Automatic Music Transcription based on Prior Knowledge from Musical Acoustics. Application to the repertoires of the Marovany zither of Madagascar

Dorian Cazau

▶ **To cite this version:**

Dorian Cazau. Automatic Music Transcription based on Prior Knowledge from Musical Acoustics. Application to the repertoires of the Marovany zither of Madagascar. Acoustics [physics.class-ph]. Université Pierre et Marie Curie - Paris VI, 2015. English. NNT : 2015PA066640 . tel-01343960

HAL Id: tel-01343960

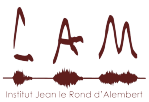
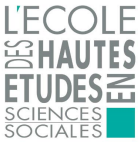
https://theses.hal.science/tel-01343960

Submitted on 11 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



École Doctorale de Science Mécanique, Acoustique, Electronique et Robotique

PhD Thesis

Discipline : Acoustics and Signal Processing

presented by

Dorian CAZAU

Automatic Music Transcription based on Prior Knowledge from Musical Acoustics. Application to the repertoires of the Marovany zither of Madagascar.

---

supervised by Olivier ADAM and Marc CHEMILLIER

Defended the 12^{th} October 2015 in front of the jury :

Pr. Laurent Daudet	University Paris Diderot	Rapporteur
Pr. Tillman Weyde	City University London	Rapporteur
Pr. Claude Depollier	University of the Maine	Examiner
Dr. Andre Holzapfel	Boğaziçi University	Examiner
Pr. Gael Richard	Telecom ParisTech	Examiner
Pr. Jean-Luc Zarader	Université Pierre and Marie Curie	Examiner
Pr. Olivier Adam	University Pierre and Marie Curie	Supervisor
Pr. Marc Chemillier	School for Advanced Studies in the Social Sciences	Co-Supervisor

Institut Jean-le-Rond-d'Alembert
Equipe Lutherie-Acoustique-Musique
11, rue de Lourmel
75015 PARIS

UPMC
Ecole Doctorale de Sciences Mécanique,
Acoustique, Electronique et Robotique
4 place Jussieu
75252 Paris Cedex 05
Boite courrier 290

Remerciements

Mes premiers mots de remerciement iront spontanément à Olivier, mentor de Master puis directeur de thèse, et j'espère collaborateur et ami pour de nombreuses années à venir. Je tiens ensuite à remercier Marc Chemillier, brillant co-directeur de ma thèse, qui a été à l'origine de ce projet de recherche passionant, amenant ainsi dans ma vie la cithare *marovany* et la musique malgache.

Ce projet de thèse a fait l'objet de nombreuses collaborations fructueuses et excitantes. Avec tout d'abord la Team : GuiGui, la Juju et Wang, merci à vous les gars pour votre boulot sur ce projet. J'espère pouvoir continuer à travailler avec vous dans les années à venir. L'équipe des ingés-sons "métalleux", menée par John Mary. Gregory Nuel, redoutable professeur de statistiques, que j'ai eu de la chance de rencontrer sur mes derniers mois de thèse. Toute l'équipe du LAM, notamment Laurent Quartier, pour son énorme travail à mes côtés sur les capteurs et le bricolage d'instruments, et bien sûr Hugues Genevois, pour m'avoir accueilli dans son labo et souvent conseillé sur l'aspect capteurs de ma thèse.

Il me faut maintenant remercier les nombreux musiciens que j'ai eu la chance de rencontrer durant ce projet. Dont l'incroyable "Papa" François Essindi alias Abakuya, pour la foi et la force qu'il porte dans des causes difficiles. Kilema, pour ta joie de vivre et l'enthousiasme que tu as mis systématiquement dans nos quelques séances de travail. Charles Kely, pour t'être rendu disponible pour mon projet, et tous les autres.

Enfin, je tiens à remercier mes proches pour leur soutien journalier. Mes parents, pour m'avoir supporté durant ces trois années. Caroline, pour son aide précieuse dans la fastidieuse tâche de recherche et de gestion bibliographique.

Abstract

Abstract

Ethnomusicology is the study of musics around the world that emphasize their cultural, social, material, cognitive and/or biological. This PhD subject, initiated by Pr. Marc CHEMILLIER, ethnomusicolog at the laboratory CAMS-EHESS, deals with the development of an automatic transcription system dedicated to the repertoires of the traditional *marovany* zither from Madagascar. These repertoires are orally transmitted, resulting from a process of memorization/transformation of original base musical motives. These motives represent an important culture patrimony, and are evolving continually under the influences of other musical practices and genres mainly due to globalization. Current ethnomusicological studies aim at understanding the evolution of the traditional repertoire through the transformation of its original base motives, and preserving this patrimony. Our objectives serve this cause by providing computational tools of musical analysis to organize and structure audio recordings of this instrument.

Automatic Music Transcription (AMT) consists in automatically estimating the notes in a recording, through three attributes: onset time, duration and pitch. On the long range, AMT systems, with the purpose of retrieving meaningful information from complex audio, could be used in a variety of user scenarios such as searching and organizing music collections with barely any human labor. One common denominator of our different approaches to the task of AMT lays in the use of explicit music-related prior knowledge in our computational systems. A first step of this PhD thesis was then to develop tools to generate automatically this information. We chose not to restrict ourselves to a specific prior knowledge class, and rather explore the multi-modal characteristics of musical signals, including both timbre (i.e. modeling of the generic "morphological" features of the sound related to the physics of an instrument, e.g. intermodulation, sympathetic resonances, inharmonicity) and musicological (e.g. harmonic transition, playing dynamics, tempo and rhythm) classes. This prior knowledge can then be used in computational systems of transcriptions. The research work on AMT performed in this PhD can be divided into a more "applied research" (axis 1), with the development of ready-to-use operational transcription tools meeting the current needs of ethnomusicologs to get reliable automatic transcriptions, and a more "basic research" (axis 2), providing deeper insight into the functioning of these tools.

Our first axis of research requires a transcription accuracy high enough (i.e. average F-measure superior to 95 % with standard error tolerances) to provide analytical supports for musicological studies. Despite a large enthusiasm for AMT challenges, and several audio-to-MIDI converters available commercially, perfect polyphonic AMT systems are out of reach of today's algorithms. In this PhD, we explore the use of multichannel capturing sensory systems for AMT of several acoustic plucked string instruments, including the following traditional African zithers: the *marovany* (Madagascar), the Mvet (Camer-

oun), the N'Goni (Mali). These systems use multiple string-dependent sensors to retrieve discriminatingly some physical features of their vibrations. For the AMT task, such a system has an obvious advantage in this application, as it allows breaking down a polyphonic musical signal into the sum of monophonic signals respective to each string. Then, we come back to a monophonic transcription problem, which is considered as practically solved. For sake of flexibility and robustness, various sensor types (optical, piezoelectric and electromagnetic) have been comparatively tested. After experimentation, piezoelectric sensors, although quite invasive, prove to provide the best signal-to-noise ratio and multichannel separability. The development of this technology has allowed the constitution of a new sound dataset dedicated to AMT evaluation for plucked-string instrument repertoires. We gathered in these datasets audio recordings, MIDI-like transcripts and sound samples over the instrument pitch ranges. We also performed field recordings in Madagascar with local musicians during two missions (in July 2013 and 2014), using our multi-sensor retrieval systems. Such systems were also explored towards applications of human-machine interaction (through the project ImproteK) and musical creativity (through "MIDIfication" of traditional acoustic instruments).

Our second axis of research tackles the AMT task on audio recordings with more fundamental investigations on the use of prior knowledge in transcription performance, in regards to different plucked-string instrument repertoires. AMT is often divided into different processing stages, generally a multi-pitch estimation stage, followed by note segmentation and post-processing stages. In this PhD thesis, we mainly build our AMT framework from two methods, namely Probabilistic Latent Component Analysis (PLCA) for multi-pitch estimation and Hidden Markov Models (HMMs) for note segmentation and sequential post-processing. PLCA belongs to a spectrogram-factorization class of methods which is based on the modeling of a signal as a sum of basic elements. HMMs are a ubiquitous statistical tool to model time series data. We then develop different configurations of these methods to provide a powerful probabilistic framework covering the time-frequency domain on different time-scales, in which we develop original integration methods of prior knowledge. Timbre prior knowledge class is used to constrain generic signal models with acoustics-based information. A pitch-dependent sparsity prior has then been developed by modifying the EM update rules of PLCA, which is informed by the phenomenon of sympathetic resonances characterized acoustically. The use of pitch-wise multi-templates corresponding to different playing modes (e.g. dynamics, plucking techniques) have also been investigated, as well as the characterization of pitch activation profiles, which can be informed by specific temporal envelop modulations (e.g. intermodulation). For its part, musicological knowledge concerns more temporal musical structure and can be integrated conveniently into the HMM framework to build informed relations between frame-wise estimations. We develop original first- and second-order HMMs to model musicological polyphonic harmonic transitions between note mixtures, as well as higher-order HMMs including note duration modeling applied to note segmentation. This research axis has first allowed achieving successful transcription enhancements in the *marovany* repertoires, by optimizing selectively the integration of prior knowledge. As an illustrative result, the traditional repertoire of the *marovany* benefits mostly from timbre-related prior knowledge, namely spectral inharmonicity and energy modulation, intermodulation and sympathetic resonances. These features have been highlighted by our acoustic characterization preceding the task of AMT, used to guide the design of the repertoire-specific priors. Also, we provide a complete framework to have a better understanding about the issues arising from the explicit inclusion of specific prior knowledge in specific instrument repertoires, and to investigate whether it is really valuable when targeting tasks such as AMT. In that

direction, we also developed the PLCA plus particle filtering framework, aiming in particular to resolve current limits of the EM algorithm in the PLCA parameter estimation, as well as proposing a more flexible unifying framework for prior integration.

Through our study of the *marovany* zither, we pave the way towards the development of AMT systems dedicated to other traditional plucked-string instruments. In future investigations, we will transpose this research framework to the repertoires of the Indian sitar and the Chinese GuQin zither.

Keywords

Automatic Music Transcription, Statistical modeling and learning, Musical Acoustics knowledge, Non-eurogenetic music, Computational Ethnomusicology, Audio Signal Processing

---

Développement d'un système de transcription automatique de musique dédié aux répertoires de la cithare *marovany*

Résumé

L'ethnomusicologie est l'étude de la musique en mettant l'accent sur les aspects culturels, sociaux, matérielles, cognitives et/ou biologiques. Ce sujet de thèse, motivé par Pr. Marc Chemillier, ethnomusicologue au laboratoire CAMS-EHESS, traite du développement d'un système automatique de transcription dédié aux répertoires de musique de la cithare *marovany* de Madagascar. Ces répertoires sont transmis oralement, résultant d'un processus de mémorisation/transformation de motifs musicaux de base. Ces motifs sont un patrimoine culturel important du pays, et évoluent en permanence sous l'influence d'autres pratiques et genres musicaux. Les études ethnomusicologiques actuelles visent à comprendre l'évolution du répertoire traditionnel, et de préserver ce patrimoine. Pour servir cette cause, notre travail consiste à fournir des outils informatiques d'analyse musicale pour organiser et structurer des enregistrements audio de cet instrument.

La transcription automatique de musique consiste à estimer les notes d'un enregistrement à travers les trois attributs : temps de début, hauteur et durée de note. Notre travail sur cette thématique repose sur l'incorporation de connaissances musicales a priori dans les systèmes informatiques. Une première étape de cette thèse fût donc de générer cette connaissance et de la formaliser en vue de cette incorporation. Cette connaissance explorer les caractéristiques multi-modales du signal musical, incluant le timbre, le langage musical et les techniques de jeu. La recherche effectée dans cette thèse se distingue en deux axes : un premier plus appliqué, consistant à développer un système de transcription de musique dédié à la *marovany*, et un second plus fondamental, consistant à fournir une analyse plus approfondie des contributions de la connaissance dans la transcription automatique de musique.

Notre premier axe de recherche requiert une précision de transcription très bonne (c.a.d une F-measure supérieure à 95 % avec des tolérances d'erreur standardes) pour faire office de supports analytiques dans des études musicologiques. Pour cela, nous utilisons une technologie de captation multicanale appliquée aux instruments à cordes pincées. Les systèmes développés à partir de cette technologie utilisent un capteur par corde, permettant de décomposer un signal polyphonique en une somme de signaux monophoniques respectifs à chaque corde, ce qui simplifie grandement la tâche de transcription. Différents types de capteurs (optiques, piézoélectriques, électromagnétiques) ont été testés. Après expérimentation, les capteurs piézoélectriques, bien qu'invasifs, se sont avérés avoir les meilleurs rapport signal-sur-bruit et séparabilité inter-capteurs. Cette technologie a aussi permis le développement d'une base de donnée dite "ground truth" (vérité de terrain), indispensable pour l'évaluation quantitative des systèmes de transcription de musique.

Notre second axe de recherche propose des investigations plus approfondies concernant l'incorporation de connaissance a priori dans les systèmes automatiques de transcription de musique. Deux méthodes statistiques ont été utilisées comme socle théorique, à savoir le PLCA (Probabilistic Latent Component Analysis) pour l'estimation multi-pitch et le HMM (Hidden Markov Models) pour le post-traitement des estimations et la segmentation de notes. La méthode PLCA appartient à la famille des méthodes de factorisation non-négative de spectrogrammes, basé sur une modélisation du signal musical comme une somme de noyaux harmoniques, dites bases spectrales. HMM permet une modélisation probabiliste du processus génératif de séries temporelles. Nous avons développé différentes extensions de ces méthodes de base, pour constuire un cadre probabiliste général couvrant les domaines fréquentielles et temporelles à différentes échelles, dans lequel nous avons pu incorporer différentes composantes de connaissance a prior. Les composantes de connaissance sur le timbre sont utilisées pour contraindre des modèles de signaux en paramétrant leurs paramètres acoustiques. Un a priori de parcimonie spécifique à chaque base spectrale a été défini en modifiant les règles d'update de l'algorithme EM, et qui est informé par de l'information sur le phénomène de résonance par sympathie. L'utilisation de templates multiples spécifique à chaque base spectrale correspondant à différents modes de jeu (par exemple, dynamique de jeu, techniques de pincement) ont été développés. Nous avons aussi proposé une caractérisation acoustique des profils d'activation temporels, basé sur la connaissance acoustique de l'enveloppe temporelle des notes. Les composantes de connaissance sur le langage musical sont utilisées pour modéliser la structure temporelle d'une pièce musicale, et les transitions harmoniques entre aggrégats de notes, par différentes méthodes basées sur des HMM. Cet axe de recherche a permis d'accomplir des gains en transcription significatifs sur des répertoires de *marovany*, en optimisant sélectivement les composantes de connaissance. Un résultat illustratif est que le répertoire de Velonjoro bénéficie principalement des composantes timbrales, telles que l'inharmonicité spectrale, la modulation énergétique de l'enveloppe, l'intermodulation et la résonance par sympathie. Nous avons aussi développé un cadre d'analyse plus général pour la compréhension de l'impact des connaissances a priori dans les systèmes de transcription automatique de musique, en relation avec différents répertoires et instruments de musique.

Mots-clés

Transcription Automatique de Musique, Modélisation acoustique et statistique, Acoustique Musicale, Musique non-eurogénétique, Analyse musicologique informatisée

Contents

Notations

All notations of terms are listed by order of appearance in the thesis.

List of acronyms

Disciplines

AMT Automatic Music Transcription

MPE Multi-Pitch Estimation

CE Computational Ethnomusicology

MCSS Multichannel Capturing Sensory System

Algorithms

FP Filtering Particle

PLCA Probabilistic Latent Component Analysis

EM Expectation-Maximization

DAEM Deterministic Annealing Expectation-Maximization

HMM Hidden Markov Model

Acronyms of notions

KCMA Knowledge Components in Musical Acoustics

Generic variables

N_{boldA} Number of elements in the set **A** ()

$\boldsymbol{\iota}$ Inter-Pitch Matrix of mutual influences between the $I \times I$ couples of different pitches

ι_{ij} Likelihood scores between pitches i and j

i Pitch index, $i \in \{1, \cdots, I\}$

I Number of pitches

$\boldsymbol{M}$ Set of note mixtures

M_{t_k} Note mixture indexed by its starting time t_k

PDF Probability Density Functions

t $\in$ **[1,T**] time-frame index, with T the number of frames

f $\in$ **[1,F**] frequency-bin index

f_t frequency of bin with time-frame index t

F_0 Fundamental frequency, or pitch

F_s Sample frequency

Sound Database

MAPS MIDI Aligned Piano Sounds (Emiya et al., 2010)

RWC Real World Computing (Goto et al., 2003)

List of method-specific variables

MCSS parameters

$\mathbf{x(t)}$ desired sensor signal

$x_n(t)$ noise sensor signal

MPE parameters

$U_{i,k}$ pitch-wise activation intervals for pitch i, indexed by k in time

$L_N\mathbf{(i,k)}$ list of binary note candidates for pitch i, indexed by k in time

$\hat{L}_N\mathbf{(i,k)}$ list of probabilistic note candidates for pitch i, indexed by k in time

PLCA parameters

$\mathbf{m}$ Playing mode index

M_n Binary mixtures of notes

N_{M_n} Number of estimated mixtures of notes

$\hat{M}_n$ Probabilistic mixtures of notes

HMM parameters

S_i state i, $S_i \in \mathbf{S}$, $\mathbf{S} = \{S_1, S_2, ..., S_{N_s}\}$, with N_s the number of states

q_t state at frame t, $\mathbf{q} \in \mathbf{S}$

$\mathbf{A}$ Transition probability matrix, with coefficients a_{ij}, i,j $\in$ [1,...,N_s]

y_t observation at frame t, $\mathbf{q} \in \mathbf{S}$

$\mathbf{B}$ Emission probability matrix, with coefficients b_{ij}, i $\in$ [1,...,N_o] and j $\in$ [1,...,N_o]

δ Score of the candidate optimal partial path

O_d order of duration for state transition

O_e order of emission

Dictionary of templates

W_n dictionary of spectrum note templates

W_s dictionary of spectrum sound state templates

W_m dictionary of motif templates

Introduction

Background

Musical signal

As a starting point, let's consider the musical chain production illustrated in figure 1. Musical Knowledge (**Step 1**) conditions any musical production. Musical signals have very rich temporal and spectral structures, and it is natural to think of them as being organized in a hierarchical way. At the lowest level of this organization, two universal processes in temporal organization of auditory sequences have been identified (Drake, 1998) : (1) segmentation sequence into groups of events; and (2) the extraction of an underlying pulse. The first, based on changes in pitch duration and salience, is already present in early infancy (Krumhansl, 1990). The second involves the extraction of temporal regularities and likewise appears early in infants (Baruch and Drake, 1997), and is assumed to be culture transcending (Carterette and Kendall, 1999). At the highest "symbolic" level of this organization, we have "prescribed rules" for music execution, which may be the score of a piece, as intended by a composer, or more generally standard codes characterizing musical practices of orally transmitted repertoires. Then, on the basis of these rules, the performers add their interpretation to music (**Step 2**), and render the score into a collection of "control signals", which are both characteristics of their instrument timbre and playing style. Let's now move towards the "signal" aspect of music, which should first be captured by some recording device and digitalized (**Step 3**). For example, most music recordings from CDs are recorded with a microphone, and digitalized using the PCM sampling method, with a sampling frequency F_s of 44.1 kHz and 16-bit resolution. Musical signal can then be observed for analysis through specific computational representation methods (**Step 4**), the most popular one being the FFT-based time-frequency plane called spectrogram, where time flows from left to right and different F_0 are arranged in an ascending order on the vertical axis.

