of the peak within $n \in [1, N_r]$ the rank of the partial. Then, the reference parameters $(B^{\star}_r, F^{\star}_{0r})$ have been estimated by minimizing the absolute deviation (which should reduce the influence of potential outliers) between the inharmonicity relation and the extracted partial frequencies:

$$(B^{\star}_r, F^{\star}_{0r}) = \underset{(B_r, F_{0r})}{\operatorname{argmin}} \sum_{n=1}^{N_r} \sum_{p=1}^{P_{r,n}} \frac{1}{P_{r,n}} \mid f^{*}_{r,n,p} - nF_{0r}\sqrt{1 + B_r n^2} \mid . \tag{3.28}$$

The $1/P_{r,n}$ factor corresponds to a weighting that allows us to consider multiple partials as one theoretical partial in the regression. The result of the estimation of $(B^{\star}, F^{\star}_0)$ for G♯2 note of Iowa grand piano is depicted in Figure 3.11(b) by plotting $(f_n/n)^2$ as a function of n^2. Finally, the curves of $B^{\star}_r$ and $F^{\star}_{0r}$ (as deviation from ET) are presented for both datasets along the range A0 (21) - G3 (55) in Figure 3.12.

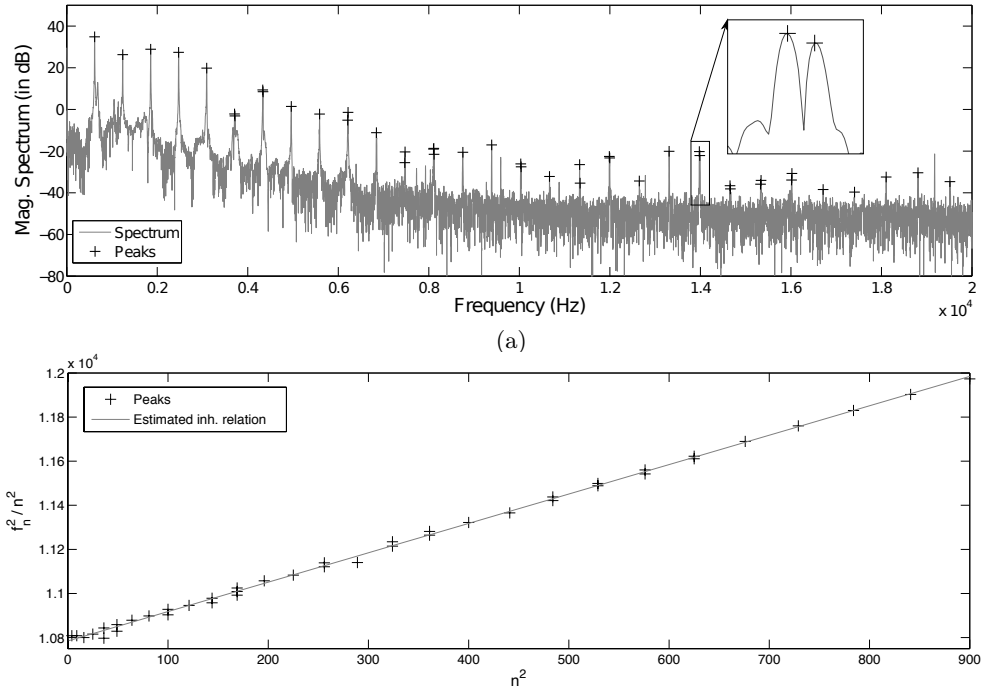
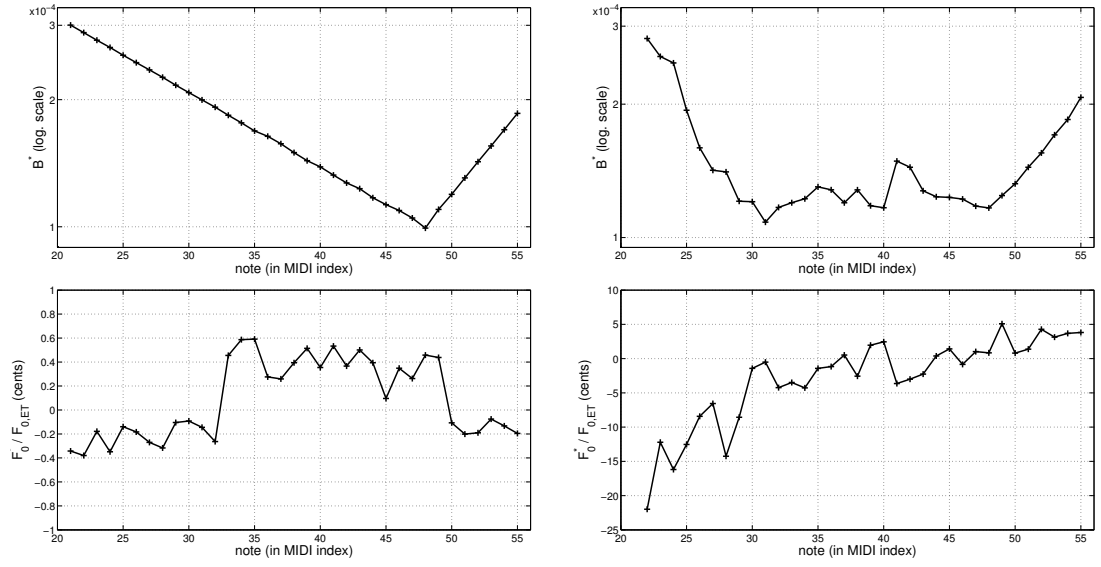


Figure 3.11: (B, F_0) reference extraction on note G♯2 of Iowa grand piano. (a) Magnitude spectrum (in gray), and selection of the peaks corresponding to transverse vibrations of the strings ('+' markers). (b) Results of the estimation of $(B^\star, F_0^\star)$ reference values in the plane $(f_n/n)^2$ as a function of n^2. '+' markers correspond to manually extracted peaks, and gray line to the estimated inharmonicity relation.



(a) Synthetic (b) Iowa

Figure 3.12: Reference curves $(B^\star, F_0^\star)$ along the range A0 (21) - G3 (55) for (a) the synthetic samples and (b) the Iowa piano tones.

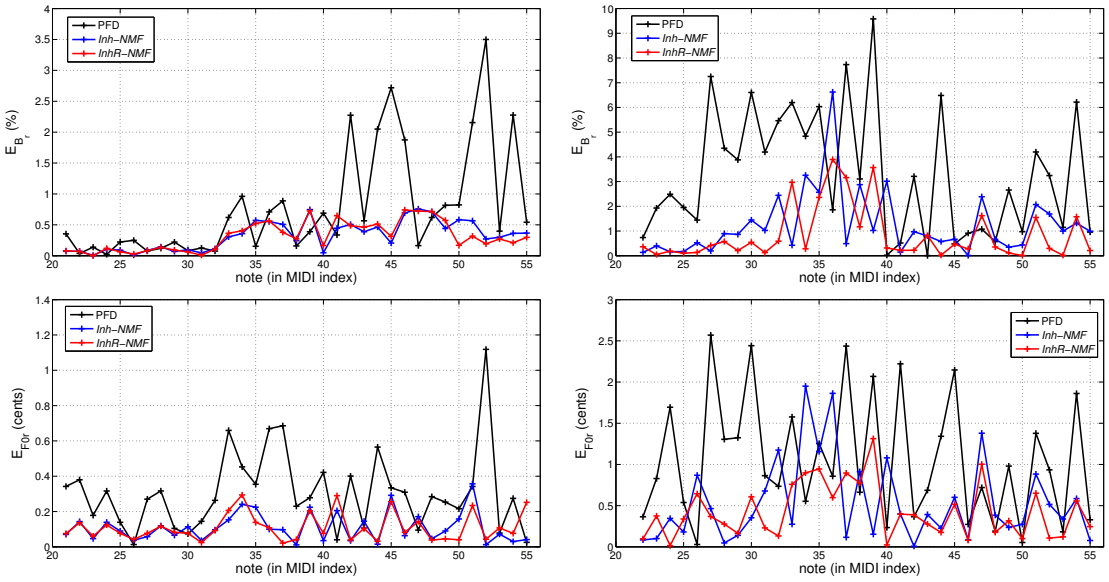
Evaluation results: The inharmonicity coefficient estimation performances are evaluated in terms of relative error with respect to the reference:

$$E_{B_r} = \left| \frac{B_r - B_r^\star}{B_r^\star} \right|. \tag{3.29}$$

F_0 estimates are evaluated in terms of deviation from the reference in cents:

$$E_{F_{0r}} = 1200 \cdot \left| \log_2 \frac{F_{0r}}{F_{0r}^\star} \right|. \tag{3.30}$$

These measures for *PFD* and the two NMF algorithms are presented on Figure 3.13 for the synthetic samples and the Iowa piano tones on the range A0 (21) - G3 (55). Table 3.2 returns these errors averaged for each piano and algorithm. Higher performances are obtained by both NMF algorithms when compared to the *PFD* algorithm. The performance obtained for the synthetic tones are similar for both NMF modelings while *InhR-NMF* returns slightly better results than *Inh-NMF* for the analysis of Iowa grand piano samples. These results on the first data set could be explained by the fact that the synthetic tones have been generated by using the theoretical inharmonicity relation (Equation (2.23)) without considering the partial frequency deviations caused by the soundboard-strings coupling. Results on the Iowa data set tend to show that it is worthwhile to take these possible deviations into account, as done by *InhR-NMF*.



(a) Synthetic (b) Iowa

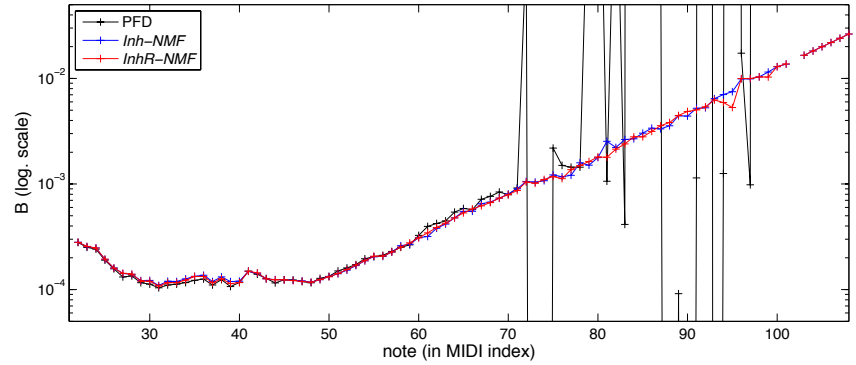
Figure 3.13: E_{B_r} (above, in %) and $E_{F_{0r}}$ (below, in cents) along the range A0 (21) -G3 (55) for *PFD* (black), *Inh-NMF* (blue) and *InhR-NMF* (red) algorithms. The evaluation is performed on (a) synthetic and (b) Iowa piano datasets.

On G♯3 (56) - C8 (108) results are not quantified because of the lack of ground truth and data for the synthetic signals. However, it can be observed graphically (*c.f.* Figure 3.14) that NMF estimates seem more consistent with typical values than those obtained by using *PFD* (not optimized there). It is worth noting that in the presented experiments,

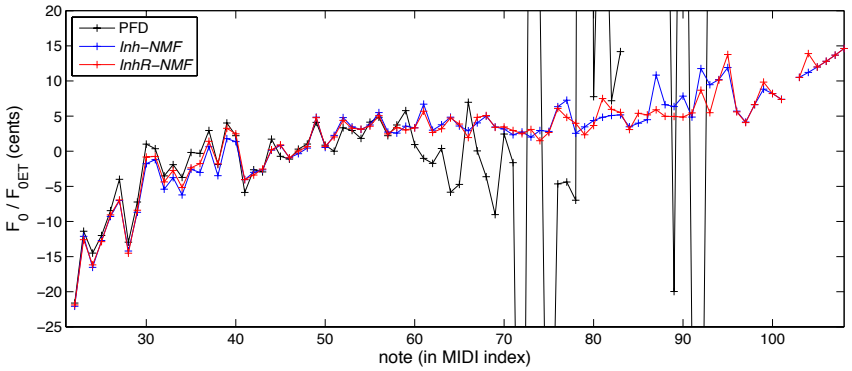
	PFD	*Inh-NMF*	*InhR-NMF*	
Synthetic	0.783	0.323	0.311	E_{B_r} (%)
	0.307	0.110	0.110	$E_{F_{0r}}$ (cents)
Iowa	3.30	1.25	0.847	E_{B_r} (%)
	1.06	0.539	0.429	$E_{F_{0r}}$ (cents)

Table 3.2: (B, F_0) estimation errors averaged on the range A0-G3, for *PFD* and NMF algorithms on synthetic and Iowa piano tones.

F_{0r} was initialized to Equal Temperament for the *PFD* algorithm (proposed as an optional input in the code). As an additional study, in order to investigate the influence of the initialization, we modified the *PFD* code so that it can take into account the same initialization of (B_r, F_{0r}) as the one we used for the NMF algorithms. This did not improve significantly the results in the high pitch range.



(a)



(b)

Figure 3.14: (a) B and (b) F_0 along the whole compass of Iowa grand piano for *PFD* (black), *Inh-NMF* (blue) and *InhR-NMF* (red) algorithms

3.1.3.4 Chord analysis

The same protocol has been applied to the analysis of 4 chords (from MAPS SptkBGCl grand piano synthesizer), respectively taken in the extreme bass, bass, middle and treble range of the compass. Each chord is composed of 5 notes. In order to have a sufficient spectral resolution the analysis window length was set to 1 second for the chords played

in the extreme bass/bass ranges and 500 ms in the medium/treble ranges. On Figure 3.15 and 3.16, the results of (B, F_0) estimates obtained respectively by *Inh-NMF* and *InhR-NMF* from isolated notes (in thin gray lines) are compared with the ones obtained from chords (one type of marker for each chord). The initialization is drawn as a dashed line. It can be observed for each algorithm that both types of estimations lead to remarkably similar results. The slight deviations in the estimation from chord recordings could be explained by the overlapping of the partials belonging to different notes, that could corrupt the estimation of the frequencies. Moreover, it has been shown in the previous section on isolated note analysis that, in the treble range, the precise estimation of (B, F_0) cannot be always guaranteed since the model of inharmonicity with one frequency peak per partial, as given by Equation (2.23), is not sufficient to explain the spectrum of the notes. The estimated spectrum by *InhR-NMF* for the chords #2 and #3 are respectively given on Figures 3.17(a) and 3.17(b), where one can see that, despite a considerable spectral overlap between the notes, the partials are well identified for every note.

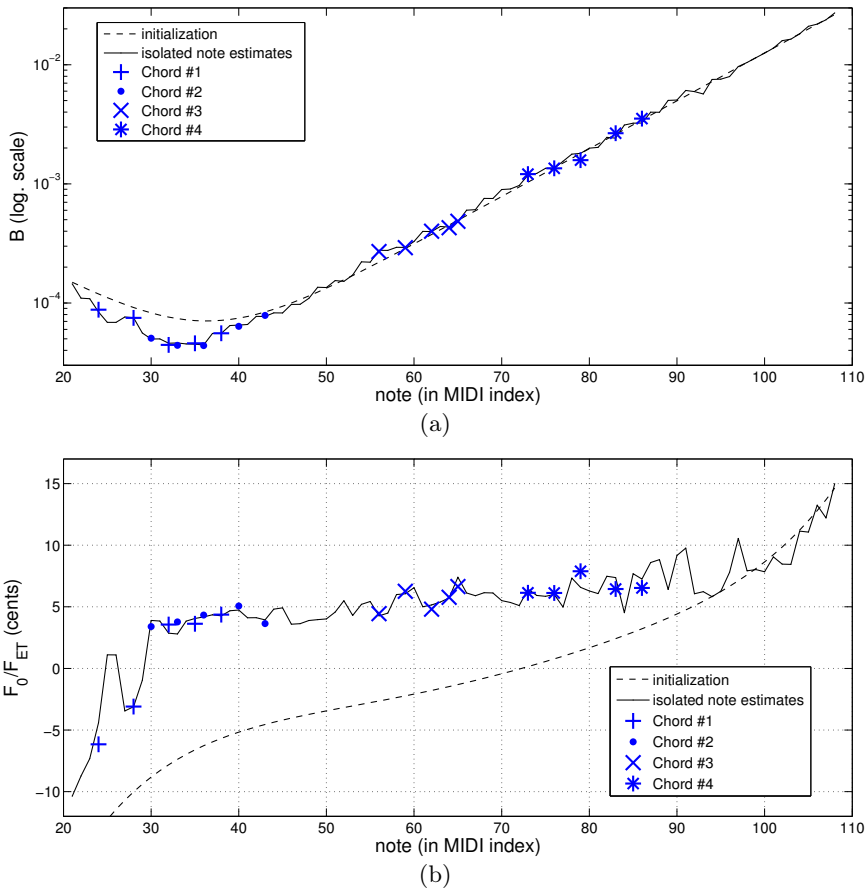
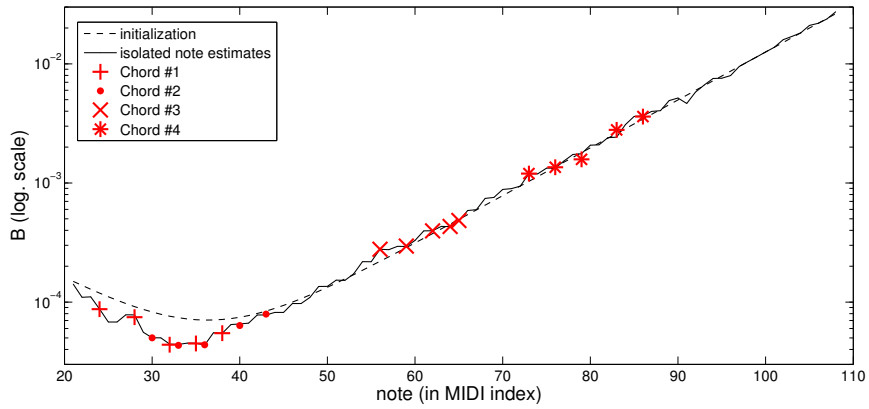
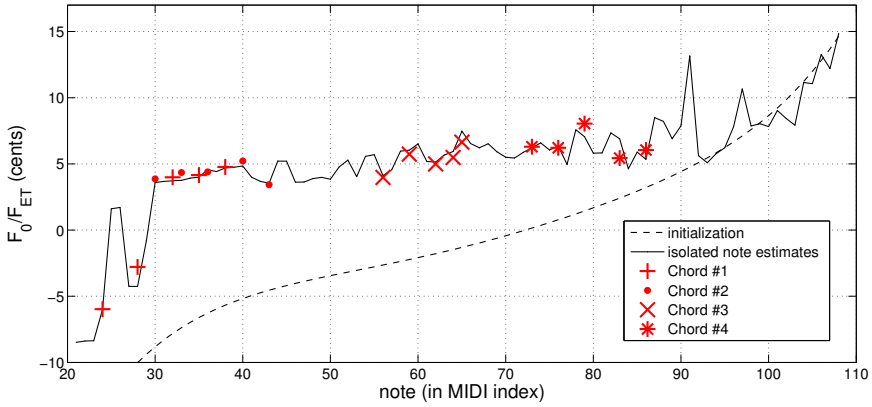


Figure 3.15: *Inh-NMF* isolated note vs. chord analysis for the MAPS SptkBGCl grand piano. (a) Inharmonicity coefficient and (b) F_0 as dev. from ET along the compass.

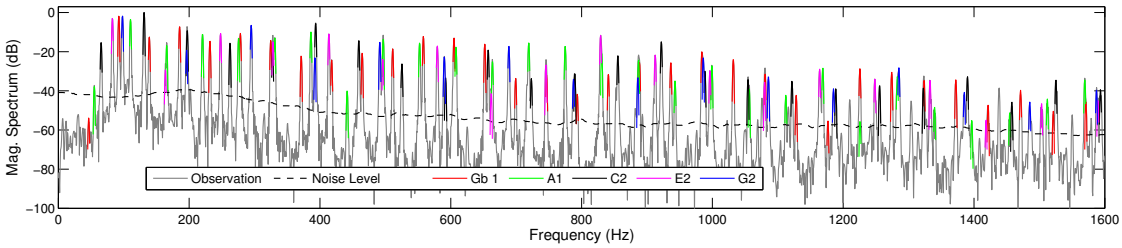


(a)

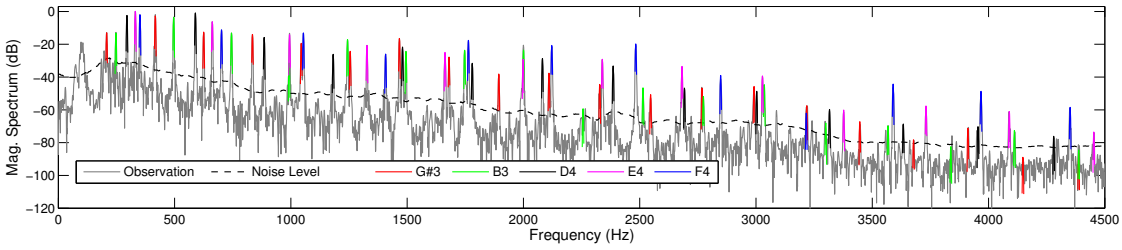


(b)

Figure 3.16: *InhR-NMF* isolated note vs. chord analysis for the MAPS SptkBGCl grand piano. (a) Inharmonicity coefficient and (b) F_0 as dev. from ET along the compass.



(a)



(b)

Figure 3.17: Result of *InhR-NMF* for the analysis of (a) the chord #2 (G♭1-A1-C2-E2-G2) and (b) the chord #3 (G♯3-B3-D4-E4-F4) of MAPS SptkBGCl piano.

3.1.3.5 Conclusion

Two NMF models based on a parameterization of the dictionary for the estimation of (B, F_0) have been presented so far in this chapter. The first model, denoted by *Inh-NMF*, forces the partial frequencies to strictly follow the theoretical inharmonicity relation, while the second model, denoted by *InhR-NMF*, relaxes this constraint by means of a weighted penalty term added to the reconstruction cost-function. An algorithm is derived for each model and special care is taken in the initialization and the optimization, in order to avoid, as much as possible, the convergence toward local optima. Both algorithms have been successfully applied to the supervised estimation of (B, F_0) along the whole compass of different pianos from isolated note and chord recordings. Their performance evaluation compares favorably with the state of the art *PFD* method, while exhibiting a benefit of the relaxed constraint of *InhR-NMF* for the analysis of real piano tones when compared to *Inh-NMF*.

In Chapter 5, both models will be evaluated in a transcription task, *i.e.* in an unsupervised context, where the activation matrix H has to be estimated as well as (B, F_0) and the partial amplitudes. A comparison of the performance with a simpler harmonic model will be performed in order to evaluate whether it is valuable to take into account inharmonicity for such applications.

3.2 Probabilistic line spectrum modeling

This section introduces a probabilistic model for the analysis of line spectra – defined here as a set of frequencies of spectral peaks with significant energy. Most algorithms dedicated to audio applications (F_0-estimation, transcription, ...) consider the whole range of audible frequencies to perform their analysis, while besides attack transients, the energy of music signals is often contained in only a few frequency components, also called partials. Thus, in a time-frame of music signal only a few frequency-bins carry information relevant for the analysis. By reducing the set of observations, *i.e.* by keeping only the few most significant frequency components, it can be assumed that most signal analysis tasks may still be performed. For a given frame of signal, this reduced set of observations is here called a line spectrum, this appellation being usually defined for the discrete spectrum of electromagnetic radiations of a chemical element. In the case of piano music analysis, most observations should correspond to frequencies related to transverse vibration of the strings and should allow estimating (B, F_0) parameters of the played notes. Thus, it should be noticed that the proposed approach is closely related to the class of methods "Iterative peak-picking and refinement of the inharmonicity coefficient" presented in Section 2.3. The probabilistic framework used here should help in giving a more rigorous derivation of the estimation algorithm and allow for performing the task in an unsupervised way.

Several studies have considered dealing with these line spectra to perform analysis. Among them, [Gómez, 2006] proposes to compute tonal descriptors from the frequencies of local maxima extracted from polyphonic audio short-time spectra. In [Doval and Rodet, 1991, 1993] a probabilistic model for multiple-F_0 estimation from sets of maxima of the Short-Time Fourier Transform is introduced. It is based on a Gaussian mixture model having means constrained by an F_0 parameter and solved as a maximum likelihood problem by means of heuristics and grid search. Such approach has been proposed more recently in [Duan et al., 2010] where special care is taken to the modeling of both peaks and non-peaks regions in order to limit the activation of spurious harmonically-related notes. The F_0 parameters are estimated jointly with a candidate note selection by means of an iterative greedy search strategy. The other parameters of the model are learned from monophonic and polyphonic training data. A similar constrained mixture model has also been proposed in [Kameoka et al., 2004] to model speech spectra (along the whole frequency range, where here a Gaussian distribution is used to model the main lobe of a peak) and solved using an Expectation-Maximization (EM) algorithm.

The Probabilistic Line Spectrum (*PLS*) model presented here is inspired by these references. The key difference is that we focus on piano tones, which have the well-known property of inharmonicity, that in turn influences tuning. This slight frequency stretching of partials should allow, up to a certain point, disambiguation of harmonically-related notes. Reversely, from the set of partial frequencies, it should be possible to estimate the (B, F_0) parameters while detecting the played notes. The model assumes that, for a time-frame of signal, the observations have been generated by a mixture of notes composed by partials (Gaussian mixture) and noise components. The Gaussian mixture for each note is constrained by (B, F_0) parameters to have means distributed according to the inharmonicity relation (2.23). Then, (B, F_0) parameters are estimated jointly with a classification of each observed frequency into partial and noise classes for each note by means of a maximum *a posteriori* Expectation-Maximization algorithm. This technique is finally applied to the unsupervised estimation of (B, F_0) along the compass of pianos,

first in a monophonic context by jointly processing isolated note recordings, and then from a musical polyphonic piece.

3.2.1 Model and problem formulation

3.2.1.1 Observations

As mentioned in the introduction of Chapter 2, the information contained in two consecutive frames of music signals is often highly redundant. This suggests that in order to retrieve the (B, F_0) parameters for the whole set of notes played in a piece of solo music, a few independent frames localized after note onset instants should contain all the information that is necessary for processing. These time-frames are indexed by $t \in [1, T]$ in the following. In order to extract peaks that contain significant energy from the magnitude spectra, a noise level estimation based on median filtering is first performed (*cf.* Appendix B). Above this noise level, local maxima (defined as having a greater magnitude than $K_{\max}$ left and right frequency bins) are extracted. An illustration of this pre-processing is given in Figure 3.18. The frequency of each maximum picked in a frame t is denoted by y_{ti}, $i \in [1, I_t]$. The set of observations for each frame is then denoted by $\mathbf{y}_t$ (a vector of length I_t), and for the whole piece of music by $Y = \{\mathbf{y}_t,\ t \in [1, T]\}$. In the following of this section, the variables denoted by lower case, bold lower case and upper case letters will respectively correspond to scalars, vectors and sets of vectors.

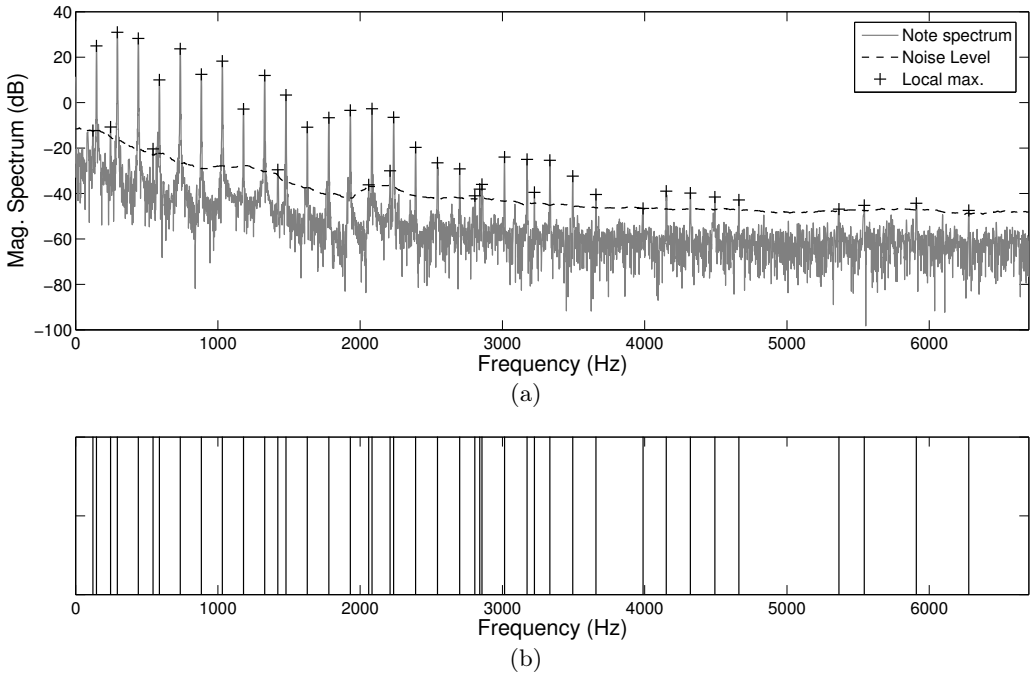


Figure 3.18: Illustration of the peak-picking pre-processing. (a) Magnitude spectrum and (b) line spectrum of a note D3 (50).

3.2.1.2 Probabilistic model

If a single note of music, indexed by $r \in [1, R]$, is present in a time-frame, most of the extracted local maxima should correspond to partials related by a particular structure

(harmonic or inharmonic for instance). These partial frequencies correspond to the set of parameters of the proposed model. It is denoted by θ, and in a general context (no information about the harmonicity or inharmonicity of the sounds) can be expressed by $\theta = \{f_{nr} | \forall n \in [1, N_r], r \in [1, R]\}$, where n is the rank of the partial and N_r the maximal rank for the note r. For the proposed model, we consider a strict inclusion of the inharmonicity relation $f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}$, as done for the *Inh-NMF* model presented in Section 3.1.1.2. Thus, the set of parameters can be rewritten as $\theta = \{F_{0r}, B_r | \forall r \in [1, R]\}$.

In order to link the observations to the set of parameters θ, the following hidden random variables are introduced:

- $q_t \in [1, R]$, corresponding to a candidate note that could have generated the observations $\mathbf{y}_t$.
- $C_t = [c_{tir}]_{(i,r)\in[1,I_t]\times[1,R]}$ gathering Bernoulli variables specifying the nature of the observation y_{ti}, for each note r. If $c_{tir} = 1$, the observation y_{ti} is assumed to correspond to a partial of the note r. If $c_{tir} = 0$, it corresponds to "noise", here defined as a non-sinusoidal component or a partial of another note.
- $P_t = [p_{tir}]_{(i,r)\in[1,I_t]\times[1,R]}$ corresponding to the rank of the partial n of the note r that could have generated the observation y_{ti} provided that $c_{tir} = 1$.

Based on these definitions, the probability that an observation y_{ti} has been generated by a note r can be expressed as:

$$\begin{aligned} p(y_{ti}|q_t = r; \theta) &= p(y_{ti}|c_{tir} = 0, q_t = r) \cdot p(c_{tir} = 0|q_t = r) \\ &+ \sum_n p(y_{ti}|p_{tir} = n, c_{tir} = 1, q_t = r; \theta) \\ &\cdot p(p_{tir} = n|c_{tir} = 1, q_t = r) \cdot p(c_{tir} = 1|q_t = r). \end{aligned} \tag{3.31}$$

It is chosen that the observations that are related to the partial n of a note r should be located around the frequencies f_{nr} according to a Gaussian distribution of mean f_{nr} and variance σ_r^2 (fixed parameter):

$$p(y_{ti}|p_{tir} = n, c_{tir} = 1, q_t = r; \theta) = \mathcal{N}(f_{nr}, \sigma_r^2), \tag{3.32}$$

$$p(p_{tir} = n|c_{tir} = 1, q_t = r) = 1/N_r. \tag{3.33}$$

On the other hand, observations that are related to noise are chosen to be uniformly distributed along the frequency axis (with maximal frequency F):

$$p(y_{ti}|c_{tir} = 0, q_t = r) = 1/F. \tag{3.34}$$

These distributions are illustrated on Figure 3.19.

Then, the probability to obtain a noise or partial observation knowing the note r is chosen so that:

• if $I_t > N_r$:

$$p(c_{tir}|q_t = r) \underset{I_t > N_r}{=} \begin{cases} (I_t - N_r)/I_t & \text{if } c_{tir} = 0, \\ N_r/I_t & \text{if } c_{tir} = 1. \end{cases} \tag{3.35}$$

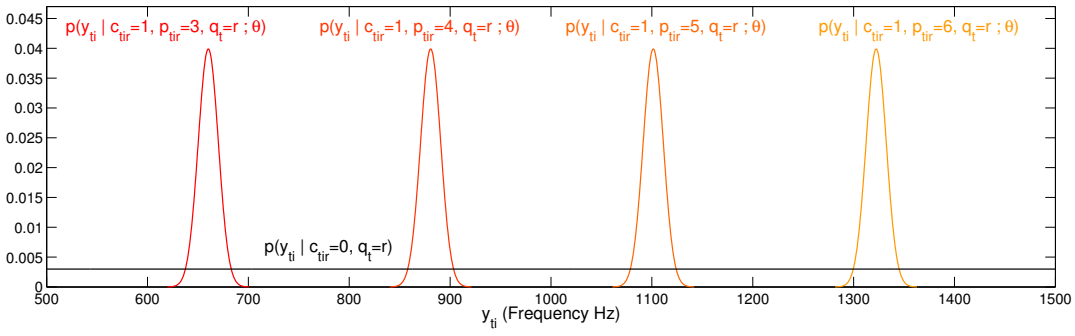


Figure 3.19: Illustration of the probabilistic model.

This should approximately correspond to the proportion of observations associated to noise and partial classes for each note.

• if $I_t \leq N_r$:

$$p(c_{tir}|q_t = r) \underset{I_t \leq N_r}{=} \begin{cases} 1-p & \text{if } c_{tir} = 0, \\ p & \text{if } c_{tir} = 1, \end{cases} \tag{3.36}$$

with $p \in [0, 1]$. This latter expression for $p < 0.5$ means that for a given note r at a frame t, most of observations should be mainly considered as noise if N_r (its number of partials), is greater than the number of observations I_t. This situation may occur for instance in a frame in which a single note from the high treble range is played. In this case, only a few local maxima are extracted and lower notes, composed of much more partials, should not be considered as present.