As illustrated in figure 2, spectrograms allow for a direct observation of the hierarchical aspect of music. In the time domain, tempo and beat specify the range of likely note transition times, drawing abrupt step-like vertical lines in the spectrogram at the time locations of these transitions. In the frequency domain, two levels of structure can be considered. First, each note is composed of a fundamental frequency (related to the pitch of the note), and partials whose relative amplitudes determine locally the timbre of the note (which is in reality a complex non-linear time-domain system (Fletcher and Rossing, 1998)). The frequencies of the partials are approximately integer multiples of the fundamental frequency, although this clearly does not apply for instruments such as bells and tuned percussion. Second, as audio signals are both additive and oscillatory (i.e. musical objects superimpose and not conceal each other), several notes played at the same time due to polyphony will merge their respective spectral structures. Eventually, when performing musical analysis, a multiple-F_0 estimator produces horizontal lines which

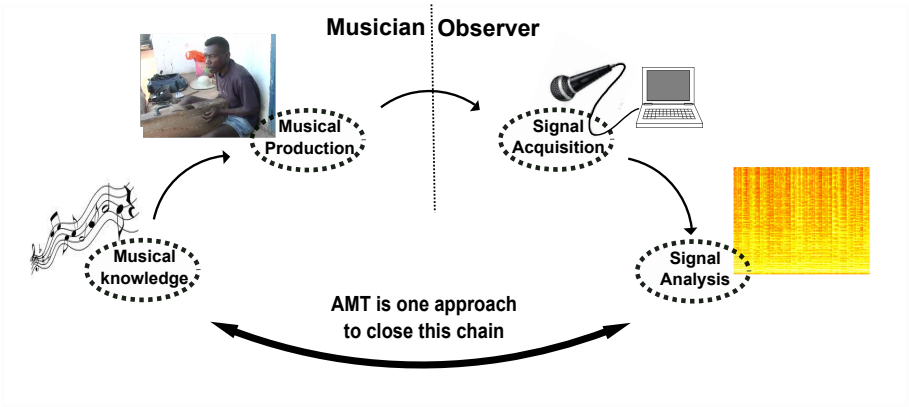


Figure 1 – A general musical chain production with the successive steps of (1) Musical Knowledge, (2) Musical Production, (3) Signal Acquisition and (4) Signal Analysis.

indicate the probabilities of different notes to be active as a function of time. It is visually clear from the spectra of a note that it will be quite easy to estimate the pitch from single-note data that is well segmented in time (so that there is not significant overlap between more than one separate musical note within any single segment). Metrical analysis, in turn, produces a framework of vertical "grid lines" which can be used to segment the note activation curves into discrete note events and to quantize their timing.

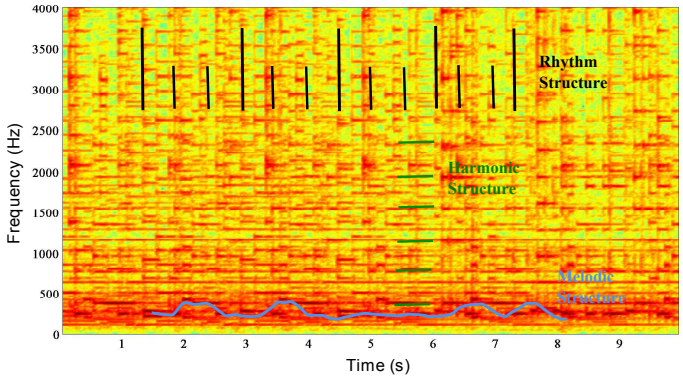


Figure 2 – Spectrogram illustrating the different hierarchical structures present in musical signal. This 10-s musical excerpt comes from the piano piece Hungarian Rhapsody N^o 2 by Liszt. Spectrogram parameters: sampling rate F_s: 44.1 kHz, frame size: 22 ms (1024 samples), 50% overlap (temporal resolution: 11 ms), FFT size: 1024 samples (spectral resolution: 11 Hz), Hamming window.

Low-level musical parameters & Piano-roll representation

Low-level representation

In describing music, we are usually interested in an abstract representation synthesizing music information, which abstracts us away from the signal details. A common approach for this is to characterize the notes played in a musical piece by the four basic parameters of Onset Location, Pitch, Duration and Loudness, through an operation called low-level transcription, and represent them graphically into a time-pitch plane. Such low-level

information is sufficient to be compiled into a MIDI [1] file. In this format standard, the time-pitch plane is called a piano-roll (an illustrative example is given in figure 3), in which continuous physical values of note parameters are discretized into different scales, as detailed in table 1.

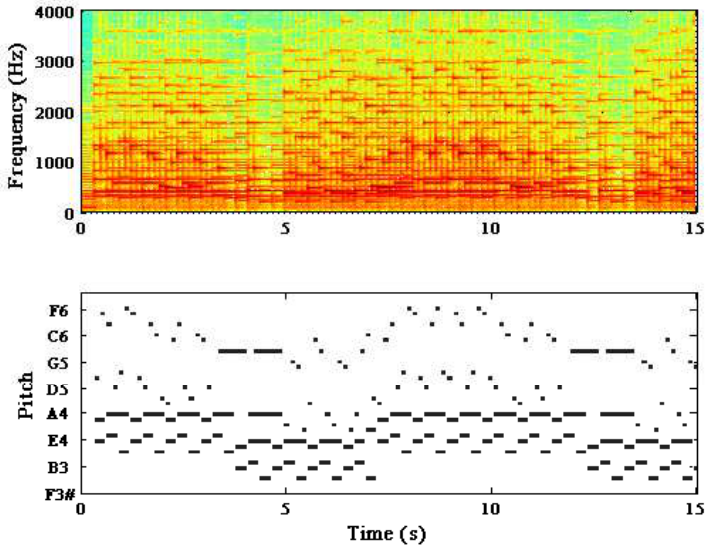


Figure 3 – Spectrogram of a 15-s musical signal, with its piano-roll plotted below.

Such such discrete scales are of course way too restrictive for general music representation, which for example does not follow the assumption of equal temperament tuning. However, the format of MIDI files can still be used with physical continuous values on Onset Location and Duration, and extensions of the MIDI standard exist with finer scales for pitch (e.g. the MIDI tuning standard (MIDIwebPage, 2015)). Furthermore, MIDI files also present the precious advantage of being readable and editable on any audio sequencer and score edition program. For these reasons, this format will be adopted to format our transcription results. We now describe in more details these four note parameters, both from the points of view of psychoacoustics and signal processing.

Note parameters	Discrete MIDI scale
Onset location	Integer multiple of T_a
Pitch	From A_0 (27 Hz) to C_8 (4186 Hz)
Duration	Integer multiple of T_a
Velocity	Arbitrary scale of 127 levels

Table 1 – The discrete MIDI scales of the four different note parameters. For onset location and duration, we defined the Tatum T_a (time quantum) as the smallest metrical unit (Klapuri et al., 2004). This notion refers to the shortest durational values in music that are still more than incidentally encountered. For pitch, this discrete scale corresponds to the semi-tone scale using an equal temperament tuning.

1. MIDI stands for Musical Instrument Digital Interface, and is a technical standard that describes a protocol, digital interface and connectors and allows a wide variety of electronic musical instruments, computers and other related devices to connect and communicate with one another (Wikipedia, 2015).

Onset

Perceptually, human ears are able to distinguish between two onsets as close as 10 ms apart (Moore, 1997). This acuity make them play an important role in our perception of music. Studies showed that onsets play a pivotal role in the perception of timbre, as it is much more difficult to recognize the timbre of tones with removed onsets (Handel, 1995; Martin, 1999). Additionally, music cognition experiments have shown that features related to the beginning of music notes can help humans to discriminate different instrument notes. There is also some evidence that the onset coding plays an important role in direction finding through representation of interaural time and level differences (Rouiller, 1997). And more globally, onsets also make it easier to detect new information in music; we can detect tones with pronounced onsets well before we can determine their pitch (Newton and Smith, 2012). At a larger scale, onset succession build the musical measure pulse, which is usually related to the harmonic change rate.

As a signal parameter to be analysed, an onset is usually defined as the exact time a note or instrument starts sounding after being played. However, this timing is hard to determine, and thus it is impossible to annotate the real onset timing in complex audio recordings with multiple instruments, voices, and effects. Thus, the most commonly used method for onset annotation is marking the earliest time point at which a sound is audible by humans. This instant cannot be defined in pure terms (e.g., minimum increase of volume or sound pressure), but is a rather complex mixture of various factors.

Pitch

Perceptually, perceived pitch is a complex function of the fundamental and all partials. Only a few cycles are needed to identify a pitched note (Robinson and Patterson, 1995). Actually, the human auditory system tries to assign a pitch to almost all kinds of acoustic signals (Meddis and Hewitt, 1991).

As a signal parameter, we define the concept of fundamental frequency, labelled F_0, which has to be limited to periodic or nearly periodic sounds.

Duration

Duration in music refers to how long or short notes are. Durations, and their beginnings and endings, can then be described as long, short, or taking a specific amount of time. A tone may be sustained for varying lengths of time. It is often cited as one of the fundamental aspects of music, encompassing rhythm, form, and even pitch. Durational patterns are the foreground details projected against a background metric structure, which includes meter, tempo, and all rhythmic aspects which produce temporal regularity or structure. Duration patterns may be divided into rhythmic units and rhythmic gestures.

As a signal parameter, the definition of a note offset has been a long-time ill-posed problem. It is often stated that for “decay instruments”, the offset time is only important in a limited period of time, because after some point the sound energy will be below the threshold of hearing (Zwicker and Fastl, 1999), even if from the musician point of view the note is still playing.

Amplitude

Perceptually, the note amplitude is correlated to the sensation of loudness. This perception is strongly non-linear and frequency-dependent.

As a signal parameter, the strong non-linear aspect of intensity perception and the absence of a precise normalization in MIDI make this parameter very hard to define. Ad-hoc relative measures, such as the root mean square value (abbreviated RMS or rms, and defined as the statistical measure defined as the square root of the mean of the squares of a sample), are generally proposed.

Automatic Music Transcription

In this section, we propose a global overview of the computational task of Automatic Music Transcription (AMT), along with more specific issues related to this research field.

Background

Due to the widespread use of the Information Technologies in the distribution and consumption of music, the topic of automatic description/annotation of audio recordings has become a research topic with many practical applications. This research is being carried out within the field that is known as Music Information Retrieval (MIR). This research field is part of a larger research area of multimedia information retrieval. Researchers working in this area focus on retrieving information from different types of media content: images, video, and sounds. Although these types of content differ from each other, separate disciplines of multimedia information retrieval share techniques like pattern recognition and learning techniques. This research field was born in the 80's, and initially focused on computer vision (Lew et al., 2006). The first research works on audio signal analysis started with automatic speech recognition and discriminating music from speech content (Typke et al., 2005).

MIR is the interdisciplinary science (bridging musicology, signal processing, psychology) of retrieving any meaningful information from complex audio, which may then be used in a variety of user scenarios such as searching and organizing music collections. Some of the problems that the MIR community attempts to solve include classification and organization of music, recommendation systems and everything up to and including complex analysis of large musical databases by musical experts. Such problems are dealt with in disciplines called (MIREX, 2011): audio identification, beat detection, prominent melody extraction, genre identification, cover song detection, or query by humming. Many of these problems have very tangible commercial premise, but most are related to the simple desire to understand basically how music functions by utilizing large databases and the power of computer processing. As in definitive, by increasing the quantity and quality of what the computer can "listen to", we can develop more effective tools for indexing and manipulating large audio collections as well as improve musician-computer interactions.

Moving now to the task of AMT, Piszczalski and Galler (1977) proposed the first monophonic music transcription system which is limited to certain types of instruments with strong fundamental frequencies. This frequency domain approach simply chooses the most pronounced spectral peak as the fundamental frequency. In the same year, Moorer (1977) proposed the first polyphonic music transcription system, focusing on duets with a limitation to two monophonic instruments playing at the same time.

Goal & Applications

AMT is the process according to which a structured symbolic representation is inferred from a musical signal. This process can be seen as reverse-engineering the source code of a music signal (Klapuri, 2004a). As illustrated in figure 1, the process of AMT can

be thought as the one closing the loop of the musical chain production. Conventionally, such symbolic representation generally results from highly time-consuming manual works, which also require a certain degree of musical education. Also, information retrieval for music transcription can be classed into two levels (Klapuri, 2004a) : the low-level (with the four parameters defined above), and the high-level (tonality, instrument recognition), which asks for more global and complex notions. In a traditional sense, transcribing a piece of music implies a number of these high-level tasks, such as estimating the tempo, metre and key of each section of the piece, identifying and labeling ornamentations and timbre information, recognising the instruments being played, and segregating "voices" according to the instrument that played them and to their function, i.e. melody, accompaniment, etc. However, performing such a high-level operation is totally out-of-reach of today AMT technology (Benetos et al., 2013b), and most proposed algorithms reduce the issue of transcribing music to identifying the four basic parameters of a note event.

For what concerns applications, in addition to the transcription application itself, computational music transcription systems, even in its low-level encoding, can be further used for a wide range of high-level applications including automatic search and annotation (e.g. retrieval of musically-similar recordings) from large audio databases, real-time interaction between musicians and computers, analysis of recordings of the same piece by different performers and audio to score alignment (also used in automatic music tutors, as it aims to match the performance of the user with the original notation in order to help the music student to align her/his performance visually) (Mayor et al., 2009).

Another AMT application of particular interest in this PhD concerns computational musicology. The term Computational Musicology comes from the research tradition of musicology, field that has focused on the study of the symbolic representations of music (scores) of the classical western music tradition. This research perspective takes advantage of the availability of scores in machine-readable format. Music theoretical models, like the one by Lerdahl and Jackendoff (1983), are very much followed and current research focuses on the understanding and modeling of different musical facets such as melody, harmony, or structure, of western classical music. Eventually, AMT systems would also aim to help amateur musicians without proper music education, or musicians whose countries do not possess a culture of music writing, such as orally transmitted traditional repertoires of Madagascar, to write down their musical compositions (Wang et al., 2003). Although this second category of music repertoires will require much more research works before automatizing the process of transcribing them, as fundamental questions should be addressed beforehand about the identification/representation of their musical codes, which may present strong singularities with standard euro-genetic [2] music language.

Challenges & perspectives

In order to fully solve the AMT problem and have a system that provides an output that is equivalent to conventional sheet music, additional issues need to be addressed, such as metre induction, rhythm parsing, key finding, note spelling, dynamics, fingering, expression, articulation and typesetting. Then, the transcription of real-world music with an unknown number of coinciding notes, arbitrary instrumentation, various musical genre and tempi, or percussive accompaniment suffers from many unsolved problems. As a con-

2. The term euro-genetic is used in this paper to avoid the misleading dichotomy of western and non-western music (Lartillot et al., 2008). It was proposed by Prof. Robert Reigle (MIAM, Istanbul) in personal communication, and was first used in specialized AMT literature by Benetos and Holzapfel (2013). Euro-genetic music refers mostly to "savant" classical music developed in Europe between the 17^{th} and 19^{th} centuries.

sequence, most AMT systems adopt simplifying processing scenarios, such as restricting their musical corpus to certain musical genres or even to only one specific class of instruments (Bello et al., 2006; Gainza et al., 2004). In particular, piano music attracted a lot of attention due to its comparatively limited spectral and temporal variations (Marolt et al., 2002; Poliner and Ellis, 2007; Emiya et al., 2010). Most current AMT systems are also limited to obtain only a partial transcription of complex musical signals, e.g. restricting the transcription to the dominant melody or bass lines (Ryynanen and Klapuri, 2008), or to pitch estimation of several concurrent sounds over short frames of a recording. This last sub-task of AMT, called Multi-Pitch Estimation (MPE), is at the core of AMT and is often used as a front-end representation to retrieve higher-level note parameters needed for AMT. Beyond these practical problems of retrieving physical features from a complex signal, the main challenge of current automatic transcription of music would be to capture all components which make the musicality of a piece, which cannot simply reduce to signal variations of pitch and amplitude. For example, it is often stated by musicologists that musical breathing guides segmentation of music, but how can such a concept be objectively captured by mathematical descriptors ? In the following, we describe two specific challenges for AMT methods which particularly interest this PhD project.