Finally, with no prior information it is chosen

$$p(q_t = r) = 1/R. \tag{3.37}$$

3.2.1.3 Estimation problem

In order to estimate the parameters of interest θ, the following maximum *a posteriori* estimation problem is solved:

$$(\theta^\star, \{C_t^\star\}_t, \{P_t^\star\}_t) = \underset{\theta, \{C_t\}_t, \{P_t\}_t}{\operatorname{argmax}} \sum_t \log p(\mathbf{y}_t, C_t, P_t; \theta), \tag{3.38}$$

where

$$p(\mathbf{y}_t, C_t, P_t; \theta) = \sum_r p(\mathbf{y}_t, C_t, P_t, q_t = r; \theta). \tag{3.39}$$

Solving problem (3.38) corresponds to the estimation of θ, jointly with a classification of each observation into noise or partial classes for each note. Note that the sum over t of Equation (3.38) arises from the time-frame independence assumption (justified in Section 3.2.1.1).

3.2.2 Optimization

Problem (3.38) has usually no closed-form solution but can be solved in an iterative way by means of an Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. The

auxiliary function at iteration $(k+1)$ is given by

$$Q(\theta, \{C_t\}_t, \{P_t\}_t | \theta^{(k)}, \{C_t^{(k)}\}_t, \{P_t^{(k)}\}_t) = \sum_t \sum_r \omega_{rt} \cdot \sum_i \log p(y_{ti}, c_{tir}, p_{tir}, q_t = r; \theta) \tag{3.40}$$

where by definition,

$$\omega_{rt} = p(q_t = r | \mathbf{y}_t, \{C_t^{(k)}\}_t, \{P_t^{(k)}\}_t; \theta^{(k)}), \tag{3.41}$$

is computed at the E-step knowing the values of the parameters at iteration (k). At the M-step, θ, $\{C_t\}_t$, $\{P_t\}_t$ are estimated by maximizing Equation (3.40). Note that the sum over i in Equation (3.40) is obtained under the assumption that in each frame the y_{ti} are independent.

3.2.2.1 Expectation

According to Equation (3.41) and model Equation (3.31)-(3.37)

$$\begin{aligned}
\omega_{rt} &\propto \prod_{i=1}^{I_t} p(y_{ti}, q_t = r, c_{tir}^{(k)}, p_{tir}^{(k)}; \theta^{(k)}) \\
&\propto p(q_t = r) \cdot \prod_{i/\ c_{tir}^{(k)}=0} p(y_{ti} | q_t = r, c_{tir}^{(k)}) \cdot p(c_{tir}^{(k)} | q_t = r) \\
&\quad \cdot \prod_{i/\ c_{tir}^{(k)}=1} p(y_{ti} | q_t = r, c_{tir}^{(k)}, p_{tir}^{(k)}, \theta^{(k)}) \cdot p(p_{tir}^{(k)} | c_{tir}^{(k)}, q_t = r) \cdot p(c_{tir}^{(k)} | q_t = r),
\end{aligned} \tag{3.42}$$

normalized so that $\sum_{r=1}^{R} \omega_{rt} = 1$ for each frame t.

As defined in Equation (3.41), ω_{rt} corresponds to the probability that a note r is active at a frame t given the observations and the values of the parameters. Thus, the matrix $\Omega = [\omega_{rt}]_{(r,t) \in [1,R] \times [1,T]}$ is similar to the activation matrix H of NMF models.

3.2.2.2 Maximization

The M-step is performed by a sequential maximization of Equation (3.40):

- First, estimate $\forall\ t, i$ and $q_t = r$ the variables c_{tir} and p_{tir}. As mentioned in Section 3.2.1.3, this corresponds to a classification step, where each observation is associated, for each note, to noise class ($c_{tir} = 0$) or partial class with a given rank ($c_{tir} = 1$ and $p_{itr} \in [1, N_r]$). This step is equivalent to a maximization of $\log p(y_{ti}, c_{tir}, p_{tir} \mid q_t = r; \theta)$ which, according to Equations (3.31)-(3.37), can be expressed as:

$$(c_{tir}^{(k+1)}, p_{tir}^{(k+1)}) = \tag{3.43}$$
$$\underset{(\{0,1\},n)}{\operatorname{argmax}} \begin{cases} -\log F + \log p(c_{tir} = 0 | q_t = r), \\ -(y_{ti} - nF_{0r}^{(k)} \sqrt{1 + B_r^{(k)} n^2})^2 / (2\sigma_r^2) - \log N_r \sqrt{2\pi}\sigma_r + \log p(c_{tir} = 1 | q_t = r). \end{cases}$$

- Then, the estimation of θ is equivalent to $(\forall r \in \{1 \ldots R\})$

$$(F_{0r}^{(k+1)}, B_r^{(k+1)}) = \underset{F_{0r}, B_r}{\operatorname{argmax}} \sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} \log p(y_{ti}, c_{tir}^{(k+1)} = 1, p_{tir}^{(k+1)}, q_t = r; \theta), \tag{3.44}$$

which, according to Equations (3.31)-(3.37), leads to the following minimization problem:

$$(F_{0r}^{(k+1)}, B_r^{(k+1)}) = \underset{F_{0r}, B_r}{\operatorname{argmin}} \sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} \left(y_{ti} - p_{tir}^{(k+1)} F_{0r} \sqrt{1 + B_r \, p_{tir}^{(k+1)^2}} \right)^2. \quad (3.45)$$

For F_{0r}, the following update rule is obtained when canceling the partial derivative of Equation (3.45):

$$F_{0r}^{(k+1)} = \frac{\sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} y_{ti} \cdot p_{tir}^{(k+1)} \cdot \sqrt{1 + B_r \, p_{tir}^{(k+1)^2}}}{\sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} p_{tir}^{(k+1)^2} \cdot (1 + B_r \, p_{tir}^{(k+1)^2})}. \quad (3.46)$$

For B_r, no closed-form solution can be obtained from the partial derivative of Equation (3.45). The minimization is thus performed by means of an algorithm based on the Nelder-Mead simplex method (as implemented in the *fminsearch* MATLAB$^{\text{TM}}$ function).

Algorithm 3 *PLS* model for the estimation of (B, F_0)

Input:
$\{\mathbf{v}_t, t \in [1, T]\}$ set of magnitude spectra

Preprocessing: $\forall t \in [1, T]$
compute the Noise Level (NL) of $\mathbf{v}_t$ (*cf.* App. B) and pick the frequencies $\mathbf{y}_t$ of the local maxima having magnitude greater than NL and $K_{\max}$ left and right frequency bins

Initialization:
(B_r, F_{0r}), $\forall\, r \in [1, R]$ according to the model of Sec. 4.4.3
N_r^{ini} and N_r^{fin}, $\forall\, r \in [1, R]$ as shown in Fig. 3.2, p. 43
compute c_{tir} and p_{tir}, $\forall\, t, i, r$ (Eq. (3.43))

Optimization:
for $it = 1$ to It **do**
  if $\text{mod}(it, 10) = 0$ **then**
    $N_r \leftarrow N_r + 1$, $\forall r \in [1, R]$ provided $N_r < N_r^{\text{fin}}$
  end if
  • **E-Step**:
    compute ω_{rt} $\forall\, r \in [1, R]$, $t \in [1, T]$ (Eq. (3.42))
    keep the 10 highest values of ω_{rt} in each frame t and set the others to 0
    normalize ω_{rt} so that $\sum_r \omega_{rt} = 1$ for each frame t
  • **M-Step**:
    compute c_{tir} and p_{tir} $\forall\, t, i, r$ (Eq. (3.43))
    compute F_{0r} $\forall\, r$ (Eq. (3.46))
      if F_{0r} is outside the limits **then** fix its value to the closest limit
    compute B_r $\forall\, r$ (Eq. (3.45) + Nelder-Mead simplex method)
      if B_r is outside the limits **then** fix its value to the closest limit
end for

Output: B_r, F_{0r}, ω_{rt}

3.2.2.3 Practical considerations

The cost-function (*cf.* maximization Equation (3.38)) is non-convex with respect to (B_r, F_{0r}) parameters. Thus, as done in Section 3.1.2.2 for the NMF-based models, the initialization of the parameters uses the model of inharmonicity and tuning of pianos along the whole compass (*cf.* Chapter 4). Also, similarly to the NMF model algorithms, the optimization is run with a few partials for each note at the initialization. Then, one partial is added for each note every 10 iterations (number determined empirically) by initializing its frequency with the current (B_r, F_{0r}) estimates.

Finally, in order to avoid situations where the algorithm optimizes the parameters of a note on data corresponding to another note (*e.g.* increasing F_0 by one semi-tone), the values of (B_r, F_{0r}) are prevented from being updated over limit curves. For B, these curves are presented in Section 4.4.3 and can be seen depicted as gray dashed-line in Figures 3.21(a) and 3.22(a). The limits curves for F_0 are set to $+/-$ 40 cents of the initialization. Overfitting issues are also tackled by applying a threshold to ω_{rt} that limits the polyphony level to 10 notes for each frame t. Each step of the optimization of the *PLS* model is summarized in Algorithm 3.

3.2.3 Results

In this section, we investigate the ability of the model and its algorithm to provide correct estimates of (B, F_0) from the unsupervised joint analysis of a set of single note recordings, as from the analysis of a piece of polyphonic music.

The observation set is built according to the description given in Section 3.2.1.1. The time-frames are extracted after note onsets and their length is set to 500 ms in order to have a sufficient spectral resolution. The FFT is computed on 2^{15} bins and the maxima are extracted by setting $K_{\max} = 20$. Note that for the presented results, the knowledge of the note onsets is taken from the ground truth (MIDI aligned files). For a complete blind approach, an onset detection algorithm should be first run. It can be assumed that this should not significantly affect these results, since onset detection algorithms usually perform well on percussive tones. The parameters of the model are chosen as follows : N_r is set to $\arg\min_{N_r}(30, f_{N_r,r} < F_s/2)$, the parameter of the Bernoulli distribution is empirically set to $p = 0.1$, and σ_r is empirically set to 3 and 2 Hz, respectively for the applications "isolated note estimation" and "estimation from a piece of music".

3.2.3.1 Supervised vs. unsupervised estimation from isolated notes jointly processed

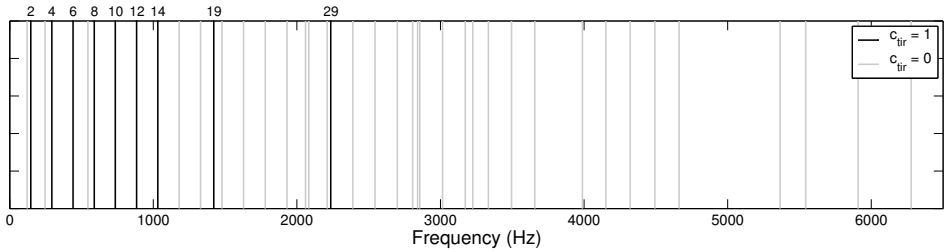
Here, we apply the algorithm to the estimation of (B_r, F_{0r}) parameters from isolated note recordings covering the whole compass of pianos. The set of observations is composed of 88 frames (jointly processed), one for each note of the piano (from A0 (MIDI index 21) to C8 (108)) and processed in both supervised and unsupervised ways.

In the supervised case, the activation matrix Ω is fixed to the identity matrix (as done in Section 3.1.3.2 for the activation matrix of the NMF-based models) and only the M-step is performed during the optimization. It is worth noting for this supervised case that the algorithm can be seen as part of the "Iterative peak-picking and refinement of the inharmonicity coefficient" class of methods presented in Section 2.3. Indeed, partials corresponding to transverse vibrations of the string are iteratively selected in the classification step and (B_r, F_{0r}) are updated by minimizing the quadratic error between the partial frequencies and the parameters. In the unsupervised case, the note activations are also estimated by the algorithm.

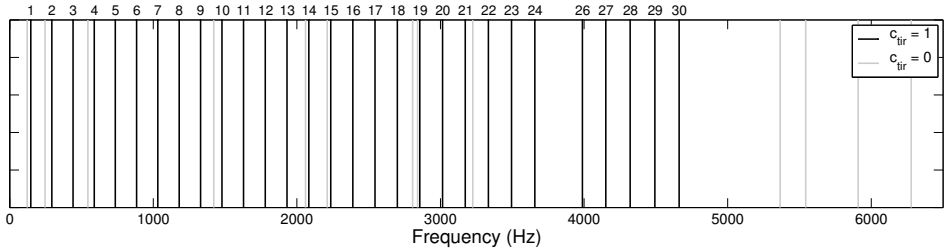
The results for 2 different pianos are presented on Figure 3.21 (Iowa piano) and 3.22 (MAPS SptkBGCl piano). Subplots (a) and (b) depict respectively the estimation of B and F_0. When comparing the supervised (blue curves) and the unsupervised (red curves) curves one can see that similar results are obtained, approximately up to note $C\sharp 6$ (85). In this range, one can see on subplot (c) that the matrix Ω estimated in the unsupervised case exhibits the expected diagonal structure, up to one mistake at time-frame $t = 53$ for Iowa

piano. This corresponds to an octave error where the note $C\sharp 6$ (85) is detected instead of $C\sharp 5$ (73). Above, the detection is not correct and logically leads to different estimates of (B, F_0) when comparing to the supervised case. As seen in Section 3.1.3.2 for the NMF-based modelings the estimation of (B, F_0) in this range may encounter difficulties because notes are composed of 3 coupled strings that produce multiple partials that do not fit well into the inharmonicity model Equation (2.23). As also shown in Section 3.1.3.2, the analysis in this range for both supervised and unsupervised cases may suffer from initialization issues and lead to (B, F_0) estimates frozen to their initial values (as seen above note F7 (101) on Figure 3.22(b)).

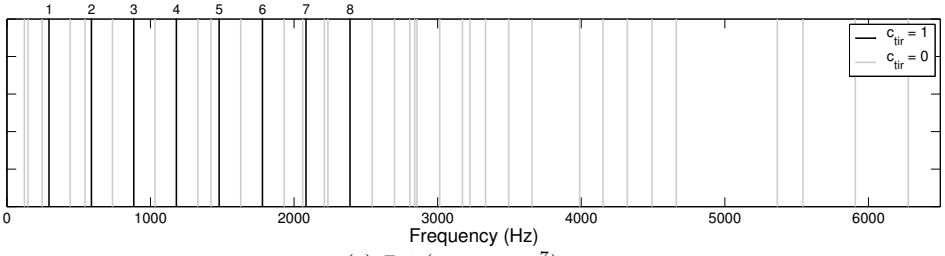
Interestingly, below note $C\sharp 6$ (85) the algorithm performs as expected, with very few harmonically-related errors in Ω. Taking for instance frame $t = 30$ of Iowa piano, and the notes D2 (38), D3 (50) and D4 (62) of the model (*cf.* Figure 3.20), one can see that only D3 is detected because most observations have been classified as partials (black vertical bars). On the contrary, a very low value of ω_{rt} is obtained for D2 and D4 notes because a large amount of observations have been classified as noise (gray vertical bars).



(a) D2 ($\omega_{rt} \approx 10^{-23}$)

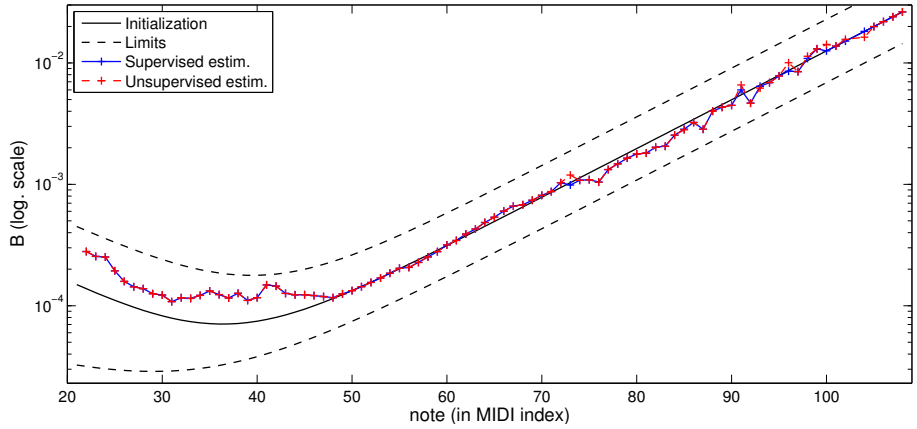


(b) D3 ($\omega_{rt} \approx 1$)

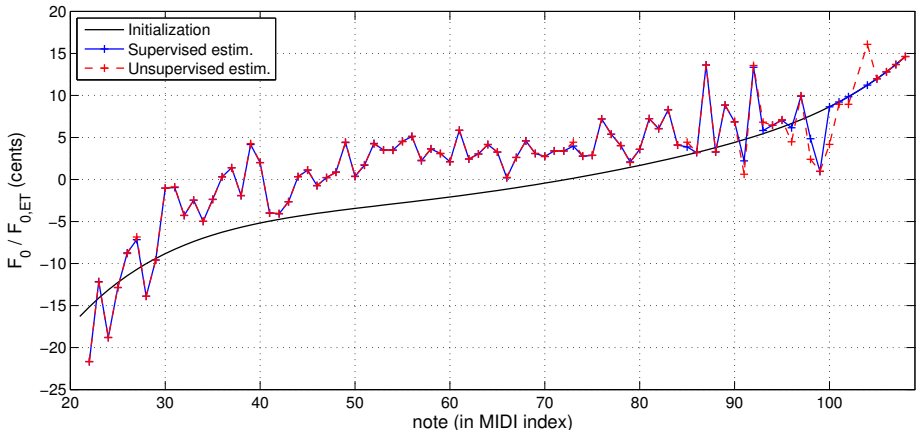


(c) D4 ($\omega_{rt} \approx 10^{-7}$)

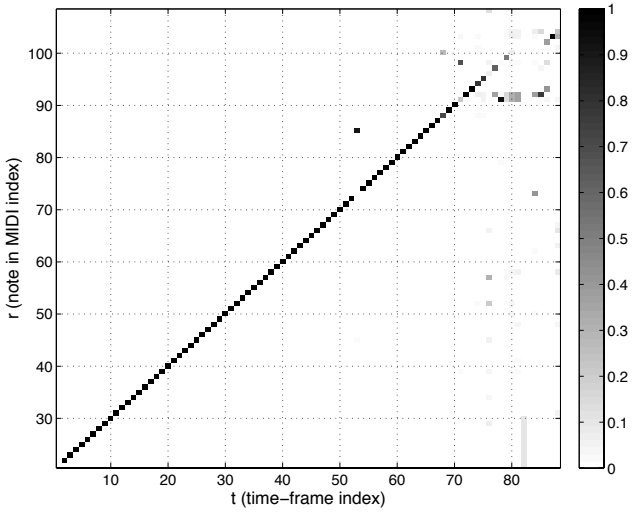
Figure 3.20: Result of the algorithm for the analysis of the frame $t = 30$ from Iowa grand piano corresponding to the line spectrum of the note D3 (50). Black and gray vertical bars correspond to the observed frequencies respectively classified as partial ($c_{tir} = 1$) and noise ($c_{tir} = 0$) for 3 octave-related notes (D2-D3-D4). The partial rank for observations with $c_{tir} = 1$ is reported above each graph.



(a)

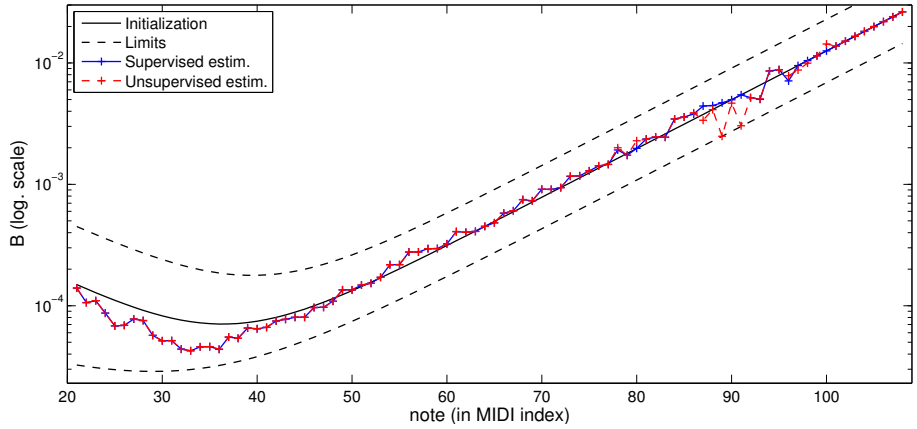


(b)

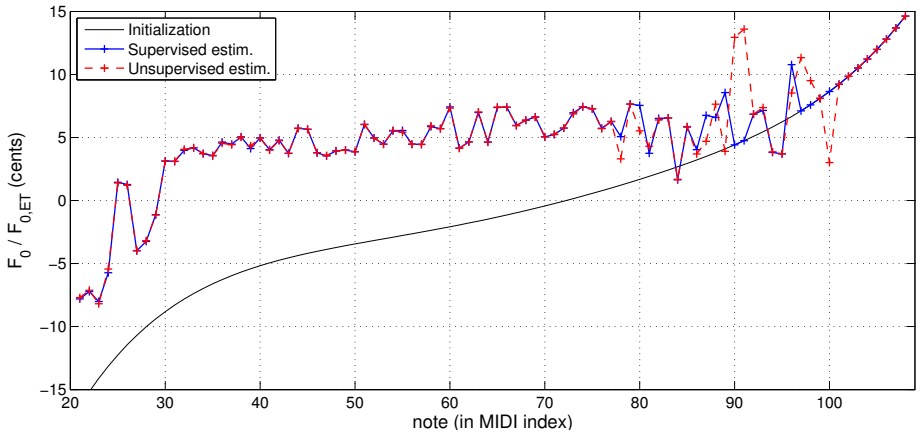


(c)

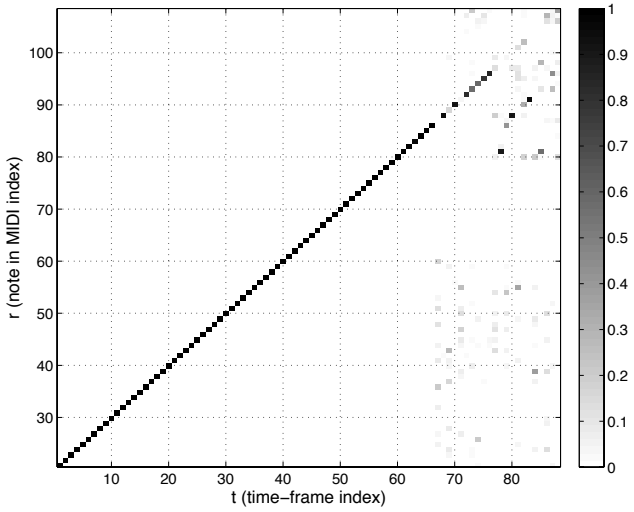
Figure 3.21: Analysis on the whole compass from isolated note recordings of Iowa piano. (a) B in log. scale and (b) F_0 as dev. from ET (in cents) along the whole compass. (B, F_0) estimates are depicted as blue and red '+' markers, respectively for the supervised and the unsupervised cases. The initialization is plotted as gray lines and the limits for the estimation of B as gray dashed-lines. (c) Ω returned by the unsupervised analysis.



(a)



(b)



(c)

Figure 3.22: Analysis on the whole compass from isolated note recordings of MAPS SP-tkBGCl piano. (a) B in log. scale and (b) F_0 as dev. from ET (in cents) along the whole compass. (B, F_0) estimates are depicted as blue and red '+' markers, respectively for the supervised and the unsupervised cases. The initialization is plotted as gray lines and the limits for the estimation of B as gray dashed-lines. (c) Ω returned by the unsupervised analysis.

Finally, the precision of the estimation of (B, F_0) is quantitatively evaluated according to the protocol described in Section 3.1.3.3. The results are presented in table 3.3 and compared with *PFD* and *Inh-NMF* algorithms. As qualitatively noticed above, the results for both supervised and unsupervised cases are similar, and comparable to those obtained by *Inh-NMF*.

	PFD	*Inh-NMF*	*PLS* sup.	*PLS* unsup.	
Synthetic	0.783	0.323	0.307	0.307	E_{B_r} (%)
	0.307	0.110	0.291	0.293	$E_{F_{0r}}$ (cents)
Iowa	3.30	1.25	1.06	1.06	E_{B_r} (%)
	1.06	0.539	0.607	0.601	$E_{F_{0r}}$ (cents)

Table 3.3: (B, F_0) estimation errors averaged on the range A0-G3, for both supervised and unsupervised *PLS* (Probabilistic Line Spectrum) models applied on the synthetic and Iowa piano datasets. *PFD* and *Inh-NMF* results are given for comparison. The metrics are defined in Section 3.1.3.3.

3.2.3.2 Unsupervised estimation from musical pieces

Finally, the algorithm is applied to an excerpt of polyphonic music (25 s of *MAPS_MUSmuss_3_SptkBGCl* file) containing notes in the range $D\sharp1$ (27) - $F\sharp6$ (90) from which 46 frames are extracted. The mean polyphony level by frame is approximately equal to 10.5. This high value results from the high time-frame duration which is set to 500 ms in order to obtain a sufficient resolution, as mentioned above. 76 notes are considered in the model and initialized in order to take into account notes from A0 (21) to C7 (96). This corresponds to a reduction of one octave in the high treble range where the notes cannot be properly processed, as seen in Section 3.2.3.1. Luckily, these notes are rarely used in a musical context.

The proposed application is here the learning of (B, F_0) along the compass of a piano from a generic polyphonic piano recording. After the optimization, a post-processing is performed in order to keep the most reliable estimates. First, a threshold is applied to the matrix Ω so that elements having values lower than 10^{-3} (threshold set empirically) are discarded. Second, notes having B estimates stuck to the limits (*cf.* gray dashed lines in Figure 3.22(a)) are rejected.

Figure 3.23 depicts the result of the frame-wise note selection obtained after the post-processing for the considered piece of music. As it can be seen, the drastic post-processing leads to a greater number of True Positive (TP) than False Positive (FP). The frame-wise evaluation returns thus a high precision of 95.7 % and a weak recall of 9.28 % (these metrics are properly defined in Chapter 5). This result is suitable regarding our application for which we mainly focus on the reliable detection of a few notes in each frame in order to perform the precise estimation of their parameters (B, F_0). In a potential application of the algorithm to a transcription task, one should consider a smaller time-frame duration (leading to a lower frame-wise polyphony level) and a lower value for the threshold of the activation matrix Ω at the post-processing.

Finally, the results of the estimation of (B, F_0) for the detected note is presented

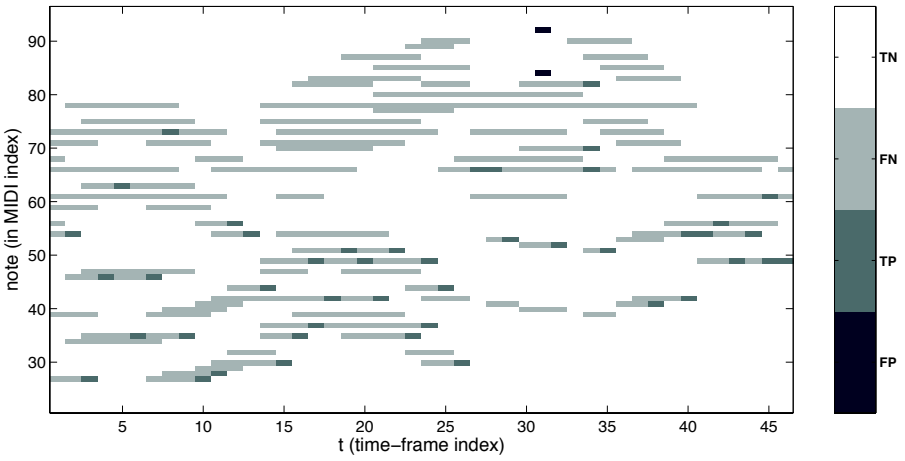


Figure 3.23: Frame-wise evaluation of the note detection.

on Figure 3.24 (red '+' markers). When comparing to the estimates obtained by the supervised analysis on isolated note recordings (black '+' markers) one can see that most estimates of notes actually present are consistent. Averaged on these 19 notes correctly detected (present below note B5 (83)) and compared to the supervised estimation from isolated note recordings, a relative error of 5.2 % for B and a mean deviation of 0.70 cents for F_0 are obtained.

Even if only a few notes are detected (19 out of 39 actually present in the presented example), the algorithm returned precise estimates of (B, F_0). By using a model of inharmonicity and tuning along the whole compass of pianos, it will be shown in the applications of Chapter 4 that (B, F_0) may be interpolated along the whole compass, thus leading to applications such as the retrieval of the tuning (related to F_0) and properties about the piano type (related to B) by the analysis of a generic polyphonic recording. Interestingly, for this task a perfect transcription of the music does not seem necessary: only a few reliable notes may be sufficient. However, an extension of this model to piano transcription could form a natural extension, but would require a more complex model taking account both temporal dependencies between frames, and spectral envelopes.

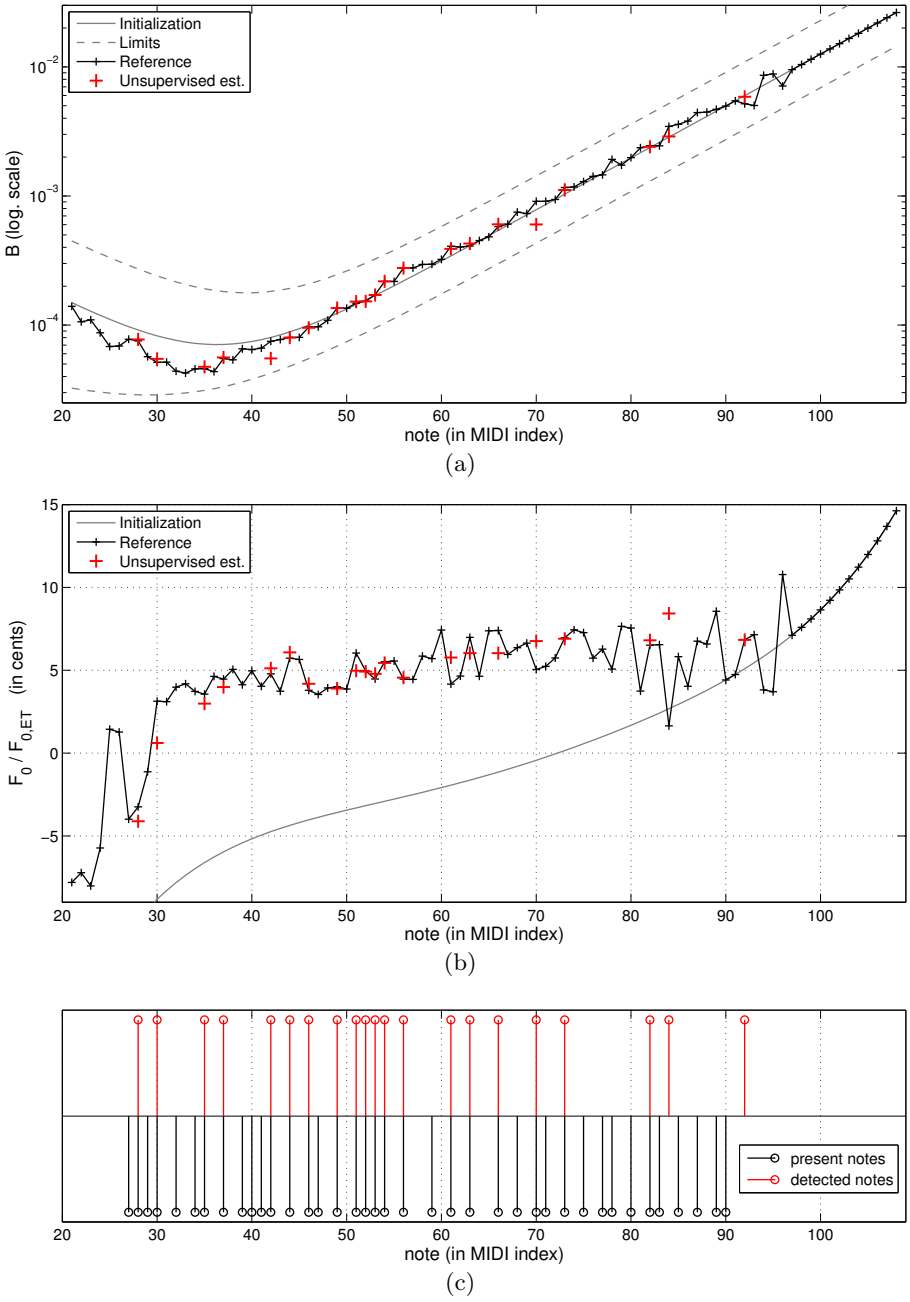


Figure 3.24: (B, F_0) estimation along the compass from a piece of music. (a) B in log. scale and (b) F_0 as dev. from ET in cents. (B, F_0) estimates are depicted as red '+' markers and compared to the supervised estimation on isolated notes (black '+' markers). The initialization is plotted as gray lines and the limits for the estimation of B as gray dashed-lines. (c) Notes detected by the algorithm (red) and notes actually present in the piece (black).

Chapter 4

A parametric model for the inharmonicity and tuning along the whole compass of the piano

This chapter introduces a model for the variations of (B, F_0) parameters along the whole compass of pianos based on instrumental design and tuning rules. Portions of this work have been published in [Rigaud et al., 2011, 2013a].

Most of the tuning rules related to a musical temperament are based on the control of beatings produced when playing different intervals, these assuming that the tones are harmonic [Rasch and Heetvelt, 1985]. In the case of the piano, these beatings cannot be exactly complied because of the partial frequency deviations inherent to the inharmonicity. The amount of inharmonicity being dependent on the physical characteristics of the strings, it is different for each note and type of piano. Thus, the fine tuning of pianos is usually done aurally and requires the expertise of a professionally-skilled tuner unlike many string instruments such as guitar or cello, for which simple devices (so-called "tuners") can be used. According to the model of piano and the choices/abilities of the tuner, the resulting tuning is unique, but within some physically-based constraints. From a musical acoustics perspective its modeling is hence an interesting challenge that has been tackled by different viewpoints. A simulation of aural piano tuning has been proposed in [Lattard, 1993] to help pianists in tuning their own pianos, replicating the tuner's work by iteratively tuning different intervals. The method is based on a mathematical computation of the beat rates, and requires the frequencies of the first 5 partials of each note. More recently, an approach using psychoacoustic principles has been introduced in [Hinrichsen, 2012]. This algorithm adjusts the 88 notes at the same time, by an optimization procedure on modified spectra of the notes according to psycho-acoustic laws and tuning updates.

Our approach is different, as it jointly models tuning and inharmonicity laws for the whole compass from a reduced set of parameters (6 mid-level parameters instead of 88×2 parameters corresponding to (B, F_0) along the whole compass). Although such an interpolated model can only capture the main trends in the inharmonicity and tuning of a given piano, it should be reminded that one of the objectives of piano manufacturing and tuning is precisely to have a timbre that is as homogeneous as possible, smoothing out as much as possible the discontinuities of physical origins: bass break (transition between treble and bass bridge), change in the number of strings per note, change of string diameter and winding. Therefore, it is not only realistic, but also relevant, to try to globally parametrize the inharmonicity and tuning with only a few parameters, at least as a first-order approximation.

The obtained synthetic description of a particular instrument, in terms of its tuning

/ inharmonicity pattern, can be useful to assess its state and also provides clues on some of the tuner's choices. In the field of musical acoustics, the use of such a model could be helpful for instance for the tuning of physically-based piano synthesizers, where we are otherwise faced with the problem of having to adjust a large number of parameters, all of them being inter-dependent. Here a higher-level control can be obtained, with a few physically meaningful parameters. In the fields of audio signal processing and Music Information Retrieval, including *a priori* knowledge is often done when trying to enhance the performance of the algorithms (*cf.* Section 2.3) and such models could be used to provide a good initialization of (B, F_0) parameters (as done in Chapters 3 and 5) and to constrain their estimation. Also, the models may be used for interpolating the inharmonicity and tuning along the whole compass of a given piano from the estimation of (B, F_0) of a few notes, for instance obtained by the analysis of a piece of music, as done in Section 3.2.3.2.

4.1 Aural tuning principles

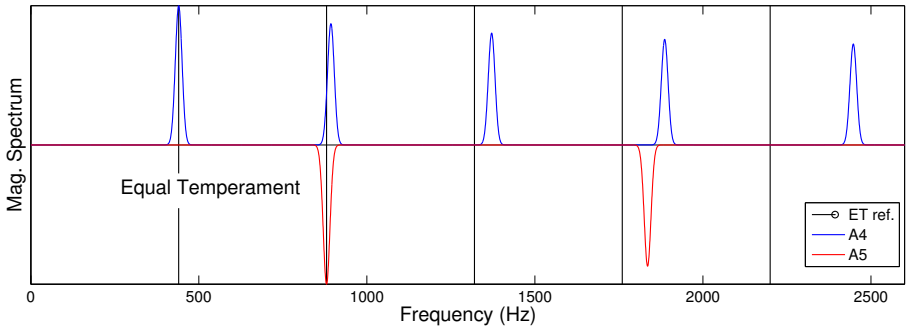
Aural tuning is based on the perception and the control of beatings between partials of two different tones simultaneously played [Bremmer, 2007b]. It always begins by the tuning of a reference note, in most cases the A4 (69 in MIDI index) at 440 Hz (sometimes 442 Hz). To do so, the tuner adjusts the tension of the strings to cancel the beatings produced by the difference of frequency between the tuning fork and the first partial of the note. Thus, $f_1(m = 69) = 440$ Hz. Even if there are different methods, skilled tuners usually begin by the scale tuning sequence: the F3-F4 octave is set by approximate Equal Temperament [Capleton, 2007; Bremmer, 2007b]. The rest of the keyboard is tuned by adjusting beatings between the partials of two different notes, typically octave-related.

Because of the inharmonicity, simply adjusting the first partial of each note on Equal Temperament (ET) would produce unwanted beatings, in particular for octave intervals (*cf.* Figure 4.1(a)). When tuning an octave interval by canceling the beatings produced by the second partial of a note indexed by m and the first partial of a note indexed by $m + 12$, the resulting frequency ratio $\frac{f_1(m+12)}{f_1(m)}$ is higher than 2 because $f_2(m) > 2f_1(m)$ (*cf.* Figure 4.1(b)). This phenomenon is called octave stretching. Depending on where the notes are in the range of the compass, the amount of stretching can be different. This fact is linked to the underlying choice of the octave type (related to perceptual effects and tuner's personal choices) during the tuning [Bremmer, 2007a]. For instance, in a 4:2 type octave, the 4th partial of the reference note is matched to the 2nd partial of its octave (*cf.* Figure 4.1(c)). Depending on the position in the compass, the piano can be tuned according to different octave types: 2:1, 4:2, 6:3, 8:4 ... or a trade-off between two (*cf.* Figure 4.1(d)). This means that the tuner may not focus only on cancelling beatings between a pair of partials, but that he controls an average beating generated by a few partials of the two notes.

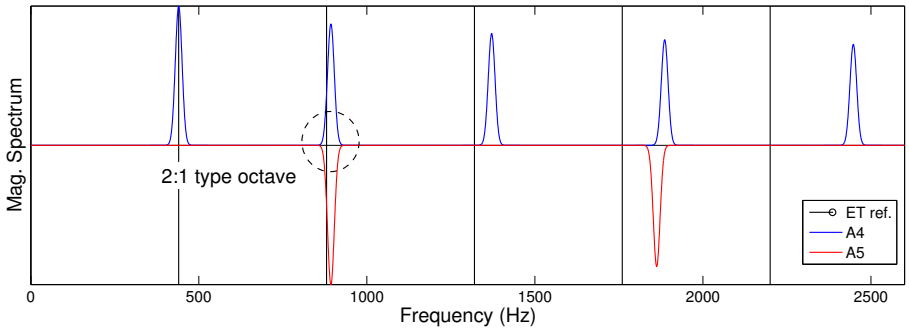
In order to highlight this stretching, the tuning along the compass (from A0 to C8, or for $m \in [21, 108]$ in MIDI index) is usually depicted as the deviation, in cents, of the first partial frequency of each note from ET:

$$d(m) = 1200 \cdot \log_2 \frac{f_1(m)}{F_{0,\mathrm{ET}}(m)}, \tag{4.1}$$

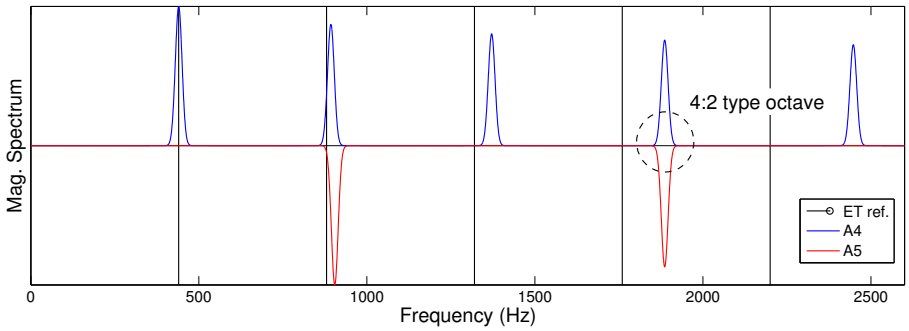
where $F_{0,\mathrm{ET}}(m)$ is the theoretical fundamental frequency given by the ET (*cf.* Appendix A).



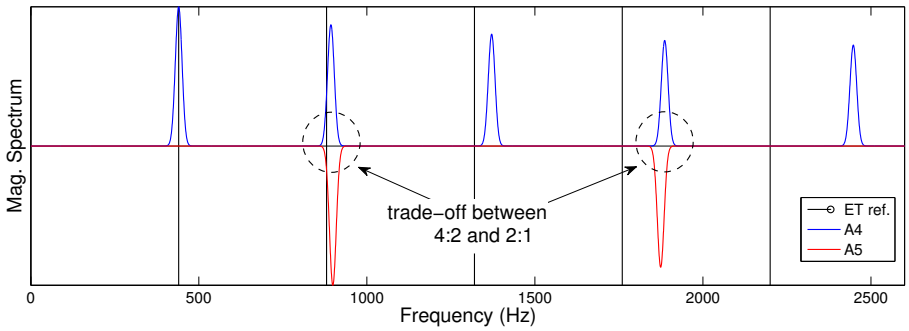
(a)



(b)



(c)



(d)

Figure 4.1: Influence of the inharmonicity and the octave type choice on the tuning of the octave A4-A5. Note: the inharmonicity coefficients have been significantly increased from typical values in order to highlight the deviations of the first partial frequencies from the harmonic reference.

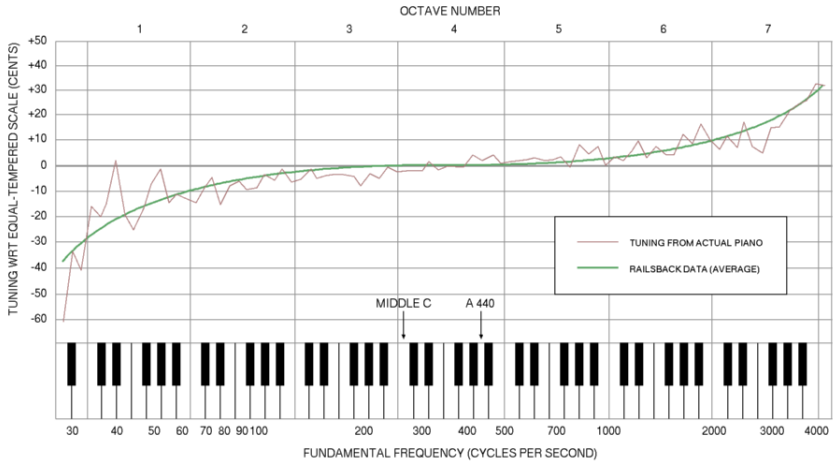


Figure 4.2: Typical measures of deviation from ET along the compass of pianos and Railsback curve. Brian Tung, from data in [Martin and Ward, 1961].

Usually [Fletcher and Rossing, 1998; Martin and Ward, 1961] the stretching increases gradually from the mid-range (deviation about $\pm$ 5 cents) to the extreme parts of the keyboard, producing deviations down to -30 cents in the low bass and up to $+30$ cents in the high treble. The goal of the proposed model is to explain the main variations of $d(m)$ along the compass (also known as the Railsback curve, *cf.* Figure 4.2) by taking into account the piano string set design characteristics (model of $B(m)$ along the compass) and the tuner's choices (model related to the octave type).

4.2 Parametric model of inharmonicity and tuning

The proposed model which simulates aural tuning on the whole compass is based on octave interval tunings. Its successive steps are a simplified version of those actually performed by a tuner, but the most important global considerations (stretching inherent to the inharmonicity and the octave type choice) are taken into account. The model starts by tuning all the octave intervals relatively to a reference note (for example the A4 at 440 Hz). From these notes, the tuning is then interpolated on the whole compass. Finally, the possibility of a global deviation is added, in order to allow for different tuning frequencies for the reference note.

4.2.1 Octave interval tuning

When tuning an "upper" octave interval (for instance A5 from A4), the cancellation of the beatings produced by the 2ρ-th partial ($\rho \in \mathbb{N}^*$) of a reference note, indexed by m (A4), and the ρ-th partial of its octave, indexed by $m+12$ (A5), can be done by tuning $F_0(m+12)$ such as:

$$F_0(m+12) = 2\ F_0(m)\sqrt{\frac{1+B(m)\cdot 4\rho^2(m)}{1+B(m+12)\cdot \rho^2(m)}}. \tag{4.2}$$

This equation clearly shows the influence of the note-dependent inharmonicity coefficient (B) and of the octave type (related to ρ) in the stretching of the octave. In the case of

"lower" octave tuning (for instance A3 from A4), the same relation can be inverted and applied by considering $m+12$ (A4) as the reference note and m (A3) as the note to tune. The next sections describe parametric models for B and ρ along the whole compass.

4.2.2 Whole compass model for the inharmonicity

4.2.2.1 String set design influence

In order to keep an homogeneous timbre along the compass, the strings are designed in such a way that discontinuities due to physical parameters variations are smoothed [Conklin, Jr., 1996b; Engelbrecht et al., 1999; Stulov, 2008]. Three main design considerations might produce such discontinuities in B along the keyboard: the bass break between the bass and treble bridges (jump in L, the speaking length of the strings, *cf.* Figure 4.3), the transitions between adjacent keys having a different number of strings (jump in T, the tension of each string [Engelbrecht et al., 1999; Fletcher and Rossing, 1998]), and the transition between plain strings to wrapped strings (jump in d, the diameter of the piano wire [Engelbrecht et al., 1999]).

On the treble bridge, from C8 note downwards, B is decreasing because of the increase of L. Down to middle C (C4 note, $m = 60$), the values of B are roughly the same for all the pianos and B follows a straight line in logarithmic scale [Young, 1952]. This result is mainly due to the fact that string design in this range is standardized, since it is not constrained by the limitation of the piano size [Conklin, Jr., 1996b].

In the low pitch range, the strings use a different bridge (the bass bridge) to keep a reasonable size of the instrument. Then, the linear mass of the strings is increased in order to adjust the value of F_0 according to Equation (2.21). Instead of increasing only the diameter d, which increases B (*cf.* Equation (2.24)), the strings are wound with a copper string wire, which increases the linear mass. Thus, on the bass bridge, B is increasing from sharpest notes downwards. Note that the number of keys associated to the bass bridge and the design of their strings are specific to each piano.

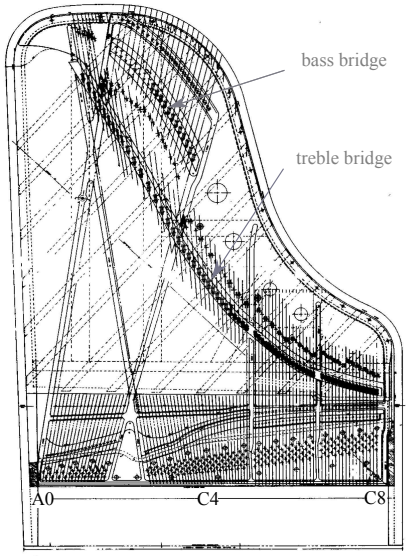


Figure 4.3: Grand piano string and bridge design. From [Conklin, Jr., 1996b].

4.2.2.2 Parametric model

According to the string design considerations, B could be modeled by two distinct functions corresponding to the two bridges, and could present discontinuities at the bass break or at the changes single-doublets and doublets-triplets of strings. The difficulty when modeling B on the whole compass is to know the position of these possible discontinuities, because they are specific to each piano model. Therefore, we propose a "continuous" additive model on the whole compass, discretized for $m \in [21, 108]$. We denote it by $B_\xi(m)$, ξ being the set of modeling parameters.

Usually, the evolution of B along the keyboard is depicted in logarithmic scale and presents two linear asymptotes. We denote by $b_T(m)$ (resp. $b_B(m)$) the treble bridge (resp. the bass bridge) asymptote of $\log B_\xi(m)$. Each asymptote is parametrized by its slope and its Y-intercept:

$$\begin{cases} b_T(m) = s_T \cdot m + y_T, \\ b_B(m) = s_B \cdot m + y_B. \end{cases} \tag{4.3}$$

According to *Young et al.* [Young, 1952], $b_T(m)$ is similar for all the pianos so s_T and y_T are fixed parameters. Then, the set of free (piano dependent) parameters reduces to $\xi = \{s_B, y_B\}$. $B_\xi(m)$ is set as the sum of the contributions of these two curves (4.3) in the linear scale:

$$B_\xi(m) = e^{b_B(m)} + e^{b_T(m)} \tag{4.4}$$

It should be emphasized that this additivity does not arise from physical considerations, but it is the simplest model that smoothes discontinuities between the bridges. Experimental data will show that it actually describes well the variations of B in the transition region around the two bridges.

The model is presented on Figure 4.4(a) for three different typical values of the set of parameters: ξ_1, ξ_2 and ξ_3, corresponding to low, medium and highly inharmonic pianos, respectively. The asymptotes corresponding to the bass and treble bridges are also drawn for $B_{\xi_2}(m)$.

4.2.3 Whole compass model for the octave type parameter

The octave tuning relation, given in Equation (4.2), considers the cancellation of the beatings produced by a single pair of partials. In practice, the deviation $\frac{F_0(m+12)}{2F_0(m)}$ could be a weighted sum of the contribution of two pairs of partials, because the amount of stretching may result from a compromise between two octave types [Bremmer, 2007a]. An alternative model to take into account this weighting is to allow non-integer values for $\rho \in [1, +\infty[$. For example, if the octave tuning is a compromise between a 2:1 and 4:2 type octaves, ρ will be in the interval $[1, 2]$. This model loses the physical meaning because ρ is not anymore related to a partial rank ; it will however be shown in Section 4.3 that it allows the inversion of Equation (4.2), in order to estimate ρ from the data.

We choose arbitrarily to model the evolution of ρ along the compass as follows:

$$\rho_\phi(m) = \frac{\kappa}{2} \cdot \left(1 - \mathrm{erf}\left(\frac{m - m_0}{\alpha}\right)\right) + 1, \tag{4.5}$$

with erf the error function, and $\phi = \{\kappa, m_0, \alpha\}$ the set of parameters. Note that ρ_ϕ is indexed by the note m, and not by the note $m + 12$ (*cf.* Equation (4.2)). It is then

defined for $m \in [21, 96]$. κ is related to the value of the asymptote in the low bass range. m_0 is a parameter of translation along m and α rules the slope of the decrease. This model expresses the fact that the amount of stretching inherent to the octave type choice is decreasing from the low bass to the high treble range.

It may be justified by the fact that the perception of the pitch of complex tones is not only based on the first partial of the notes, but on a set of partials contained in a "dominant region" of the human hearing [Moore et al., 1984; Ritsma, 1967; Plomp, 1967]. For bass tones (with fundamental frequencies around 100 to 400 Hz, *i.e.* in the range G2-G4, $m \in [43, 67]$), this dominant region covers the third to fifth partials [Ritsma, 1967]. While going up to the treble part of the compass, the dominant region tends to be localized on the partials with a lower rank. For tones having a first partial frequency above 1400 Hz (*i.e.* for a higher note than F6, $m = 89$) the perception of the pitch is mainly linked to the first partial [Plomp, 1967]. Further research on the pitch perception of inharmonic tones consistently shown that the pitch perception is based on one dominant partial whose rank is decreasing from the bass (6 for A1) to the treble range (2 for C♯6) [Järveläinen et al., 2002]. Moreover, the perceived pitch frequency has been found equal to the frequency of the dominant partial divided by its rank (such as the sixth partial frequency divided by six for the A1). In accordance with those studies, in the presented octave type model the high treble asymptote is set to 1. It corresponds to the minimal octave type (2:1), and means that the tuner focuses on the first partial of the highest note. In the low bass range, the asymptote is set by the value of the parameter $\kappa + 1$.

The model is represented on Figure 4.4(b) for three different values of the set of parameters: ϕ_1, ϕ_2 and ϕ_3, respectively corresponding to a low, mid and high octave type choice in the low bass range.

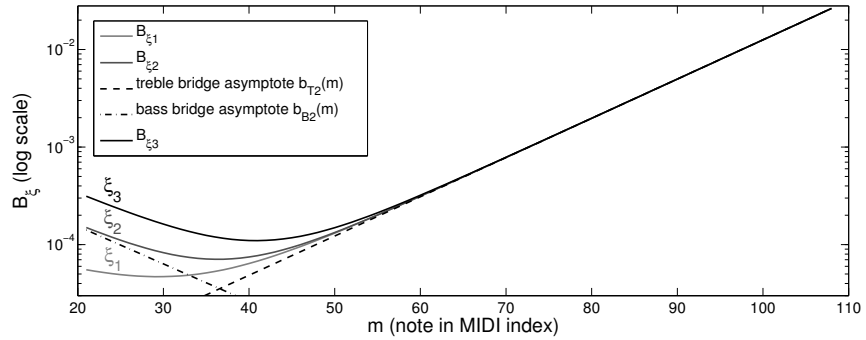
4.2.4 Interpolation of the tuning along the whole compass

From the estimation of the sets of parameters, ξ related to the design of the strings, and ϕ related to the choices of the tuner, it is possible to tune all the octaves of a reference note. If A4 is tuned such that $f_1(m = 69) = 440$ Hz, all the A notes of the keyboard can be iteratively tuned by using Equation (4.2). To complete the tuning on the whole compass, a Lagrange polynomial interpolation is performed on the deviation from ET of the tuned notes of the model (computed by using Equation (4.1)). The interest of this method is that the interpolated curve is constrained to coincide with the initial data. The interpolated model of deviation from ET is denoted by $d_{\xi,\phi}(m)$.

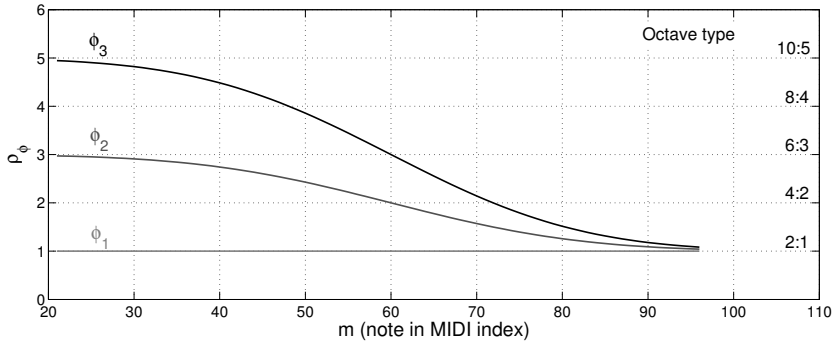
4.2.5 Global deviation

Finally, in order to take into account the fact that the reference note is not necessarily a A4 at 440 Hz (other tuning forks exist, for instance A4 at 442 Hz or C5 at 523.3 Hz) we add in the model the possibility of a global "detuning". In the representation of the deviation from ET in cents, it corresponds to a vertical translation of the curve. Then, the deviation from ET of the model is set to $d_{\xi,\phi}(m) + d_g$, where d_g is an extra parameter of the model, corresponding to the global deviation.

The whole compass tuning model is depicted on Figure 4.5 for different values of the sets of parameters ξ and ϕ (corresponding to those used in Figure 4.4), and for $d_g = 0$. The tuning of the A notes from A4 at 440 Hz is indicated with black dots on the middle curves. Sub-figure 4.5(a) corresponds to the influence on the tuning of B_ξ (for ξ_1, ξ_2 and

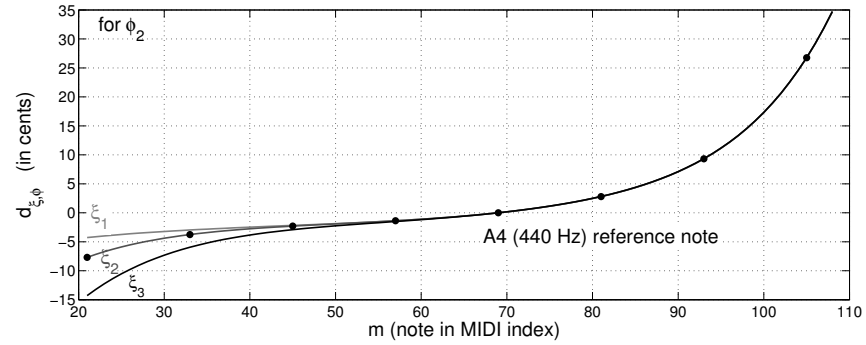


(a)

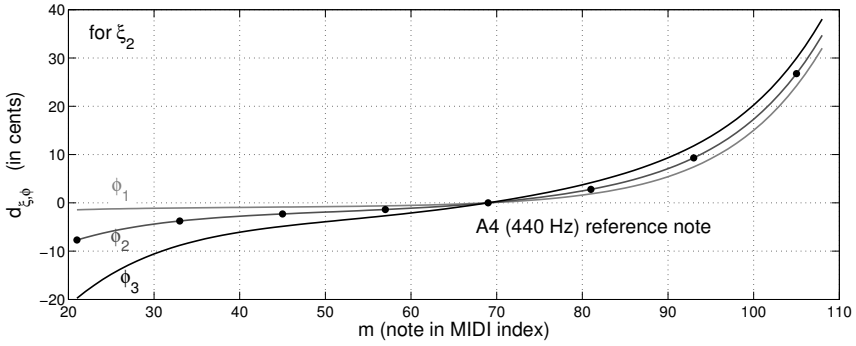


(b)

Figure 4.4: Model for (a) the inharmonicity coefficient $B_\xi(m)$ and (b) octave type parameter $\rho_\phi(m)$ along the compass for different values of the sets of parameters.



(a)



(b)

Figure 4.5: Model for the deviation of tuning from ET along the compass. (a) Influence of ξ in the tuning for ϕ fixed. (b) Influence of ϕ in the tuning for ξ fixed. The different values for ξ and ϕ correspond to those used to generate the curves of Figure 4.4.

ξ_3), for ϕ_2 fixed. Since the string design is standardized in the range C4-C8, the tuning changes significantly only in the bass range. Sub-figure 4.5(b) represents the influence on the tuning of ρ_ϕ (for ϕ_1, ϕ_2 and ϕ_3), for ξ_2 fixed. Its influence is visible on the whole compass but it is mainly important in the bass range, where it can produce a deviation up to -20 cents.

4.3 Parameter estimation

The estimation of the above presented models considers as input $(B(m), F_0(m))$ values that have been previously estimated, for instance by using the algorithms proposed in Chapter 3.

B_ξ estimation: We first estimate the fixed parameters $\{s_T, y_T\}$, corresponding to the string set design on the treble bridge and being almost equal for all the models of pianos, by using $B(m)$ estimates of 6 different pianos (the databases are presented in Section 4.4) in the range C4 (m=60) - C8 (m=108). These are obtained by an L1 regression (in order to reduce the influence of potential outliers), *i.e.* by minimizing the absolute deviation, between the model and the average of the estimated inharmonicity curves over the 6 different pianos. We find $s_T \simeq 9.26 \cdot 10^{-2}$, $y_T \simeq -13.64$. These results are in accordance with estimates based on physical considerations [Young, 1952]: $s_{T[\text{Yo52}]} \simeq 9.44 \cdot 10^{-2}$, $y_{T[\text{Yo52}]} \simeq -13.68$.

Each piano is then studied independently to estimate the particular parameters $\xi = \{s_B, y_B\}$ on a set of notes M. ξ is estimated minimizing the absolute deviation between $\log B(m)$ and $\log B_\xi(m)$:

$$\hat{\xi} = \arg\min_{\xi} \sum_{m \in M} |\log B(m) - \log B_\xi(m)|. \tag{4.6}$$

ρ_ϕ estimation: For each piano, the data $\rho(m)$ is estimated for $m \in [21, 96]$ from $(B(m), F_0(m))$ values by inverting Equation (4.2):

$$\rho(m) = \sqrt{\frac{4F_0(m)^2 - F_0(m+12)^2}{F_0(m+12)^2 B(m+12) - 16F_0(m)^2 B(m)}}. \tag{4.7}$$

Then, the set of parameters ϕ is estimated by minimizing the least absolute deviation distance between $\rho_\phi(m)$ and $\rho(m)$ on a set M of notes:

$$\hat{\phi} = \arg\min_{\phi} \sum_{m \in M} |\rho(m) - \rho_\phi(m)|. \tag{4.8}$$

d_g estimation: Once the ξ and ϕ sets of parameter have been estimated, the octaves of the reference note are tuned according to Equation (4.2). Then, the deviation from ET of the model $d_{\xi,\phi}$ is obtained on the whole compass after the Lagrange interpolation stage. Finally, d_g is estimated by minimizing the absolute deviation, on the reference octave F3-F4 ($m \in [53, 65]$) between $d(m)$, the deviation from ET estimated on the data (see Eq. (4.1)), and $d_{\xi,\phi}(m) + d_g$:

$$\hat{d}_g = \arg\min_{d_g} \sum_{m=53}^{65} \big|d(m) - (d_{\xi,\phi}(m) + d_g)\big|. \tag{4.9}$$

4.4 Applications

The results presented in this section are obtained from 6 different pianos (the Iowa and the 3 RWC grand pianos, and the MAPS ENSTDkCl upright piano and SptkBGCl grand piano synthesizer) of the database presented in Section 3.1.3.1, page 48. The estimates of (B, F_0) are obtained by using *InhR-NMF* algorithm on isolated note recordings in a supervised way (*cf.* Section 3.1).

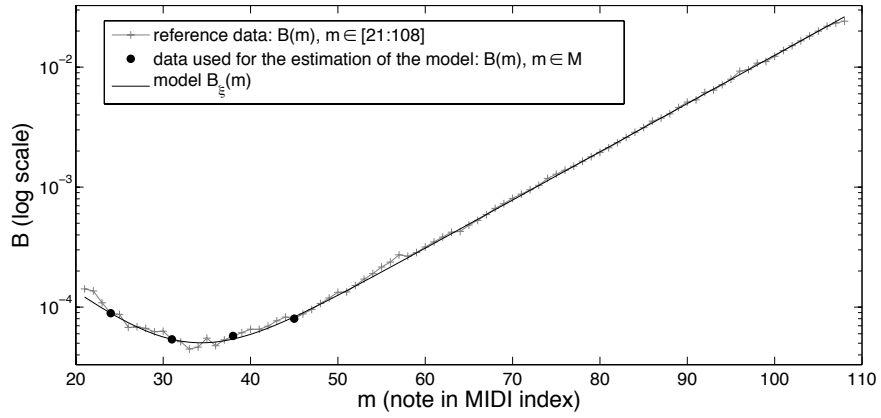
4.4.1 Modeling the tuning of well-tuned pianos

The results of the estimation of the whole compass tuning model for four different pianos are presented on Figures 4.6 (RWC3), 4.7 (RWC2), 4.8 (Iowa) and 4.9 (MAPS SptkBGCl). Sub-figures (a), (b) and (c), correspond to the inharmonicity coefficient B, the octave type parameter ρ and the deviation from ET curves along the whole compass, respectively. The data corresponding to the estimation of (B, F_0) from isolated note recordings is depicted as '+' markers, and the model as black lines. The values of the parameters for each piano are given in Table 4.1.

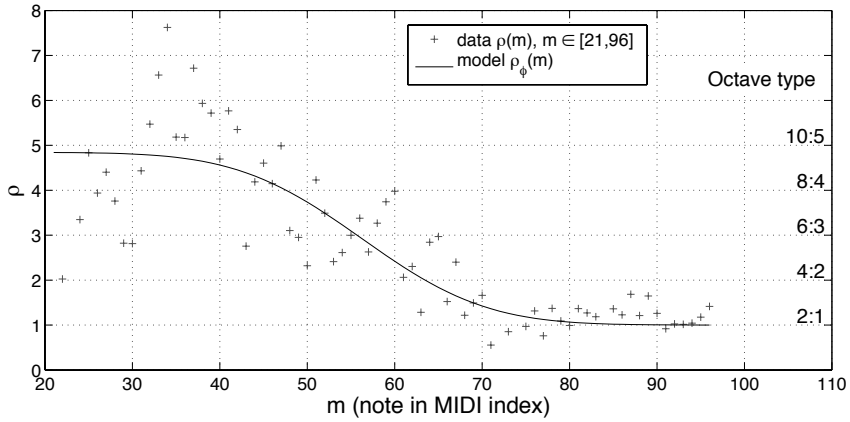
B along the compass (sub-figures (a)): The estimation of the parameters has been performed from a limited set of 4 notes (black dot markers), taken in the bass range and equally spaced by fifth intervals. As the string set design on each bridge is quite regular, a few notes can be used to correctly estimate the model. In the case where an important discontinuity is present in the variations of $B(m)$ (for instance between C2 ($m = 37$) and D2 ($m = 38$) notes, on Figure 4.7(a)) the 2-bridges additive model produces a smooth curve. It is worth noting from RWC2 grand piano design characteristics that the slight jump between D♯1 ($m = 27$) and E1 ($m = 28$) might be explained by the single string to doublet of strings transition, and the important jump between C2 ($m = 37$) and D2 ($m = 38$) by the bridge change, jointly with the transition from doublet of strings to triplet.

ρ along the compass (sub-figures (b)): The curves of $\rho(m)$ can present a significant dispersion around the mean model $\rho_\phi(m)$, but the global variations are well reproduced. In the medium range, the estimated octave types are a trade-off between 6:3 and 4:2, which is common in piano tuning [Bremmer, 2007a]. The variations, more important in the bass range, could be explained by the fact that the model of the partial frequencies (*cf.* Equation (2.23)) does not take into account the frequency shifts caused by the bridge coupling, mainly appearing in the low frequency domain. Moreover, the proposed tuning model is a simplification of a real tuning procedure, it is based on octave interval tuning, while an expert tuner would jointly control different intervals along the keyboard and can do local readjustments after a global tuning. Note that some values of $\rho(m)$ can be missing when the quantity under the square root of Equation (4.7) is negative. This happens if the corresponding octave interval is compressed instead of being stretched.

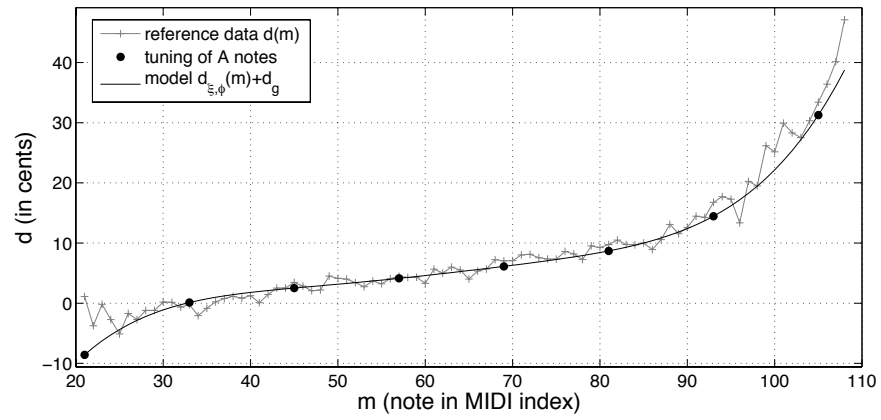
Deviation from ET along the compass (sub-figures (c)): The curves demonstrate that the model reproduces the main variations of the tuning in a satisfactory manner. This confirms that, besides the well-known influence of the inharmonicity on the tuning, perceptual effects (taken into account through the octave type consideration) can take part in the stretching, mainly in the bass range. Note that the tuning of A notes is marked with black dot markers.