Facing music diversity

The meeting between Music Information Retrieval and Ethnomusicology [3] has given way to a new scientific discipline called Computational Ethnomusicology (CE) (Tzanetakis et al., 2007), which aims to adapt MIR tools and develop specific ones to the corpus of ethnic music. Following the conferences of the International Society of Music Information Retrieval, we can see this progress and identify the current trends (Downie et al., 2009). The community is quite conscious of the limitations of the current approaches (Lidy et al., 2010) and advancements are being explored by increasing the sizes of the audio collections and the variety of the data types used. Also, the expansion of the Folk Music Analysis (FMA) workshop (with a 5^{th} edition organized by our team at the University of Pierre and Marie Curie in Paris) also largely confirms this tendency.

On the contrary to MIR, which focuses more on the fundamental development of generic computational tools, quite regardless of the evaluation sound datasets, CE projects rather start from musical corpus to choose/adapt/develop their computational tools. Several PhD projects have already been completed in this field, we can mention Gomez (2006), who worked on methods of tonality induction, including in her test corpus a great variety of musical genres which extend the "usual" classical sound dataset. Kranenburg (2010) developed content-based retrieval systems for a vast corpus of folk song melodies. Gedik (2012) was interested in automatic transcription of traditional Turkish art music. More major research projects have also been initiated, such as the CompMusic (2011-2017) and DIADEMS (2012-2015) projects.

This research field of CE has already known the pitfall of opposing too radically between "western" and "non-western" (terms used in Gomez and Herrera (2008)) repertoires in their computational studies (see interesting review of (Lartillot et al., 2008) about Gomez and Herrera (2008)'s paper), without addressing explicitly the fundamental question of knowing to what extent and regarding which features the computational analysis, especially AMT, of non-eurogenetic music could be seen as more complex than euro-genetic

3. Blacking (1979) proposed the following definition of ethnomusicology : "The main task of ethnomusicology is to explain music and music making with reference to the social, but in terms of the musical factors involved in performance and appreciation".

repertoires. Several papers have raised such concerns in those terms (Moelants et al., 2006, 2007; Lidy et al., 2010).

Incorporating musical knowledge

To overcome current limitations of AMT systems, a practical engineering solution was to use computational techniques from statistics and digital signal processing allowing the insertion of prior knowledge from cognitive science, musicology and musical acoustics (Engelmore and Morgan, 1988; Ellis, 1996). This approach is close to human experience, in which the perception of sounds is embedded with prior knowledge, using a collection of global properties such as musical genre, tempo, and orchestration, as well as more specific properties, such as the timbre of a particular instrument.

An important aspect of current AMT programs in regards of musical knowledge is that they are strongly oriented towards eurogenetic instrument repertoires. While the classical solo piano is indeed the most represented instrument in AMT studies (Emiya et al., 2010), most of them also follow a series of assumptions based on cultural concepts. These assumptions apply to structural aspects (e.g. equal temperament, tonal key, assumption of octave equivalence, instrumentation), social organisation of the music (e.g. composers, performers, audience) and technical aspects (e.g. record company, release date). To address this gap between current AMT technologies and actual diversity of musical repertoires around the world, the Music Information Retrieval (MIR) and Ethnomusicology communities have met to give way to a new scientific discipline called Computational Ethnomusicology (CE) (Tzanetakis et al., 2007). This new discipline aims to adapt MIR tools to the multiple corpus of non-eurogenetic music. As a result, this global tendency of AMT systems has raised many concerns (Tzanetakis et al., 2007; Moelants et al., 2007; Lidy et al., 2010) about the appropriateness of current AMT systems to efficiently transcribe repertoires from different musical cultures. Investigations on the subject have already been initiated, either in a broad perspective with a radical eurogenetic / non-eurogenetic opposition (Gomez and Herrera, 2008; Lidy et al., 2010), or through a single traditional instrument repertoire (e.g. Benetos and Holzapfel (2013) in Turkish Makam music). In this PhD project, we will focus on the instrument repertoires of the *marovany* zither from Madagascar.

Motivations, Contributions & Plan

Motivations

Dr. Marc Chemillier in the department of ethnomusicology at the EHESS, contacted researchers from the d'Alembert Institute of University of Pierre and Marie Curie to collaborate with him on the development of computational tools to analyse the *marovany* zither repertoires from Madagascar. Traditional music repertoire of the *marovany* is evolving continually under the influences of other musical practices and genres due to globalization. Current ethnomusicological studies carried out by Dr. Chemillier aim at understanding the evolution of the traditional repertoire through the transformation of its original base motives. The question of computational representation and notation of orally transmitted music is also addressed, as well as the ways on how preserving this patrimony.

Context & Objectives

Within this PhD project, we will deal with the task of automatic transcription of polyphonic music from solo instruments [4]. All instruments study in this PhD, including the *marovany* zither as main instrument of investigation, are decay instruments, i.e. producing sounds characterized by a sudden transient attack with a rapid rise to its peak amplitude. This attack is followed by a long decay envelop. For plucked string instruments and piano, these two successive phases respectively result from the sudden excitation of a string, and then from the free-resonating behaviour of the soundboard.

The two underlying objectives of this PhD project are to 1. develop efficient automatic music transcription systems dedicated to different repertoires of pluck string instruments, with a case study on the *marovany* zither from Madagascar, and 2. understand the contributions of musical knowledge in the optimization of these systems to the different repertoires. These two goals call for both "practical research" with short-term developments of operational tools, and "fundamental research" providing deeper insight into the functioning of these tools.

Our long-term objectives would serve ongoing ethnomusicological studies by providing computational tools able to generate automatically ready-to-use robust transcriptions, so they can be used as analytical supports for ethnomusicologs. They would further help in organizing and structuring audio recordings of this instrument, and thus in better understanding the evolution of the traditional repertoire.

Methodological approach

Our methodological framework is at the crossroads of different research fields. The science of instrumental acoustics (1) study the mechanisms that produce the sound using precise physical modeling of the vibrations and couplings which take place in musical instruments when subject to a player excitation (Fletcher and Rossing, 1998; Rossing, 2010). This physics-based approach allows relating the signal characteristics of a tone to the physical properties of the sound producing components of an instrument (e.g. the materials and the geometry of strings and soundboard, the string-sound board coupling, etc.). In our PhD project, we will not perform complex physical modeling and modal analysis of our instruments, but rather use generic physics formula and principles to guide and validate them through our signal analysis. This approach actually consists in mixing instrumental acoustics with audio signal processing (2), which rather aims at modeling musical sounds according to their "morphological" attributes, without necessarily paying due regard to the particular instruments that are played. Eventually, the science of statistics learning and modeling (3), such as sparse coding (Mallat and Zhang, 1993), Bayesian modeling (Gelman et al., 2003), or rank reduction methods (Cichocki et al., 2009), is also used, which allows building probabilistic signal modeling with the incorporation of instrument-specific musical knowledge.

The *marovany* zither of Madagascar

In Malagasy, *marovany* means "with many strings". This zither is derived from the *valiha*, a tubular zither made of bamboo that is considered as the national instrument of Madagascar. In figure 4, we give a photo of the *valiha*, and a map of Madagascar, on which

4. We remind that music from solo instruments is not in any way equivalent to monophonic music, meaning that the performer is playing only one note at a time, as solo instruments frequently have notes overlapping in both time and frequency domains.

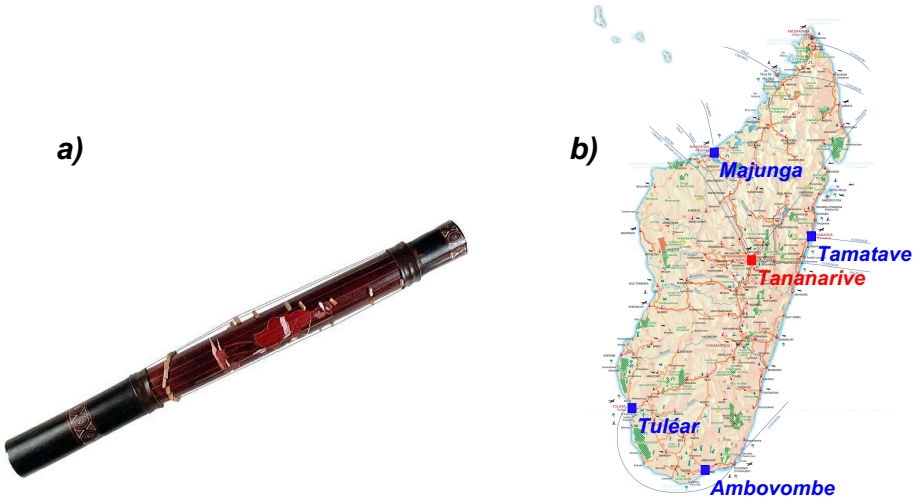


Figure 4 – a) Photo of a *valiha*; b) Map of Madagascar with the most important cities, indicated in colors, for *marovany* music

we indicated the most important cities for *marovany* music. Especially, the emblematic *marovany* player Rakoutzav was born in Tamatave. During this PhD, Marc Chemillier and I performed in June 2013 a field mission in the city of Majunga during two weeks, where we recorded the musician Velonjoro. Marc Chemillier performed two other missions during my PhD, in Tuléar in June 2012 and in Tananarive in 2014, during which he also recorded the musician Velonjoro.

The *marovany* is a tall zither in the form of a rectangular box built from recycled wood products (the box is commonly made in plywoods, and the easels in rosewoods), as illustrated with four different models of this instrument in figure 5. Table 2 shows the physical dimensions of our different *marovany* models. The metallic strings, measuring up to 1 m 20, and mostly coming from brake cables type motorcycle, are stretched on each side of the box. They are nailed at each end on an easel, made of wood or metal, and are raised by battens whose places along a string determines its pitch. Then, during a song, to each string corresponds a single pitch. However, musicians may switch pitches and tuning of their *marovany* between different songs. Wood type, sizing and number of strings of a zither are not fixed. One can indeed find zithers made of light or heavy weight wood, measuring from 1 to 2 meter in length, possessing from 10 to 12 strings on each side, with battens also ranging from 2 cm to 0.5 cm in height (it is known that bringing strings closer to the soundboard produce more powerful sounds, according to an effect called *mafo be*, i.e. “louder”). Each set of strings of the *marovany* forms, like the famous tubular zither *valiha*, an alternating diatonic scale. The minimal harmonic interval of this instrument is then the semi-tone, although tuning deviations from the well-tempered scale are often observed, as described later.

Originally, the *marovany* is a traditional instrument of Madagascar, whose musicians often take part to trance rituals called *tromba*. The *marovany* repertoires are mostly orally transmitted, resulting from a process of memorization/transformation of original base musical motives. These motives represent an important culture patrimony. Nowadays, it has been exported outside the country, mainly by native musicians participating to World Music festivals, subject to the influences of different musical genres and practices.

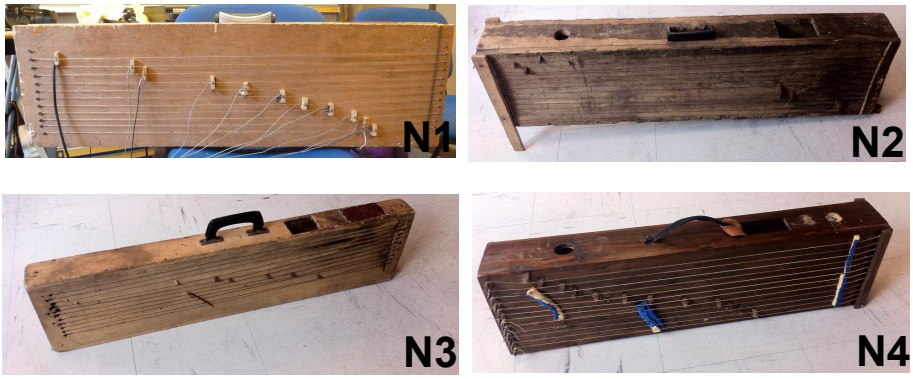


Figure 5 – Photos of the four *marovany* models used in this PhD, labelled from N_1 to N_4.

Although the *valiha* zither has already been subjected by past research studies (Razafindrakoto, 1999; Domenichini, 1984), there is currently no large-scale systematic analysis and classification of the *marovany* repertoires, based on precise musical (rhythmic, modal, structural properties) criteria. Especially, the total absence of manuscript support for the traditional repertoire of the instrument has further impeded the development of such a work. The most common feature of CE studies so far is indeed the use of audio recordings instead of symbolic data (as many traditional repertoires are transmitted orally, without well-established musical codes), which calls for the need of transcribing sampled recordings of music into a piano-roll type representation. The relevance of this endeavour to the CE community is straightforward: piano-roll provides a cartoon-like representation of musical data that is extremely compact when compared to the sampled audio data, yet retains the necessary information for many content based analyses and queries. The automation of this process is made imperative as manual transcriptions are cumbersome and time-consuming. Also, the complexity of this transcription (due to speed of playing, polyphonic characteristics, noisy environment) implies a great variability in hand-made results, making them prone to errors, without systematic estimation of transcription accuracy.

Instrument model label	Instrument maker	Dimensions (L × H × W) in cm	String Number	Pitch range
N_1	Velonjoro (Madagascar)	114 × 37 × 9	23	$[F_3 : D_6]$
N_2	Velonjoro (Madagascar)	108 × 32 × 8	22	$[F_3 : C_6]$
N_3	Velonjoro (Madagascar)	81 × 26 × 8	20	$[G_3 : B_5]$
N_4	Charles Kely (France)	76 × 25 × 7	22	$[A_3 : G_5]$

Table 2 – Physical characteristics (L=Length, H= Height, W=Width) of the four *marovany* models used in this PhD, labelled from N_1 to N_4.

Contributions

As interdisciplinarity has been the hallmark of this research project, it is completely in character that its main contributions divide into different scientific fields. In the following, we list the contributions brought by our PhD project, along with our research communications supporting them, in the fields of:

Automatic Music Transcription,

1. with the development of an engineering system dedicated based on multichannel capturing sensory systems and used to generate AMT ground truths of differ-

ent plucked-string instruments, including non-eurogenetic ones (Conference 1, Conference 2, Conference 3) ;

2. with the development of an automatic music system with original priors specially conceived and optimized for the *marovany* instrument (Article 1) ;
3. by investigating the impact of priors on different instrument repertoires, and addressing the broader question of the appropriateness of current state-of-the-art AMT algorithms for traditional instruments ;

Computational General Musicology,

1. with an extensive analytical work dedicated to the *marovany* zither, including the multifacets of its repertoires and instrument makings (Conference 4, Conference 5);
2. by making the *marovany* a benchmark instrument on test/validation experiments of MIR tools (Article 4) ;

Statistical modeling & Audio Signal Processing,

1. with the theoretical development of a particular filter based system for PLCA parameter estimation, aiming to resolve current limits of the EM algorithm (Article 3) ;

Musical Creativity,

1. with the integration of traditional instrument in the ImproteK project of human-machine interaction (Book 1) ;
2. using the creative potential of our retrieval system in musical performances, through the collaboration with three musicians playing different instruments, each one having his own personal project on how using this "new" instrument ;

Despite a few shortcomings, we think the general methodology described in this thesis is of high interest, and is undoubtedly useful as a brief tutorial on the use of computational methods in comparative music research with large corpora. As such, the thesis is rather technical, although we tried adding more explanation, examples, and illustrations to increase its accessibility to empirical musicologists.

Plan

This PhD thesis is organized as follows. Chapter 1 presents the different Knowledge Components of Musical Acoustics (KCMA) used to help in transcription, and develops the methods used to identify and extract these components. Chapter 2 details the different baseline statistical methods used for transcription, which also constitute the statistical background for KCMA incorporation, with the methods detailed in Chapter 3. Chapter 4 presents the methods we developed to create our ground truth sound database, which is mainly based on the Multi-Channel Sensory System (MCSS) technology. We eventually present results in Chapter 5 and discuss some specific questions around the notion of KCMA in AMT systems.

Chapter 1

Generating knowledge from Musical Acoustics

Abstract

This chapter aims to define different Knowledge Components in Musical Acoustics (KCMA) covering the multi-facets of musical signal, in view of assisting the task of Automatic Music Transcription (AMT). We will develop the computational processes of knowledge generation, which has been automatized in order to process a vast corpus quickly and reliably. The descriptors used in this process have all been validated on reduced labelled datasets to ensure a good reliability in knowledge generation. Figure 1.1 illustrates the relation of this chapter with the others.

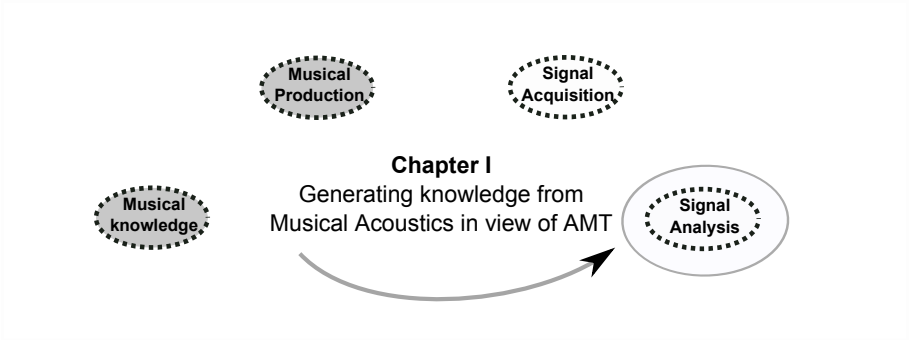


Figure 1.1 – Schematic diagram of the PhD organization for chapter 1.