(a)

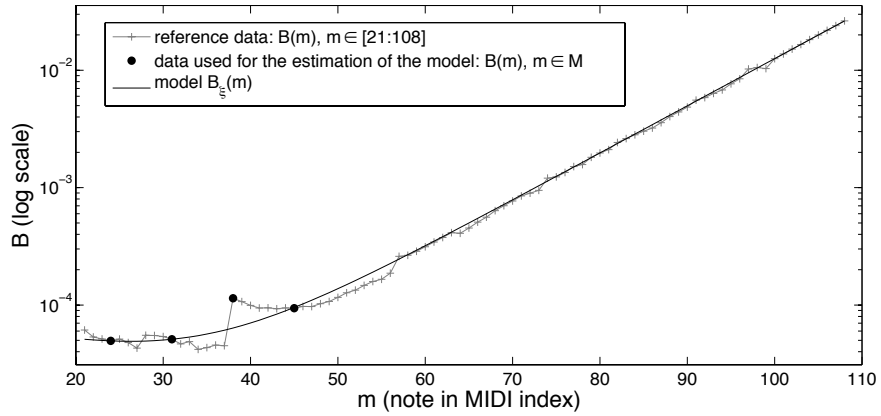


(b)

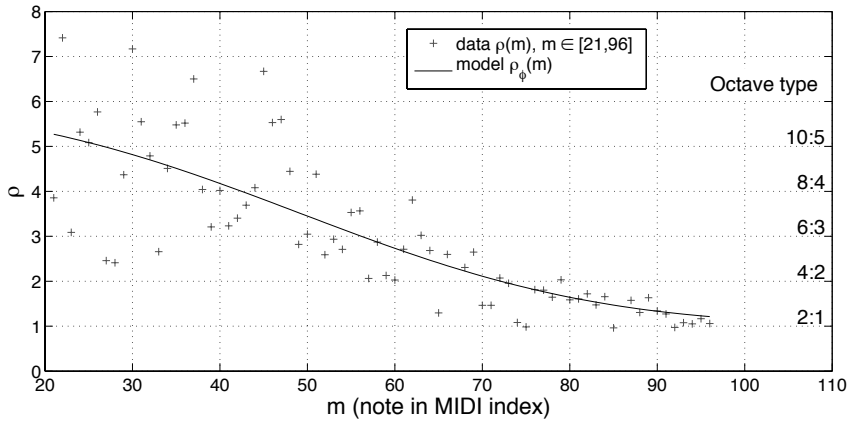


(c)

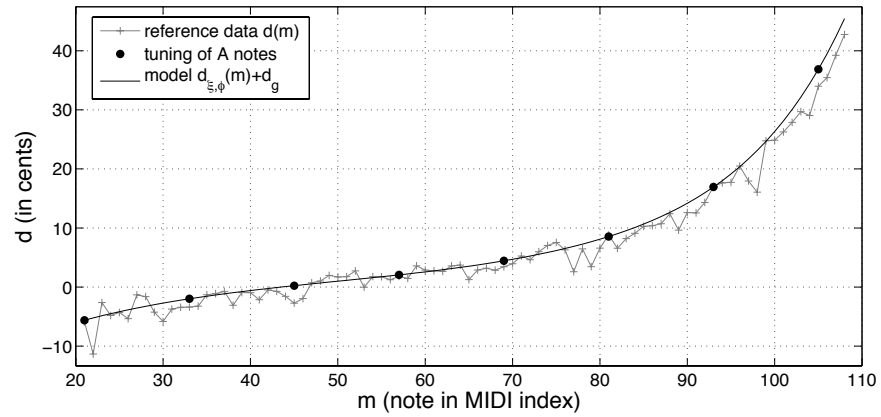
Figure 4.6: RWC3 grand piano. (a) Inharmonicity coefficient B, (b) octave type parameter ρ, (c) deviation from ET along the whole compass. The data are depicted as gray '+' markers and the model as black lines.



(a)

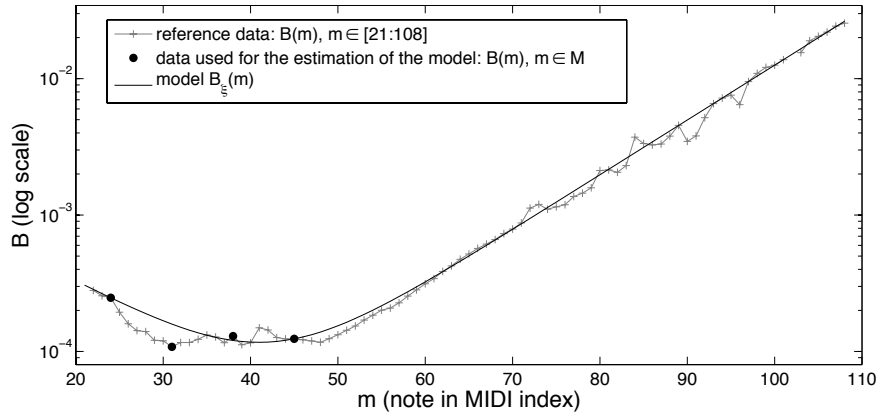


(b)

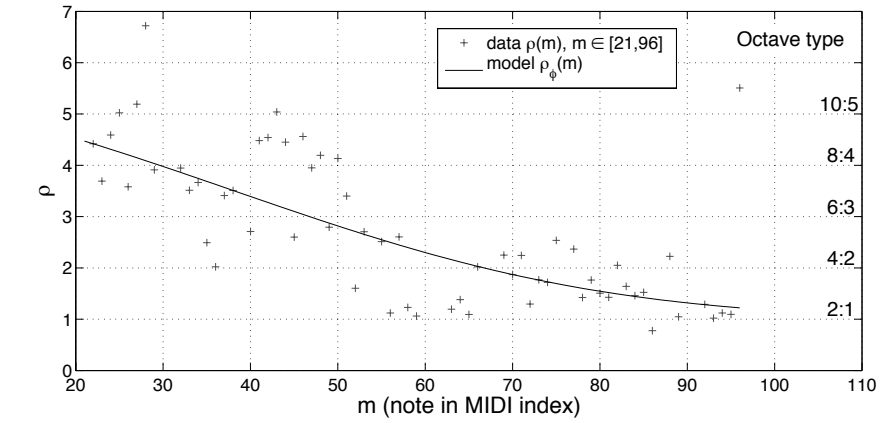


(c)

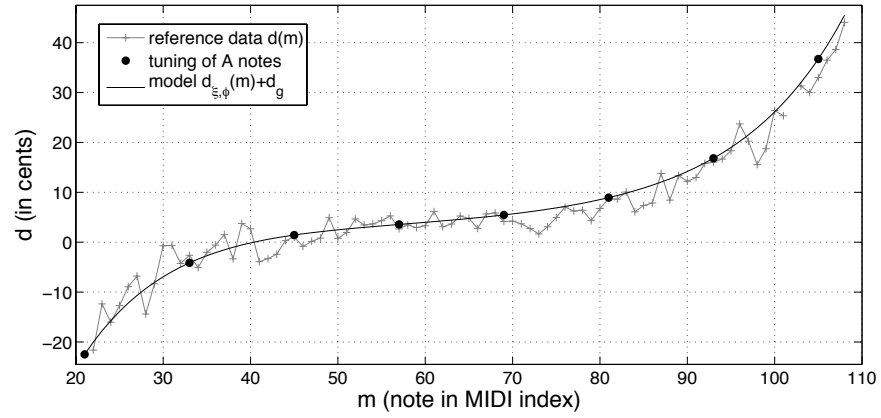
Figure 4.7: RWC2 grand piano. (a) Inharmonicity coefficient B, (b) octave type parameter ρ, (c) deviation from ET along the whole compass. The data are depicted as gray '+' markers and the model as black lines.



(a)

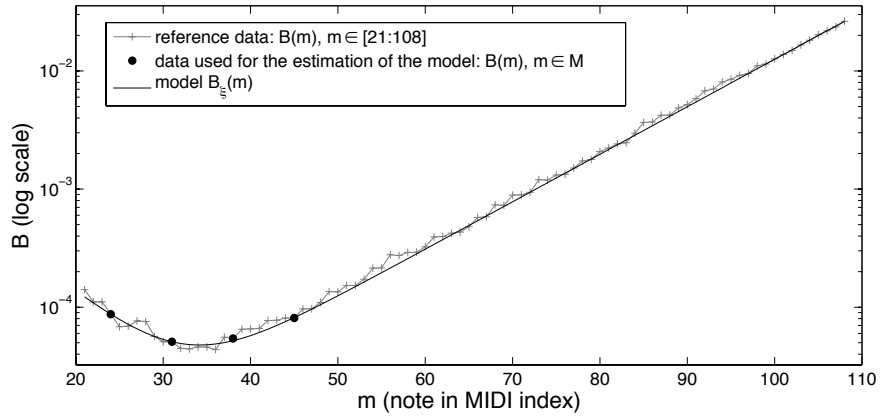


(b)

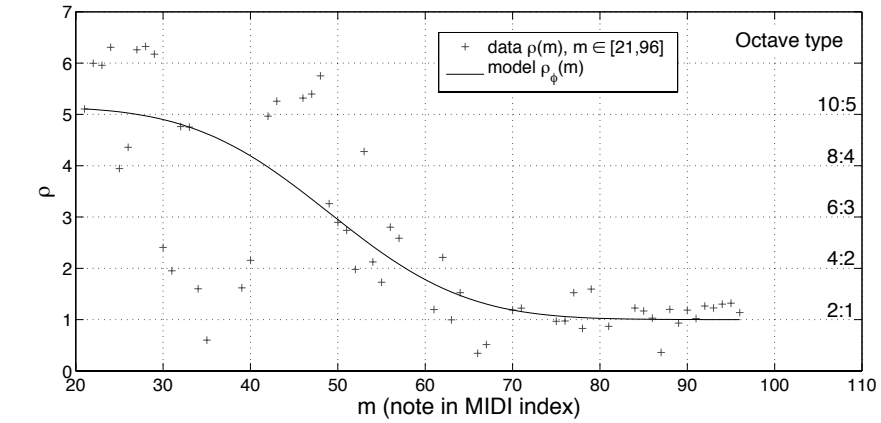


(c)

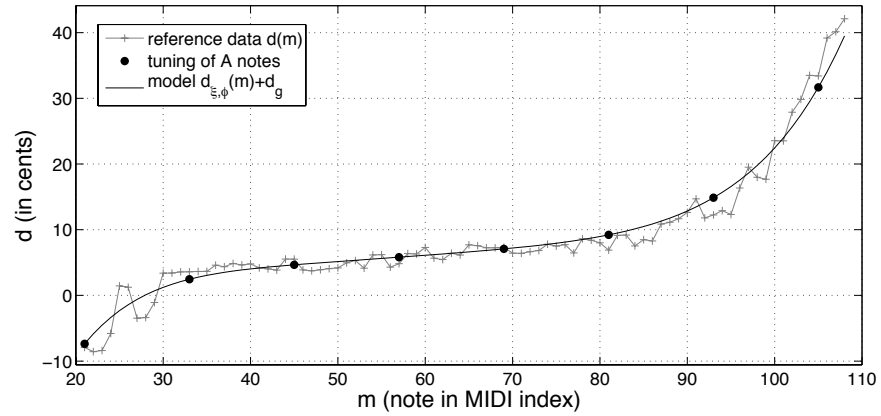
Figure 4.8: Iowa grand piano. (a) Inharmonicity coefficient B, (b) octave type parameter ρ, (c) deviation from ET along the whole compass. The data are depicted as gray '+' markers and the model as black lines.



(a)



(b)



(c)

Figure 4.9: MAPS SptkBGCl. (a) Inharmonicity coefficient B, (b) octave type parameter ρ, (c) deviation from ET along the whole compass. The data are depicted as gray '+' markers and the model as black lines.

	RWC3	RWC2	Iowa	SptkBGCl
s_B	-0.1239	-0.0354	-0.0773	-0.1340
y_B	-6.486	-9.319	-6.497	-6.265
κ	3.470	4.887	13.26	6.296
m_0	58.03	50.22	-13.57	37.76
α	15.93	36.50	80.32	27.40
d_g	6.067	4.491	5.354	6.907

Table 4.1: Values of the parameters for the 4 well-tuned pianos.

4.4.2 Tuning pianos

Because the model of octave type choice $\rho_\phi(m)$ is defined for well-tuned pianos (the stretching of the octaves is implicitly assumed to be higher than 2), it cannot be used to study the tuning of strongly out-of-tune pianos. In this case, we generate tuning curves deduced from a mean model of octave type choice. The model is obtained by averaging the curves $\rho(m)$ over three pianos (RWC2, RWC3 and Iowa grand pianos), that were assumed to be well-tuned, by looking at the shape of their deviation from ET curves. From this averaged data, a mean model $\bar{\rho}_\phi(m)$ is estimated. In order to give a range of fundamental frequencies in which the pianos could be reasonably re-tuned, we arbitrarily define a high (respectively low) octave type choice as $\bar{\rho}_{\phi,H}(m) = \bar{\rho}_\phi(m) + 1$ (resp. $\bar{\rho}_{\phi,L}(m) = \min(\bar{\rho}_\phi(m) - 1, 1)$) . These curves are shown on Figure 4.10.

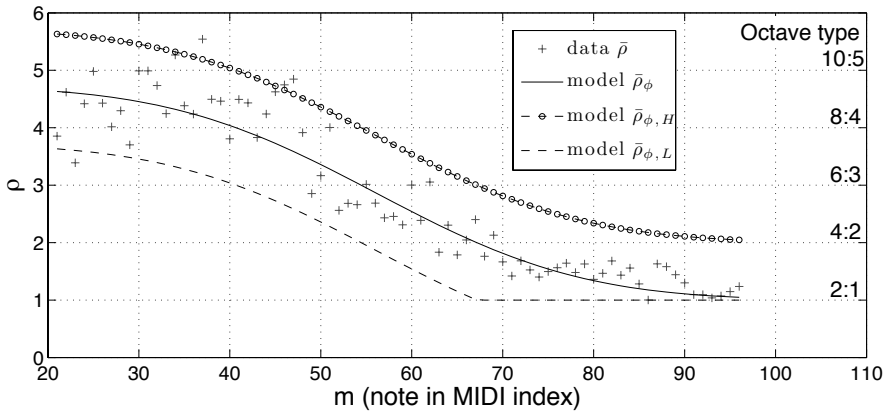


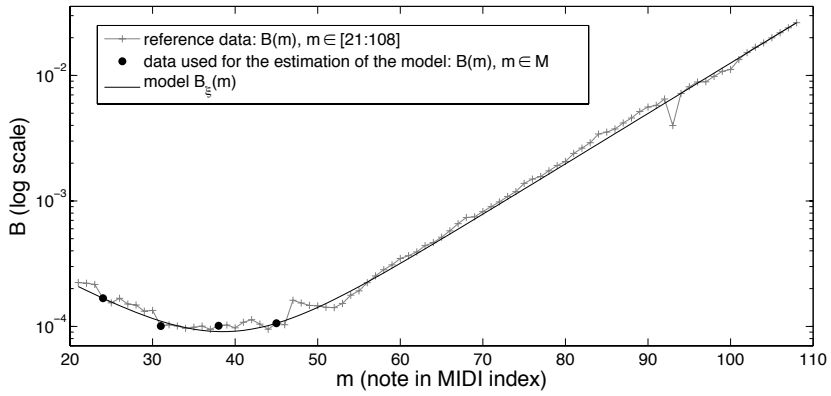
Figure 4.10: Mean octave type choice for tuning application. '+' gray markers correspond to an average of $\rho(m)$ over 3 different pianos. The black line corresponds to the estimated model. Circle markers (resp. in dashed line) represent the high (resp. low) octave type choice model.

Tuning curves are then computed from the estimation of ξ and $\bar{\rho}_\phi(m)$. The global deviation parameter d_g is set to 0. The values of the parameters for each piano are given in Table 4.2. The results are presented on Figure 4.11 and 4.12, respectively for RWC1 grand piano and MAPS ENSTDkCl upright piano. The current tuning is depicted as '+' gray markers and clearly shows that the piano is not well-tuned, mainly in the bass range where the tuning is "compressed". The space between the tuning curves obtained

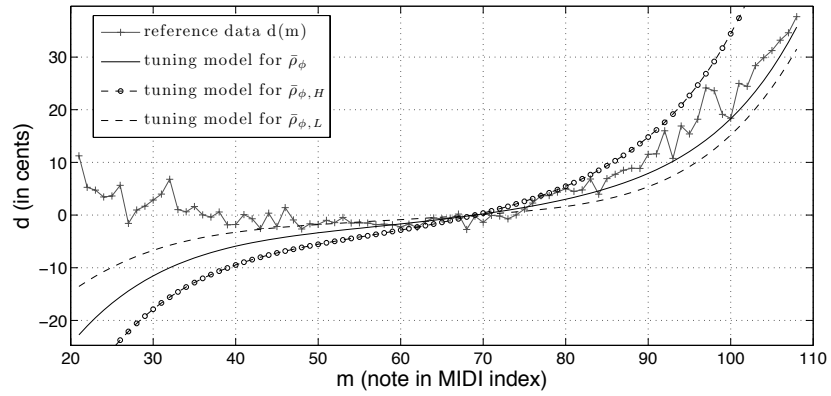
from $\bar{\rho}_{\phi,H}(m)$ and $\bar{\rho}_{\phi,L}(m)$ corresponds to a range in which we assume the piano could be well-tuned. For a quantitative interpretation, it will be interesting to compare our curves with those obtained after a re-tuning done by a professional tuner.

	RWC1	ENSTDkCl
s_B	-0.0808	-0.0864
y_B	-6.823	-6.336
$\bar{\kappa}$	3.715	
$\bar{m_0}$	56.59	
$\bar{\alpha}$	25.15	
d_g	0 (fixed)	

Table 4.2: Values of the parameters for the 2 detuned pianos.



(a)



(b)

Figure 4.11: RWC1 grand piano. (a) Inharmonicity curves along the compass. (b) Actual tuning and proposed tuning. '+' gray markers correspond to the data. The model corresponding to the octave type choice $\bar{\rho}_{\phi}(m)$ (resp. $\bar{\rho}_{\phi,L}(m)$, $\bar{\rho}_{\phi,H}(m)$) is depicted as black line (resp. black dashed line, black dashed line with circle markers).

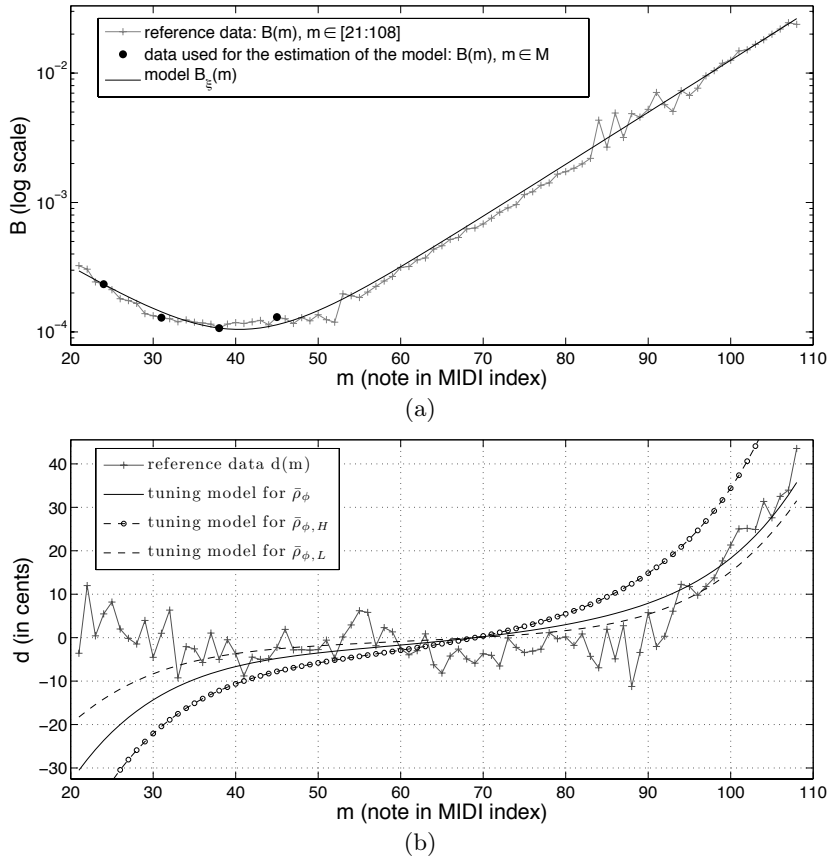


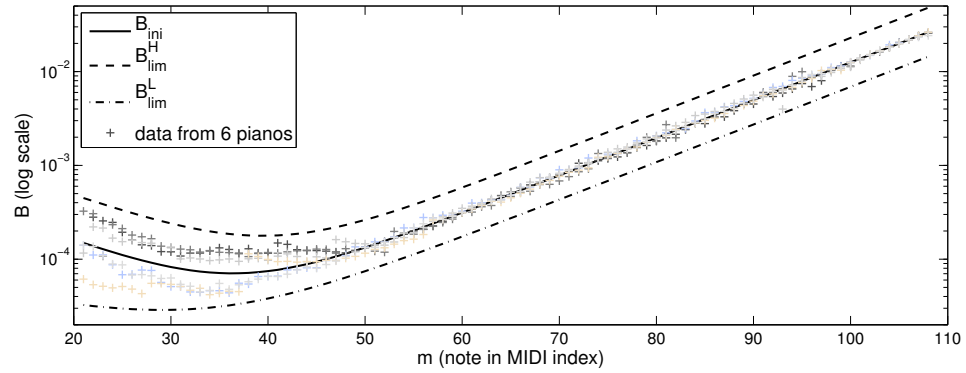
Figure 4.12: MAPS ENSTDkCl upright piano. (a) Inharmonicity curves along the compass. (b) Actual tuning and proposed tuning. '+' gray markers correspond to the data. The model corresponding to the octave type choice $\bar{\rho}_\phi(m)$ (resp. $\bar{\rho}_{\phi,L}(m)$, $\bar{\rho}_{\phi,H}(m)$) is depicted as black line (resp. black dashed line, black dashed line with circle markers).

4.4.3 Initializing algorithms

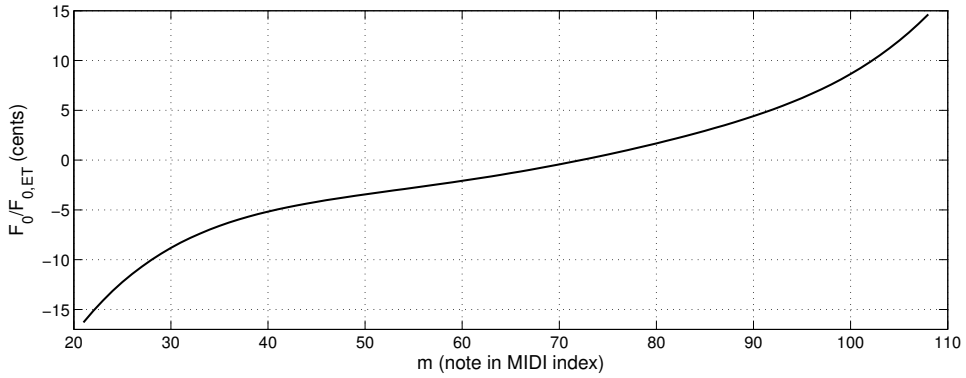
As show in Chapter 3, a good initialization of (B, F_0) parameters in analysis algorithms is a matter of importance in order to avoid the convergence of the algorithms toward local optima. The curves for the initialization of these parameters are here obtained by considering a mean model, obtained from the 6 pianos analyzed in this chapter. These are presented on Figure 4.13. Subplot 4.13(a) depicts in black line the initialization curve B_{ini}, estimated from the data averaged over the 6 pianos ('+' markers). The low and high limit curves (respectively denoted by B^L_{lim} and B^H_{lim}) used in Section 3.2.2.3 for the *PLS* algorithm are obtained by means of a manual tuning of ξ so that all estimates are contained in the delimited area (values given in Table 4.3). Then, the initialization curve of F_0 is computed from B_{ini} and $\bar{\rho}_\phi$, the mean octave type obtained in the previous section.

	B_{ini}	B^L_{lim}	B^H_{lim}
s_B	-0.0889	-0.047	-0.083
y_B	-7.0	-9.5	-6.0

Table 4.3: Values of the parameters for the initialization and limit curves of B.



(a)



(b)

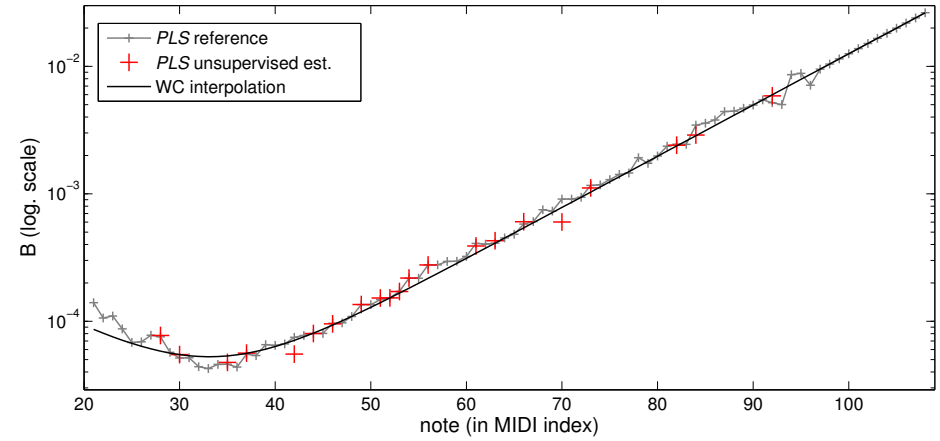
Figure 4.13: Mean model for the initialization of (B, F_0) in piano analysis algorithms. (a) '+' markers correspond to the values of B for the 6 different pianos. In black line, B_{ini}, the model estimated from the estimates averaged over the 6 pianos. In dashed lines, the limit curves for B values. (b) Mean model for F_0 computed from B_{ini} and $\bar{\rho}_\phi$.

4.4.4 Learning the tuning on the whole compass from the analysis of a piece of music

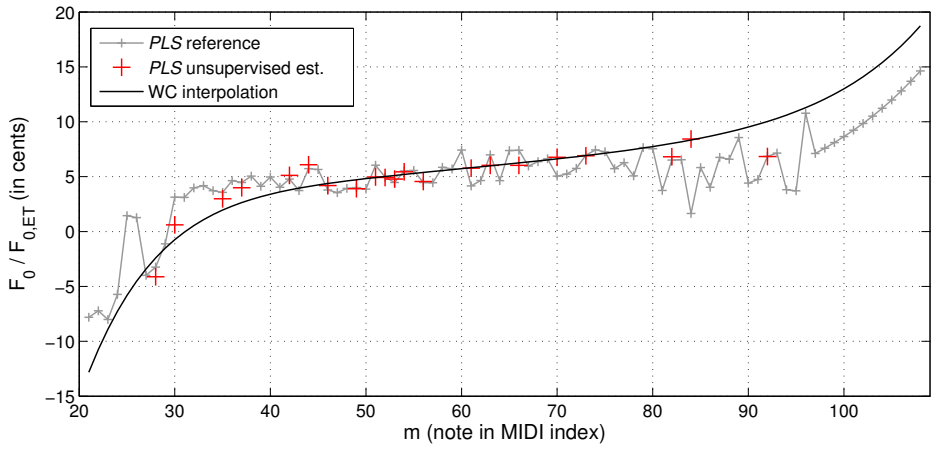
Finally, the whole compass model is applied to the interpolation of (B, F_0) estimates obtained on a restricted set of notes from the analysis of a piece of music. The results are depicted in Figure 4.14 for the example presented in Section 3.2.3.2. The input data is here depicted as '+' red markers and the interpolated curves are plotted as black lines. The gray curve, displayed for comparison purposes, corresponds to the estimates along the whole compass obtained by the *PLS* algorithm from isolated notes processed in a supervised way (the algorithm knows which notes are being played). When comparing the interpolated model and the reference curve, one can see that the main trends are fairly well reproduced. Averaged below note $C\sharp 7$ (97) (*i.e.* for the range where the reference data is reliable, as discussed in Section 3.2.3.1), a mean relative error of 9.48 % is obtained for B and a mean deviation of 2.20 cents for F_0.

It is worth noting that a few requirements are needed for this interpolation application. First, since the variations of B between different pianos are mainly present for the notes

associated to the bass bridge, the interpolation requires estimates in the bass range. One can see on Figure 4.14(a), that 4 notes were present below note E2 (40) in the input data. Here, one can assume that the interpolation may have been improved with a few more estimates below note E1 (28). Second, since the tuning model, given in Equation (4.2), is based on the tuning of octave intervals, the input data has to contain, as much as possible, (B, F_0) estimates for octave-related notes. In the presented example, 6 octave intervals were present. Thus, in order to increase the number of input data, that should lead to a higher precision of the interpolation, one could consider processing several pieces of music played by the same instrument.



(a)



(b)

Figure 4.14: (B, F_0) interpolation along the whole compass from a few estimates obtained on a piece of music with *PLS* algorithm (*cf.* Section 3.2.3.2). (a) B in log. scale and (b) F_0 as dev. from ET in cents. The whole compass models (black lines, denoted as "WC interpolation") are estimated from the data obtained by the unsupervised analysis of a piece of music (red markers). These are compared with the reference curves (gray markers) obtained from the supervised analysis on isolated note recordings.

Chapter 5

Application to the transcription of polyphonic piano music

This chapter presents the evaluation on a transcription task of the two inharmonic NMF-based models introduced in Chapter 3. In order to quantify the benefits or losses that may result from the inclusion of the inharmonicity, the performances are compared with those obtained by a simpler harmonic model. Influence of the model design, the initialization and the optimization are then investigated. Portions of this work have been published in [Rigaud et al., 2013c].

5.1 Automatic polyphonic music transcription

This section introduces briefly the music transcription task, and defines the metrics commonly used when assessing the performance of the methods. For a complete description and an exhaustive state of the art of this domain, the interested reader may refer to [Klapuri and Davy, 2006; Emiya, 2008; Bertin, 2009; Benetos, 2012].

5.1.1 The task

Automatic music transcription is the process of recovering a symbolic representation of a musical piece, such as for instance a score, a chord grid or a guitar tablature, from the unsupervised analysis of a performance recording. According to the degree of information which is targeted, the problem is usually decomposed into several sub-tasks, each focusing on the extraction of a specific feature related to the musical structure. These can be for instance the recovery of the played notes (their pitch, loudness, onset time and duration), but also higher-level information, such as the instruments, the rhythmic structure (*e.g.* tempo and time signature), the tonality, or the detection of musical sequence repetitions.

Automatic music transcription, and more generally tasks related to Music Information Retrieval, have become a very active field of research this last decade due to their wide range of applications. Beyond straightforward uses as supports for musical practice or musicological analysis, applications include automatic music data indexing for classification/search purposes. Also, when processed in real-time during a musical performance, the extracted information can be used as control inputs in human-machine interactive systems (*e.g.* score following or rhythm tracking for automatic accompaniment).

In the case of the analysis of a Western tonal piece of music, the most complete transcription task corresponds to the retrieval of the score. However as mentioned above, this complex representation requires the processing of a large number of sub-tasks. Thus, in this thesis as in most works related to music transcription using NMF-based models, we restrict the transcription problem to the detection of the notes (pitch) that are played, jointly with their onset and duration. Thus the obtained representation is usually displayed in the form of a piano roll (*i.e.* binary activation of the notes according to the time axis, as shown in Figure 5.1).

As seen in Chapter 2, NMF-based methods seem particularly relevant for the task of *audio to piano roll* transcription as the activation matrix H naturally presents a structure similar to the desired representation. However, a binary decision (note active/inactive) along the lines of H is required after the NMF optimization. Such post-processing may include additional temporal modeling of notes, based for instance on Hidden Markov Models [Poliner and Ellis, 2007; Emiya et al., 2010a], or heuristics based on thresholding and gathering of adjacent activated frames [Vincent et al., 2010; Dessein et al., 2010].

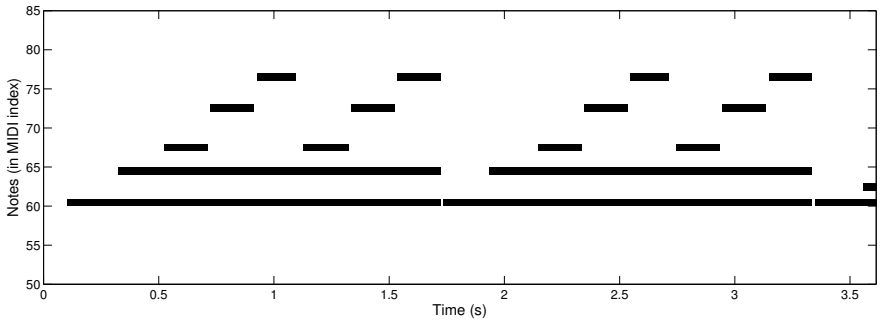


Figure 5.1: Piano roll of the first bar of the score given in Figure 2.1.

5.1.2 Performance evaluation

The performance evaluation of transcription algorithms requires a ground truth, for instance in the form of a synchronized MIDI file. Such reference may be obtained after a manual annotation of the data (as for the RWC database), or by synthesizing the music signals directly from the MIDI reference file. This latter case presents the advantage of ensuring the near-perfect alignment between the audio and the ground truth and of avoiding errors inherent in the manual annotation process. This is the case for the MAPS database used for our experiments, and for which the generation of the audio is using either piano synthesizers based on high quality sampling, or a MIDI controlled acoustic piano (*cf.* details in Section 3.1.3.1).

Thus, when evaluating the performance, each note detected by the algorithm can be classified into two classes. If the note is actually present in the ground truth, it is labeled as True Positive (TP). If not, it corresponds to a false detection and it is labeled as False Positive (FP). Finally, the notes that are missing (*i.e.* not detected by the algorithm, although present in the ground truth) are labeled as False Negative (FN). These classes, common in classification task evaluations, are summarized in Table 5.1. Note that a fourth category True Negative (TN) exists but it is not used in a transcription context since it is not relevant. Indeed, it corresponds to a non-detection when the note is not present in the ground truth.

Such classification of the results may be performed according to two evaluation configurations. In a frame-wise evaluation, the notes detected by the algorithm are compared with the reference in each time-frame, independently (as for instance performed in Section 3.2.3.2). However, this type of evaluation is more adapted to multi-pitch algorithms that do not target the retrieval of the note onset and duration. Thus, a note-wise evaluation is often preferred for transcription tasks. In that case, a note is considered as TP if its onset is contained in a time interval centered in the ground truth onset (usually +/− 50 ms).