1.1 Introduction

Musically, the occurrence of simultaneous notes can result either from “acoustic polyphony”, or from “musical polyphony”. ”Acoustics polyphony” is strongly related to the timbre of the instrument, and more precisely to the physical phenomena of mutual resonances and note persistence. Although this type of polyphony is an integral part of instrument timbre, it represents a noise signal added to the actual played note from the point of view of music transcription. For what concerns “musical polyphony”, it corresponds to the note combinations played by the musician and intended by a composer with a proper polyphonic writing. It directly provides useful information about which notes are commonly played simultaneously in a musical piece. In this PhD thesis, we will formalize all knowledge one can collect on these two types of music polyphony through the notion of Knowledge Components of Musical Acoustics.

1.1.1 Sources of KCMA

We will begin this chapter by presenting the different data sources used in this PhD project to generate our different Knowledge Components in Musical Acoustics (KCMA). These components cover the multi-facets of musical signal, from timbre to music language. Based on their different sources, the KCMA extracted can be roughly categorized into three types of musical acoustics knowledge, as described in table 1.1. Although these different categories have been designed for sake of clarity, they should not be understood as exclusive categories, as they are of course strongly correlated in musical practice. In the following, we briefly present the different data sources of our KCMA.

Theoretical musical codes

Two different sciences have been studied to compose this first data source of KCMA. From the science of pyschoacoustics, many theories have been put forward about the degree of acceptance and pleasure when listening to certain acoustic events. For example, studies show that concurrent harmonic sounds (sharing harmonic partials in simple algebraic relation) are preferred, in comparison to noisy or inharmonic sounds. Specific knowledge about psychoacoustic similarities between chords can also be deduced from these theoretical concepts. From the science of musicology, euro-genetic is well-known to have a long history of theorization, from which many different musicological concepts have been created, such as the notions of tonality and equal temperament.

Isolated note samples

This second data source of KCMA consist of instrument isolated sound samples. Musical knowledge extracted from these samples allows extracting various components of musical acoustics knowledge. It basically allows learning timbre features, which can be used to constrain generic signal models with acoustics-based information in different instrument-dedicated AMT systems (e.g. Emiya et al. (2010) and Rigaud et al. (2013) using inharmonicity measures for the piano instrument, Benetos and Holzapfel (2013) using pitch-wise spectral acoustic signatures for the ney instrument Turkish makam music). But it can also allows to extract the precise tuning of an instrument, which is more of a musicological information, and also to identify the impact of specific playing techniques on instrumental note acoustics.

Transcripts from playing

This fourth data source is commonly obtained from a technology called Multichannel Capturing Sensory System, which allows a physical measure of the musical playing directly on the instrument being played. Such a technology will be developed in Sec. 4 of this PhD thesis. Such transcripts encompass both the "prescribed" musical features, as found in written scores, as well as the "interpreted" musical features, i.e. to the particular playing style of a musician, and include all knowledge informing the question on how an instrument is played (e.g. the use of the soft pedal in piano, or palm mute in guitar), which modifies intrinsic acoustic properties of the instrument, as characterized by isolated note samples.

Written Scores

Musical knowledge extracted from our third data source of KCMA, namely written scores, is rather related to music language itself. This language encompasses "prescribed' features, which refer more to standard codes characteristic of a musical piece, and possibly to its belonging musical genre. "Prescribed" features include note-to-note harmonic transitions, tonality, rhythmical figures, base musical motives.

Index	Data sources		Categories of KCMA
I	Theoretical musicological codes	→	Musical codes
II	Isolated note samples	→	Musical codes / Timbre / Playing style
III	Transcripts from playing	→	Musical codes / Playing style
IV	Written scores	→	Musical codes

Table 1.1 – Description of data sources from which we extracted our KCMA. Musical codes refer to "Prescribed features", while playing style refers to "Interpreted features".

1.1.2 Preliminary discussions of notions

Temporal scaling: frames / note events

Most MPE algorithms do not focus on note objects and their features but generally refer to a data granularity of a lower abstraction level: frames (labelled $t \in \mathbf{T}$). A frame is a short chunk (typically on the order of 10 ms) of audio, from which both time and frequency domain features can be computed. Consecutive frames are usually considered with some overlap for smoother analyses. Furthermore, note segmentation is commonly done by a simple energy-based aggregation of successive frames (see Sec. 2.3.3), without considering longer observation time scales. However, using notes as the basic analysis units, instead of pitched frames, enables a more profound melody modeling, since the musicological relationships between these units are well-defined when examining them as constituents of melodies and musical context.

Then, when formatting our KCMA, the question of their time scale is of first importance. Indeed, although frame-based knowledge is more easily fused with algorithm estimations, as sharing the same time scale, note event-based knowledge is much more pertinent for music language modeling than frame-based information, especially for time-dependent sequential knowledge (e.g. a same melody played at different tempi will provide different frame-based modeling on harmonic transitions). Furthermore, note event-based information can be directly interpreted and informed by musicologists, and allows inter-repertoire comparisons based on explicit musical features. In literature, the work of Ryynanen and

Klapuri (2005) is one of the very few examples which adopted such a note event-based approach for music knowledge incorporation in their AMT system.

Analysis musical event: chords / note mixtures

Studies interested in chord segmentation and recognition (e.g. Bello and Pickens (2005); Papadopoulos and Peeters (2007); Lee and Slaney (2008)) naturally build their sequential musicological knowledge on the musical object of chords. The chord is a fundamental unit of music structure, and corresponds to attributes of Western chord notation such as "minor", "major", "diminished" ... It is determined both by its constituent pitch classes, and also by the syntactical context in which the chord occurs. Contextual significance is then a consequence both of the chord's position relative to other chords within a temporally organized sequence and also to its functional relationships to other chords and (ultimately) to an underlying "tonic". Furthermore, because the chord as a structure comprised essentially of pitches, its contextual identity (and the perceptibility thereof) is typically strong enough to shine through the addition or deletion of individual notes.

However the musical object of chords has limited applications in music retrieval tasks other than chord recognition, such as AMT, as it is not flexible enough to allow for a general modeling of music language. To answer this concern, we introduce in this PhD thesis the concept of the note mixture event (which will be defined mathematically in Sec. 2.4.5). This notion of mixture events refer to any grouping of notes simultaneous in time.

1.1.3 Definition of the matrix of inter-pitch influence $\boldsymbol{\iota}$

This 12-dimensional matrix $\boldsymbol{\iota}$ quantifies the likelihood that the 12 pitch classes within an octave range either combine with each other, or transit from one to another. This matrix can easily be generalized to a size of I^2, with I the number of pitches. The matrix $\boldsymbol{\iota}$ is then defined as follows, $\forall (i,j) \in \{1, \cdots, I\}^2$,

$$\boldsymbol{\iota} = \begin{pmatrix} \iota_{11} & \cdots & \iota_{1I} \\ \vdots & \iota_{ij} & \vdots \\ \iota_{I1} & \cdots & \iota_{II} \end{pmatrix} \tag{1.1}$$

where ι_{ij} denotes the likelihood that pitch i is combined with or is followed by pitch j. It is noteworthy that simpler modeling of prior knowledge, such as a simple pitch-dependent vector, can also take the form of a diagonal matrix $\boldsymbol{\iota}$ of size I^2, with the vector values put into this diagonal (the zero-coefficients of S provide an unitary prior value which does not affect particle weights).

1.1.4 Context and objectives

In this PhD, our different KCMA are listed in table 1.2. Although we have reviewed in our Introduction all data sources of KCMA for sake of completeness, in our context of orally transmitted music, i.e. without any written supports, a priori knowledge can only be extracted from the first three data sources from table 1.1.

1.2 Knowledge source I : from theoretical concepts

We propose in the following knowledge-based musicological modeling, where the output distribution parameters are fixed based on an expert's music theoretical knowledge. In

Knowledge sources	KCMA name	Likelihood princple	Polyphony dimension	KCMA category
Theoretical concepts	Theo1	Psychoacoustic emphasis of chord acoustic properties	Transition	Musical codes
	Theo2 / Theo3	Minor and Major keys given key	Combination / Transition	Musical codes
	Theo4 / Theo5	Minor and Major keys	Combination / Transition	Musical codes
	Theo6	Circle of fifths	Transition	Musical codes
	Theo7	Krumhansl tone profiles	Combination	Musical codes
Isolated note samples	IsoNo0	Signal samples	Note object	Timbre / Playing techniques
	IsoNo1-IsoNo4	Acoustic descriptors	Note object	Timbre
	IsoNo5	Acoustic phenomena of sympathetic resonances between strings	Combination	Timbre
	IsoNo6	Non-tempered tuning, i.e. deviation in cents from equal temperament	Combination	Musical codes
Transcripts from playing	TraPla1 / TraPla2	Frequency counting	Combination / Transition	Musical codes
	TraPla3	Pitch-wise frame-to-frame transition	Transition	Playing style
	TraPla4	Pitch-wise note duration	Transition	Playing style
	TraPla5	Melodic motive dictionnary	Transition	Musical codes

Table 1.2 – Table detailing the characteristics of the different KCMA.

typical tonal music, most note mixture progressions are repeated in a cyclic fashion as the piece unfolds. Also, notes comprising a note mixture act as central polarities for the choice of notes at the next note mixture in a musical piece. Furthermore, given that a particular temporal region in a musical piece is associated with a certain note mixture, notes comprising that mixture or sharing some harmonics with notes of that mixture are more likely to be present. These "universal" musical tendencies provide a strong prediction on note succession and superposition. Such a "universal" musical tendency will be modeled in the following with two different approaches, from pyschoacoustics and musicology. In our following section titles, we report the name of the KCMA (like e.g. ($KCMA \sim Theo1$)), as listed in table 1.2, described in the given section.

1.2.1 By Psychoacoustics ($KCMA \sim Theo1$)

In this first section, psychoacoustic considerations on acoustic properties of note mixtures will be used to compute a likelihood on note transition. This likelihood will quantify perceptual acoustic similarity in the transition from one mixture to another. This modeling starts with the assignment of a single timbre for each group of observed notes forming a note mixture (Vassilakis, 1999).

Parametric note model The frequency content of an idealized musical note i is composed of a fundamental frequency $F_{0,i}$ and integer multiples of that frequency. The amplitude of the h-th harmonic $F_{h,i} = \text{h } F_{1,i}$ of note i can be modeled with geometric decaying ρ^h, with $0 < \rho < 1$. A slightly more complex signal modeling will be used here, which includes a parameter of inharmonicity β in the relation between fundamental and harmonics. For real strings, the frequencies of the partials obey the formula

$$F_{h,i} = hF_{0,i}\sqrt{1 + \beta(h^2 - 1)} \tag{1.2}$$

where F is the fundamental frequency, h the harmonic index (partial number ≥ 1), and β the inharmonicity factor (Fletcher and Rossing, 1998). This inharmonicity phenomenon is due to the stiffness of real strings and causes the higher-order partials to be slightly shifted upwards in frequency. Furthermore, to take into account the spectral energy spreading around each harmonic location, a harmonic spectrum is modeled as the sum of Gaussians representing the partials, similarly to the HTC model (Kameoka et al., 2007). Indeed, due to the convolutive nature of this kind of source-filter model, the harmonic partials can be designed independently of the pitch of a note and the spectral spreading of its partials.

Perceptual bin-wise loudness We then map this harmonic model with frequency f to a discrete pitch scale with a 20-cents resolution [1], and define a perceived loudness of each spectral bin b_f present in a note mixture M_{t_k} (this notion will be defined later in Sec. 2.4.5) as

$$l_k(b_f) = \max_{h\in\mathbb{N}, i\in M_{t_k}} (\rho^h | m(f_{h,i}) = b_f) \tag{1.3}$$

The max function is used instead of a sum in order to account for the masking effect (Moore, 1997). For each note mixture M_{t_k}, we then have $\boldsymbol{l_k} = \{l_1(i_1), \cdots, l_d(i_d)\}$ corresponding to the perceived strength of the harmonics related to every note i_d of the well-tempered scale.

Perceptual loudness within an octave We can eventually use octave invariance to give a measure v_k of the relative strength of each spectral bin in a given note mixture. We then obtain $\frac{1200}{20} = 60$ bins b_{fo} within an octave, i.e. 5 bins per pitch class, forming the note mixture representation $\boldsymbol{v_k} = \{v_k(0), \cdots, v_k(59)\}$, whose elements are defined as

$$v_k(b_{fo}) = \sum_{(b_f mod 60)=i} l(b_f) \tag{1.4}$$

where i is the pitch. Figure 1.2 shows an example of a GMM-based parametric model (top graph) for the note mixture $C_4, F_4, G_4, A_4\#$, with its corresponding perceptual bin-wise loudness $\boldsymbol{l_k}$ (middle graph) and perceptual loudness wrapped within an octave (bottom graph).

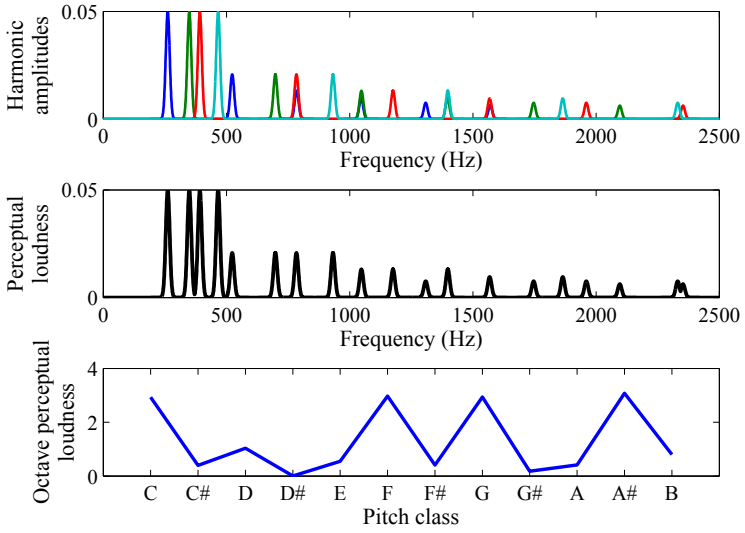


Figure 1.2 – Example of a GMM-based parametric note mixture model (top graph, each color representing a note model) for the note mixture $C_4, F_4, G_4, A_4\#$, with its corresponding perceptual bin-wise loudness $\boldsymbol{l_k}$ (middle graph) and perceptual loudness within an octave $\boldsymbol{v_i}$ (bottom graph).

1. This spectral resolution higher than the twelve-tone equal temperament commonly used in chord modeling allows us taking into account tuning deviations and inharmonicity.

Perceptual similarity From the previous timbre-related note mixture modeling $\boldsymbol{v_k}$, we need to define a transition probability between two note mixtures. Perceptually similar mixtures tend also to be close in Euclidian distance (Krumhansl, 1990), which motivate us to define our probability as

$$Ed = \| \boldsymbol{v_i} - \boldsymbol{v_j} \|^2 \tag{1.5}$$

and from this timbre information, we can obtain a transition probability between two mixtures by computing

$$Theo1 \sim P(i,j) = \frac{e^{-Ed(i,j)}}{\sum_j e^{-Ed(i,j)}} \tag{1.6}$$

1.2.2 By Musicology

Major and minor keys given the key ($KCMA \sim Theo2/Theo3$)

Key is an architectonic system of contextual relationships, and has been foundational to Western music, both classical and popular, since the 17^{th} century. The term key (or tonality) is usually defined as the relationship between a set of pitches having a tonic as its main tone, after which the key is named (Kennedy and Bourne, 1996). A key is then defined by both its tonic and its mode, and generates expectations favouring certain pitch sequences. The tonic identifies to one of the 12 semitones of the chromatic scale within an octave range. The mode is usually minor or major, depending on the used scale. The major and minor keys then rise to a total set of 24 different tonalities. Figure 1.3 represents the key signatures of the 24 major and minor keys. Euro-genetic music is mainly governed by the chord templates (e.g. see http://www.piano-keyboard-guide.com/major-chords.html)), defined as the theoretical chroma vectors corresponding to the 24 Major and minor triads.

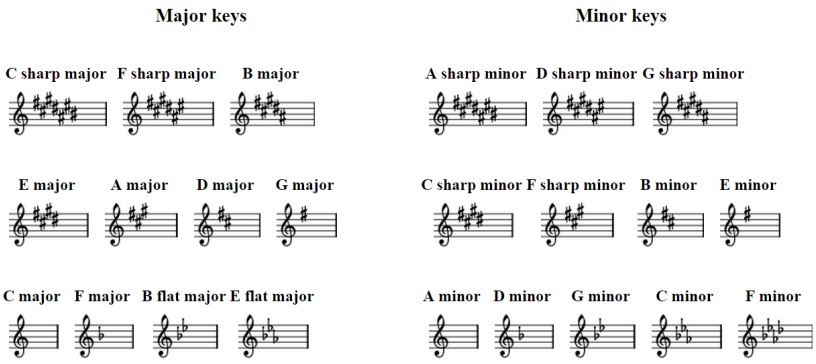


Figure 1.3 – Major and minor chords (from http://hymns.reactor-core.org/keysignatures.html).

Knowing the key of a piece of course provides very valuable information about the chords as well, as in Euro-genetic tonal music, a key and chords are very closely related. For instance, if a musical piece is in the key of C major, then we can expect frequent appearances of chords such as C major, F major, and G major, which correspond to the tonic, subdominant, and dominant chord, respectively. On the other hand, minor or major chord do not appear, since neither has any harmonic function in a C major key. Based on the key information, we can then compute 24 ι matrices of harmonic transitions,

one for each key, as illustrated in figures 1.4 for the keys A major and C sharp major. The amplitude of a note in the key is set to 1 if the note belongs to the considered key. These matrices can then be used as KCMA for both note combination, i.e. $\sim Theo2$, and transition, i.e. $\sim Theo3$.

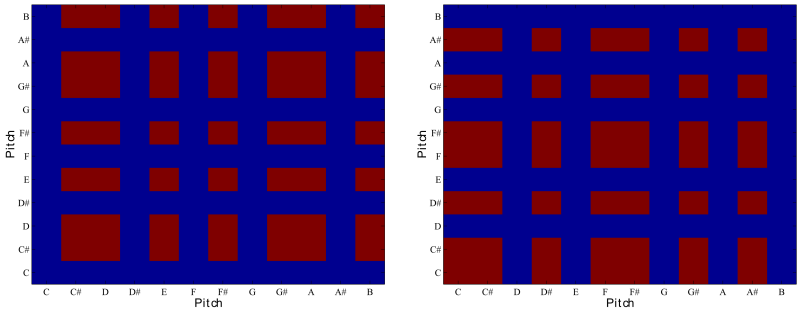


Figure 1.4 – Transition probability matrices of musicological modeling. It gets its knowledge from the theoretical keys described in Sec. 1.2.2. One state transition matrix is created for each key, as here for example, are represented matrices associated with keys A major and C sharp major.

Major and minor keys ($KCMA \sim Theo4/Theo5$)

When key information is not available, a more global modeling can be performed using knowledge of all major and minor keys, by summing and normalizing element-wise the 24 ι matrices respective to each key. As previously, this matrix can then be used as KCMA for both note combination, i.e. $\sim Theo4$, and transition, i.e. $\sim Theo5$.

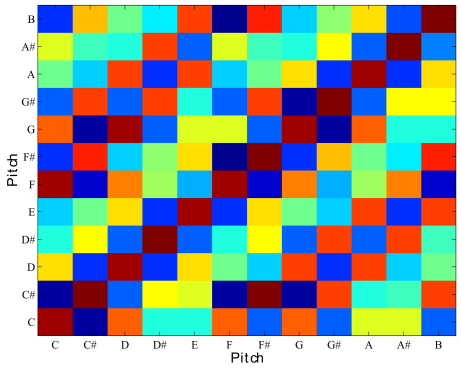


Figure 1.5 – State transition matrices of musicological modeling 1. It gets its knowledge from the theoretical keys described in Sec. 1.2.2.

The circle of fifths ($KCMA \sim Theo6$)

Chords succeed to one another following certain rules. The transition probability between two chords can be derived from musical knowledge: their distance in the doubly-nested circle of fifths, as proposed by Bello and Pickens (2005) (see this reference for details) and represented in figure 1.6. The doubly-nested circle of fifths depicts relationships among the 12 equal-tempered pitch classes comprising the chromatic scale. Although we do not know which state is going to follow another one, musical rules allow us to make hypotheses that are more probable than others. For instance, especially in popular western

music, an A major chord is more likely to be followed by a F# minor or D major chord than by a G# Major chord. It is also assumed through this model that most music tends not to make large and quick harmonic shifts, but rather one might gradually wander from the C to the F#, but not immediately.

*Theo*6 is computed by forming a ι matrix obtained from this chord-based modeling by decomposing them into their constitutive notes, and report to the individual pitches their different likelihoods from the chords they belong.

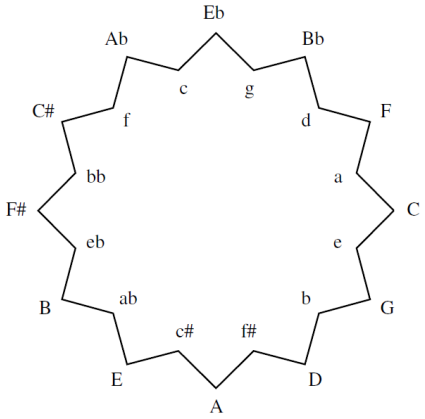


Figure 1.6 – Doubly-nested circle of fifths, with the minor triads (lower case) staggered throughout the major triads (upper case). Triads closer to each other on the circle are more consonant, and thus receive higher initial transition probability mass than triads further away. From Bello and Pickens (2005).

Krumhansl (1990)'s tone profiles ($KCMA \sim Theo7$)

Krumhansl and Shepard (1979) have shown that stable pitch distributions give rise to mental schemas that structure expectations and facilitate the processing of musical information, making individuals intuitively expect and perceive basic patterns of tonal organization, almost as if they were born with this understanding (Krumhansl, 1990). Using the now famous probe-tone method, Krumhansl (1990) showed that listeners' ratings of the appropriateness of a test tone in relation to a tonal context is directly related to the relative prevalence of that pitch-class in a given key. The Krumhansl tone profiles for major and minor keys (Krumhansl, 1990, p. 67), represented in figure 1.7 have been widely used in key recognition research (e.g. Temperley (2002); Hu and Saul (2009)). *Theo*7 is directly identified to these tone profiles in order to predict most likely note combination according to this theoretical knowledge, given the key.

1.3 Knowledge source II : from isolated note samples

1.3.1 Signal samples ($KCMA \sim IsoNo0$)

In our knowledge incorporation, we will use isolated note samples $x(t)$ to extract probabilistic templates which directly fit the spectral characteristics of these notes. Figure 1.8 reveals that even from these simple representations, clear differences can be observed between notes with different pitches. The spectra of these note samples are classically computed using the short-time Fourier transform of the signal $x(t)$, defined as

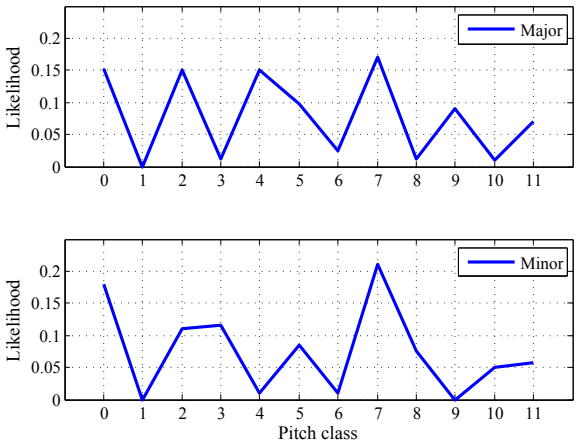


Figure 1.7 – Krumhansl tone profiles for major and minor keys.

$$X(f,t) = \int_{-\infty}^{+\infty} [x(\tau)w(t-\tau)]e^{-j2\pi f\tau}d\tau \tag{1.7}$$

with $w(t-\tau)$ a gliding hamming-window, $f \in \boldsymbol{F}$ the frequency bin and $t \in \boldsymbol{T}$ the time frame index.

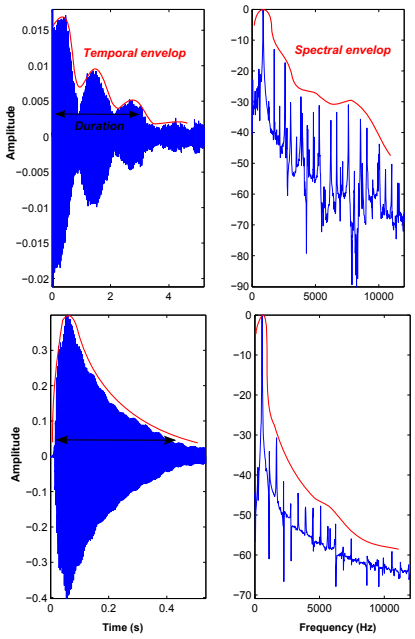


Figure 1.8 – Illustration in the temporal waveforms and spectra of two different notes, with pitches C_4 (on the top) and D_2 (on the bottom).

Furthermore, acoustic analysis of musical signals (Fletcher and Rossing, 1998) reveal that the timbre of an instrument varies over its pitch range, as well as over different playing dynamics. This dynamics refers to the volume of a note, whose variations, called nuances, are crucial for expressiveness in music. Also, in plucked string instruments the strings may be either plucked or rubbed at truly different distances from the bridge on which they pass, involving also a great spectral envelop variety for each different pitch. As a result, this greatly varies the intrinsic acoustic signature of an instrument inducing a local

timbre alteration. As it will be explained later (see Sec. 4.4), the recording of isolated note samples allows including different instrument models and playing techniques, with their corresponding modifications in acoustic features.

1.3.2 Basic acoustic descriptors

Any instrument has a specific acoustic signature, which makes them recognizable among different plucked string instruments playing a same pitch [2]. It is well known that timbre is multidimensional, i.e. it is not correlated with a single acoustic property. Among these properties, attack and decay transients, inharmonicity and changes in the distribution (i.e. amplitude and shape) of spectral energy contribute to the perception of timbre (Iverson and Krumhansl, 1993; Handel, 1995; Godsmark and Brown, 1999). We propose in the following a low-modeling of this timbre through the computation of four acoustic descriptors on isolated note samples (see Annex A for computational details).

Note duration ($KCMA \sim IsoNo1$)

Physical duration of a note is related to both the strength of the excitation and the vibratory properties of the resonating structure, e.g. shorter string with strong stiffness (i.e. higher pitch) tends to have a shorter duration (Fletcher and Rossing, 1998). We use the descriptor $EffDur$, detailed in Annex A, taken from Peeters et al. (2011).

Amplitude of Energy Modulation ($KCMA \sim IsoNo2$)

Amplitude of Energy Modulation aims to quantify a signal modulation depth (i.e. peak-to-valley difference) in the temporal envelop of a sound. The phenomenon of intermodulation in musical sound is characterized by a strong amplitude of energy modulations with most often the formation of several valleys in the envelop. The *marovany* timbre presents a certain emphasis of this feature, depending on pitches and playing modes. We use the descriptor $AmpMod$, detailed in Annex A, taken from Peeters et al. (2011).

Spectrum Inharmonicity ($KCMA \sim IsoNo3$)

Spectrum Inharmonicity quantifies small departure from exact harmonicity, particularly the overtone series is slightly stretched in the high-frequency band, as illustrated in figure 1.9. To quantify automatically this feature, we use the descriptor HD, detailed in Annex A.

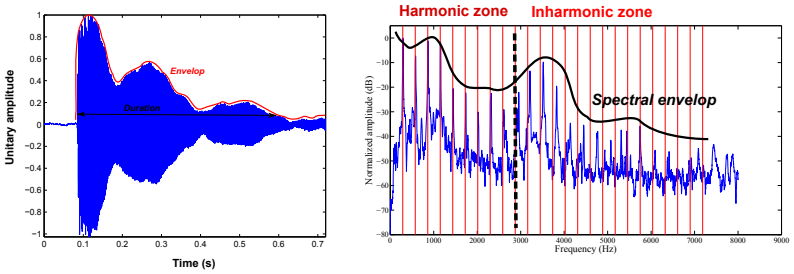


Figure 1.9 – Illustration of the phenomena of inharmonicity in the note D_2 of *marovany*.

2. Definition of the timbre (Fletcher and Rossing, 1998).

Spectrum Variability ($KCMA \sim IsoNo4$)

Spectrum Variability quantifies frame-to-frame differences in spectral envelops. We use the descriptor $SpecVar$, detailed in Annex A.

In certain instruments, this acoustic signature encompassed by these different acoustic descriptors can be very variable through the different pitches, as illustrated in figure 1.10 for the *marovany*. As we study decay instruments, with small variations on the onset part, the two KCMA $\sim IsoNo1$ and $IsoNo2$ encompass most of the temporal timbre signature characteristic of each instrument.

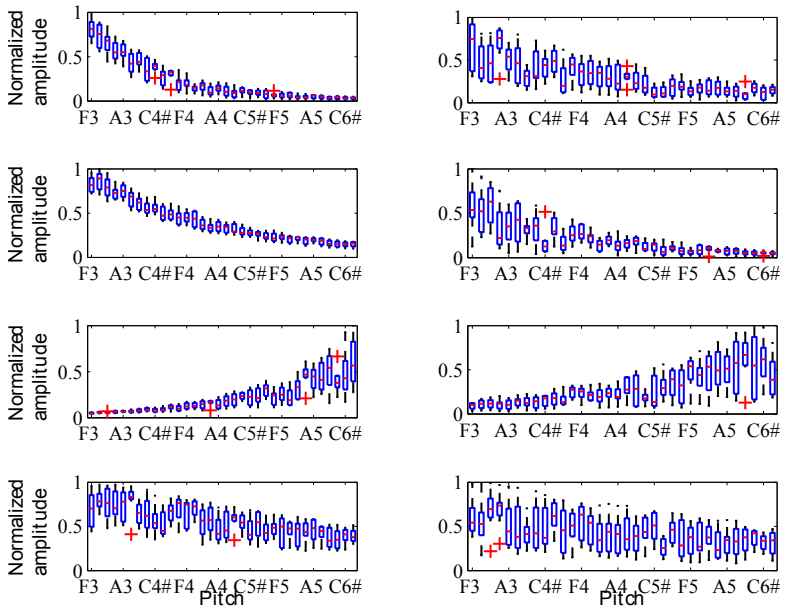


Figure 1.10 – Box plots of different timbre descriptors, computed on an isolated note dataset of a single *marovany* model (on the left), and on four different models of *marovany* (on the right). From top to bottom, these descriptors are the duration, the intermodulation, the spectrum inharmonicity and the spectrum variability.

1.3.3 Sympathetic resonances ($KCMA \sim IsoNo5$)

Background

Mutual resonances (also called sympathetic resonance or sympathetic vibration) result from a harmonic phenomenon wherein a formerly passive string or vibratory body responds (or modes) to external vibrations to which it has a harmonic likeness (Rossing, 2010). For strings on a bowed, plucked, or hammered instruments, mutual resonances result from sympathetic strings, which vibrate (and thereby sound a note) in sympathetic resonance with the note sounded near them by some other agent (Kennedy and Bourne, 1996). Unison or octave will provoke the largest response as there is maximum likeness in vibratory motion.

Measurement protocol & **Quantified measure**

To generate knowledge for this prior, we quantify the acoustic residuals in pitches j due to mutual resonances from a played pitch $i \neq j$. It takes the form of the matrix $\boldsymbol{\iota}$. To compute each coefficient ι_{ij}, we used two datasets of isolated notes, a first one composed of free-resonating notes, and a second one in which all strings were muted excepting the played one. For each note sample of these two datasets, the spectrum was computed with a FFT using a 4096-sample Hamming window after the onset, unitary normalized and labelled $X_{(d,i)}$ for pitch i and dataset d (equal to 1 or 2). We then used the algorithm 1 to get the different scores ι_{ij}.

Algorithm 1 Computation of coefficients ι_{ij} from sympathetic resonances.

for For each pitch i $\in \{1, \cdots, N_I\}$ **do**
  $\tilde{X}_i = ||X_{(1,i)} - X_{(2,i)}||$
  Binary thresholding of $\tilde{X}_i$, i.e.
$$\tilde{X}_i(f) = \begin{cases} 1 & \text{for } f = arg(\tilde{X}_i \geq 0.5) \\ 0 & \text{otherwise} \end{cases}$$
  for Each pitch j $\in \{1, \cdots, N_I\}, j \neq i$ **do**
    $\iota_{ij} = X_{(2,j)} \cdot \tilde{X}_i$, with $[\cdot]$ the element-wise product
  end for
end for

It is noteworthy that due to different acoustic behaviors of mutual resonances depending on the excited pitch, the matrix $\boldsymbol{\iota}$ corresponding to $KCMA \sim IsoNo6$ is not symmetrical. Figure 1.11 provides an example of such a matrix for the *marovany*, using the instrument model N_1.

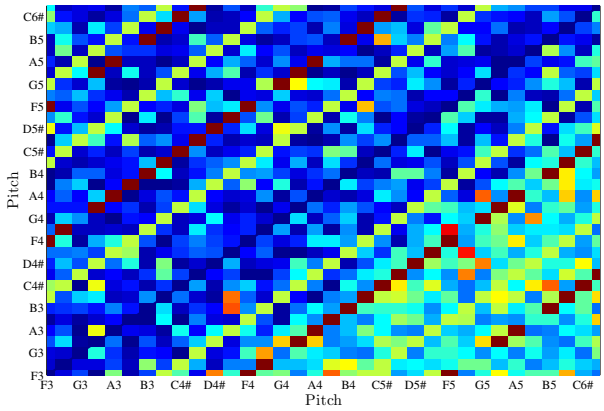


Figure 1.11 – Illustration of sympathetic resonances through an inter-pitch influence matrix $\boldsymbol{\iota}$, computed on a *marovany* model.

1.3.4 Non-tempered tuning ($KCMA \sim IsoNo6$)

Background

In a "non-tempered" scale, harmonic intervals within each octave are not multiple of a common unity, the semi-tone. Tempered scale, as instaured in the MIDI scale, has been introduced in the XVIII century. Octave are correctly tuned in the sense of natural

resonances, i.e. frequencies are in a rapport of 2, but smaller harmonic intervals are multiple of the semi-tone, which results in the equal temperament, where neither thirds and fifths are correctly tuned in the sense of natural resonances, i.e. their frequencies are not in a simple division of 3 or 5. In the XVIIe century, other musical temperament were used to tune harpsichord for example, such as bare fifths (Pythagore temperament) or bare thirds (Zarlino temperament).

Although theoretical pitch scale with equal temperament can be assigned to *marovany* repertoires for convenience of analysis and music transcription, as illustrated in figure 1.12, it appears from quantitative measurements that the players do not systematically adopt this standard tuning. Especially traditional *marovany* players, without any influences from eurogenetic music concepts, tend to tune their instrument with the "just" intervals for thirds, fifths and, minor seventhes, i.e. at an exact sub-multiple of the chord fundamental note frequency.