		Ground truth (presence of note)	
		True	False
Algorithm (detection of note)	Positive	TP	FP
	Negative	FN	TN

Table 5.1: Classification of the results provided by a transcription algorithm when comparing to the ground truth.

Finally in order to evaluate the performance for a whole piece of music, commonly used metrics are the precision $\mathcal{P}$ and the recall $\mathcal{R}$. These are respectively defined as [Rijsbergen, 1979]:

$$\mathcal{P} = \frac{\#\{\text{TP}\}}{\#\{\text{TP}\} + \#\{\text{FP}\}}, \tag{5.1}$$

$$\mathcal{R} = \frac{\#\{\text{TP}\}}{\#\{\text{TP}\} + \#\{\text{FN}\}}, \tag{5.2}$$

where $\#\{\}$ denotes the cardinality of each set of class obtained over the considered piece of music. As defined by Equation (5.1) the precision evaluates the ability of the algorithm to provide correct detections, no matter the number of forgotten notes (FN). The recall, defined in Equation (5.2), acts as a complementary metric since it gauges the ability of the model to detect all the notes actually present, no matter the number of false notes added (FP). Thus, a combination of both metrics is often used in order to assess the global performance, as for instance the F-measure $\mathcal{F}$ which corresponds to the harmonic mean between $\mathcal{P}$ and $\mathcal{R}$:

$$\mathcal{F} = 2 \cdot \frac{\mathcal{P} \cdot \mathcal{R}}{\mathcal{P} + \mathcal{R}}. \tag{5.3}$$

5.1.3 Issues arising from the inharmonicity inclusion in NMF-based transcription model

As mentioned in Section 2.3, in the case of piano music transcription, taking into account the inharmonicity of the instrument tones in an NMF-based transcription model has been proposed [Vincent et al., 2008], but surprisingly the results were found slightly below those obtained by a simpler harmonic model. These results seem in contradiction with a naive intuition that inharmonicity should help lifting typical ambiguities such as with harmonically-related notes (octave or fifth relations, for instance). The goal of the study presented in this chapter is to have a better understanding about this issue. Although we found in Chapter 3 that the two inharmonic NMF-based models (namely *Inh-NMF* and *InhR-NMF*) were performing well when applied to the supervised (*i.e.* the played

notes were known) estimation of (B, F_0), it is not straightforward that such models may provide good results in the context of unsupervised analysis, *i.e.* when the matrix H has to be estimated jointly with the inharmonicity relation parameters. Since it is difficult to gauge intrinsically the quality of NMF decompositions, these are here evaluated on a transcription task, on a large database of piano recordings from MAPS where we have a ground-truth transcription at hand. Performances of both inharmonic models are evaluated and compared with a similar parametric harmonic model. Influence of the model design (harmonicity vs. inharmonicity and number of partials considered), of the initialization (naive initialization vs. mean model of inharmonicity and tuning along the whole compass) and of the optimization process (number of partials fixed at the initialization vs. iterative inclusion) are then investigated. The parameter of the β-divergence used to define the reconstruction cost-function is here fixed to $\beta = 1$ (KL divergence).

It should be emphasized that the proposed algorithms do not target to be competitive with state-of-the-art fully dedicated piano transcription algorithms, since the only information that is taken into account is the inharmonicity of piano tones (for instance, no model of smooth spectral envelope or temporal continuity of the activations is considered). Here, the use of a simple threshold-based post-processing of H should allow one to better highlight the differences in the core model.

5.2 Does inharmonicity improve an NMF-based piano transcription model?

5.2.1 Experimental setup

5.2.1.1 Database

The dataset consists of 45 pieces from MAPS database (*cf.* details of the pianos in Section 3.1.3.1), randomly chosen (5 out of 30 for each of the 9 pianos) and re-sampled to 22050 Hz. For each piece, 30 second excerpts are taken, starting from $t_0 = 5$ s. The mean polyphony level by time-frame is about 3.23. Then, the magnitude spectrograms are computed with a Hann window of length $\tau = 90$ ms, a hop-size of $\tau/8$ and a 2^{13}-point Fast Fourier Transform.

The number of spectra in the dictionary is fixed to $R = 64$, and initialized for notes having MIDI note number in $[33, 96]$ (A1 to C7). This choice corresponds to a reduction of one octave in the extreme bass (where the spectral resolution is not sufficient to perform the analysis) and one octave in the high treble range (where, as highlighted in Section 3.1.3.2, the non-linear coupling between triplets of strings at the soundboard produces complex spectra with multiple partials that cannot be fully explained by a simple harmonic or inharmonic model). However, these notes in the extreme parts of the keyboard are rarely played. Over the complete MAPS dataset (352710 notes for 159 different pieces), they only account for 1.66% of the notes (*cf.* distribution of notes for the whole MAPS database on Figure 5.2).

5.2.1.2 Harmonicity vs. Inharmonicity

In order to quantify the benefits that may results from the inharmonicity inclusion in NMF-based model, the transcription performances are compared with a simpler harmonic model (thereafter denoted by *Ha-NMF*) based on a similar parametric additive model of

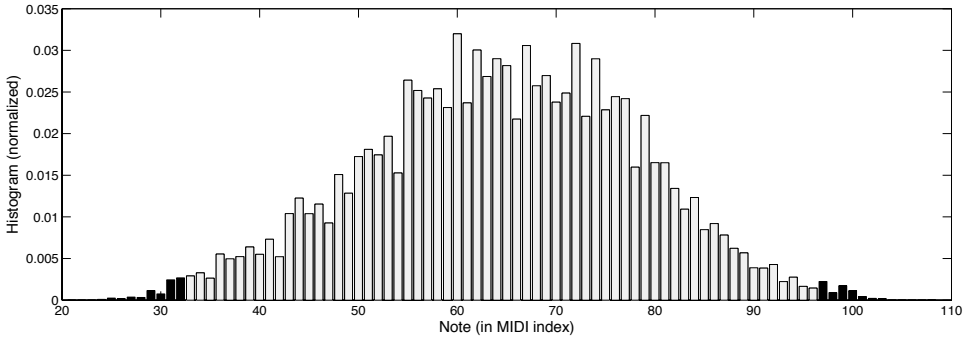


Figure 5.2: Normalized histogram of note occurrences for the complete MAPS database. Black bars corresponds to the notes that are not considered in the analysis.

spectrum (*cf.* Equation (3.1), page 38).

- *Strictly harmonic / Ha-NMF*:
  We enforce harmonicity of the spectra of the dictionary by fixing

$$f_{nr} = nF_{0r}, \quad n \in \mathbb{N}^*, \tag{5.4}$$

directly in the parametric model (Equation (3.1)). Then, the set of parameters for a single atom corresponds to $\theta_r^{\text{Ha}} = \{a_{nr}, F_{0r} \mid n \in [1, N_r]\}$. Similarly to *Inh-NMF*, the cost-function is defined as:

$$C^{Ha}(\theta^{Ha}, H) = \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} d_\beta \left(V_{kt} \mid \sum_{r=1}^{R} W_{kr}^{\theta_r^{Ha}} \cdot H_{rt} \right), \tag{5.5}$$

where, as explained in Section 3.1.1.1, the set of frequency-bins for which the modeled spectrogram is defined is denoted by $\mathcal{K} = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau], \forall n \in [1, N_r], \forall r \in [1, R]\}$.

Then, the following update is obtained for F_{0r} parameters when deriving a similar decomposition of the reconstruction cost-function as for *Inh-NMF*:

$$F_{0r} \overset{Ha}{\leftarrow} F_{0r} \cdot \frac{Q_0^{Ha}(F_{0r})}{P_0^{Ha}(F_{0r})}, \tag{5.6}$$

with

$$\begin{aligned} P_0^{Ha}(F_{0r}) = \sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} \Bigg[& \left(\sum_{n=1}^{N_r} a_{nr} \frac{-n.f_k.g'_\tau(f_k - nF_{0r})}{f_k - nF_{0r}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-1} \\ & + \left(\sum_{n=1}^{N_r} a_{nr} \frac{-n^2.F_{0r}.g'_\tau(f_k - nF_{0r})}{f_k - nF_{0r}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-2}.V_{kt} \Bigg], \end{aligned} \tag{5.7}$$

$$\begin{aligned} Q_0^{Ha}(F_{0r}) = \sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} \Bigg[& \left(\sum_{n=1}^{N_r} a_{nr} \frac{-n.f_k.g'_\tau(f_k - nF_{0r})}{f_k - nF_{0r}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-2}.V_{kt} \\ & + \left(\sum_{n=1}^{N_r} a_{nr} \frac{-n^2.F_{0r}.g'_\tau(f_k - nF_{0r})}{f_k - nF_{0r}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-1} \Bigg], \end{aligned} \tag{5.8}$$

and $\mathcal{K}_r = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau], \forall n \in [1, N_r]\}$.

• Updates for all parametric NMF models:

For all NMF parametric models, the decomposition of the partial derivative of the reconstruction cost-function with respect to H_{rt} parameter leads to the following update:

$$H_{rt} \leftarrow H_{rt} \cdot \frac{Q_0(H_{rt})}{P_0(H_{rt})}, \tag{5.9}$$

with

$$P_0(H_{rt}) = \sum_{k \in \mathcal{K}_r} W_{kr}^{\theta_r} \cdot \widehat{V}_{kt}^{\beta-1}, \tag{5.10}$$

$$Q_0(H_{rt}) = \sum_{k \in \mathcal{K}_r} W_{kr}^{\theta_r} \cdot \widehat{V}_{kt}^{\beta-2} \cdot V_{kt}. \tag{5.11}$$

Note that this latter decomposition is similar to the one commonly used for standard NMF models (*cf.* Equation (2.17)). The difference lies here in the sum over k which is limited to the set of frequency-bins $\mathcal{K}_r$ for which the spectrum model $W_{kr}^{\theta_r}$ is defined. Note that in the following experiments, the activation matrices are initialized with random positive values, as commonly done in NMF applications.

For a_{nr} parameters, the update rules are the same for all parametric models and given in Equation (3.8). In order to obtain a unique decomposition for each method (*i.e.* no scaling transformations of H when running the same algorithm twice on the same data), the a_{nr} are normalized to a maximal value of 1 for each atom indexed by r. Thus, after each update $\forall r \in [1, R]$ and $\forall n \in [1, N_r]$ of the a_{nr}, the following steps are applied:

$$A_r = \max_{n \in [1, N_r]} (a_{nr}), \quad \forall r \in [1, R], \tag{5.12}$$

$$a_{nr} = a_{nr} / A_r, \quad \forall r \in [1, R], \ \forall n \in [1, N_r], \tag{5.13}$$

$$H = \mathrm{diag}(A) \cdot H, \tag{5.14}$$

where $\mathrm{diag}(A)$ is a diagonal matrix of dimension $R \times R$ containing the A_r, $\forall r \in [1, R]$.

5.2.1.3 Post-processing

In order to obtain a list that contains the detected notes, their onset and offset time, a post-processing is applied to the activation matrix H. Each line is processed by a low-pass differentiator filter. The obtained matrix dH is then scaled so that its maximal element is 1. Finally, in each line, an onset is detected if dH increases above a threshold $10^{T_{\mathrm{on}}/20}$, and the corresponding offset found when dH crosses (from negative to positive values) a second threshold $-10^{T_{\mathrm{off}}/20} < 0$ (*cf.* illustration on Figure 5.3). If the same note is found to be repeatedly played at less than 100 ms of interval it is then considered as a unique note. As discussed in Section 5.1.3, this very simple post-processing has been chosen in order to better highlight the differences in the model itself.

It should be noticed that when targeting a transcription algorithm that is competitive with state-of-the-art methods, the optimal values of these thresholds should be first estimated on a learning dataset. As mentioned above, our goal here is to study whether inharmonicity may improve an NMF-based piano transcription model rather than proposing a new competitive algorithm. Then, performances of all methods will be studied on a grid $T_{\text{on}} \in [-50, 1]$ dB, with T_{off} arbitrarily fixed to -80 dB.

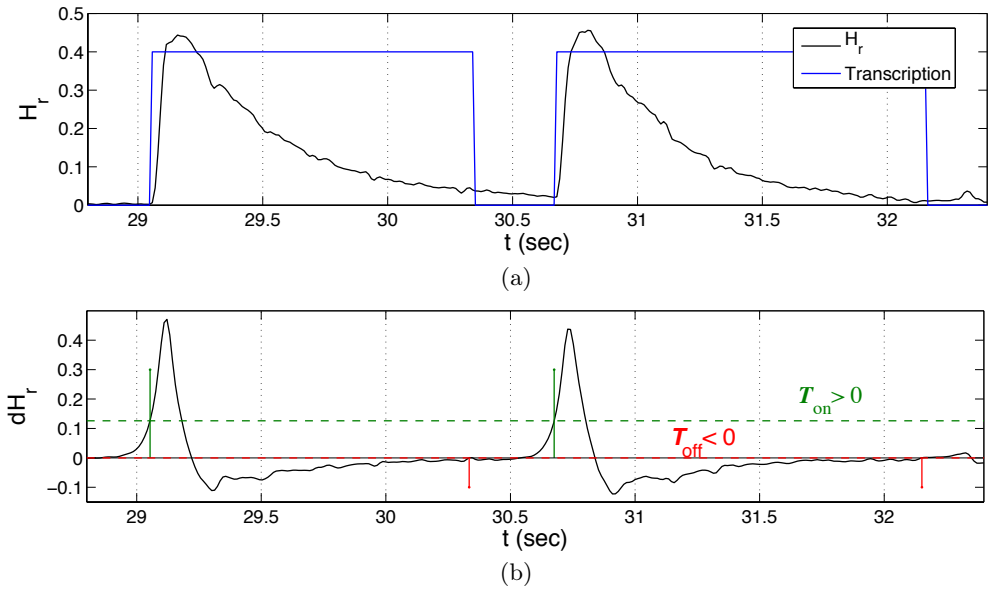


Figure 5.3: Post-processing of the activation matrix H. (a) Results of the note activation decision (blue) for a line of H (black). (b) Illustration of the thresholding (detection of onsets in green and offsets in red) applied to the derivative of the line of H (black).

5.2.2 Supervised transcription

In order to separate the influence of the design of the model from the initialization / optimization process, we compare as a preliminary study all three parametric models (respectively *Ha-NMF*, *Inh-NMF*, and *InhR-NMF*) on a supervised transcription task, *i.e.* by learning first the dictionary on isolated note recordings for each piano of the database and keeping it fixed during the transcription optimization. This study should thus exhibit the performance bounds that may be obtained using such models, in case the amplitude and frequency parameters of the dictionary are properly estimated.

5.2.2.1 Protocol

As a first step, for each piano and each NMF model, the parameters of the dictionary are learned according to the protocol defined for the supervised estimation of (B, F_0) from isolated note recordings (*cf.* Section 3.1.3). The analysis parameters used for the computation of the isolated note spectra are then the same as those used for the computation of the spectrograms to transcribe. Finally, only the activation matrices H are updated according to Equation (5.9) during the optimization of the transcription task (arbitrarily fixed to 50 iterations).

We also add a fourth supervised NMF model, denoted by *Oracle-NMF*, for which the dictionary is composed of the isolated note short-time spectra (H is thus optimized using Equation (2.17)). Since all the parametric dictionaries are limited to the modeling of a restricted set of partials (*e.g.* no consideration of the partials related to longitudinal vibrations of the strings or partials resulting from couplings, but also of other effects such as the hammer stroke noise) this latter supervised model should give a performance bound in case the tones are modeled along the whole frequency axis. Such spectra of the dictionary are illustrated on Figure 5.4 for all four methods.

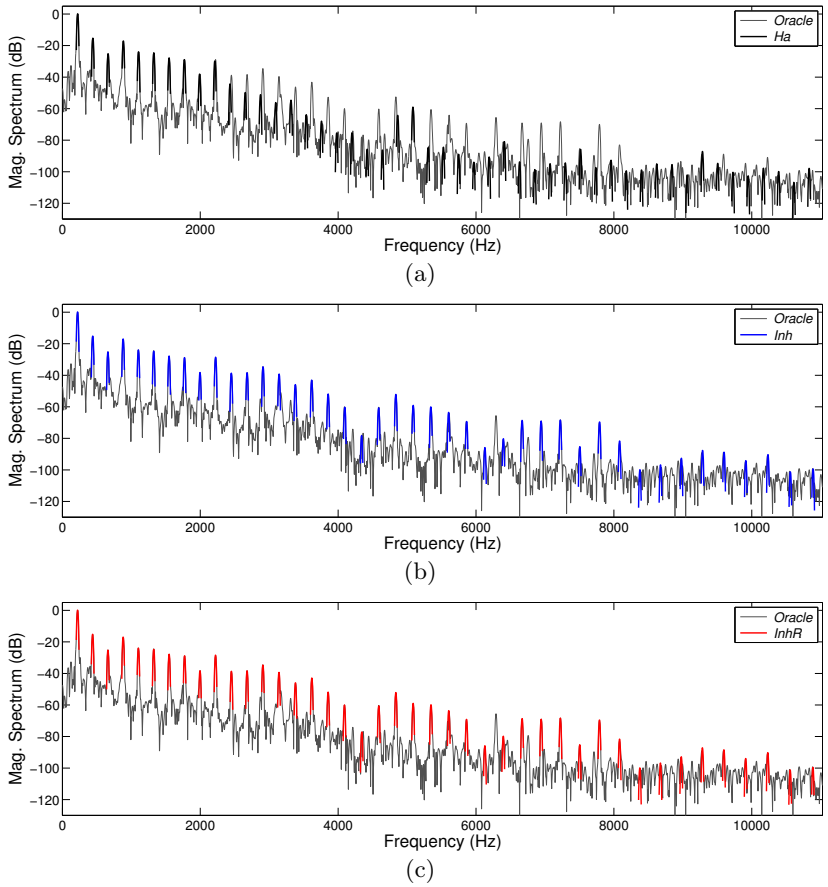


Figure 5.4: Spectrum of the note A3 (MAPS ENSTDkAM piano) for the dictionary of (a) *Ha-NMF*, (b) *Inh-NMF*, and (c) *InhR-NMF*. The original spectrum (in gray) is used for the dictionary of *Oracle-NMF*.

5.2.2.2 Results

The four methods are then applied to the supervised transcription of the 45 excerpts of pieces presented in Section 5.2.1.1. For the 3 parametric models (*Ha/Inh/InhR-NMF*) the influence of N_r, the maximal number of partials for each note, is studied on a grid having values in $\{5, 10, 20, 30, 40, 50\}$.

Precision, Recall and F-measure curves (averaged over the 45 pieces) as a function of the onset detection threshold T_{on} are depicted in Figure 5.5, for $N_r = 20$. When comparing the four methods, one can see that refining the model of piano tone spectra leads to a

higher Precision but does not improve the Recall (consistent for all methods). Thus, it seems that including more information in the dictionary model mainly helps in avoiding the detection of FP notes. This finally results in an increase of the F-measure.

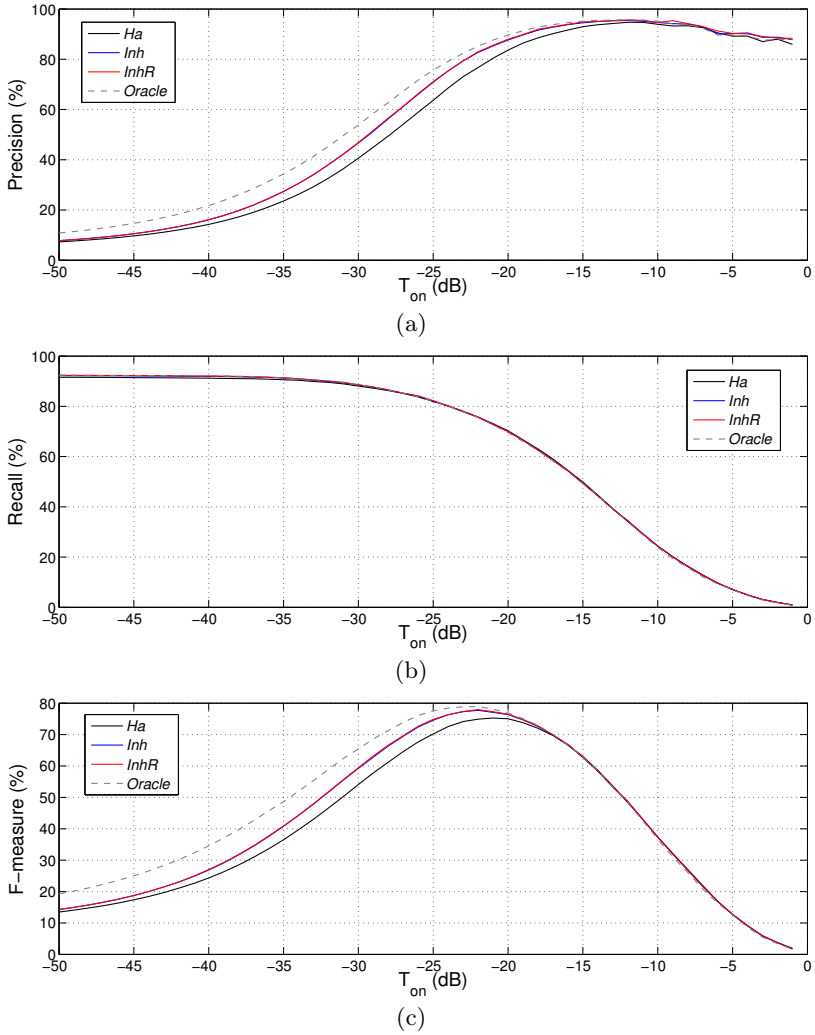


Figure 5.5: Supervised transcription performances. (a) Precision, (b) Recall, and (c) F-measure (in %) as a function of the onset detection threshold T_{on} (in dB) for *Ha/Inh/InhR-NMF* with $N_r = 20$, and *Oracle-NMF*.

The influence of N_r for the three parametric models is also presented on Figure 5.6 for a detection threshold fixed to $T_{\text{on}} = -22$ dB (that corresponds approximately to the value leading to optimal F-measure performances for all methods). Again, increasing the number of partials of the spectra tends to improve the Precision but does not seem to have a significant influence on the Recall. Logically, when increasing the number of partials, the performance gap between the two inharmonic models and the harmonic model is increasing. However, for all parametric models a limit is reached for $N_r = 20$. This result is surprising (particularly for the inharmonic models for which the partials of transverse vibration were accurately learned up to rank 50 in the first step) since notes below $A\sharp 4$ (70) are usually composed of more partials, up to more than a hundred in the bass range. A possible explanation for this may be that these experiments have been performed using

a KL divergence ($\beta = 1$) for the reconstruction cost-function. As highlighted in Section 2.1.1.3, the KL divergence is not scale invariant and thus favors the reconstruction of the components having highest magnitudes in the spectrogram. As it can be seen on Figure 5.4, partials with a rank greater than 20 have magnitudes below -50 dB when compared to the first partial. These may thus be neglected in the optimization of the activation matrix H. It should be then interesting to study the influence of N_r on the performance, jointly with a grid of β values (for instance $\in [0, 1]$)[1]. Indeed, such an application of parametric NMF-based models to transcription has shown optimal performances for $\beta = 0.5$ in [Vincent et al., 2010].

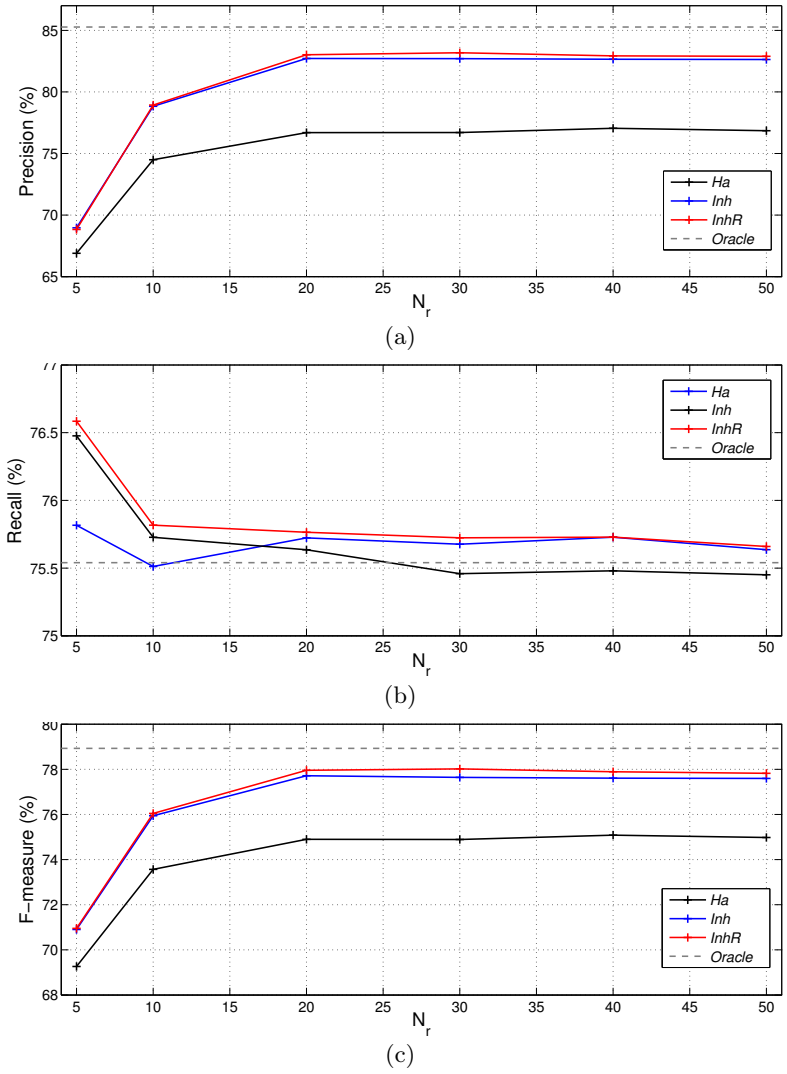


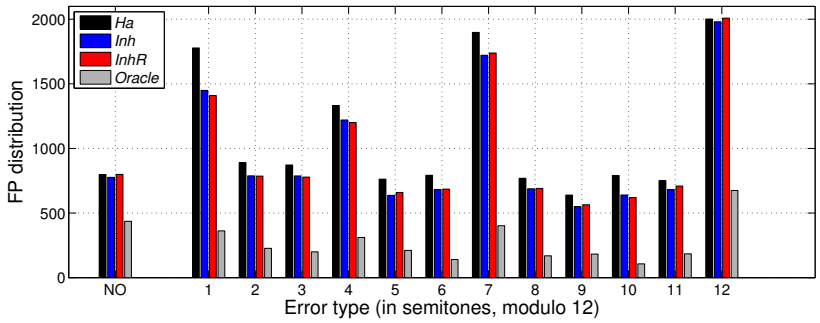
Figure 5.6: (a) Precision, (b) Recall and (c) F-measure (in %) obtained for all methods as a function of the number of partials N_r for $T_{\text{on}} = -22$ dB.

Finally, the FP errors made by all methods are analyzed in detail. Each FP is classified

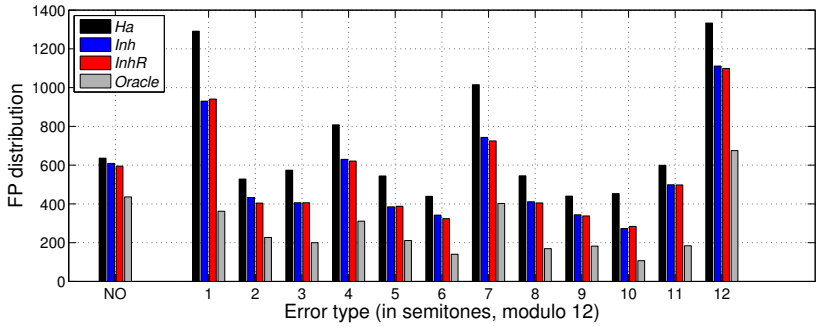
[1]Further experiments have been conducted after the writing of this thesis. When setting $N_r = 50$ and computing the supervised transcription for a grid of β having values linearly distributed in $[0, 2]$ (step of 0.1), we found that $\beta = 0.9$ leads to the highest F-measure for *Inh/InhR/Oracle-NMF* and $\beta = 1$ for *Ha-NMF*.

according to its distance (in semitones, modulo 12) with respect to the notes that are actually present in the ground truth within +/− 50 ms. Another class (denoted by 'NO') is considered for FP detected while no onset is present in the ground truth. Histograms of these errors are presented on Figure 5.7 for all methods with $N_r \in \{5, 10, 20\}$ and $T_{\text{on}} = -22$ dB.

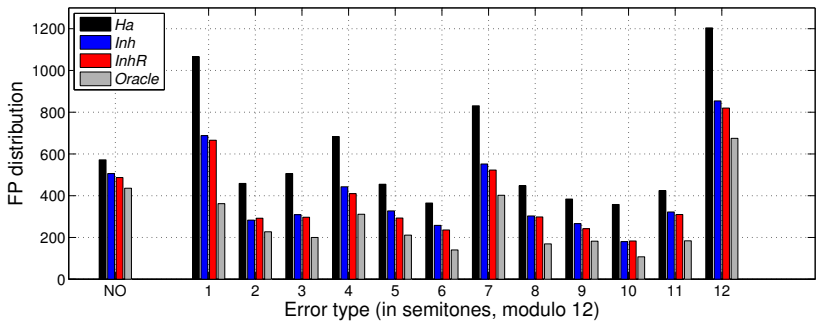
In accordance with the results presented above, the improvement of Precision (from *Ha-NMF* to *Oracle-NMF*, with N_r increasing) is related to a decrease of the number of FP. The inharmonicity inclusion is relevant for $N_r > 5$ and interestingly, as expected, tends to mainly reduce harmonically-related FP (*e.g.* major thirds, fifths and octaves corresponding respectively to 4, 7 and 12 semitones).



(a) $N_r = 5$



(b) $N_r = 10$



(c) $N_r = 20$

Figure 5.7: Histogram of FP errors returned by all methods for different values of N_r and $T_{\text{on}} = -22$ dB. Error types are given in semitones (modulo 12), except for 'NO' which corresponds to FP errors of the algorithms while no onset was present in the reference.

5.2.2.3 Conclusion

This preliminary study shows that increasing the precision of the piano tone spectra model may really improve the performance of an NMF-based transcription system, in the case when the parameters of the dictionary are properly estimated. In particular it helps in reducing FP detections. However, there is no guarantee that similar results are obtained in the case of an unsupervised transcription. As mentioned in Section 2.1.2, and highlighted in Section 3.1.2.2 for piano tone analysis, the NMF cost-functions are often non-convex and special care has to be taken for the initialization and the optimization scheme, in such a way that the algorithm converges toward the desired optimum.

5.2.3 Unsupervised transcription

Finally all three parametric models (*Ha/Inh/InhR-NMF*) are evaluated on the same dataset in an unsupervised transcription task. Thus, at each loop of the optimization algorithms H and W^{θ} matrices are iteratively updated.

5.2.3.1 Protocol

Initialization: For both inharmonic models, the influence of the initialization of (B, F_0) parameters is investigated. Two configurations are then proposed. First, a naive initialization for which F_{0r} is set to Equal Temperament (no "octave stretching") and B_r to 5.10^{-3}, $\forall r \in [1, R]$ is used. Second, (B_r, F_{0r}) parameters are initialized according to the mean model of inharmonicity and tuning along the whole compass of pianos presented in Section 4.4.3. These two initialization configurations are respectively denoted by 'ini_1' and 'ini_2' in the following. For *Ha-NMF*, F_{0r} is simply initialized to exact Equal Temperament.

Optimization scheme: The optimization is detailed in Algorithms 4, 5 and 6, respectively for *Ha-NMF*, *Inh-NMF* and *InhR-NMF*.

It is worth noting here that the optimization for the transcription task slightly differs from the one presented in Section 3.1.2.3 for the supervised estimation of (B, F_0). Besides the inclusion of the update of H and the normalization of a_{nr} values, a few modifications of the steps of the algorithms are done. First, we did not retain the noise level as discussed in Section 3.1.2.2, where it is used to remove the influence of partials drowned in noise; such a component cannot be simply adapted to a transcription application. Second, the number of partials for each note N_r is here fixed at the initialization instead of being initialized with a small number and iteratively increased (as done in lines 14-16 of Algorithms 1 and 2, page 47). However, further work will investigate the influence of such an iterative optimization scheme on transcription performance as it has been shown in Section 3.1.2.2 that it should help in avoiding the convergence of (B, F_0) parameters toward local optima.

Learning of the regularization parameter of *InhR-NMF*: In order to estimate an appropriate value for the regularization parameter λ of the *InhR-NMF* method, a learning set composed of 9 pieces is built (1 piece for each piano, none of them in the test set). The influence of λ is then studied for $N_r = 10$ on a grid covering the range $[10^{-11}, 10^{-3}]$ with values logarithmically distributed. Also, as mentioned in Section 3.1.2.2, the spectrograms are normalized to a maximal value of 1 in order to limit the influence of the scaling property of KL divergence in the tuning of λ. The optimal F-measure is obtained for $\lambda = 2 \cdot 10^{-5}$,

and it should be noted that the performance did not depend on a fine tuning of this parameter[2].

Algorithm 4 *Ha-NMF* unsupervised transcription

Input:
V spectrogram (normalized to a max. of 1)
β

Initialization: $\forall\ r \in [1, R]$, $n \in [1, N_r]$,
F_{0r} according to ET
$f_{nr} = nF_{0r}$, $a_{nr} = 1$,
W^θ computation (*cf.* Eq. (3.1))
H with random positive values

Optimization:
for $it = 1$ to It **do**
  • H_{rt} update $\forall\ r \in [1, R]$, $t \in [1, T]$ (Eq. (5.9))
  • a_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.8))
    a_{nr} normalization $\forall\ r \in [1, R]$ (Eq. (5.12)-(5.14))
    W^θ update (Eq. (3.1))
  for $u = 1$ to 10 **do**
    • F_{0r} update $\forall r \in [1, R]$ (*cf.* Eq. (5.6))
      W^θ update (*cf.* Eq. (3.1))
  end for
end for

Output: H,B_r, F_{0r}, a_{nr}

Algorithm 5 *Inh-NMF* unsupervised transcription

Input:
V spectrogram (normalized to a max. of 1)
β

Initialization: $\forall\ r \in [1, R]$, $n \in [1, N_r]$,
(B_r, F_{0r}) according to ini_1 or ini_2
$f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}$, $a_{nr} = 1$,
W^θ computation (*cf.* Eq. (3.1))
H with random positive values

Optimization:
for $it = 1$ to It **do**
  • H_{rt} update $\forall\ r \in [1, R]$, $t \in [1, T]$ (Eq. (5.9))
  • a_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.8))
    a_{nr} normalization $\forall\ r \in [1, R]$ (Eq. (5.12)-(5.14))
    W^θ update (Eq. (3.1))
  for $u = 1$ to 10 **do**
    • F_{0r} update $\forall r \in [1, R]$ (*cf.* Eq. (3.13))
      W^θ update (*cf.* Eq. (3.1))
    • B_r update $\forall r \in [1, R]$ (*cf.* Eq. (3.12))
      W^θ update (*cf.* Eq. (3.1))
  end for
end for

Output: H,B_r, F_{0r}, a_{nr}

[2]The actual value found in [Rigaud et al., 2013c] is $\lambda = 1$. However, in this latter study the penalty terms of *InhR-NMF* cost-function does not account for the constant multiplicative factor $K_\tau T$ introduced in this thesis (*cf.* Equation (3.7)). Thus, the value $\lambda = 2 \cdot 10^{-5}$ given in this section is considering this scaling factor, with – according to the analysis parameters – $K_\tau = 17$ bins and $T = 2660$ time-frames.