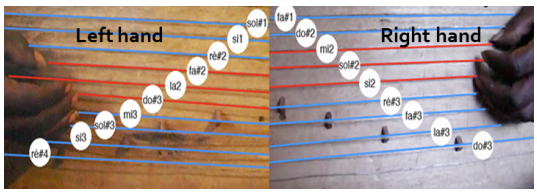


Figure 1.12 – Theoretical tuning for the *marovany* model N_3 based on equal temperament.

Measurement protocol & Quantified measure

According to our training data, this tuning deviation has a reasonably strong presence in Velonjoro repertoire in particular, with a variation from 5 cents to 45 cents depending on the pitch (i.e. almost up to half of a semi-tone, as this interval is equal to 100 cents), as shown in figure 1.13. These measures have been performed automatically using the YIN pitch tracker (de Cheveigné and Kawahara, 2002), and subtracting these physical measures to the theoretical values of the equal temperament.

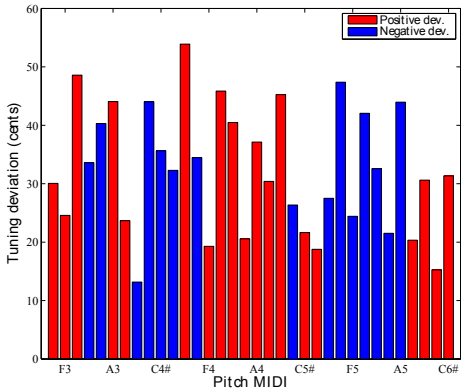


Figure 1.13 – Tuning deviation (in cents) from the equal temperament against MIDI pitches, computed for a *marovany* musical repertoire.

1.4 Knowledge source III : from playing-based transcripts

1.4.1 Data-based frequency counting of note mixture occurrences ($KCMA \sim TraPla1/TraPla2$)

Let's first model which pitches are the most likely to be played together, which is done here by a frame-wise counting of pitches j played simultaneously to pitch i from a training dataset. This counting can then be unitary normalized and identified to the coefficients of matrix $\boldsymbol{\iota}$ (eq. 1.1). As previously, this matrix is used as the KCMA $\sim TraPla1$. Figure 1.15 provides an example of such a matrix.

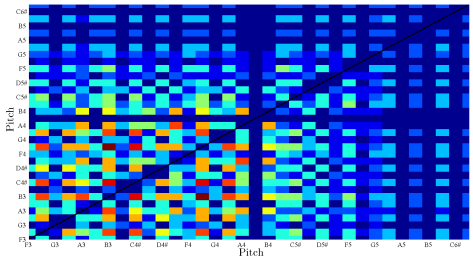


Figure 1.14 – Pitch co-occurrence probability matrix corresponding to a data-based modeling with a simple frequency counting of note mixture observations.

Similarly, to compute the KCMA $\sim TraPla2$ on note transitions, we perform a frequency counting of successive note mixture events M_k and M_{k+1}. To train M_k transitions, we count the number of occurrences of each M_k transition in the training set. By anticipation on our AMT methods (see Chapter 2), this data-based approach allows a direct modeling of harmonic transitions on note mixtures, also learned from the training dataset. Then, in this KCMA, we keep the musical object of note mixtures as the modeling unit. Figure 1.15 provides an example of a dictionary of note mixtures, along with the matrix $\boldsymbol{\iota}$ modeling their likely harmonic transitions.

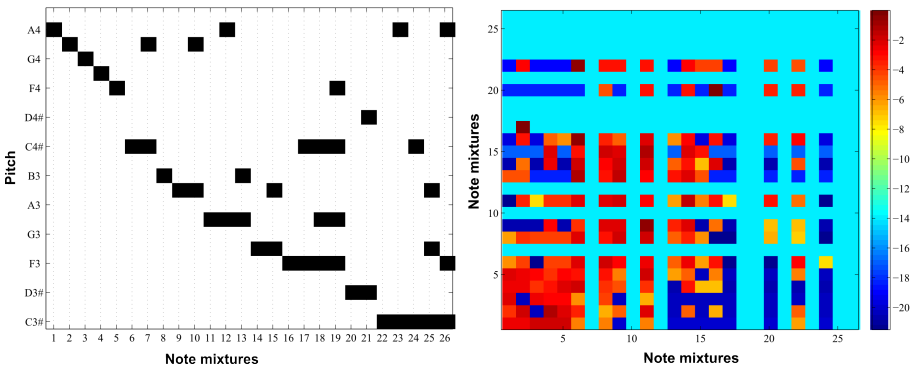


Figure 1.15 – Representation of the different note mixtures (on the left), and log-transition probability matrix between these mixtures obtained with a simple frequency counting (on the right).

1.4.2 Pitch-wise frame-to-frame transition ($KCMA \sim TraPla3$)

Special playing techniques on an instrument repertoire, such as palm muting, pizzicato, staccato and pedal effects, mainly deviates the temporal waveform of a note from its free-

resonating behaviour. These transitions are determined by sampling the training MIDI transcripts at the precise times corresponding to the analysis frames of the activation matrix, and just checking for the presence of a note in each frame.

1.4.3 Duration-informed pitch-wise frame-to-frame transition ($KCMA \sim TraPla4$)

A more refined modeling of ($KCMA \sim TraPla3$) is obtained by including a duration information in the previous pitch-wise frame-to-frame transition model. In contrast to the physical duration of each pitch, as already measured for $KCMA \sim IsoNo1$ in Sec. 1.3.2, we extract a pitch-wise duration distribution from notes under musical playing, which are directly readable from our transcript data by counting successive frames corresponding to active notes.

1.4.4 Motive structure ($KCMA \sim TraPla5$)

Background

In music theory, Schenker (1954) asserted that repetition is what gives rise to the concept of the motive, which is defined as the smallest structural element within a musical piece. Ockelford (2005) argued that repetition/imitation is what brings order to music, and order is what makes music aesthetically pleasing. Melodic progressions then constitute a fixed, non-dynamic structure in time and thus can be used to aid in describing long-term musical structure. Ruwet and Everist (1987) used repetition as a criterion for dividing music into small parts, revealing the syntax of the musical piece. A ground-truth example of the motive structure in a *marovany* musical piece is given in figure 1.16.

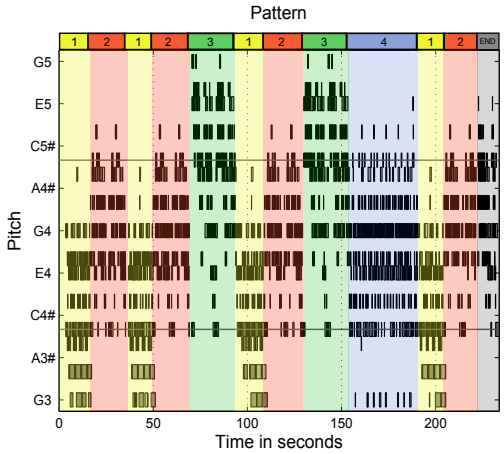


Figure 1.16 – Ground-truth example of the motive structure in the *marovany* musical piece called Folera by Kilema.

Constitution of a motif dictionary ($KCMA \sim TraPla4$)

The learning phase generates a dictionary of the N_{max} most recurrent of D-long note patterns, with d $\in [D_{min} : D_{max}]$. This dictionary has a size of $(D_{max} - D_{min}).N_{max}$. The parameters $(D_{min}; D_{max}; N_{max})$ are set arbitrarily by the user, depending on the quantity of information he wants to integrate in the modeling. A maximal probability is assigned

to each different pattern with a simple counting strategy as performed previously. Figure 1.17 illustrates a dictionary with ($D_{min} = 4; D_{max} = 8; N_{max} = 4$).

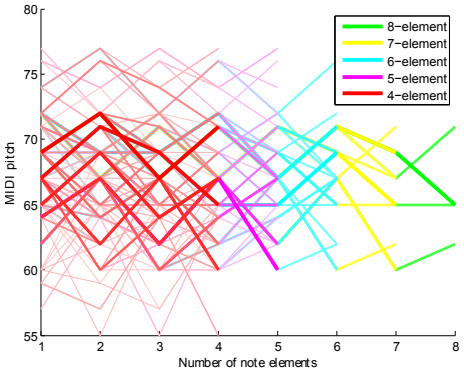


Figure 1.17 – Example of a motif dictionary computed from the instrument repertoire of classical piano.

1.5 Conclusion

In this chapter, we have developed several Knowledge Components of Musical Acoustics (KCMA) in order to characterize an instrument repertoire in view of its automatic transcription.

Chapter 2

Baseline statistical methods for AMT

Abstract

This chapter presents the baseline statistical methods we used to perform Automatic Music Transcription (AMT), which are mainly based on Probabilistic Latent Component Analysis (PLCA) and Hidden Markov Models (HMMs). PLCA is a probabilistic subspace analysis method which can be used for decomposing audio spectrograms. HMMs belong to dynamic Bayesian networks, and are especially known for their applications in modeling and recognition of time series. These two statistical methods use Bayesian probabilistic frameworks, which allow powerful modeling with the incorporation of prior knowledge. PLCA is frame-based, performing the bottom level of analysis against individual spectral slices derived from short time frames. Since note events typically last for many frames, temporal continuity is introduced by some higher-level processing modeled by HMM. These two methods are then complementary by covering both the time and frequency domains of music signals. Figure 2.1 illustrates the relation of this chapter with the others.

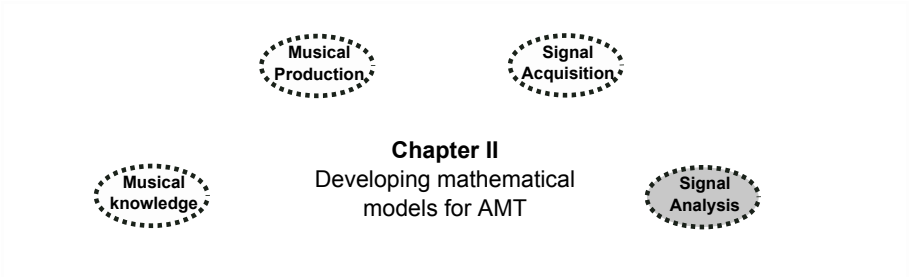


Figure 2.1 – Schematic diagram of the PhD organization for chapter 2.

2.1 Introduction

2.1.1 AMT methods in three families

Classically, the four processing steps of the AMT process are as follows

1. Multi-pitch estimation: within the MIR framework, the goal of Multi-Pitch Estimation (MPE) is to extract the fundamental frequencies of all (possibly concurrent) notes within a polyphonic musical piece.
2. Harmonic (vertical) prior / post-processing
3. Sequential (horizontal) prior / post-processing
4. Note segmentation

The methods proposed so far for AMT perform these steps in different ways, and can be roughly categorized into three main families, according to which signal representation or modeling methods they are based on.

Feature-based

Methods from this family mainly use signal feature properties in a blind way to extract information. Klapuri (2003) estimates the notes present in each frame using an algorithm that iteratively estimates and removes the fundamental frequencies of notes present. Bello et al. (2006) report a novel method for multiple-F_0 estimation using both frequency and time domain information. The signal frames considered are a linearly weighted sum of waveforms in a database of individual piano notes. In Emiya et al. (2010), the likelihood maximization principle is accounted for, assuming a sinusoidal model with coloured background noise and an autoregressive spectral envelope for the overtones. Based on this principle, a F_0 estimator is developed to find the fundamental frequency that maximally flattens both the noise spectrum and the sinusoidal spectrum. Klapuri (2008)'s system uses a computational model of the human auditory periphery, followed by a periodicity analysis mechanism where fundamental frequencies are iteratively detected and cancelled from the mixture signal. Yeh et al. (2005) present a method for evaluating together multiple F_0 hypotheses based on three physical principles: harmonicity, spectral smoothness and synchronous amplitude evolution within a single source. Evaluation of each possible harmonic sound is performed using a score function, which formulates the guiding principles in a mathematical way. In Goto (2004), Goto proposes a F_0 estimation method, called PreFEst, which obtains the most predominant F_0 by estimating the relative dominance of each possible F_0 (represented as a probability density function of the F_0). The method uses MAP (maximum a posteriori probability) estimation and accounts for temporal continuity by using a multiple-agent architecture.

Spectrogram factorization-based

As audio signals are both additive and oscillatory, it is not possible for example to look for energy and harmonic changes simply by differentiating the original signal in the time domain; this has to be done on an intermediate signal that reflects, in a simplified form, the local structure of the original. Hence, methods from ranking reduction and source separation methods have been employed. A powerful method for MPE is then to represent spectra as a linear combination of vectors from a dictionary. Such models take advantage of the inherent low-rank nature of magnitude spectrograms to provide compact and informative descriptions.

Musical signals are highly structured, and it is then possible to state the whole multiple-F_0 estimation problem in terms of a signal model, the parameters of which should be estimated. The parametric model paradigm (Itoyama, 2008; Heittola et al., 2009; Ewert and Muller, 2011) allows to obtain a note-wise parametrization of the spectrogram. Event-driven feature analysis has been shown to give more accurate musical feature extraction than more traditional approaches based on frames of equal length (Bello et al., 2005).

In Leveau et al. (2008), the audio signal is decomposed into a small number of sound atoms or molecules bearing explicit musical instrument labels. Each atom consists of a sum of windowed harmonic partials, whose relative amplitudes are specific to one instrument. Each molecule is composed of several atoms from the same instrument spanning successive time windows. The Specmurt technique is proposed by Saito et al. (2008). This technique consists in searching the fundamental frequency distribution by deconvolving the observed spectrum with an assumed common harmonic structure. In Canadas et al. (2008), harmonic decompositions are modified in order to maximize the spectral smoothness for those atom amplitudes that belong to the same harmonic structure.

Statistic and model-based

Here, expert systems incorporating models of sound characteristics or musical properties are employed. These experts allow for solving otherwise ambiguous situations and obtaining meaningful transcription results. Poliner and Ellis (2007) treat transcription as a classification problem, using support vector machines to classify individual frames as to whether they contain particular notes or not. Kameoka et al. (2007) have proposed the Harmonic Temporal structured Clustering (HTC) method. This method decomposes the energy patterns diffused in the time-frequency space into distinct clusters, grouping patterns from the same source.

Ryynanen and Klapuri (2005) use two probabilistic models: a note event model, used to represent note candidates, and a musicological model, which controls the transitions between note candidates by using key estimation and computing the likelihoods of note sequences. In the note event model, a three-state HMM is allocated to each MIDI note number in each frame. The states in the model represent the temporal regions of note events, comprising namely an attack, a sustain and a noise state, and therefore taking into consideration the dynamic properties and peculiarities of musical performances. State observation likelihoods are determined with recourse to features such as the pitch difference between the measured F_0 and the nominal pitch of the modelled note, pitch salience and onset strength. The observation likelihood distributions are modelled with a four-component Gaussian Mixture Model (GMM) and the HMM parameters are calculated using the BaumâĂŞWelch algorithm. The note and the musicological models then constitute a probabilistic note network, which is used for the transcription of melodies by finding the most probable path through it using a token-passing algorithm. Tokens emitted out of a note model represent note boundaries.

2.1.2 Context and objectives

The objectives of this chapter are to expose theoretically our baseline methods for AMT, which are based on Probabilistic Latent Component Analysis (PLCA) method 2.2 and Hidden Markov Models (HMMs) 2.4. These baseline frameworks will allow to cover the first three processing steps of the AMT process. At the end of this chapter, we also present our baseline methods to perform note segmentation.

2.2 Probabilist Latent Component Analysis

2.2.1 Background

Probabilistic Latent Component Analysis (PLCA) belongs to a class of probabilistic models, known as latent class models, that attempt to explain the observed histograms as having been drawn from a set of latent classes, each with its own distribution. In other words, latent component models enable one to attribute the observations as being due to hidden or latent factors. The PLCA model is then probabilistic in the sense that the objective is to develop a structured representation for an empirically developed probability mass function characterizing the probability distribution for vibration amplitude "quanta" as a function of frequency f and time t. The main characteristic of these models is conditional independence, i.e. multivariate data are modeled as belonging to latent classes such that the random variables within a latent class are independent of one another.

Originally, PLCA is a straightforward extension of Probabilistic Latent Semantic Indexing (Hofmann, 1999) which deals with an arbitrary number of dimensions and can exhibit various features such as sparsity or shift-invariance. PLCA can also be seen as a direct probabilistic extension of Non-Matrix Factorization (NMF) to obtain a better semantic interpretation, leading to enhanced modeling.

Unlike NMF which tries to characterize the observed data directly, latent class models characterize the underlying distribution $P(x_1, x_2)$. This subtle difference of interpretation preserves all the advantages of NMF, while overcoming some of its limitations by providing a framework that is easy to generalize, extend, and interpret. From past studies, the following advantages of PLCA have been highlighted, in particular from the NMF method:

1. It is efficient for modeling non-stationary sound since the learnt dictionary accommodates invariant characteristics of input sound, for this reason, linear combinations of its dictionary elements (basis) are sufficient for representing the time-varying patterns in audio signal ;
2. It can separate the non-negative components without assumptions about their orthogonality ;
3. Its probabilistic nature makes it possible to utilize additional a priori information, in particular to follow sparseness assumptions

2.2.2 General formulation

In more technical terms, PLCA decomposes a multi-dimensional distribution as a mixture of latent components where each component is given by the product of one-dimensional marginal distributions. Recently, it has been shown that PLCA is numerically identical to NMF for two-dimensional input, and non-negative tensors for arbitrary dimensions (Smaragdis et al., 2008). The basic model is defined as

$$P(x) = \sum_{z} P(z) \prod_{j=1}^{J} P(x_j|z) \tag{2.1}$$

where $P(x)$ is an J-dimensional distribution of the random variable $x = (x_1, \ldots, x_J)$, z is a latent variable and the $P(x_j|z)$ are one dimensional distribution with $j \in \{1, \ldots, J\}$.