Algorithm 6 *InhR-NMF* unsupervised transcription

Input:
V spectrogram (normalized to a max. of 1)
β, λ

Initialization: $\forall\ r \in [1, R]$, $n \in [1, N_r]$,
(B_r, F_{0r}) according to ini_1 or ini_2
$f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}$, $a_{nr} = 1$,
W^θ computation (*cf.* Eq. (3.1))
H with random positive values

Optimization:
for $it = 1$ to It **do**
  • H_{rt} update $\forall\ r \in [1, R]$, $t \in [1, T]$ (Eq. (5.9))
  • a_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.8))
    a_{nr} normalization $\forall\ r \in [1, R]$ (Eq. (5.12)-(5.14))
    W^θ update (Eq. (3.1))
  • f_{nr} update $\forall\ r \in [1, R]$, $n \in [1, N_r]$ (Eq. (3.18))
    W^θ update (Eq. (3.1))
  for $v = 1$ to 30 **do**
    F_{0r} update $\forall r \in [1, R]$ (*cf.* Eq. (3.20))
    B_r update $\forall r \in [1, R]$ (20 times) (*cf.* Eq. (3.19))
  end for
end for

Output: H,B_r, F_{0r}, a_{nr}, f_{nr}

5.2.3.2 Results

The influence of N_r, the maximal number of partials is here studied on the grid $\{5, 10, 20, 30\}$ for $T_{\text{on}} \in [-50, 1]$ dB.

The Precision, Recall and F-measure performances are presented on Figure 5.8 for $N_r = 20$. For the first initialization, *Inh-NMF* and *InhR-NMF* do not perform as well as *Ha-NMF* (this is consistent with the observation in [Vincent et al., 2008]). Conversely, for the second initialization with the mean model of inharmonicity and piano tuning, these methods perform significantly better than *Ha-NMF* (ANOVA p-values lower than 0.05 for $N_r < 30$ and $T_{on} = -18$ dB). Furthermore, both inharmonic models give comparable mean F-measures (p-values higher than 0.5). Standard deviations are not reported in the plots but are around 10 to 14 %.

The influence of N_r on the performances, for $T_{\text{on}} = -18$ dB, is presented on Figure 5.9. Again, for all different values of N_r, both inharmonic models return highest performances than *Ha-NMF* in the case where the initialization of (B, F_0) parameters was performed according the mean model of inharmonicity and tuning. In accordance with the results obtained for the supervised case, increasing the number of partials tends to improve the performances of *Inh/InhR-NMF* up to $N_r = 20$. Surprisingly, the performance of *Ha-NMF* keeps increasing for $N_r = 30$. It seems here that adding more partials avoids a situation where high notes explain high rank partials belonging to lower notes. It is also interesting to notice that *Ha-NMF* results for $N_r = 30$ are comparable to those obtained by the NMF under the harmonicity constraint presented in [Vincent et al., 2010] (Section V.B) for similar experimental setups.

These experiments tend to demonstrate that such parametric NMF models with inharmonicity constraints are highly dependent on the initialization. Indeed, the reconstruction cost function is non-convex with respect to f_{nr}, F_{0r} or B_r parameters, and present a large number of local minima. Hence, multiplicative update rules (as well as other optimization

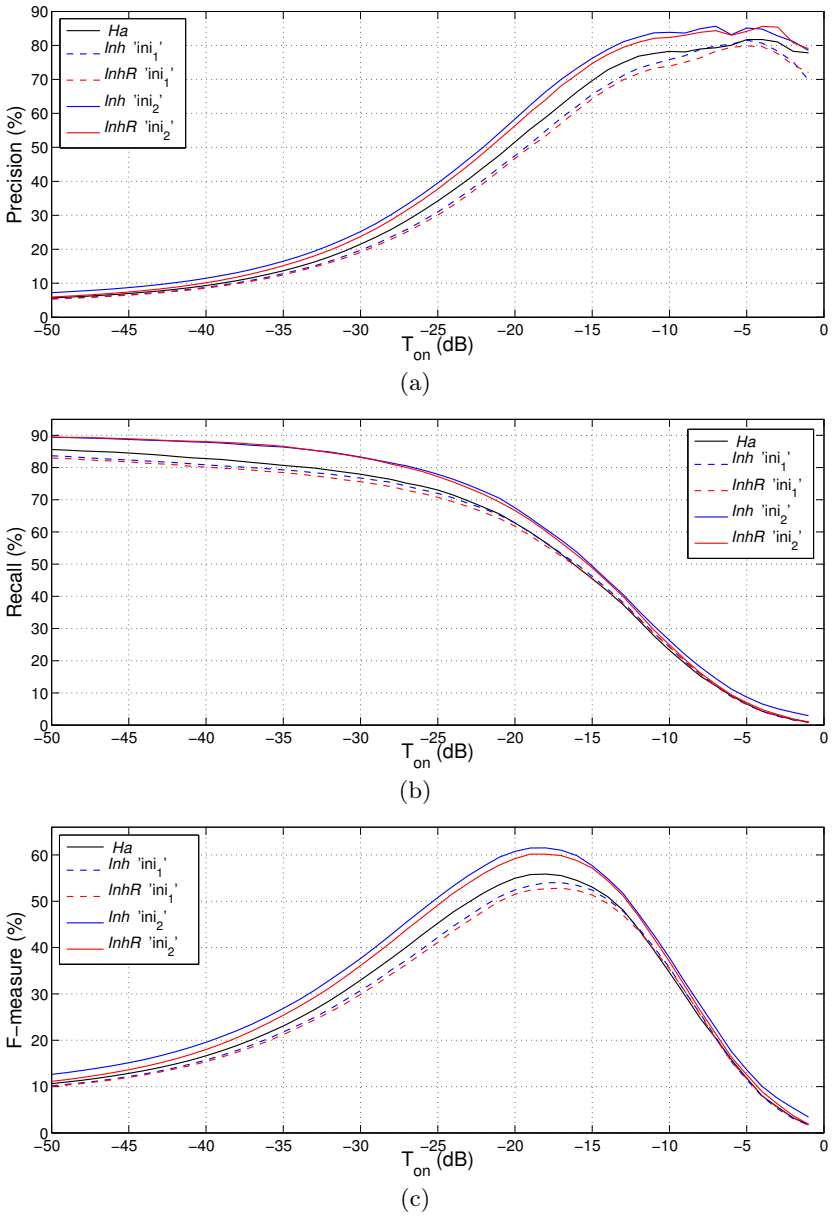
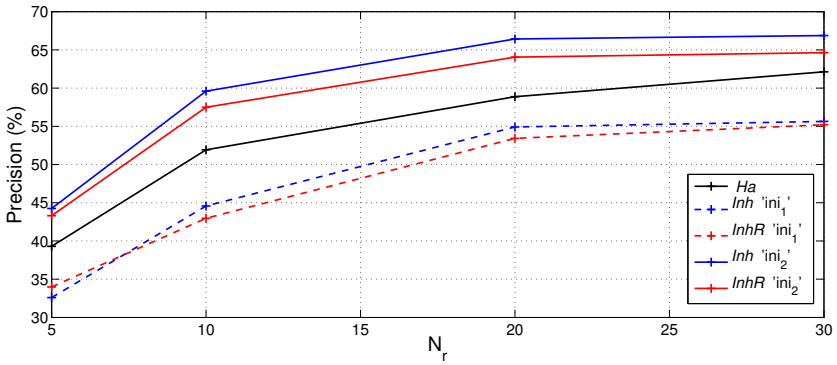


Figure 5.8: (a) Precision, (b) Recall and (c) F-measure (in %) as a function of the detection onset threshold T_{on}. Performances of *Ha*, *Inh* and *InhR* are respectively depicted as black, blue and red curves. For *Inh* and *InhR*, dashed and plain lines respectively correspond to the results obtained for 'ini$_1$' and 'ini$_2$'.

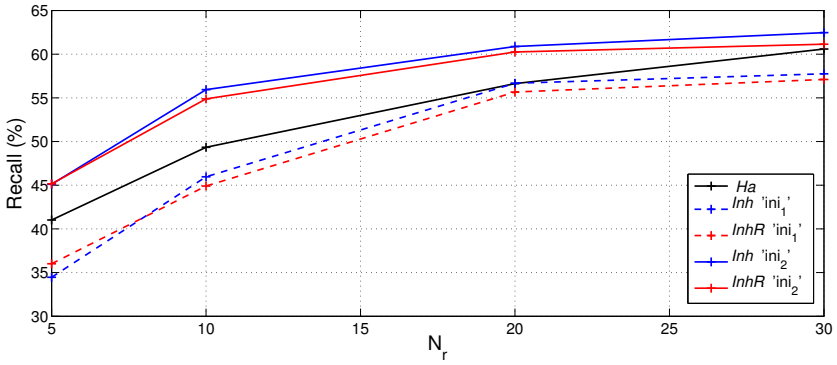
methods based on gradient descent) cannot ensure that these parameters will be correctly estimated. This results in the fact that the initialization of such parameters requires special care. Also, an alternative optimization scheme that would consider initializing the notes with a few partials and increasing the number iteratively (as proposed in Section 3.1.2.2) may be an efficient mean for avoiding the convergence toward local optima.

In contrast with the results for the supervised estimation of (B, F_0) presented in Chapter 3, taking into account the dispersion of the partial frequencies from a theoretical inharmonic relation in *InhR-NMF* does not seem valuable for a transcription task, when

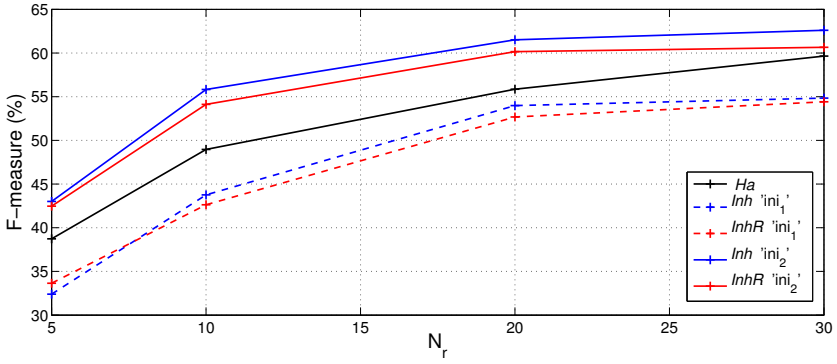
compared to *Inh-NMF*. A possible explanation for this result may be that *InhR-NMF* was found particularly useful for the analysis of low-bass tones, as the deviations related to the string-bridge couplings are mainly present in the very low pitch range. However, as shown in Figure 5.2, these notes are rarely played in a musical context. Also, introducing a larger number of free parameters in the model may not be valuable when targeting complex tasks such as transcription.



(a)



(b)



(c)

Figure 5.9: (a) Precision, (b) Recall and (c) F-measure (in %) as a function of the number of partials N_r for $T_{\text{on}} = -18$ dB. Performances of *Ha*, *Inh* and *InhR* are respectively depicted as black, blue and red curves. For *Inh* and *InhR*, dashed and plain lines respectively correspond to the results obtained for 'ini_1' and 'ini_2'.

5.3 Conclusion

Including inharmonicity in parametric NMF models has been shown to be relevant in a piano transcription task, provided that the inharmonicity and tuning parameters are sufficiently well estimated, this being highly dependent on the initialization. More precisely, an initialization with the same average value for the inharmonicity of all notes, and Equal Temperament for the tuning, turns out to provide worse estimates than the simpler purely harmonic model. However, a note-dependent inharmonicity law, and the corresponding "stretched" tuning curves, provide a good initialization to our models, that lead to significant improvement in the transcription results. Further work will investigate if an optimization based on an iterative addition of partials in the model may still increase the performance. Also, in order to build a competitive transcription system, additional studies will examine how these models on partials frequencies can be combined with amplitude models (smooth spectral envelopes), or frame dependencies in time.

Chapter 6
Conclusions and prospects

6.1 Conclusion

In this thesis, we have tackled the issue of the modeling and the representation of musical sounds by an approach inspired by both acoustics and signal processing perspectives. We focused our studies on the analysis of piano music, with particular attention paid to the inclusion of the inharmonic structure of the tones (as given by the model of transverse vibration of stiff strings) in signal-based modelings. To this end, two frameworks have been presented in Section 3.

We first considered enforcing the inharmonicity in the dictionary of spectra of NMF-based models. Two different ways of including such a constraint have been studied. The first one strictly enforces the partial frequencies of the model to follow the inharmonicity relation (*Inh-NMF* model). The second one includes inharmonicity by considering a regularized problem, *i.e.* by adding a penalization term to the NMF reconstruction cost-function (*InhR-NMF* model). Optimization algorithms as well as practical considerations have been given for both models in order to perform the estimation of the parameters.

We then introduced a probabilistic line spectrum model (*PLS*). From a prior peak-picking in time-frequency representations, the model assumes that the observed frequencies have been generated by a mixture of notes, each being composed of partial and noise components. The partial components are modeled by Gaussian mixtures having means constrained by the inharmonicity relation parameters. The proposed optimization algorithm returns a classification of each observation in partial and noise components for each note as well as their inharmonicity relation parameters and the probability for each note to have generated the observations of each time-frame.

Such models have been successfully applied to both acoustics and signal processing applications giving answers to the issues raised in the introduction:

• Can such classes of models be used to efficiently learn physics-related parameters (*e.g.* information about the design and tuning) of a specific instrument?

In Chapter 3 we focused on the precise estimation of the inharmonicity coefficient B and the F_0 of piano tones. All models have been tested on the whole compass of 11 different pianos in order to access the robustness with respect to the variability of the tones. These have been compared favorably to one state the art algorithm.

Both NMF-based models have been successfully applied to the supervised (*i.e.* having the knowledge of the played notes) estimation of (B, F_0) of isolated note and chord recordings. In this context, the relaxed inclusion of the inharmonicity constraint in *InhR-NMF* has shown benefits in allowing for slight deviations of the partial frequencies around the inharmonicity relation (as for instance caused by the string-bridge coupling).

On the same dataset of isolated note recordings, the *PLS* model has shown benefits in still performing well for a large range of the piano compass in the context of an unsupervised analysis. It has also been applied on a generic polyphonic piece of music, for which (B, F_0) parameters were accurately learned for a restricted set of notes that were played. Interestingly for the proposed application a perfect transcription of the music did not seem necessary.

In Chapter 4, a model for the inharmonicity and the tuning variations along the whole piano compass have been presented. While considering a small set of high-level parameters (only 6 global parameters to account for the main trends of 88×2 string parameters) these models have been found useful for various applications. In particular these have been applied to the retrieval of parameters highlighting some tuner's choices on different piano types, the generation of tuning curves for out-of-tune pianos or piano synthesizers, the initialization of the inharmonicity coefficient and the F_0 of analysis algorithms, and finally to the interpolation of inharmonicity and tuning along the whole compass of a piano from a prior estimation of these parameters on a restricted set of notes played in a polyphonic piece of music.

• Does refining/complexifying generic signal models actually improve the performance of analysis tasks targeted by the MIR community?

In Chapter 5, the two inharmonic NMF-based models have been applied to a transcription task and compared to a simpler harmonic model. The results have shown that inharmonicity inclusion was relevant in transcription tasks, in particular because it helps in reducing the detection of False Positives corresponding to harmonically-related notes (thirds, fifths and octaves). However, the performances of such models are highly dependent on the initialization of (B, F_0) parameters because of the non-convexity, with respect to these parameters, of the cost-functions. Thus, it has been found, in accordance with previous studies, that a naive initialization of such parameters can lead to a degradation of the results, when compared to a simpler harmonic model. Conversely, the use of a mean model of inharmonicity and tuning along the whole compass (introduced in Chapter 4) for the initialization has shown here a significant improvement in the transcription performances. The investigation of alternative optimization schemes that should be less sensitive to local optima will be investigated in further studies.

6.2 Prospects

6.2.1 Building a competitive transcription system

All the models that have been presented in this thesis focus on the inclusion of the inharmonic structure of the piano tones. However, when targeting the design of a competitive transcription system, a few more elements of sound modeling have to be taken into ac-

count. This could consist in constraining the shape of the spectral envelopes of the spectra (possibly varying over time in order to account for the fact that high rank partials are decaying faster than low ranks in piano tones) or the temporal activations. Such modelings have been widely investigated in the literature on NMF (as detailed in Section 2.1.2) and should be easily adapted to our model formulation. Including such additional information in the unsupervised transcription algorithm should also help obtaining more robust estimates of (B, F_0) parameters. While performing the transcription of a piece of music, one could imagine retrieving the model of piano that has been played by comparing the B estimates with reference curves of inharmonicity of several pianos.

Future research may also be conducted on the extension of the model to poly-instrumental music composed of harmonic and inharmonic instruments (possibly combining *Ha-NMF* and *Inh-NMF* models). This could be of interest in order to investigate whether inharmonicity is able to lift ambiguities of harmonically-related notes played by different instruments (*e.g.* C on piano and G on trumpet) while providing a clustering of all detected notes in the different instruments of the model.

6.2.2 Extension to other string instruments

The models that have been presented in this thesis are sufficiently generic to be applied to other struck but also plucked string instruments, such as for instance the harpsichord, the guitar or the harp. However, as it has been shown for the case of the piano, a prior study of the variations of (B, F_0) parameters along the compass seems essential for the initialization of the algorithms.

We propose here some prospects for an application to the guitar. As for the piano, some invariants in the design of the instruments can be used to build parametric models for the inharmonicity along the whole compass. Figure 6.1 presents the evolution of the inharmonicity coefficient along the register covered by the 6 strings of an electric guitar (e, B, G and D, A, E respectively correspond to plain and wrapped set of strings). These have been estimated on isolated note recordings using *InhR-NMF* model according to the protocol detailed in Section 3.1.3.2, with a manual tuning for the initialization of the parameters. One can notice here that the evolution of the inharmonicity coefficient along the range covered by each string has a linear behavior in logarithmic scale. For each string, the slope is related to the position of the frets that modify the vibrating length of the string. For most guitars the fret position is designed so that the guitar is tuned according to Equal Temperament. Thus, in the case where the tension is constant (no string bending), according to the expression of B (Equation (2.24)), the slope is equal to $\log 2^{2m/12}$, where m denotes the MIDI note index (as depicted in gray dashed lines in Figure 6.1). Then, in order to model the inharmonicity variations along the compass of a guitar, one could consider 6 different parameters corresponding to the Y-intercept of the straight lines (the jump between the 6 lines being mainly affected by the changes in the string diameters).

As the guitar allows us to play the same note with different combinations of string / neck position, such a parametric model of inharmonicity may potentially be used in *audio to tablature* transcription tasks [Barbancho et al., 2012], where beyond retrieving the notes that are played, inharmonicity could help here in assessing the position of the player's fingers along the guitar neck (as illustrated on Figure 6.2).

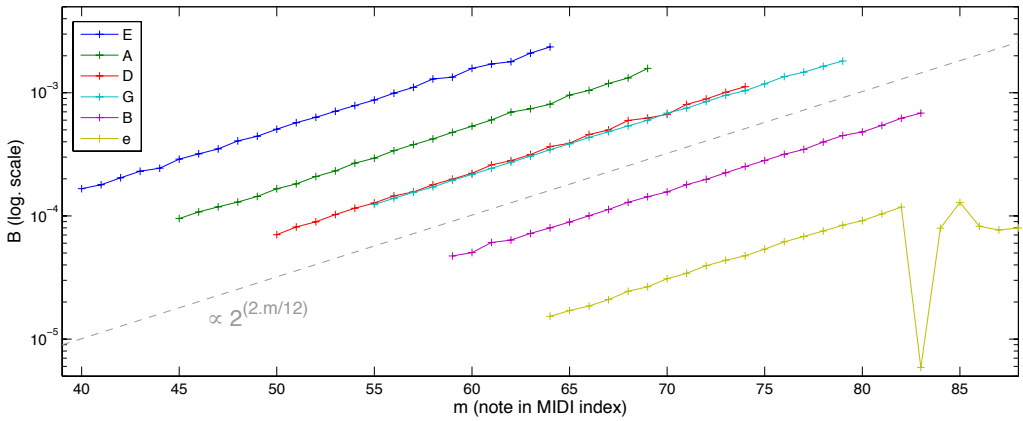
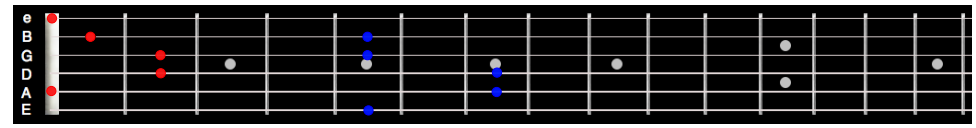
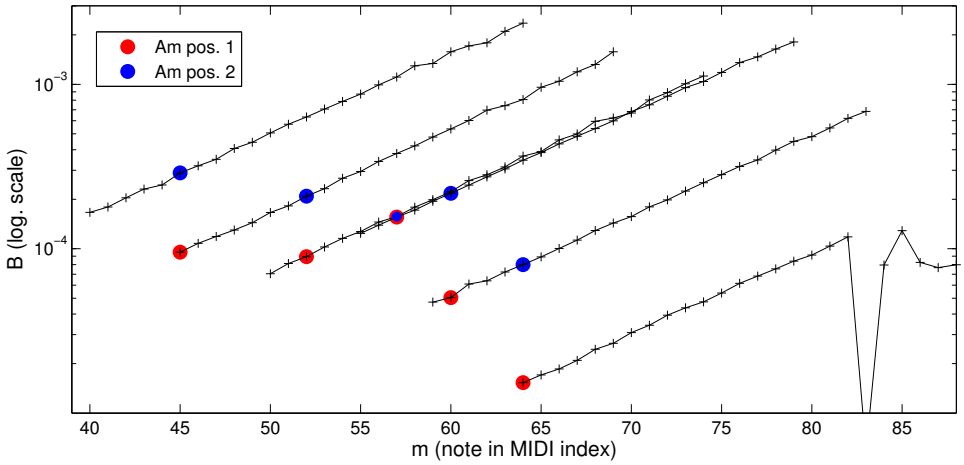


Figure 6.1: Inharmonicity curves along the range covered by the 6 strings of an electric guitar.



(a)



(b)

Figure 6.2: (a) A Minor (Am) chord played for two different positions on the guitar neck. (b) Inharmonicity patterns for the two positions.

Appendix A
Piano compass / MIDI norm

The register of pianos is usually composed of 88 notes spanning 9 octaves, from A0 to C8 (*cf.* Figure A.1). In MIDI norm, theses notes are indexed by $m \in [21, 108]$ and their fundamental frequency is given by the Equal Temperament (ET), this latter considering the A4 ($m = 69$) at 440 Hz as a reference note for the tuning, and a constant ratio of $2^{1/12}$ (100 cents) for every semitone:

$$F_{0,\mathrm{ET}}(m) = 440 \cdot 2^{(m-69)/12}. \tag{A.1}$$

In practice, pianos are never exactly tuned to ET because of the inharmonic structure of their tones and the actual F_0 values may deviate from Equation (A.1), up to 30 cents in the high and low register (*cf.* Chapter 4).

A summary of the piano notes, their MIDI index and fundamental frequency according to ET is given in Table A.1.

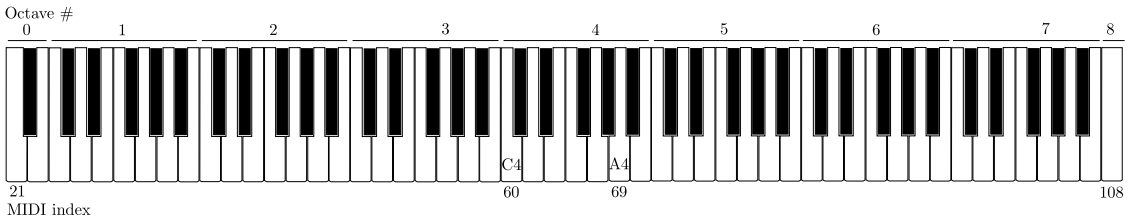


Figure A.1: Keyboard of the piano.

Octave #	C (Do)	C ♯ (Do ♯)	D (Ré)	E ♭ (Mi ♭)	E (Mi)	F (Fa)	F ♯ (Fa ♯)	G (Sol)	G ♯ (Sol ♯)	A (La)	B ♭ (Si ♭)	B (Si)	
0 (-1)										21 27.50	22 29.13	23 30.86	MIDI index F_{0ET} (Hz)
1 (0)	24 32.70	25 34.64	26 36.70	27 38.89	28 41.20	29 43.65	30 46.24	31 48.99	32 51.91	33 55.00	34 58.27	35 61.73	MIDI index F_{0ET} (Hz)
2 (1)	36 65.40	37 69.29	38 73.41	39 77.78	40 82.40	41 87.30	42 92.49	43 97.99	44 103.82	45 110.00	46 116.54	47 123.47	MIDI index F_{0ET} (Hz)
3 (2)	48 130.81	49 138.59	50 146.83	51 155.56	52 164.81	53 174.61	54 184.99	55 195.99	56 207.65	57 220.00	58 233.08	59 246.94	MIDI index F_{0ET} (Hz)
4 (3)	60 261.62	61 277.18	62 293.66	63 311.12	64 329.62	65 349.22	66 369.99	67 391.99	68 415.30	**69** **440**	70 466.16	71 493.88	MIDI index F_{0ET} (Hz)
5 (4)	72 523.25	73 554.36	74 587.32	75 622.25	76 659.25	77 698.45	78 739.98	79 783.99	80 830.60	81 880	82 932.32	83 987.76	MIDI index F_{0ET} (Hz)
6 (5)	84 1046.50	85 1108.73	86 1174.65	87 1244.50	88 1318.51	89 1396.91	90 1479.97	91 1567.98	92 1661.21	93 1760.00	94 1864.65	95 1975.53	MIDI index F_{0ET} (Hz)
7 (6)	96 2093.00	97 2217.46	98 2349.31	99 2489.01	100 2637.02	101 2793.82	102 2959.95	103 3135.96	104 3322.43	105 3520.00	106 3729.31	107 3951.06	MIDI index F_{0ET} (Hz)
8 (7)	108 4186.01												MIDI index F_{0ET} (Hz)

Table A.1: Correspondence between notes, MIDI indices and fundamental frequencies given by ET along the compass of the piano. French notations for the notes and octave numbers are given in parenthesis.

Appendix B
Noise level estimation

This appendix presents the noise level estimation method used in the pre-processing step of the (B, F_0) estimation algorithms introduced in Chapter 3. The method assumes an additive colored noise, i.e. generated by the filtering of a white Gaussian noise and added to the signal of interest [Yeh and Röbel, 2006]. In a given narrow band, if the noise filters have a quasi flat frequency response, the noise can be considered as white Gaussian and its spectral magnitude follows a Rayleigh distribution:

$$p(x_f; \sigma_f) = \frac{x_f}{\sigma_f^2} \cdot e^{-x_f^2/(2\sigma_f^2)} \ , \quad x_f \in [0, +\infty[. \tag{B.1}$$

In this pre-processing stage, we want to estimate the noise distribution in each band without removing the partials. To do so, a good estimator for σ_f is the median $\text{med}_f = \sigma_f\sqrt{\log 4}$. Indeed, when the length of the analysis window is adapted to the resolution, in a given narrow band, there are much less bins corresponding to partials than bins corresponding to noise, so partials have a reduced influence on the estimate of the noise median. The tradeoff sits in the choice of the bandwidth: the bands have to be narrow enough so that the white noise approximation holds, but wide enough so that most of the bins correspond to noise. We chose a 300Hz median filtering on the magnitude spectrum $S(f)$ to estimate σ_f. Finally, we define the noise level in each band $\text{NL}(f)$ as the magnitude such that the cumulative distribution function is equal to a given threshold T, set to $T = 0.9999$. With this choice of T, only 6 bins corresponding to noise on average (out of 2^{16}) should be above the noise level. The cumulative density function of a Rayleigh distribution is given by:

$$c(x_f; \sigma_f) = 1 - e^{-x_f^2/(2\sigma_f^2)}. \tag{B.2}$$

Thus the noise level can be expressed as:

$$\text{NL}(f) = \frac{\text{med}_f}{\sqrt{\log 4}} \cdot \sqrt{2 \log \frac{1}{1-T}}. \tag{B.3}$$

An illustration of the noise level computation for a note C4 is given in Figure B.1.

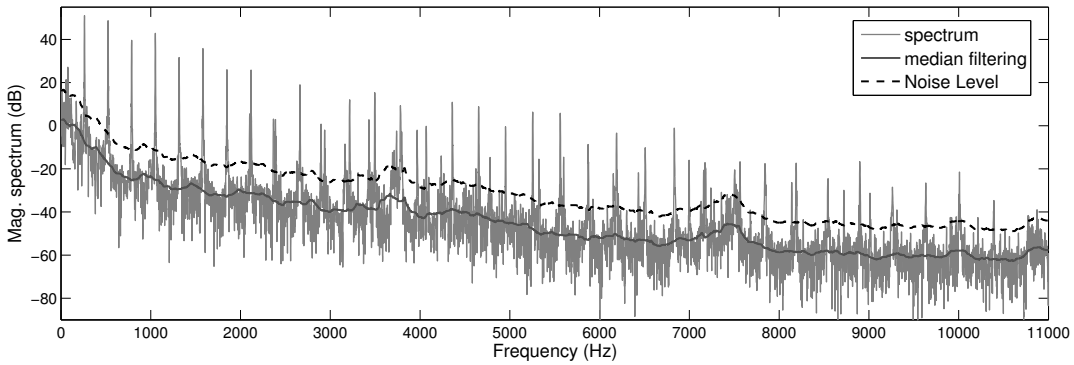


Figure B.1: Noise level computation for the magnitude spectrum of a note C4.

Appendix C

Derivation of *Inh/InhR/Ha-NMF* update rules

This appendix presents the mathematical derivation of the update rules used in the optimization of *Inh/InhR-NMF* (*cf.* Section 3.1.2.1) and *Ha-NMF* (*cf.* Section 5.2.1.2) models.

C.1 Multiplicative update rules

As presented in Section 2.1.1.5, multiplicative update rules can be obtained by decomposing the partial derivative of a cost-function, with respect to a given parameter θ^*, as a difference of two positive terms:

$$\frac{\partial C(\theta^*)}{\partial \theta^*} = P(\theta^*) - Q(\theta^*), \qquad P(\theta^*),\, Q(\theta^*) \geq 0. \tag{C.1}$$

In the case the cost-function includes a regularization term, for instance in the form

$$C(\theta) = C_0(\theta) + \lambda \cdot C_1(\theta, \gamma), \tag{C.2}$$

the same kind of decomposition can be performed independently for each term so that

$$\frac{\partial C(\theta^*, \gamma)}{\partial \theta^*} = \underbrace{\left(P_0(\theta^*) + \lambda \cdot P_1(\theta^*)\right)}_{P(\theta^*)} - \underbrace{\left(Q_0(\theta^*) + \lambda \cdot Q_1(\theta^*)\right)}_{Q(\theta^*)}, \tag{C.3}$$

Then, the parameter is updated as follows:

$$\theta^* \leftarrow \theta^* \times Q(\theta^*)/P(\theta^*) \tag{C.4}$$

C.2 Partial derivatives of the reconstruction cost-function

C.2.1 Problem reminder

For all three parametric models, the reconstruction cost-function is given by (*cf.* Section 3.1.1):

$$C_0(\theta, H) = \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} d_\beta \left(V_{kt} \mid \hat{V}_{kt}\right), \tag{C.5}$$

where

$$\hat{V}_{kt} = \sum_{r=1}^{R} W_{kr}^{\theta_r} \cdot H_{rt}, \tag{C.6}$$

and

$$W_{kr}^{\theta_r} = \sum_{n=1}^{N_r} a_{nr} \cdot g_\tau(f_k - f_{nr}), \tag{C.7}$$

with $\theta_r = \{a_{nr}, f_{nr} \mid n \in [1, N_r]\}$, and $\mathcal{K} = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau], \forall n \in [1, N_r], \forall r \in [1, R]\}$.