Such a general model has been successfully applied to audio signal, with a theoretical framework developed by Smaragdis et al. (2006). PLCA method is then based on the assumption that a suitably normalized magnitude spectrogram, V, can be modeled as a joint distribution over time and frequency, $P(f, t)$. This quantity can be factored into

a frame probability $P(t)$, which can be computed directly from the observed data, and a conditional distribution over frequency bins $P(f|t)$; spectrogram frames are treated as repeated draws from an underlying random process characterized by $P(f|t)$. We can model this distribution with a mixture of latent factors as follows:

$$P(f,t) = \sum_z P(z,t)P(f|z) \tag{2.2}$$

where $P(f|z)$ is a multinomial distribution of frequencies for latent component z. It can be viewed as a spectral vector from a dictionary. $P(z,t)$ is a multinomial distribution of weights for the aforementioned dictionary elements at time t, i.e. time activations.

Using an asymmetric factorization, which treats f and t differently, we can decompose $P(f,t)$ as a product of a spectral basis matrix and a component activity matrix, as follows

$$P(f,t) = P(t)\sum_z P(f|z)P(z|t) \tag{2.3}$$

where z is the component index, $P(t)$ is the energy of the input spectrogram (known quantity), $P(f|z)$ is the spectral template that corresponds to the z^{th} component, and $P(z|t)$ is the activation of the z^{th} component.

Note that when there is only a single latent variable z this is identical to NMF. The latent variable framework, however, as already said, has the advantage of a clear probabilistic interpretation which makes it easier to introduce additional parameters and constraints. It is worth emphasizing that the distributions in PLCA are all multinomials. This can be somewhat confusing as it may not be immediately apparent that they represent the probabilities of time and frequency bins rather than specific values; it is as if the spectrogram were formed by distributing a pile of energy quanta according to the combined multinomial distribution, then seeing at the end how much energy accumulates in each time-frequency bin. This subtle yet important distinction is at the heart of how and why these factorization-based algorithms work.

2.2.3 Formulation for Automatic Music Transcription

We now propose a reformulation of this generic method for our task of AMT of polyphonic solo instrument. The magnitude spectrogram of a sound source can be viewed as a histogram of "sound quanta" across time and frequency. With this view, probabilistic factorization, which is a type of non-negative factorization, has been used to model a magnitude spectrogram as a linear combination of spectral vectors from a dictionary.

The model then takes as input a log-frequency spectrogram $X_{f,t}$ and, as stated above, approximates it as a joint distribution over time and log-frequency $P(f,t)$ (f is the log-frequency index and t the time index), that is

$$X_N(f,t) \approx P(f,t) \tag{2.4}$$

with $X_N = \frac{|X(f,t)|}{\sum_{f,t}|X(f,t)|}$ the normalized spectrogram making it a distribution of acoustic energy across the time-frequency plane. This quantity can be factored into a frame probability $P(t)$, which can be computed directly from the observed data (i.e. energy spectrogram), and a conditional distribution over frequency bins $P(f|t)$, as follows

$$P(f,t) = P(t)P(f|t) \tag{2.5}$$

Spectrogram frames are then treated as repeated draws from an underlying random process characterized by $P(f|t)$. Let's first define the latent variable $z = \{i, m\}$, representing respectively pitch and playing mode, with $i \in \mathbf{I}$, $\mathbf{I}$ being the set of pitches, and $m \in \{1, \cdots, M\}$, M being the number of playing modes considered. We can then model this distribution with a mixture of latent factors related to music transcription as follows:

$$P(f|t) = \sum_{i,m} P(f|i,m)P(t|i,m) \tag{2.6}$$

$$= \sum_{i,m} P(f|i,m)P(m|i,t)P(i|t) \tag{2.7}$$

where $P(f|i,m)$ are the spectral templates (also called kernel distribution) for pitch i and playing mode m. For the selected pitch i, the frequency f is selected in a probability distribution $P(f|i,m)$. No orthogonality between the different spectral basis $P(f|i)$ can be assumed as several of these basis are mixed into the same frequency f, or in other words, these basis overlap in frequency. $P(m|i,t)$ is the playing mode activation, and $P(i|t)$ is the pitch activation (i.e. the transcription, also called impulse distribution). The playing mode m will refer to different dynamics of instrument playing (i.e. note loudness). The constant-Q Transform (CQT) is usually used to compute the logarithm spectrogram. Annex A.3.2 provides details about this representation. When employed on a constant-Q transform of a mixture of notes, PLCA can be used to decompose the input data into a summation of convolutions of pitch spectra and the pitch track corresponding to their temporal activation.

2.2.4 Shift-Invariant PLCA

For what concerns its algorithmic implementation, we use the Shift-Invariant PLCA (Smaragdis et al., 2008) extension of the PLCA, exploiting the fact that in a CQT, a change of fundamental frequency is traduced by a simple frequency translation of its partials, resulting in a shift invariance over log-frequency. SI-PLCA then performs a multi-pitch detection with a frequency resolution higher than MIDI scale. Shifting of templates are performed by re-writing eq. 2.7 as

$$P(f|t) = \sum_{i,\delta_f,m} P(f - \delta_f|i,m)P(m|i,t)P(i|t)P(\delta_f|i,t) \tag{2.8}$$

where δ_f is the pitch shifting factor. To constrain δ_f so that each sound state template is associated with a single pitch, the shifting occurs in a semitone range around the ideal position of each pitch. Thus because we are using in this paper a log-frequency representation with a spectral resolution of 60 bins/octave, i.e. a 20 cent resolution, we have $\delta_f \in$ [-2:2].

2.2.5 EM-based model parameter estimation

Classical EM-based estimation

To estimate the model parameters $P(m|i,t)$ and $P(i|t)$, since there is usually no closed-form solution for the maximization of the log-likelihood or the posterior distributions, iterative update rules based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) are employed. We then follow the likelihood principle (McLachlan and Basford, 1988), which leads us to maximize the log-likelihood function of eq. 2.5, i.e.

$$L = \sum_{f \in \boldsymbol{F}} \sum_{t \in \boldsymbol{T}} n(f,t) log(P(f,t)) \tag{2.9}$$

where $n(f,t)$ denotes the term frequency, i.e., the number of times f occurred in t. As stated by Hofmann (1999), "a standard procedure for maximum likelihood estimation in latent variable models is the EM-algorithm (Dempster et al., 1977)". EM alternates two steps: (i) an expectation (E) step where posterior probabilities are computed for the latent variables z, based on the current estimates of the parameters, (ii) an maximization (M) step, where parameters are updated for given posterior probabilities computed in the previous E-step.

For the Expectation step, we compute the contribution of latent variables i, δ_f and m over the complete model reconstruction using Bayes' theorem as follows

$$P(i, \delta_f, m | f, t) = \frac{P(f - \delta_f | i, m) P(\delta_f | i, t) P(i|t) P(m|i,t)}{\sum\limits_{i, \delta_f, m} P(f - \delta_f | i, m) P(\delta_f | i, t) P(i|t) P(m|i,t)} \tag{2.10}$$

For the Maximization step, we utilize the posterior of eq. 2.10 for maximizing the log-likelihood of eq. 2.9, resulting in the following update equations:

$$P(\delta_f | i, t) = \frac{\sum\limits_{f,m} P(i, \delta_f, m | f, t) X_N(f,t)}{\sum\limits_{\delta_f, f, m} P(i, \delta_f, m | f, t) X_N(f,t)} \tag{2.11}$$

$$P(m | i, t) = \frac{\sum\limits_{f, \delta_f} P(i, \delta_f, m | f, t) X_N(f,t)}{\sum\limits_{m, f, \delta_f} P(i, \delta_f, m | f, t) X_N(f,t)} \tag{2.12}$$

$$P(i | t) = \frac{\sum\limits_{f, \delta_f, m} P(i, \delta_f, m | f, t) X_N(f,t)}{\sum\limits_{i, f, \delta_f, m} P(i, \delta_f, m | f, t) X_N(f,t)} \tag{2.13}$$

$$P(f | m, i) = \frac{\sum\limits_{\delta_f, t} P(i, \delta_f, m | f + \delta_f, t) X_N(f + \delta_f, t)}{\sum\limits_{f, t, \delta_f} P(i, \delta_f, m | f + \delta_f, t) X_N(f + \delta_f, t)} \tag{2.14}$$

Alternating eq. 2.10 with eq. 2.11-2.13 defines a convergent procedure that approaches a local maximum of the log-likelihood in eq. 2.9. Typically 15-20 iterations are sufficient. This convergence towards a fixed point can be monitored by calculating the log-likelihood (eq. 2.9) associated with the model for each iteration.

Although this set of EM equations can be commonly found in specialized literature (Hofmann, 1999; Smaragdis et al., 2008; Benetos et al., 2013a), this EM-solved optimization problem coming with the parameter estimation of PLCA-based models has never been properly posed and justified. We then propose in our Technical note 1 to re-define the theoretical framework of this problem, with the motivation of making it clearer to understand, and more admissible for further developments of PLCA-based computational systems. The pitch activity matrix $P(i,t)$ is deduced from $P(i|t)$ with the Baye's rule

$$P(i,t) = P(t) P(i|t) \tag{2.15}$$

Limits of EM-based estimation

The major limitation of current PLCA models lies in the inherent problems of the EM algorithm. This algorithm was originally introduced by Dempster et al. (1977) to overcome the difficulties in maximizing likelihoods of missing data models. The main advantage of that method is its easy implementation, consisting of initializing the parameters and iterating expectation and maximization likelihoods in a step-by-step process until convergence. Its major drawback, besides the requirement of convex likelihoods, lies in its sensitiveness to initialization, which increase the risks to local convergences (Robert and Casella, 1999). That issue is exacerbated in the case of multimodal likelihoods. Indeed, the increase of the likelihood function at each step of the algorithm ensures its convergence to the maximum likelihood estimator in the case of unimodal likelihoods, but implies a dependence on initial conditions for multimodal likelihoods. Alternative techniques have also been proposed to optimize the search of global maxima, such as running the algorithm a number of times with different, random starting points, or using variants from the basic EM algorithm such as Deterministic Annealing EM (DAEM) algorithm (Ueda and Nakano, 1998). These theoretical issues have reached research fields working on audio signals.

To tackle the problem of dependency to initialization, some authors (Grindlay and Ellis, 2010; Benetos and Dixon, 2013) perform a training of the instrument templates, which has proved to be an effective way to initialise the spectral bases. Indeed, dependency to initialization should be a problem in the basic PLCA model when both basis functions and activations are estimated from a given magnitude spectrogram. If, however, one of the two is fixed, the estimation of the other is effectively a gradient descent method with a convex cost function which should converge against a global optimum. In other words, by fixing them without data-driven updating, we obtain a stable output for the gain function, independent of its initialisation. However, when the model becomes more complex with for example the introduction of different instrument variables, performing robust initialization is more difficult. For what concerns the local convergence problem, some works (Hoffman et al., 2009; Grindlay and Ellis, 2010; Cheng et al., 2013) have used the DAEM algorithm based on a temperature parameter. This limitation becomes particularly critical when integrating priors into the PLCA framework. Generally speaking, this integration introduces generic problems in optimization convergence to global maxima, especially when the prior has a multi-modality form. Indeed, when a prior is injected, the maximization step becomes a maximum a posteriori step and the log posterior probability needs to have the good properties for maximization. Fuentes et al. (2013) used of a numerical fixed point algorithms to solve the modified EM equations with a sparsity prior, whose convergence is only theoretically supposed, but "observed in practice" (although the sensitivity of the algorithm convergence to the evaluation sound dataset is not detailed). Benetos and Dixon (2013) privileged the use of pre-defined templates, which allows them to skip computing the EM update equation of templates, and just to apply a sparsity constraint on the pitch activity matrix and the pitch-wise source contribution matrix. Also, the simultaneous use of several priors on a same model parameter leads to some difficulties in terms of mathematical calculation and increases convergence problems Fuentes et al. (2013).

To overcome such limitations, two estimation algorithms are now proposed as alternatives of the classical EM algorithm, namely the Deterministic Annealing EM (DAEM) and the Filtering Particle (FP) algorithms.

2.2.6 Deterministic Annealing EM (DAEM) -based estimation

Deterministic Annealing EM (DAEM) is an extension of the algorithm EM, proposed by Ueda and Nakano (1998), and already applied to PLCA by Cheng et al. (2013). EM equations are modified to integrate a temperature coefficient β as follows

$$P(i,\delta_f,m|f,t) = \frac{P(f-\delta_f|i,m)P(\delta_f|i,t)P(i|t)P(m|i,t)^{\beta}}{\sum\limits_{i,\delta_f,m} P(f-\delta_f|i,m)P(\delta_f|i,t)P(i|t)P(m|i,t)^{\beta}} \tag{2.16}$$

2.2.7 Filtering particle -based model parameter estimation

State space representation

The PLCA model can be expressed as :

$$\begin{aligned} P(x,t) &= \sum_z P(z,t)P(x|z) \\ &= \sum_{z_1,\dots,z_K} P(z_1,\dots,z_K,t)\prod_{j=1}^{J} P(x_j|z_1,\dots,z_K) \\ &= \sum_{z_1,\dots,z_K} P(z_K,t)\prod_{k=1}^{K-1} P(z_k|z_{k+1},\dots,z_K,t) \\ &\qquad \prod_{j=1}^{J} P(x_j|z_1,\dots,z_K) \end{aligned} \tag{2.17}$$

with :

— $z \in Z_1 \times \ldots \times Z_K$ is a vector of K latent components $(z_1,\dots,z_K)$ associated to a finite subset $Z_k = \{1,\dots,L_k\}$

— $t \in \{0,\dots,T\}$ is the time variable

— $x \in X_1 \times \ldots \times X_J$ is a vector of J features $(x_1,\dots,x_J)$ where $X_j = \{1,\dots,F_j\}$

In this decomposition, $P(z_K,t)$ can be seen as the activation distribution of the latent variable z_K, $P(z_k|z_{k+1},\dots,z_K,t)$ as the weight of the variable z_k conditionally to $(z_{k+1},\dots,z_K)$ and $P(x_j|z_1,\dots,z_K)$ as the J features basis.

To estimate the set of parameters $p_t = \{P(z_K,t), P(z_k|z_{k+1},\dots,z_K,t)\forall k \in \{1,\dots,K-1\}\}$ at each time $t \in \{0,\dots,T\}$, the model can be rearranged as a state space process

$$p_t \sim f(p_t|p_{t-1}) \tag{2.18}$$

$$y_t \sim g(y_t|p_t) \tag{2.19}$$

where f is the transition state density function for p_t defined above and g the observation function of y_t.

Transition and observation densities

Transition density Assuming that each latent variable $z_k \in Z_k$ is i.i.d, each marginal vector $P(z_K,t)$ and $P(z_k|z_{k+1},\dots,z_K,t)$ can be independently estimated. Recalling that at a given time t, $P(z_K,t)$ and $P(z_k|z_{k+1},\dots,z_K,t)$ represent distributions, Dirichlet priors are injected to ensure that their elements belong to $[0,1]$, as follows, $\forall(z_2,\dots,z_K) \in Z_2 \times \ldots \times Z_K, \forall k \in \{1,\dots,K-1\}$

$$P(z_K, t) \sim Dir(\theta_t^1, \ldots, \theta_t^{L_K}) \tag{2.20}$$

$$P(z_k|z_{k+1}, \ldots, z_K, t) \sim Dir(\delta_k(z_k, \ldots, z_{K-1})_t^1, \ldots, \delta_k(z_k, \ldots, z_{K-1})_t^{L_K}) \tag{2.21}$$

where θ and δ_k are random variables representing the weight of each component of z_K in $P(z_K, t)$ and $P(z_k|z_{k+1}, \ldots, z_K, t)$. That injection leads to the following hierarchical model

$$\begin{array}{ccc} H_t & \rightarrow & H_{t+1} \\ \downarrow & & \downarrow \\ P_t & & P_{t+1} \\ \downarrow & & \downarrow \\ Y_t & & Y_{t+1} \end{array} \tag{2.22}$$

with $H_t = (\Theta_t, \Delta_t)$ the new states defined by

$$\Theta_t = \{\theta_t^{z_K}, \forall z_K \in Z_K\} \tag{2.23}$$

$$\Delta_t = \{\delta_k(z_k, \ldots, z_{K-1})_t^{z_K}, \forall z_K \in Z_K\} \tag{2.24}$$

where we have defined

$$\theta_{t+1}^{z_K} = \theta_t^{z_K} \times \alpha_t^{z_K}, \alpha_t^{z_K} \sim \phi \tag{2.25}$$

$$\delta_k(z_k, \ldots, z_{K-1})_{t+1}^{z_K} = \delta_k(z_k, \ldots, z_{K-1})_t^{z_K} \times \gamma_t^{z_K}, \gamma_t^{z_K} \sim \psi_k \tag{2.26}$$

with ϕ and ψ_k are positive distributions.

Observation density y_t has been defined as a representation of x at time t. In that state space approach, each component of y_t is represented by the sum of the PLCA model and a white noise, $\forall x \in X_1 \times \ldots \times X_J$,

$$y_t(x) = P(x, t) + V_t = \sum_{z_1, \ldots, z_K} P_t(z_1, \ldots, z_K) \prod_{j=1}^{J} P_t(x_j|z_1, \ldots, z_K) + V_t \tag{2.27}$$

where $V_t \sim N(0, \sigma^2)$. Denoting $\hat{y_t}$ the vector of components $P(x, t)$, the observation density g follows a normal distribution, i.e. $g \sim N(\hat{y_t}, \sigma^2)$.

2.2.8 Definition of spectral templates $P(f|i, m)$

Fixed parametric note model

Here, the spectral templates $P(f|i, m)$ are extracted from isolated note spectra of each pitch using a one component PLCA, and are not updated during parameter estimation. In our baseline PLCA system, labelled B_0 in the following, note spectra are built synthetically with parametric harmonic note models. We model a note spectrum as a weighted sum

of fixed narrowband harmonic spectra, called kernels, convolved with a pitch impulse distribution. Amplitudes of kernels are fixed, and set to decrease as 1 over the square of the harmonic number. This acoustic model intends to be as generic as possible in modeling plucked-string instrument sounds.

EM-updated parametric note model

In the PLCA model developed by Fuentes et al. (2013), at a given time t, the spectrum of the m^{th} polyphonic source is decomposed as a weighted sum of different harmonic notes, each one having its own fundamental frequency and spectral envelope. These authors then use the algorithm EM to adapt these parametric kernel amplitudes, i.e. spectral envelop, to the input audio data.

Pre-recorded note samples

Kernels can be learned from pre-recorded isolated notes using a one component PLCA model, as performed by Benetos et al. (2013a). This prior makes use of a single template per pitch, selected as the one with the highest amplitude. These templates can be adaptively updated with input data, or kept fixed.

Data-based template adaptation & Conservative transcription

As a last strategy to define PLCA templates, the concept of conservative transcription for data-based template adaptation has been lately proposed (Tidhar et al., 2010; Benetos et al., 2014a), using the EM update eq. 2.14. When updating templates on unlabelled data, one has to avoid that one latent variable is scattered into $P(i|t)$ and $P(f|i,m)$ for multiple bases, which consequently degrades the accuracy of these magnitudes, leading to an incorrect solution. To answer this problem, some authors (Tidhar et al., 2010; Benetos et al., 2014a) have then introduced the concept of conservative transcription. This particular transcription first consists of identifying only those detected note events for which we have a high degree of confidence (i.e. the system returns few false alarms but might miss several notes present in the recording), and omitting any unsure candidates. This is performed with a PLCA model using a generic fixed template dictionary.

Note events are then extracted by thresholding the event activation matrix $P(i,t) = P(t)P(i|t)$, and select only the time intervals $\boldsymbol{Tp}$ whose activity values exceed the threshold $Thres_{fix}$ (see Sec. 2.3). We will now use these intervals to learn the template dictionaries, adaptively to the input data. As before, a one-component PLCA is used for each time interval, taking as input $X(f,t_k)$ with $t_k \in \boldsymbol{Tp}$. The output for each latent component z is a spectral template $P(f|z)$ which can be used in order to expand the present dictionary. Following Benetos et al. (2014a), this template adaptation can be controled by a parameter ε through the following equation

$$P(f|z) = \frac{\sum_t \varepsilon P_t(z|f)X(f,t) + (1-\varepsilon)\boldsymbol{w_{theo}}}{\sum_{f,t} \varepsilon P_t(z|f)X(f,t) + (1-\varepsilon)\boldsymbol{w_{theo}}} \tag{2.28}$$

with $\boldsymbol{w_{theo}} = P(f|z)$ corresponding to some initial template dictionaries. $P_t(z|f)$ is the posterior of the model (defined in Benetos et al. (2013a)). Higher values of the parameter ε, which will be set to 0.02 here, allows increasing template adaptation. This parameter controls the levels of precision/recall: a low threshold has a high recall and low precision; the opposite occurs with a high threshold, which is done for our conservative segmentation.

We can now use these intervals to learn the template dictionaries corresponding to the different pitches, adaptively to the input data.

2.3 Note segmentation

2.3.1 Notations

This note segmentation stage allows estimating a subset of played notes $\tilde{\boldsymbol{I}} \subset \mathbf{I}$, and the time intervals $\boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})} = [t_k, \cdots, t_k + length(\boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})})]$ on which a note is played, for each pitch $i \in \tilde{\boldsymbol{I}}$. Each interval $\boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})}$ is indexed by its starting time frame t_k. Based on these different intervals, we can compute the binary piano-roll transcription output $\hat{P}(i,t) \in \{0,1\}$, defined as

$$\hat{P}(i,t) = \begin{cases} 1 & \text{for t} \in \boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})} \\ 0 & \text{otherwise} \end{cases}$$

2.3.2 Monophonic transcription

A first simple note segmentation is to choose the most likely pitch in each frame by using the MAP estimator, $\forall t$

$$\hat{P}(i,t) = \begin{cases} 1 & \text{for } i = \underset{i}{\text{argmax}} \{P(i|t)\} \\ 0 & \text{otherwise} \end{cases}$$

2.3.3 Simple thresholding

Eventually, as in most spectrogram factorization-based transcription or pitch tracking methods (Grindlay and Ellis, 2011; Mysore and Smaragdis, 2009; Dessein et al., 2010), we use a simple threshold-based detection of the note activations from the pitch activity matrix $P(i,t)$, followed by a minimum duration pruning. The threshold on note activations, is labelled $Thres_{fix}$ and set to 0.6 by default. The threshold for minimum duration for pruning was set to 130 ms as in Dessein et al. (2010). In comparative studies of AMT systems, the use of this simple thresholding method allows one to better highlight the differences between these different systems, or brought by the incorporation of different KCMA. The binary matrix $\hat{P}(i,t)$ is given by

$$\hat{P}(i,t) = \begin{cases} 1 & \text{for } \{i,t\} = \{i,t|P(i,t) \geq Thres_{fix}\} \\ 0 & \text{otherwise} \end{cases}$$

2.3.4 Adpative thresholding

In this paper, we first extracted onsets from $P(i,t)$ using a pitch-wise energy-based Onset Detection Function (ODF), which computes a spectral flux on each activation band to emphasize local energy changes, followed by a half-wave rectification (Bello et al., 2005) so that negative peaks marking the offset of a musical event are discarded and only positive values are taken into account. Most ODFs are post-processed with a dynamic thresholding $Thres_{dyn}(t)$ to take into account the loudness variations of a music piece, and in our case the corresponding activation probabilities, which can be computed (adapted from Bello et al. (2005)) as

$$Thres_{dyn}(t) = Thres_{fix} + median(max(P(:, t-M)), ..., max(P(:, t+M))) \quad (2.29)$$

with $Thres_{fix}$ an offset coefficient. All these peaks are then stored, and unitary normalized over all peak values. An onset activation score is itself defined as the likelihood averaged over four temporal frames just after the frame where the onset has been located. Based on this score, we define the offset time of each note as the time it takes to the onset activation score to decrease by 80 % of its value. The binary matrix $\hat{P}(i,t)$ is given by

$$\hat{P}(i,t) = \begin{cases} 1 & \text{if } \{i,t\} = \{i,t|P(i,t) \geq Thres_{dyn}(t)\} \\ 0 & \text{otherwise} \end{cases}$$

2.4 Hidden Markov Models

2.4.1 Theoretical Background

The two main questions to be addressed when developing Hidden Markov Models (HMMs) are

1. What are your observations ? From which one will compute observation probabilities ;
2. What is the supposed generative process of these observations ? From which one will compute transition probabilities between different states to explain the observations.

Markov Chain

Consider a system that is described by a set of N distinct states S_k, where $S_k \in \mathbf{S} = \{S_1, S_2, ..., S_{N_s}\}$. The states of the system may change with time, and at the time instants t, they are denoted by q_t, a discrete-time random variable taking value in the finite set **S**. The dynamics of the system is described by a homogeneous first-order Markov chain, that is, when at time t the system is in state S_i , there is a fixed probability that at time t, it will be in state S_j, where the probability depends only on the state at time t-1. This Markov chain is then defined as follows:

$$P(q_t = S_j|q_{t-1} = S_i, q_{t-2} = S_k, ...) = P(q_t = S_j|q_{t-1} = S_i) = a_{ij} \quad (2.30)$$

where $\mathbf{A} = [a_{ij}]$ is an $N_s \times N_s$ transition matrix, with

$$a_{ij} \geq 0 \quad \text{and,} \quad \sum_{j=1}^{N} a_{ij} = 1$$

Training Markov chains (of any order) from data is fairly straightforward if state values are directly observable, as they are for symbolic data such as text and musical scores. To obtain the Markov chain under which the observed data are most likely to have occurred, one simply sets the transition probability vector from each state to match the relative frequencies of each observed transition.

Hidden Markov Chain

Although simple Markov chains lend themselves well to applications involving symbolic data, to model continuous data such as feature vector sequences describing audio we need to add another layer of complexity. In such modeling scenarios, the state sequence is not directly observable, but hidden from the observer. Then Hidden Markov Model (HMM) assumes that each state is associated with an emission probability density function that generates our observed data. A HMM can then be seen as a doubly stochastic generative process, with two components: a set of hidden variables that can not be observed directly from the data, and a Markov property that is usually related to some dynamical temporal behaviour of the hidden variables.

In mathematical terms, at every time instant t, the system generates an observation y_t according to a probability distribution that depends on the underlying state q_t. If the number of distinct observations is N_o and the set of observation symbols is $\boldsymbol{\upsilon} = \{\upsilon_1, \upsilon_2, ..., \upsilon_{N_o}\}$, the probability distributions of observed symbols are given by an $N_s \times N_o$ matrix $\mathbf{B}$, whose elements b_{jk} are known as emission probabilities and are defined according to

$$b_{jk} = P(y_t = v_k | q_t = S_j), \qquad 1 \leq j \leq N_s \quad \text{and,} \quad 1 \leq k \leq N_o \tag{2.31}$$

with

$$b_{jk} = \sum_{k=1}^{N_o} b_{jk} = 1 \tag{2.32}$$

Finally, to complete the specification of the model, one needs to provide the initial state distribution defined by $\mathbf{\Pi} = (\Pi_1 \dots \Pi_{N_s})$, where $\Pi_i = P(q_1 = S_i)$. The three probability distributions described by $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{\Pi}$ are, in short, denoted by $\boldsymbol{\lambda} = \{\mathbf{A}, \mathbf{B}, \mathbf{\Pi}\}$. Typically, a common assumption for an observed sequence $\mathbf{y}^T = [y_1, y_2, ..., y_T]$ is that its joint probability mass function conditioned on the state sequence $\mathbf{q}^T = [q_1, q_2, ..., q_T]$ and the parameters $\boldsymbol{\lambda}$ is given by

$$P(\mathbf{y}|\mathbf{q}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} P(y_t | q_t, \boldsymbol{\lambda}) \tag{2.33}$$

which means conditional independence of the observations.

Sigmoid projection

The following sigmoid function f is applied to this observation distribution, in order to project non-negative data into the range $[0 : 1]$,

$$f(x) = \frac{1}{1 + \exp^{-x-\lambda}} \tag{2.34}$$

where the parameter λ controls the smoothing strength, i.e. the higher its value, the more low probability values will be discarded. The use of this sigmoid function has already been proposed in the literature (Benetos and Dixon, 2013). We propose in our Technical note 2 a theoretical development generalizing the use of this sigmoid function to format observation distributions before note segmentation, relating its parameters to explicit music characteristics. For parameter estimation, we employ a GEM algorithm.

HMM problems

There are three basic problems related to HMMs (Rabiner, 1989), and in order of increasing complexity, they are the following:

1. Given a set of observations $\mathbf{y}^T = [y_1, y_2, ..., y_T]$ and the model parameters $\boldsymbol{\lambda}$, the objective is to find the probability of the observed sequence $\mathbf{y}$, $\mathrm{P}(\mathbf{y}|\boldsymbol{\lambda})$;
2. Given a set of observations $\mathbf{y}^T = [y_1, y_2, ..., y_T]$ and the model parameters $\boldsymbol{\lambda}$, the objective is to find the corresponding state sequence $\mathbf{q}$;
3. Given a set of observations $\mathbf{y}^T = [y_1, y_2, ..., y_T]$, the objective is to find the state sequence $\mathbf{q}$ as well as the model parameters $\boldsymbol{\lambda}$

The solutions to these three problems are well known (Rabiner, 1989). The first problem can be solved efficiently by the forward-backward procedure, the second, by the Viterbi algorithm, and the third by the iterative method of BaumâĂŞWelch.

Viterbi resolution

We develop here the well-known resolution of the problem 2 described above, using the Viterbi algorithm Dempster et al. (1977). It is very complex and difficult to directly model the joint probability density function of the observation sequence. A reasonable approach is to group nearby observations of similar characteristics as being produced by the same state and then consider how do the states progress and how does a state sequence produce the observation sequence. Hence, the model is split into two parts, the first part considers the probability of a state sequence and the second part considers the observation sequence based on a state sequence, and the joint probability distribution given by the model is written as, given a sequence of measurements $y_1, ..., y_T$ and assuming a certain sequence of hidden states $q_1, ..., q_T$,

$$P(q_1, ..., q_T, y_1, ..., y_T) = P(q_1)P(y_1|q_1) \prod_{t=2}^{T} P(y_t|q_t) \cdot P(q_t|q_{t-1}) \tag{2.35}$$

From this equation 2.35, we define the score of the candidate optimal partial path that, at time t, reaches state i

$$\delta(i,t) = \max_{q_1,...,q_t} ln(P(y_1, ..., y_T, q_1, ..., q_t = i)) \tag{2.36}$$

The most likely state sequence Q is then given by

$$Q = arg \max_{i,t} \{\delta(i,t)\} \tag{2.37}$$

which can be estimated by the Viterbi algorithm (Dempster et al., 1977). Once we have the initial model parameters, this algorithm is used to decode the training data and obtain the new optimal state sequences. Then, we can re-estimate the model parameters based on the new optimal state sequences. The decoding and re-estimation procedure is iterated until the likelihood score of the data converges or a prescribed number of iterations are reached.

Higher-Order HMM

So far, we have made the assumption that the future state of a process depends only on its single most recent state, in terms of state transition and output observation. Consequently, the feature vectors of consecutive sound frames belonging to the same state are independent identically distributed, and trajectory modeling (i.e., frame correlation) in the frame space is not included. Such modeling characteristics are often unreasonable and reveal limits of first-order HMMs, especially in regards to state duration modeling as they violate the high correlations among successive frame-wise states. Indeed, the piecewise stationary assumption does not precisely match the non-stationary nature of the actual sound process.

In regards to such limitations, it sometimes makes sense to take more than one previous state into account using a higher-order Markov chain to get a higher prediction accuracy. HO-HMMs can be defined by making transition probabilities depend on the last k states in addition to the most recent state. Ching (2004) and Ching et al. (2008) proposed the high-order (k^{th}-order) Markov models. To a k^{th}-order ($k > 1$) Markov chain, the present state relies not only upon the last one state, but also on the k-1 previous states. We can define a k^{th}-order Markov chain as follows

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...) = P(q_t = S_j | q_{t-1} = S_i, ..., q_{t-k} = S_l) = a_{i...l} \quad (2.38)$$

HO-HMMs have received a great deal of interest to model frame state duration (Mari et al., 1997; Ching et al., 2008). By definition, one can actually build more "memory" into states by using a high-order Markov model with a trajectory modeling within state sequences.

Mathematically, if we consider that state transition depends only on the previous O_d states and output depends only on the previous O_e states, then the model can be thought as an HO-HMM of orders (O_d, O_e). The joint probability of state and observation sequences of eq. 2.35 can then be re-written as

$$\begin{aligned} P(q_1, ..., q_T, y_1, ..., y_T) = P(q_1)P(y_1|q_1) \\ \cdot P(q_2|q_1)P(y_2|q_1, q_2) \\ \cdot P(q_t|q_{t-O_d}, ..., q_{t-1})P(y_t|q_{t-c+1}, ..., q_t) \\ \cdot P(q_T|q_{T-O_d}, ..., q_{T-1})P(y_T|q_{T-c+1}, ..., q_T) \end{aligned} \quad (2.39)$$

From eq. 2.39, we can see that there is an enormous number of parameters to be estimated (about N^{O_d} transition probabilities and N^{O_e} probability density functions for an N-state HO-HMM). The corresponding transition matrix $\mathbf{A}$ is an $N^{O_d} \times N$ matrix. Obviously, the total number of independent parameters is $N^{O_d} \times (N-1)$, which grows exponentially with the order O_d, so that it's impossible to achieve effective parameters estimation perfectly.

It is noteworthy that some researchers (Dubnov, 2003) have shown that at very low orders - such as the second order or so-called bigram Markov models - generate strings that do not recognizably resemble strings in the corpus, while at very high orders, the model simply replicates strings from the corpus.

2.4.2 General applications in audio

In audio recognition systems, especially in speech, based on HMMs are very popular not only because of its modeling power but also because there are powerful mathematical

tools that can be used to efficiently estimate the model parameters, calculate the likelihood scores, and guide the design of systems. Applied to music, the HMM paradigm is used to solve three main tasks, namely classification, segmentation and learning, which can be related to these three HMM problems mentioned above (see Sec. 2.4.1).

Problem 1 Given several candidate HMM models representing different acoustic sources (musical instruments in our case), the classification problem computes the probability that the observations came from these models. The model that gives the highest probability is chosen as the likely source of the observation.

Problem 2 The segmentation problem means finding the most likely sequence of the hidden states given an observation $o_1, ..., o_T$. HMMs have been used to address musical segmentation problems by several researchers (e.g. Raphael (1999); Aucouturier and Sandler (2001)). These works dealt with segmentation of a sound into large-scale entities such as characterizing repetitive patterns (Logan and Chu, 2000), with the purpose of performing tasks such as score following or identification of texture changes in a musical piece. In the same vein, HMMs have also been used for music structure analysis at