C.2.2 Derivative with respect to θ

According to the partial derivative of β-divergences ($\partial d_\beta(x \mid y)/\partial y$, *cf.* Equation (2.10)), the partial derivative of the reconstruction cost-function with relation to a specific parameter $\theta^* \in \theta$ is given by:

$$\frac{\partial C_0(\theta, H)}{\partial \theta^*} = \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} \frac{\partial \hat{V}_{kt}}{\partial \theta^*} \cdot \hat{V}_{kt}^{\beta-2} (\hat{V}_{kt} - V_{kt}), \tag{C.8}$$

with

$$\frac{\partial \hat{V}_{kt}}{\partial \theta^*} = \sum_{r=1}^{R} \frac{\partial W_{kr}^{\theta_r}}{\partial \theta^*} \cdot H_{rt}, \tag{C.9}$$

which can be decomposed as a difference of two positive terms

$$\frac{\partial \hat{V}_{kt}}{\partial \theta^*} = \underbrace{\sum_{r=1}^{R} \frac{\partial W_{kr}^{\theta_r}{}^{\oplus}}{\partial \theta^*} \cdot H_{rt}}_{\frac{\partial \hat{V}_{kt}^{\oplus}}{\partial \theta^*}} - \underbrace{\sum_{r=1}^{R} \frac{\partial W_{kr}^{\theta_r}{}^{\ominus}}{\partial \theta^*} \cdot H_{rt}}_{\frac{\partial \hat{V}_{kt}^{\ominus}}{\partial \theta^*}}. \tag{C.10}$$

Finally, the quantity $\partial C_0(\theta, H)/\partial \theta^*$ can also be expressed as a difference of two positive quantities:

$$\begin{aligned} \frac{\partial C_0(\theta, H)}{\partial \theta^*} = & \underbrace{\sum_{k \in \mathcal{K}} \sum_{t=1}^{T} \left[\frac{\partial \hat{V}_{kt}^{\oplus}}{\partial \theta^*} \cdot \hat{V}_{kt}^{\beta-1} + \frac{\partial \hat{V}_{kt}^{\ominus}}{\partial \theta^*} \cdot \hat{V}_{kt}^{\beta-2} \cdot V_{kt} \right]}_{P_0(\theta^*) \geq 0} \\ & - \underbrace{\sum_{k \in \mathcal{K}} \sum_{t=1}^{T} \left[\frac{\partial \hat{V}_{kt}^{\oplus}}{\partial \theta^*} \cdot \hat{V}_{kt}^{\beta-2} \cdot V_{kt} + \frac{\partial \hat{V}_{kt}^{\ominus}}{\partial \theta^*} \cdot \hat{V}_{kt}^{\beta-1} \right]}_{Q_0(\theta^*) \geq 0}. \end{aligned} \tag{C.11}$$

C.2.2.1 Derivative with respect to a_{nr}

$\forall r \in [1, R]$ and $n \in [1, N_r]$

$$\frac{\partial \hat{V}_{kt}}{\partial a_{nr}} = g_\tau(f_k - f_{nr}) \cdot H_{rt} > 0 \tag{C.12}$$

It is then chosen,

$$\begin{cases} \dfrac{\partial \hat{V}_{kt}^{\oplus}}{\partial a_{nr}} = g_\tau(f_k - f_{nr}) \cdot H_{rt}, \\ \dfrac{\partial \hat{V}_{kt}^{\ominus}}{\partial a_{nr}} = 0. \end{cases} \tag{C.13}$$

Finally, by replacing in Equation (C.11), Equations (3.9) and (3.10) are obtained:

$$\begin{cases} P_0(a_{nr}) = \displaystyle\sum_{k \in \mathcal{K}_{nr}} \sum_{t=1}^{T} \left[(g_\tau(f_k - f_{nr}).H_{rt}) \, .\widehat{V}_{kt}^{\beta-1} \right], \\ Q_0(a_{nr}) = \displaystyle\sum_{k \in \mathcal{K}_{nr}} \sum_{t=1}^{T} \left[(g_\tau(f_k - f_{nr}).H_{rt}) \, .\widehat{V}_{kt}^{\beta-2}.V_{kt} \right], \end{cases}$$

with $\mathcal{K}_{nr} = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau]\}$.

C.2.2.2 Derivative with respect to f_{nr}, F_{0r} and B_r

- ***InhR-NMF*:**
  $\forall r \in [1, R]$ and $n \in [1, N_r]$

$$\frac{\partial \hat{V}_{kt}}{\partial f_{nr}} = -a_{nr} \cdot g'_\tau(f_k - f_{nr}) \cdot H_{rt} \tag{C.14}$$

Regardless the analysis window that has been used, the quantity $g'_\tau(f_k - f_{nr})$ changes its sign on each lobe of g_τ. In order to obtain a satisfying expression (*i.e.* a difference of two positive terms), the spectral support of $g_\tau(f_k - f_{nr})$ is limited to its main lobe (so the sign of its derivative is changing once) and its derivative is expressed as:

$$g'_\tau(f_k - f_{nr}) = (f_{nr} - f_k) \cdot \frac{-g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}. \tag{C.15}$$

The quantity $\frac{-g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}$ stays positive on the main lobe, for every kind of analysis window (an illustration can be found in [Hennequin et al., 2010]). Thus,

$$\begin{cases} \dfrac{\partial \hat{V}_{kt}^{\oplus}}{\partial f_{nr}} = a_{nr} \cdot f_k \cdot \dfrac{-g'_\tau(f_k - f_{nr})}{f_k - f_{nr}} \cdot H_{rt}, \\ \dfrac{\partial \hat{V}_{kt}^{\ominus}}{\partial f_{nr}} = a_{nr} \cdot f_{nr} \cdot \dfrac{-g'_\tau(f_k - f_{nr})}{f_k - f_{nr}} \cdot H_{rt}. \end{cases} \tag{C.16}$$

And finally, by replacing in Equation (C.11), Equations (3.21) and (3.22) are obtained:

$$\begin{cases} P_0^{InhR}(f_{nr}) = \sum_{k\in\mathcal{K}_{nr}} \sum_{t=1}^{T} \left[\left(a_{nr} \frac{-f_k.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-1} \right. \\ \qquad\qquad \left. + \left(a_{nr} \frac{-f_{nr}.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-2}.V_{kt} \right], \\ Q_0^{InhR}(f_{nr}) = \sum_{k\in\mathcal{K}_{nr}} \sum_{t=1}^{T} \left[\left(a_{nr} \frac{-f_k.g'_\tau(f_k - f_{nr})}{f_k - f_{nr}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-2}.V_{kt} \right. \\ \qquad\qquad \left. + \left(a_{nr} \frac{-f_{nr}.g'_\tau(f - f_{nr})}{f_k - f_{nr}}.H_{rt} \right).\widehat{V}_{kt}^{\beta-1} \right], \end{cases}$$

Very similar decompositions can be performed for *Inh-NMF* and *Ha-NMF*. We just give here the expression of the partial derivatives of $\hat{V}_{kt}$.

- ***Inh-NMF*:** $f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}$

$\forall r \in [1, R]$

$$\frac{\partial \hat{V}_{kt}}{\partial B_r} = -\sum_{n=1}^{N_r} C_{nr} \cdot a_{nr} \cdot g'_\tau(f_k - f_{nr}) \cdot H_{rt}, \tag{C.17}$$

$$\frac{\partial \hat{V}_{kt}}{\partial F_{0r}} = -\sum_{n=1}^{N_r} D_{nr} \cdot a_{nr} \cdot g'_\tau(f_k - f_{nr}) \cdot H_{rt}, \tag{C.18}$$

with

$$C_{nr} = \frac{\partial f_{nr}}{\partial B_r} = \frac{n^3 F_{0r}}{2\sqrt{1 + B_r n^2}}, \tag{C.19}$$

$$D_{nr} = \frac{\partial f_{nr}}{\partial F_{0r}} = n\sqrt{1 + B_r n^2}. \tag{C.20}$$

- ***Ha-NMF*:** $f_{nr} = nF_{0r}$

$\forall r \in [1, R]$

$$\frac{\partial \hat{V}_{kt}}{\partial F_{0r}} = -\sum_{n=1}^{N_r} n \cdot a_{nr} \cdot g'_\tau(f_k - f_{nr}) \cdot H_{rt}, \tag{C.21}$$

$$\tag{C.22}$$

C.2.3 Derivative with respect to H_{rt}

Similarly to Equation (C.8), the partial derivative of C_0 with respect to H_{rt} can be expressed as:

$$\frac{\partial C_0(\theta, H)}{\partial H_{rt}} = \sum_{k\in\mathcal{K}_r} \sum_{t=1}^{T} W_{kr}^{\theta_r} \cdot \hat{V}_{kt}^{\beta-2}(\hat{V}_{kt} - V_{kt}), \tag{C.23}$$

with $\mathcal{K}_r = \{k \mid f_k \in f_{nr} + [-2/\tau, 2/\tau], \forall n \in [1, N_r]\}$. A straightforward decomposition in the form

$$\frac{\partial C_0(\theta, H)}{\partial H_{rt}} = \underbrace{\sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} W_{kr}^{\theta_r} \cdot \hat{V}_{kt}^{\beta-1}}_{P_0(H_{rt})} - \underbrace{\sum_{k \in \mathcal{K}_r} \sum_{t=1}^{T} W_{kr}^{\theta_r} \cdot \hat{V}_{kt}^{\beta-2} V_{kt}}_{Q_0(H_{rt})}, \tag{C.24}$$

leads thus to the update rule given in Equation (5.9).

C.3 Partial derivatives of the regularization term of *InhR-NMF*

$$C_1(f_{nr}, \gamma_r) = K_\tau T \cdot \sum_{r=1}^{R} \sum_{n=1}^{N_r} \left(f_{nr} - n F_{0r} \sqrt{1 + B_r n^2} \right)^2, \tag{C.25}$$

where $K_\tau = \text{Card}\{f_k \in [-2/\tau, 2/\tau]\}$ is the number of frequency-bins for which the partials of the model are defined and T is the number of time-frames T.

C.3.1 Derivative with respect to f_{nr}

$\forall r \in [1, R]$, $n \in [1, N_r]$

$$\frac{\partial C_1}{\partial f_{nr}} = K_\tau T \cdot 2 \left(f_{nr} - n F_{0r} \sqrt{1 + B_r n^2} \right) \tag{C.26}$$

Then Equations (3.23) and (3.24) are directly obtained:

$$\begin{cases} P_1(f_{nr}) = K_\tau T \cdot 2 f_{nr}, \\ Q_1(f_{nr}) = K_\tau T \cdot 2 n F_{0r} \sqrt{1 + B_r n^2}. \end{cases}$$

C.3.2 Derivative with respect to B_r

$\forall r \in [1, R]$:

$$\begin{aligned} \frac{\partial C_1}{\partial B_r} &= K_\tau T \cdot \sum_{n=1}^{N_r} 2 \left(f_{nr} - n F_{0r} \sqrt{1 + B_r n^2} \right) \cdot \frac{-n^3 F_{0r}}{2\sqrt{1 + B_r n^2}} \\ &= K_\tau T \cdot F_{0r} \sum_{n=1}^{N_r} \left(n^4 F_{0r} - \frac{n^3 f_{nr}}{\sqrt{1 + B_r n^2}} \right) \end{aligned} \tag{C.27}$$

Thus, Equations (3.25) and (3.26) are obtained:

$$\begin{cases} P_1(B_r) = F_{0r} \sum_{n=1}^{N_r} n^4, \\ Q_1(B_r) = \sum_{n=1}^{N_r} \frac{n^3 f_{nr}}{\sqrt{1 + B_r n^2}}. \end{cases}$$

C.3.3 Derivative with respect to F_{0r}

$\forall r \in [1, R]$:

$$\frac{\partial C_1}{\partial F_{0r}} = K_\tau T \cdot \sum_{n=1}^{N_r} 2\left(f_{nr} - nF_{0r}\sqrt{1 + B_r n^2}\right) \cdot \left(-n\sqrt{1 + B_r n^2}\right). \tag{C.28}$$

An exact analytic solution allows to cancel the partial derivative (corresponding to the update rule (3.20)):

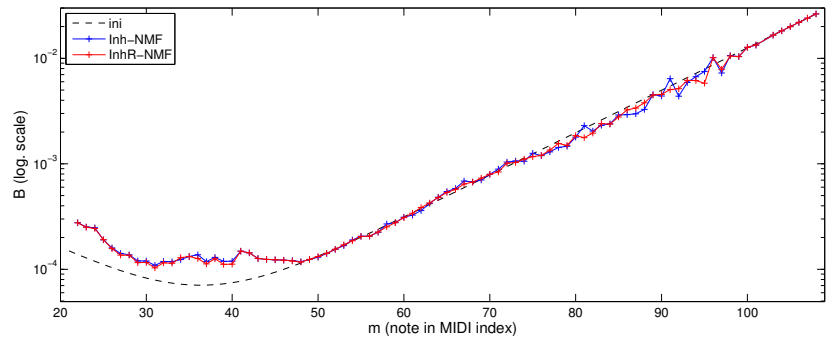
$$F_{0r} = \frac{\sum_{n=1}^{N_r} f_{nr} n\sqrt{1 + B_r n^2}}{\sum_{n=1}^{N_r} n^2(1 + B_r n^2)}.$$

Appendix D

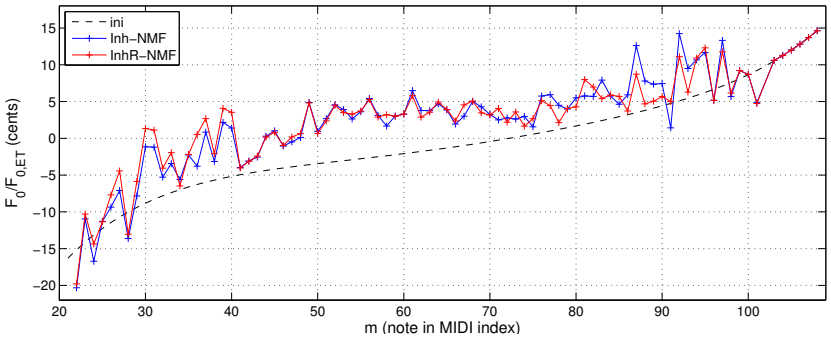
(B, F_0) curves along the whole compass estimated by the NMF-based models

This appendix presents the results for the estimation of (B, F_0) from isolated note recordings along the whole compass of the 11 considered pianos. The details of the experimental protocol are given in Section 3.1.3.2. The initialization of (B, F_0) is depicted as black dashed lines and the blue and red curves respectively correspond to the estimates obtained by *Inh-NMF* and *InhR-NMF* algorithms. Note that some estimates may be missing when the notes were not given in the databases.

- **Iowa database:**



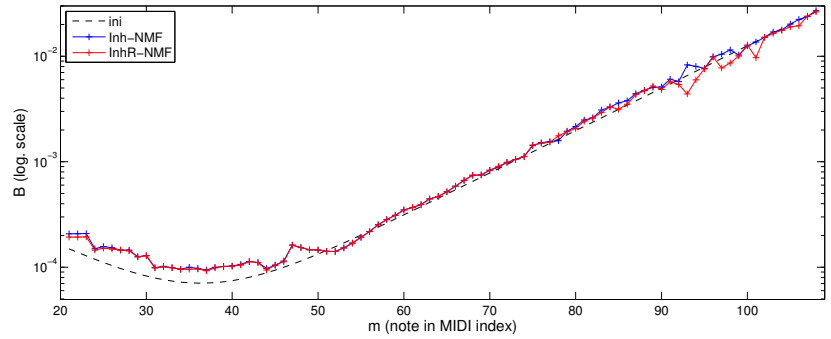
(a)



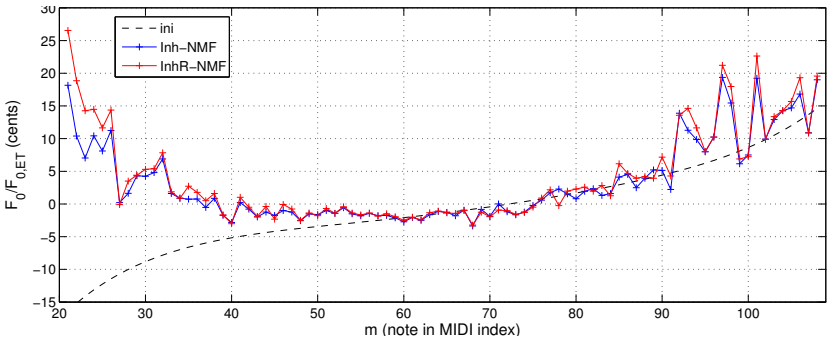
(b)

Figure D.1: Iowa (a) B and (b) F_0 as dev. from ET along the whole compass.

- **RWC database:**

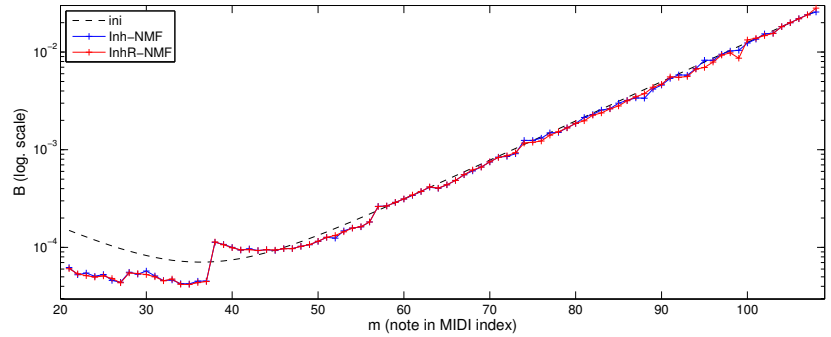


(a)

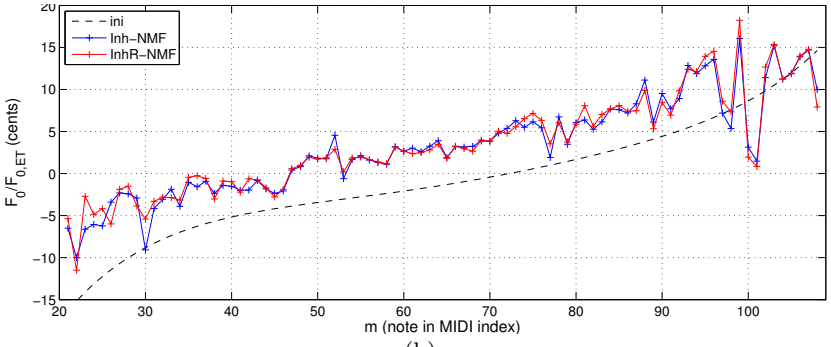


(b)

Figure D.2: RWC1 (a) B and (b) F_0 as dev. from ET along the whole compass.

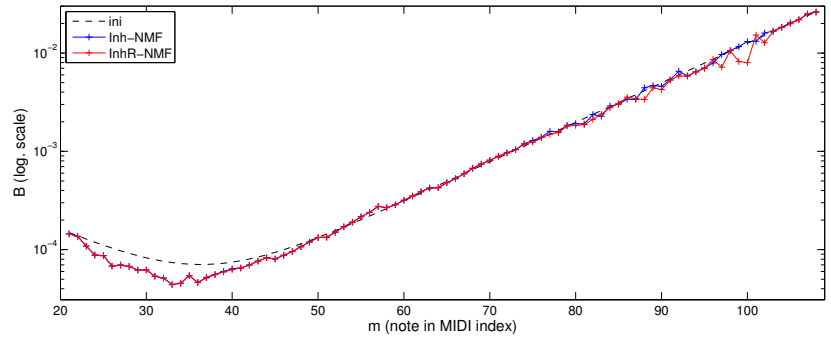


(a)

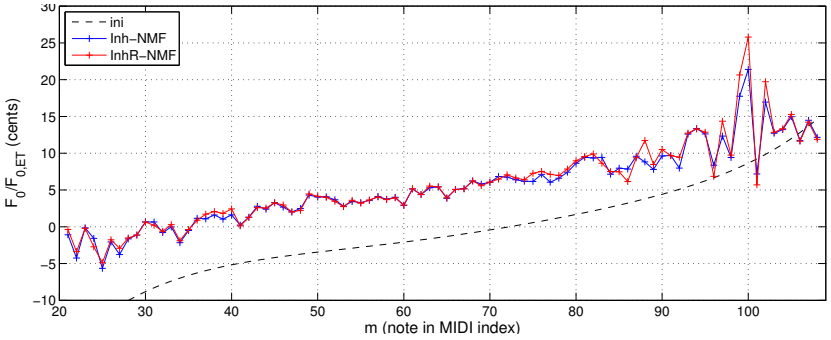


(b)

Figure D.3: RWC2 (a) B and (b) F_0 as dev. from ET along the whole compass.



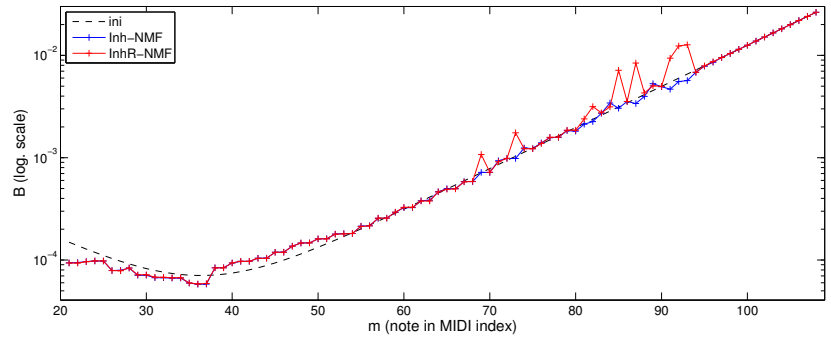
(a)



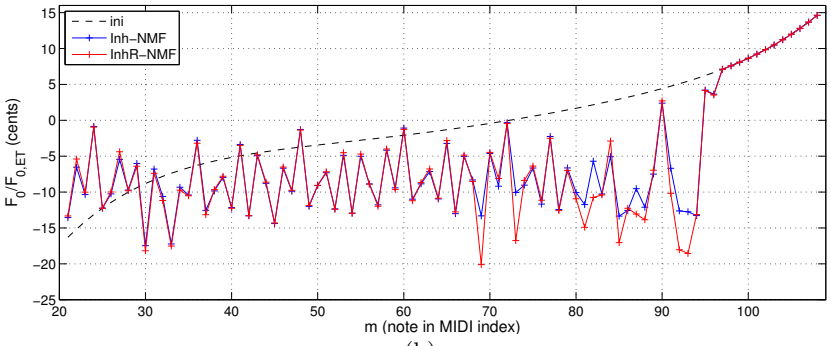
(b)

Figure D.4: RWC3 (a) B and (b) F_0 as dev. from ET along the whole compass.

- **MAPS database:**



(a)



(b)

Figure D.5: AkPnBcht (a) B and (b) F_0 as dev. from ET along the whole compass.

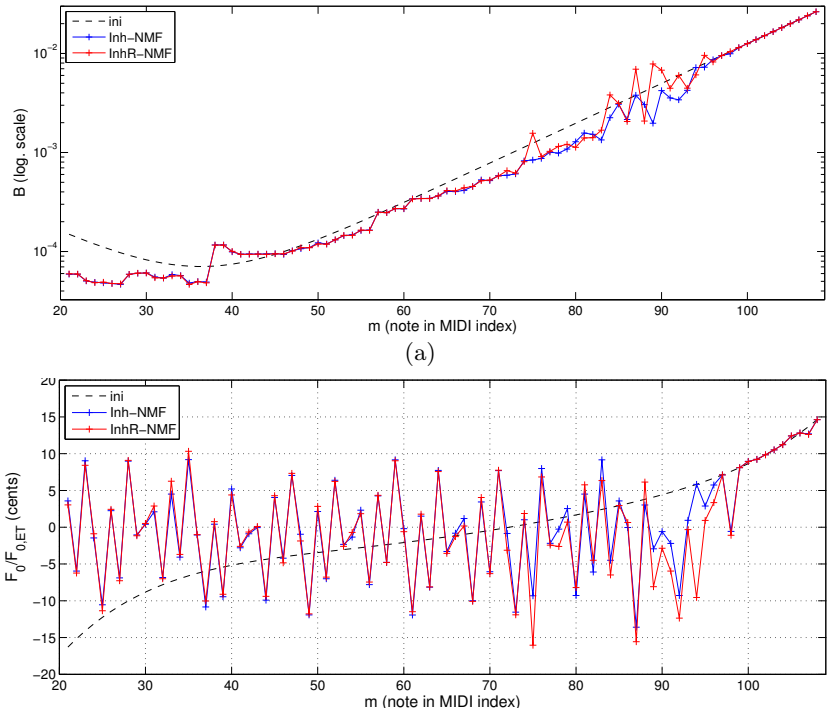


Figure D.6: AkPnBsdf (a) B and (b) F_0 as dev. from ET along the whole compass.

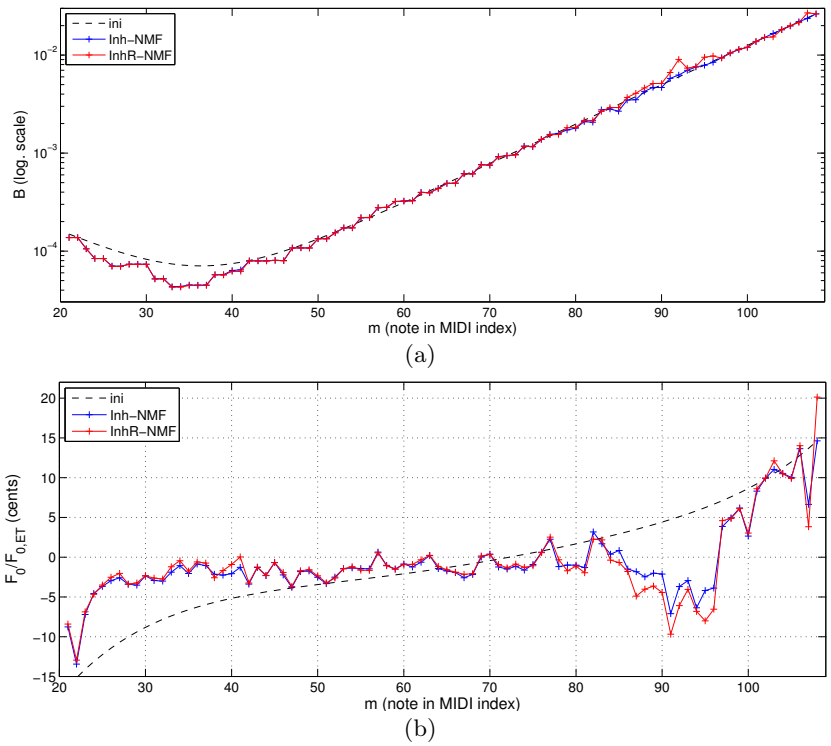


Figure D.7: AkPnCGdD (a) B and (b) F_0 as dev. from ET along the whole compass.

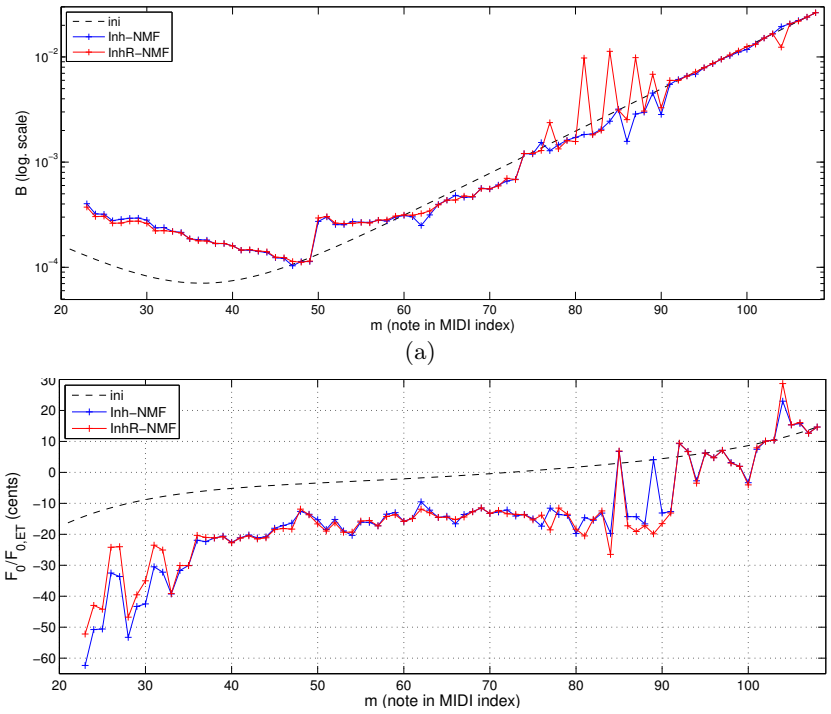


Figure D.8: AkPnStgb (a) B and (b) F_0 as dev. from ET along the whole compass.

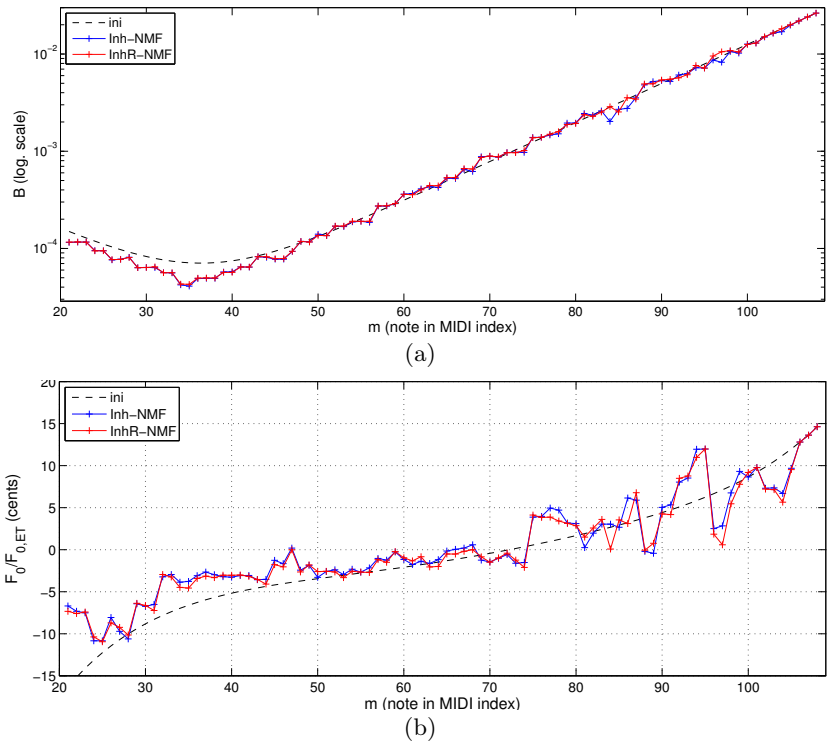
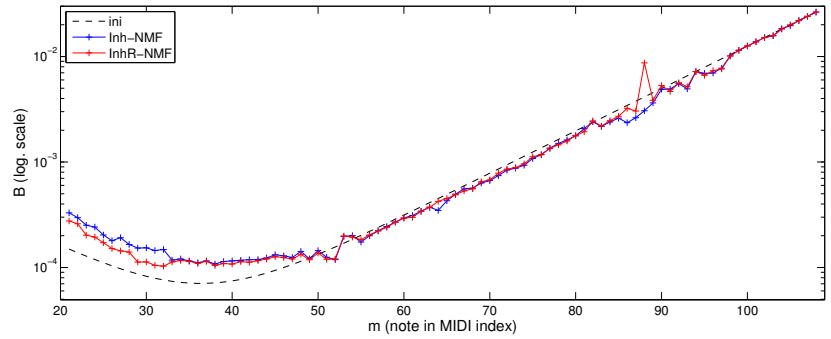
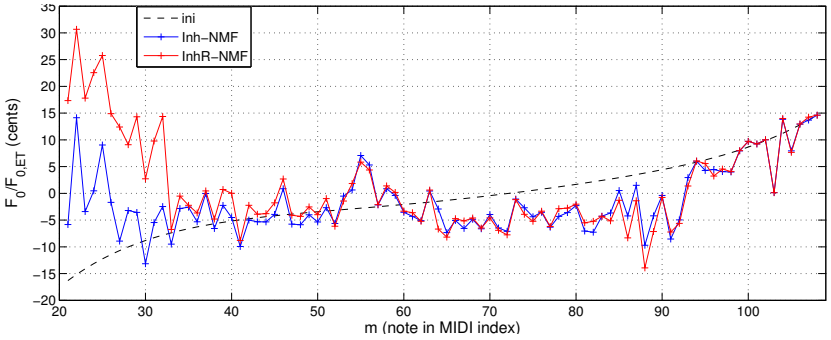


Figure D.9: StbgTGd2 (a) B and (b) F_0 as dev. from ET along the whole compass.

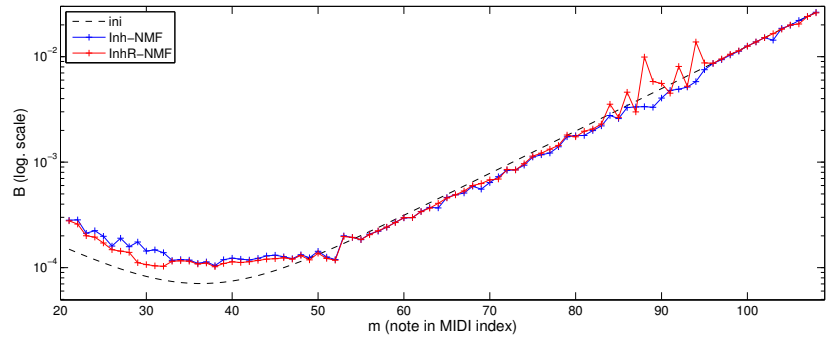


(a)

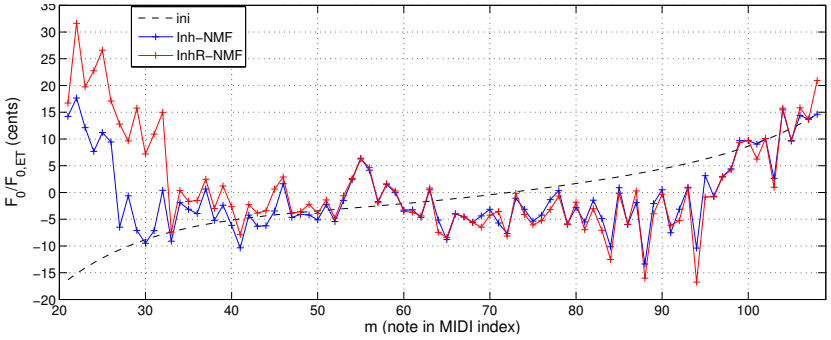


(b)

Figure D.10: ENSTDkAm (a) B and (b) F_0 as dev. from ET along the whole compass.



(a)



(b)

Figure D.11: ENSTDkCl (a) B and (b) F_0 as dev. from ET along the whole compass.

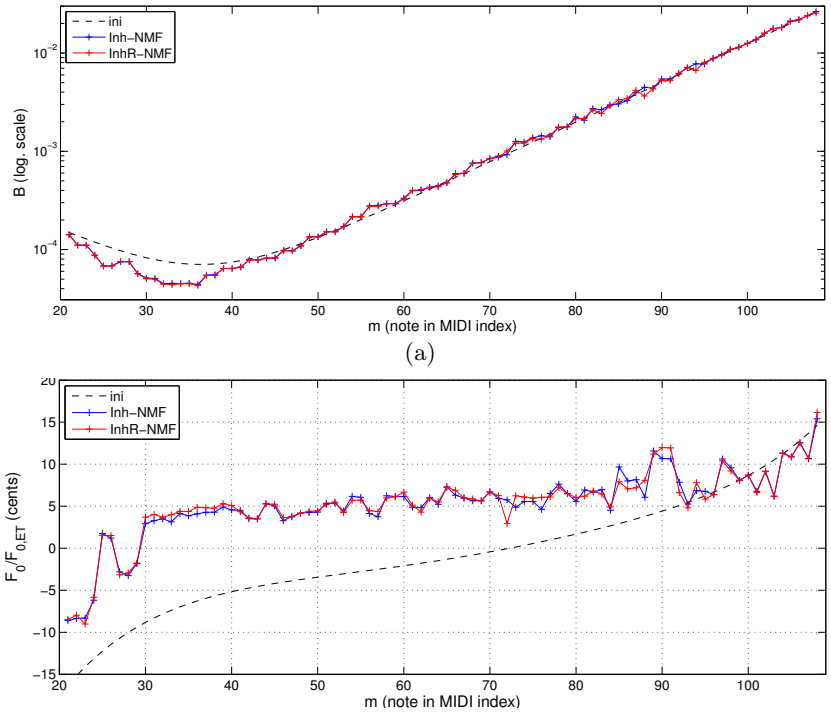


Figure D.12: SptkBGAm (a) B and (b) F_0 as dev. from ET along the whole compass.

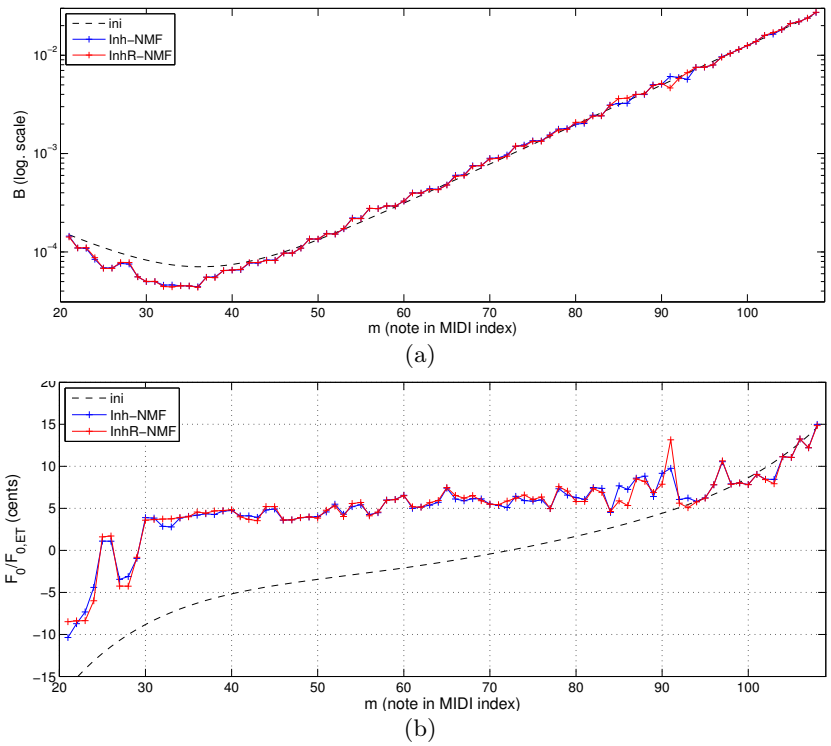


Figure D.13: SptkBGCL (a) B and (b) F_0 as dev. from ET along the whole compass.

Bibliography

Author's publications

— Peer-reviewed journal article —

Rigaud, F., David, B., and Daudet, L. (2013a). A parametric model and estimation techniques for the inharmonicity and tuning of the piano. *Journal of the Acoustical Society of America*, 133(5):3107–3118.

— Peer-reviewed conference articles —

Rigaud, F., David, B., and Daudet, L. (2011). A parametric model of piano tuning. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*, pages 393–399.

Rigaud, F., David, B., and Daudet, L. (2012). Piano sound analysis using non-negative matrix factorization with inharmonicity constraint. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 2462–2466.

Rigaud, F., Drémeau, A., David, B., and Daudet, L. (2013b). A probabilistic line spectrum model for musical instrument sounds and its application to piano tuning estimation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Rigaud, F., Falaize, A., David, B., and Daudet, L. (2013c). Does inharmonicity improve an NMF-based piano transcription model? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15.

Manuscript references

Allen, J. B. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238.

Aramaki, M., Bensa, J., Daudet, L., Guillemain, P., and Kroland-Martinet, R. (2001). Resynthesis of coupled piano string vibrations based on physical modeling. *Journal of New Music Research (Taylor & Francis Group)*, 30(3):213–226.

Badeau, R., Bertin, N., and Vincent, E. (2010). Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881.

Badeau, R., David, B., and Richard, G. (2006). High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials. *IEEE Transactions on Signal Processing*, 54(4):1341–1350.

Bank, B. and Lehtonen, H.-M. (2010). Perception of longitudinal components in piano string vibrations. *Journal of the Acoustical Society of America (JASA express letters)*, 128(3):117–123.

Bank, B. and Sujbert, L. (2005). Generation of longitudinal vibrations in piano strings: From physics to sound synthesis. *Journal of the Acoustical Society of America*, 117(4):2268–2278.

Bank, B., Zambon, S., and Fontana, F. (2010). A modal-based real-time piano synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(4):809–821.

Barbancho, I., Tardón, L. J., Sammartino, S., and Barbancho, A. M. (2012). Inharmonicity-based method for the automatic generation of guitar tablature. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 20(6):1857–1868.

Bello, J. P., Daudet, L., and Sandler, M. B. (2008). Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2241–2251.

Benetos, E. (2012). *Automatic Transcription of Polyphonic Music Exploiting Temporal Evolution*. PhD thesis, School of Electronic Engineering and Computer Science Queen Mary University of London.

Benetos, E. and Dixon, S. (2011). Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1111–1123.

Benetos, E. and Kotropoulos, C. (2008). A tensor-based approach for automatic music genre classification. In *Proceedings of the 16th European Signal Processing Conference (EUSIPCO)*.

Benetos, E., Kotti, M., and Kotropoulos, C. (2006). Musical instrument classification using non-negative matrix factorization algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–224.

Bensa, J. (2003). *Analyse et synthèse de sons de piano par modèles physiques et de signaux*. PhD thesis, Université de la méditerranée Aix-Marseille II.

Bensa, J., Bilbao, S., Kroland-Martinet, R., and Julius O. Smith, I. (2003). The simulation of piano string vibration: From physical models to finite difference schemes and digital waveguides. *Journal of the Acoustical Society of America*, 114(2):1095–1107.

Bertin, N. (2009). *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* PhD thesis, EDITE Télécom-Paristech.

Bertin, N., Badeau, R., and Vincent, E. (2009). Fast bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 29–32.

Bertin, N., Badeau, R., and Vincent, E. (2010). Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):538–549.

Blackham, E. (1965). The physics of the piano. *Scientific American*, 213:88–99.

Blanco-Martín, E., Casajus-Quiros, F. J., and Ortiz-Berenguer, L. I. (2008). An improved pattern-matching method for piano multi-pitch detection. In *124th Convention of the Audio Engineering Society (AES)*.

Bremmer, B. (2007a). Aural tuning tests for 2:1, 4:2 and 6:3 type octaves. Technical report, Piano Technicians Guild. www.billbremmer.com/articles/ (date last viewed 11/06/12).

Bremmer, B. (2007b). Midrange piano tuning. Technical report, Piano Technicians Guild. www.billbremmer.com/articles/ (date last viewed 11/06/12).

Capleton, B. (2007). *Theory and Practice of Piano Tuning. A Manual on the Art, Techniques and Theory.*, pages 1–626. Amarilli Books, Malvern, UK.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):669–696.

Chabassier, J. (2012). *Modélisation et simulation numérique d'un piano par modèles physiques*. PhD thesis, Ecole Polythechnique.

Chaigne, A. and Askenfelt, A. (1993). Numerical simulations of piano strings. II. Comparisons with measurements and systematic exploration of some hammer-string parameters. *Journal of the Acoustical Society of America*, 95(3):1631–1640.

Chaigne, A. and Kergomard, J. (2008). *Acoustique des instruments de musique*. Belin, Paris, France.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Cichocki, A., Zdunek, R., and Amari, S.-I. (2006). Csiszár's divergence for non-negative matrix factorization: Family of new algorithms. In 6^{th} *International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 32–39.

Cichocki, A., Zdunek, R., and Amari, S.-I. (2008). Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, 25(1):142–145.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.-I. (2009). *Nonnegative Matrix and Tensor Factorizations. Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley.

Cohen, J. E. and Rothblum, U. G. (1993). Nonnegative ranks, decompositions, and fatorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168.

Comon, P. (1994). Independent component analysis, a new concept. *Signal Processing special issue Higher-Order Statistics*, 36:287–314.

Conklin, Jr., H. A. (1996a). Design and tone in the mechanoacoustic piano. Part II. Piano structure. *Journal of the Acoustical Society of America*, 100(2):695–708.

Conklin, Jr., H. A. (1996b). Design and tone in the mechanoacoustic piano. Part III. Piano strings and scale design. *Journal of the Acoustical Society of America*, 100(3):1286–1298.

Conklin, Jr., H. A. (1999). Generation of partials due to nonlinear mixing in a stringed instrument. *Journal of the Acoustical Society of America*, 105(1):536–545.

Cont, A. (2006). Realtime audio to score alignment for polyphonic music instruments, using sparse non-negative constraints and hierarchical HMMs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Davy, M., Godsill, S. J., and Idier, J. (2006). Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Serie B*, 39(1):1–38.

Dessein, A., Cont, A., and Lemaitre, G. (2010). Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proceedings of the 11th International Society for Music Information Retrieval conference (ISMIR)*, pages 489–494.

Dhillon, I. S. and Sra, S. (2006). Generalized nonnegative matrix approximations with Bregman divergences. In *Proceedings of the Neural Information Processing Systems (NIPS) Conference*, pages 283–290.

Dixon, S., Mauch, M., and Tidhar, D. (2012). Estimation of harpsichord inharmonicity and temperament from musical recordings. *Journal of the Acoustical Society of America*, 131(1):878–887.

Donoho, D. and Stodden, V. (2004). When does non-negative matric factorization give a correct decomposition into parts? In *Proceedings of the conference Advances in neural information processing systems*.

Doval, B. and Rodet, X. (1991). Estimation of fundamental frequency of musical sound signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 3657–3660.

Doval, B. and Rodet, X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 221–224.

Downie, J. S. (2006). The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12).

Duan, Z., Pardo, B., and Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133.

Durrieu, J.-L., David, B., and Richard, G. (2011). A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180–1191.

Durrieu, J.-L., Richard, G., David, B., and Févotte, C. (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575.

Ege, K. (2009). *La table d'harmonie du piano. Etudes modales en basses et moyennes fréquences.* PhD thesis, Ecole Polytechnique.

Ege, K., Boutillon, X., and David, B. (2009). High-resolution modal analysis. *Journal of Sound and Vibration (Elsevier)*, 325(4-5):852–869.

Eggert, J. and Körner, E. (2004). Sparse coding and NMF. In *Joint Conference on Neural Networks*, volume 4, pages 2529–2533.

Elie, B., Gautier, F., and David, B. (2013). Estimation of mechanical properties of panels based on modal density and mean mobility measurements. *Mechanical Systems and Signal Processing (Elsevier)*, 40(2):628–644.

Emiya, V. (2008). *Transcription automatique de la musique de piano.* PhD thesis, EDITE Télécom-Paristech.

Emiya, V., Badeau, R., and David, B. (2007a). Multipitch estimation of quasi-harmonic sounds in colored noise. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)*, pages 93–98.

Emiya, V., Badeau, R., and David, B. (2010a). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing*, 18:1643–1654.

Emiya, V., Bertin, N., David, B., and Badeau, R. (2010b). MAPS - A piano database for multipitch estimation and automatic transcription of music. Technical report, Télécom ParisTech.

Emiya, V., David, B., and Badeau, R. (2007b). A parametric method for pitch estimation of piano tones. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pages 249–252.

Engelbrecht, J., Mägi, A., and Stulov., A. (1999). Grand piano manufacturing in Estonia: the problem of piano scaling. *Proceedings of the Estonian Academy of Sciences*, 5(2):155–167.

Essid, S. (2012). A single-class SVM based algorithm for computing an identifiable NMF. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2053–2056.

Ewert, S. and Müller, M. (2012). Using score-informed constraints for NMF-based source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 129–132.

Ewert, S., Müller, M., and Sandler, M. (2013). Efficient data adaption for musical source separation methods based on parametric models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Fessler, J. A. and Hero, A. O. (2010). Space-alternating generalized expectation maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677.

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis. *Neural Computation (MIT Press Journals)*, 21(3):793–830.

Févotte, C. and Cemgil, A. T. (2009). Nonnegative matrix factorization as probabilistic inference in composite models. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917.

Févotte, C. and Idier, J. (2011). Algorithm for nonnegative matrix factorization with the β-divergence. *Neural Computation (MIT Press Journals)*, 23(9):2421–2456.

FitzGerald, D., Cranitch, M., and Coyle, E. (2009). On the use of the beta divergence for musical source separation. In *Proceedings of Signals and Systems Conference (ISSC 2009), IET Irish*, pages 1–6.

Flanagan, J. L. and Golden, R. M. (1966). Phase vocoder. *Bell System Technical Journal*, 45(9):1493–1509.

Fletcher, H. (1964). Normal vibration frequencies of a stiff piano string. *Journal of the Acoustical Society of America*, 36(1):203–209.

Fletcher, H., Blackham, E. D., and Stratton, R. (1962). Quality of piano tones. *Journal of the Acoustical Society of America*, 34(6):749–761.

Fletcher, N. H. and Rossing, T. D. (1998). *The physics of musical instruments. 2nd Ed.* Springer-Verlag, New-York Inc.

Fritsch, J. and Plumbey, M. D. (2013). Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 888–891.

Galembo, A. and Askenfelt, A. (1999). Signal representation and estimation of spectral parameters by inharmonic comb filters with application to the piano. *IEEE Transactions on Speech and Audio Processing*, 7:197–203.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, 2nd edition.

Giordano, N. (1998). Mechanical impedance of a piano soundboard. *Journal of the Acoustical Society of America*, 103(4):2128–2133.

Godsill, S. and Davy, M. (2005). Bayesian computational models for inharmonicity in musical instruments. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 283–286.

Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18:294–304.

Goto, M., Nishimura, T., Hashiguchi, H., and Oka, R. (2003). RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the 4th International Society for Music Information Retrieval conference (ISMIR)*, pages 229–230.

Gough, C. E. (1981). The theory of string resonances on musical instruments. *Acta Acustica with Acustica*, 49(2):124–141.

Gribonval, R. and Bacry, E. (2003). Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111.

Hayashi, E., Yamame, M., and Mori, H. (1999). Behavior of piano-action in a grand piano. I. Analysis of the motion of the hammer prior to string contact. *Journal of the Acoustical Society of America*, 105(6):3534–3544.

Helmholtz, H. V. (1863). *On the Sensation of Tone As a Physiological Basis for the Theory of Music (Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik).*

Hennequin, R. (2011). *Décomposition de spectrogrammes musicaux informée par des modèles de synthèse spectrale. Modélisation des variations temporelles dans les éléments sonores.* PhD thesis, EDITE Télécom Paristech.

Hennequin, R., Badeau, R., and David, B. (2010). Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, pages 246–253.

Hennequin, R., Badeau, R., and David, B. (2011a). NMF with time-frequency activations to model non stationary audio events. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):744–753.

Hennequin, R., Badeau, R., and David, B. (2011b). Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*.

Hinrichsen, H. (2012). Entropy-based tuning of musical instruments. *Revista Brasileira de Ensino de Física*, 34(2):2301–1,8.

Hirschkorn, M., McPhee, J., and Birkett, S. (2006). Dynamic modeling and experimental testing of a piano action mechanism. *Journal of computational and nonlinear dynamics*, 1(1):47–55.

Hodgkinson, M., Wang, J., Timoney, J., and Lazzarini, V. (2009). Handling inharmonic series with median-adjustive trajectories. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, pages 471–477.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):471–441.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469.

Itakura, F. and Saito, S. (1968). Analysis-synthesis telephony based upon the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics*, volume 17, pages C17–C20.

Järveläinen, H., Välimäki, V., and Karjalainen, M. (2001). Auditibility of the timbral effects of inharmonicity in stringed instrument tones. *Acoustics Research Letters Online (ARLO)*, 2(3):79–84.

Järveläinen, H., Verma, T., and Välimäki, V. (2002). Perception and adjustment of pitch in inharmonic string instrument tones. *Journal of New Music Research*, 31(4):311–319(9).

Joder, C., Weninger, F., Virette, D., and Schuller, B. (2013). A comparative study on sparsity penalties for NMF-based speech separation: Beyond lp-norms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 858–862.

Kaiser, F. and Sikora, T. (2010). Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the 11th International Society for Music Information Retrieval conference (ISMIR)*, pages 429–434.

Kameoka, H., Nishimoto, T., and Sagayama, S. (2004). Multi-pitch detection algorithm using constrained gaussian mixture model and information criterion for simultaneous speech. In *Proceedings of Speech Prosody (SP2004)*, pages 533–536.

Klapuri, A. and Davy, M. (2006). *Signal Processing Methods for Music Transcription*. Springer-Verlag New York Inc.

Klapuri, A. P. (2001). Multipitch estimation and sound separation by the spectral smoothness principle. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3381–3384.

Klingenberg, B., Curry, J., and Dougherty, A. (2009). Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Patter Recognition*, 42(5):918–928.

Kobzantsev, A., Chazan, D., and Zeevi, Y. (2005). Automatic transcription of polyphonic piano music. In *Proceedings of the 4th International Symposium on Image and Signal Processing Analysis (ISPA)*, pages 414–418.

Kompass, R. (2007). A generalized divergence measure for nonnegative matrix factorization. *Neural Computation*, 19(3):780–791.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lattard, J. (1993). Influence of inharmonicity on the tuning of a piano. *Journal of the Acoustical Society of America*, 94(1):46–53.

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K., and Jensen, S. H. (2008). Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008:1–9.

Le Carrou, J. L., Gautier, F., Dauchez, N., and Gilbert, J. (2005). Modelling of sympathetic string vibrations. *Acta Acustica united with Acustica*, 91(2):277–288.

Le Roux, J., Kameoka, H., Ono, N., de Cheveigné, A., and Sagayama, S. (2011). Computational auditory induction by missing-data non-negative matrix factorization. *Speech Communication. Special issue on Perceptual and Statistical Audition*, 53(5):658–676.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562.

Lehtonen, H.-M., Penttinen, H., Rauhala, J., and Välimäki, V. (2007). Analysis and modeling of piano sustain-pedal effects. *Journal of the Acoustical Society of America*, 122(3):1787–1797.

Leveau, P., Vincent, E., Richard, G., and Daudet, L. (2008). Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):116–128.

Lin, C.-J. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10):2756–2779.

Liu, W. and Yuan, K. (2008). Sparse p-norm nonnegative matrix factorization for clustering gene expression data. *International Journal of Data Mining and Bioinformatics*, 2(3):236–249.

Liutkus, A., Badeau, R., and Richard, G. (2011). Gaussian processes for undertermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155–3167.

Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition: The Sparce Way*. Academic Press.

Mallat, S. and Zhang, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

Mamou-Mani, A., Frelat, J., and Besnainou, C. (2008). Numerical simulation of a piano soundboard under downbearing. *Journal of the Acoustical Society of America*, 123(4):2401–2406.

Marolt, M. (2004). A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449.

Martin, D. W. and Ward, W. D. (1961). Subjective evaluation of musical scale temperament in pianos. *Journal of the Acoustical Society of America*, 33(5):582–585.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754.

Monti, G. and Sandler, M. (2002). Automatic polyphonic piano note extraction using fuzzy logic in a blackboard system. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx)*, pages 39–44.

Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1984). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, 77(5):1853–1860.

Morse, P. M. (1948). *Vibration and Sound*. American Institute of Physics for the Acoustical Society of America.

Mysore, G. J., Smaragdis, P., and Raj, B. (2010). Non-negative hidden markov modeling of audio with application to source separation. In *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 140–148.

Nakano, M., Le Roux, J., Kameoka, H., Kitano, Y., Ono, N., and Sagayama, S. (2010). Nonnegative matrix factorization with markov-chained bases for modeling time-varying patterns in music spectrograms. In *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 149–156.

Niedermayer, B. (2008). Non-negative matrix division for the automatic transcription of polyphonic music. In *Proceedings of the 9th International Society for Music Information Retrieval conference (ISMIR)*, pages 544–549.

Ochiai, K., Kameoka, H., and Sagayama, S. (2012). Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 133–136.

Ortiz-Berenguer, L. I., Casajús-Quirós, F. J., Torres-Guijarro, M., and Beracoechea, J. A. (2004). Piano transcription using pattern recognition: aspects on parameter extraction. In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx-04)*, pages 212–216.

Ortiz-Berenguer, L. I. and Casajús-Quirós, F. J. (2002). Polyphonic transcription using piano modeling for spectral pattern recognition. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, pages 45–50.

Ozerov, A. and Févotte, C. (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550–563.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

Parry, R. M. and Essa, I. (2007a). Incorporating phase information for source separation via spectrogram factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 661–664.

Parry, R. M. and Essa, I. (2007b). Phase-aware non-negative spectrogram factorization. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, pages 536–543.

Parvaix, M. and Girin, L. (2011). Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. volume 19, pages 1721–1733.

Pauca, V. P., Shahnaz, F., Berry, M. W., and Plemmons, R. J. (2004). Text mining using non-negative matrix factorizations. In *Proceedings of the 4th SIAM International Conference on Data Mining*, pages 452–456.

Paulus, J. and Virtanen, T. (2005). Drum transcription with non-negative spectrogram factorization. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO 2005)*.

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America*, 41(6):1526–1533.

Poliner, G. E. and Ellis, D. (2007). A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007(1).

Raj, B., Virtanen, T., Chaudhuri, S., and Singh, R. (2010). Non-negative matrix factorization based compensation of music for automatic speech recognition. In *Proceeding of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 717–720.

Rasch, R. A. and Heetvelt, V. (1985). String inharmonicity and piano tuning. *Music Perception*, 3(2):171–190.

Rauhala, J., Lehtonen, H. M., and Välimäki, V. (2007a). Fast automatic inharmonicity estimation algorithm. *Journal of the Acoustical Society of America (Express Letters)*, 121:184–189.

Rauhala, J., Lehtonen, H. M., and Välimäki, V. (2007b). Toward next-generation digital keyboard instruments. *IEEE Signal Processing Magazine*, 24(2):12–20.

Rauhala, J. and Välimäki, V. (2007). F0 estimation of inharmonic piano tones using partial frequencies deviation method. In *Proceedings of the International Computer Music Conference (ICMC'07)*, pages 453–456.

Rijsbergen, C. (1979). *Information Retrieval*. Butterworths, London, UK, 2nd edition.

Ritsma, R. J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America*, 42(1):191–198.

Roads, C. (2007). *L'audionumérique. Musique et informatique (The Computer Music Tutorial).* Dunod, 2nd edition.

Rossing, T. D. and Fletcher, N. H. (1995). *Principles of vibration and sound.* Springer-Verlag, 2nd edition.

Serra, X. and Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24.

Smaragdis, P. and Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Application of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180.

Smaragdis, P., Raj, B., and Shashanka, M. (2008). Sparse and shift-invariant feature extraction from non-negative data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2069–2072.

Stulov, A. (1995). Hysteretic model of the grand piano hammer felt. *Journal of the Acoustical Society of America*, 97(4):2577–2585.

Stulov, A. (2008). Physical modelling of the piano string scale. *Applied Acoustics (Elsevier, Kidlington, United Kingdom)*, 69(11):977–984.

Suzuki, H. (1986a). Model analysis of a hammer-string interaction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1986)*, pages 1285–1288.

Suzuki, H. (1986b). Vibration and sound radiation of a piano soundboard. *Journal of the Acoustical Society of America*, 80(6):1573–1582.

University of Iowa (1997). Music Instrument Samples database (date last viewed: 2013/07/30). `http://theremin.music.uiowa.edu/MISpiano.html`.

Valette, C. and Cuesta, C. (1993). *Mécanique de la corde vibrante.* Hermes Science Publications.

Vincent, E., Bertin, N., and Badeau, R. (2008). Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 109–112.

Vincent, E., Bertin, N., and Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):528–537.

Vincent, E. and Plumbey, M. D. (2007). Low bitrate object coding of musical audio using bayesian harmonic models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1273–1282.

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074.

Virtanen, T., Cemgil, T., and Godsill, S. (2008). Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1825–1828.

Virtanen, T. and Klapuri, A. (2002). Separation of harmonic sounds using linear models for the overtone series. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1757–1760.

Wang, W. and Zou, X. (2008). Non-negative matrix factorization based on projected nonlinear conjugate gradient algorithm. In *Proceedings of ICA Research Network International Workshop (ICARN)*, pages 5–8.

Weinreich, G. (1977). Coupled piano strings. *Journal of the Acoustical Society of America*, 62(6):1474–1484.

Wilson, K. W., Rah, B., Smaragdis, P., and Divakaran, A. (2008). Speech denoising using nonnegative matrix factorization with priors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4029–4032.

Yeh, C. and Röbel, A. (2006). Adaptive noise level estimation. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pages 145–148.

Yoshioka, T. and Sakaue, D. (2012). Log-normal matrix factorization with application to speech-music separation. In *Proceedings of the 2012 Workshop on Statistical and Perceptual Audition Organized Jointly with the Speech Communication with Adaptive Learning Consortium (SAPA-SCALE 2012)*, pages 80–85.

Young, R. W. (1952). Inharmonicity of plain wire piano strings. *Journal of the Acoustical Society of America*, 24(3):267–273.

Zdunek, R. and Cichocki, A. (2007). Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing*, 87(8):1904–1916.

Zhang, Y. and Fang, Y. (2007). A NMF algorithm for blind separation of uncorrelated signals. In *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, pages 999–1003.

Remerciements

Ces dernières pages sont pour moi l'occasion de remercier les personnes ayant contribué de près ou de loin à l'élaboration de cette thèse.

Mes premiers remerciements vont à mes encadrants, Bertrand David et Laurent Daudet, qui sont à l'origine de ce sujet original, situé à l'intersection entre des problématiques de traitement du signal et d'acoustique musicale. Leur vision croisée de ces deux domaines et leur expertise m'auront permis de mener à bien ces recherches tout en m'aidant à prendre du recul lors des difficultés rencontrées. Une grande partie de leurs intuitions et des pistes lancées lors des premières réunions se sont avérées fructueuses tout au long de ces trois années. Enfin, leurs travaux de relecture ont été précieux, en particulier au cours de l'été "chaud" de rédaction de ce manuscrit.

Une partie des résultats de cette thèse sont le fruit d'échanges et de collaborations avec diverses personnes. En particulier, je remercie Angélique Drémeau et Antoine Falaize dont les collaborations fructueuses ont données lieu a des publications, ainsi que Romain Hennequin pour les discussions et conseils du début de thèse.

Je tiens aussi à remercier Vesa Välimäki, Emmanuel Vincent, Philippe Depalle, Simon Dixon et Cédric Févotte pour l'intérêt qu'ils ont porté à mes travaux en acceptant l'invitation au jury de soutenance. Leurs relectures minutieuses ont grandement contribué à l'amélioration de la version finale de ce manuscrit.

L'environnement de travail et la bonne ambiance du laboratoire, et en particulier de l'équipe AAO, auront aussi fortement participé au bon déroulement de cette thèse. Je souhaite ainsi remercier (autant pour leurs qualités scientifiques qu'humaines) les collègues que j'ai rencontré au cours de ces trois années : Félicien, Benoit M&F, Mounira, Romain, Antoine, Manu, Thomas, Rémi, Sébastien, Aymeric, Nicolas, Angélique, Cécilia, Xabier, Davide, Laure, Hequn, Martin, Anne-Claire, Cyril, Fabrice associé doctorant AAO, Gaël, Slim, Roland, Bertrand, Alexandre, Yves.

Enfin, impliqués de façon indirecte mais tout aussi importante à mes yeux, je souhaite remercier mes amis proches (Louis, Dinhu, Indiana, Robuzz, Papy, Ludo, Julian, Matys) et ma famille pour leur soutien et les moments de décompression.

Je conclurai en remerciant tout spécialement MM pour son soutien au cours des derniers mois ainsi que pour tous les échanges que nous avons pu avoir.

Writing is the flip side of sex – it's good only when it's over.

Dr. Hunter S. Thompson

Modèles de signaux musicaux informés par la physique des instruments. Application à l'analyse de musique pour piano par factorisation en matrices non-négatives.

RESUME : Cette thèse introduit des nouveaux modèles de signaux musicaux informés par la physique des instruments. Alors que les communautés de l'acoustique instrumentale et du traitement du signal considèrent la modélisation des sons instrumentaux suivant deux approches différentes (respectivement, une modélisation du mécanisme de production du son, opposée à une modélisation des caractéristiques "morphologiques" générales du son), cette thèse propose une approche collaborative en contraignant des modèles de signaux génériques à l'aide d'information basée sur l'acoustique. L'effort est ainsi porté sur la construction de modèles spécifiques à un instrument, avec des applications aussi bien tournées vers l'acoustique (apprentissage de paramètres liés à la facture et à l'accord) que le traitement du signal (transcription de musique). En particulier nous nous concentrons sur l'analyse de musique pour piano, instrument pour lequel les sons produits sont de nature inharmonique. Cependant, l'inclusion d'une telle propriété dans des modèles de signaux est connue pour entraîner des difficultés d'optimisation, allant jusqu'à endommager les performances (en comparaison avec un modèle harmonique plus simple) dans des tâches d'analyse telles que la transcription. Un objectif majeur de cette thèse est d'avoir une meilleure compréhension des difficultés liées à l'inclusion explicite de l'inharmonicité dans des modèles de signaux, et d'étudier l'influence de l'apport de cette information sur les performances d'analyse, en particulier dans une tâche de transcription.

Dans ce but, nous introduisons différents modèles basés sur des méthodes génériques tels que la Factorisation en Matrices Non-négatives (deux modèles de spectres inharmoniques NMF) et le cadre Bayésien (un modèle probabiliste génératif des fréquences inharmoniques du spectre). Les algorithmes d'estimation correspondant sont introduits, avec une attention particulière portée sur l'initialisation et l'optimisation afin d'éviter une convergence vers des minima locaux. Ceux-ci sont appliqués à l'estimation précise de paramètres physiques liés à la facture et à l'accord de différents pianos, à partir d'enregistrements monophoniques et polyphoniques, pour des conditions supervisées (les notes jouées sont connues) et non-supervisées.

Les variations le long de la tessiture de ces paramètres physiques sont ensuite étudiées en introduisant un modèle joint d'inharmonicité et d'accord basé sur des invariances dans les règles de facture et d'accord. Outre des applications à l'analyse, l'utilité de ces modèles est démontrée pour l'obtention de courbes d'accord de références pour les pianos désaccordés ou les synthétiseurs basés sur une modélisation physique, pour l'initialisation des paramètres des algorithmes d'analyse, et finalement pour interpoler l'inharmonicité et l'accord le long de la tessiture d'un piano à partir de l'analyse d'une pièce de musique polyphonique contenant un ensemble réduit de notes.

Finalement, l'efficacité d'un modèle NMF inharmonique pour la transcription de musique pour piano est étudiée en comparant les deux modèles inharmoniques proposés avec un modèle harmonique. Les résultats montrent que les performances sont améliorées par l'ajout de l'information d'inharmonicité, à condition que les paramètres physiques soient suffisamment bien estimés. En particulier, une augmentation significative des performances est obtenue lors de l'utilisation d'une initialisation appropriée.

Mots clés : *modèles de signaux musicaux, NMF, inharmonicité, accord des pianos, transcription*

Models of music signals informed by physics. Application to piano music analysis by non-negative matrix factorization.

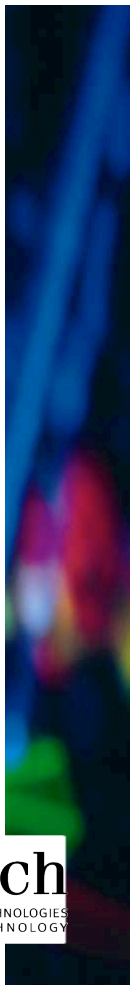
ABSTRACT: This thesis introduces new models of music signals informed by the physics of the instruments. While instrumental acoustics and audio signal processing target the modeling of musical tones from different perspectives (modeling of the production mechanism of the sound *vs* modeling of the generic "morphological" features of the sound), this thesis aims at mixing both approaches by constraining generic signal models with acoustics-based information. Thus, it is here intended to design instrument-specific models for applications both to acoustics (learning of parameters related to the design and the tuning) and signal processing (transcription). In particular, we focus on piano music analysis for which the tones have the well-known property of inharmonicity. The inclusion of such a property in signal models however makes the optimization harder, and may even damage the performance in tasks such as music transcription when compared to a simpler harmonic model. A major goal of this thesis is thus to have a better understanding about the issues arising from the explicit inclusion of the inharmonicity in signal models, and to investigate whether it is really valuable when targeting tasks such as polyphonic music transcription.

To this end, we introduce different models of piano tones built on generic signal frameworks such as those of Non-negative Matrix Factorization (NMF with two different models of inharmonic spectra) and Bayesian probability (an inharmonic line spectrum model). Corresponding estimation algorithms are derived, with a special care in the initialization and the optimization scheme in order to avoid the convergence of the algorithms toward local optima. These algorithms are applied to the precise estimation of physical parameters related to the design and tuning of different pianos from monophonic and polyphonic recordings, in both supervised (played notes are known) and unsupervised conditions.

The variations along the piano compass of such physical parameters are then modeled by introducing a joint model of inharmonicity and tuning based on invariants in design and tuning rules. Beyond analysis applications, the usefulness of this model is also demonstrated for providing tuning curves for out-of-tune pianos or physically-based synthesizers, for initializing the parameters of analysis algorithms, and finally to interpolate the inharmonicity and tuning of pianos along the whole compass from the analysis of a polyphonic recording containing only a few notes.

Finally the efficiency of an inharmonic model for NMF-based transcription is investigated by comparing the two proposed inharmonic NMF models with a simpler harmonic model. Results show that it is worth considering inharmonicity of piano tones for a transcription task provided that the physical parameters are sufficiently well estimated. In particular, a significant increase in performance is obtained when using an appropriate initialization of these parameters.

Keywords: *models of music signals, NMF, inharmonicity, piano tuning, transcription*



TELECOM ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY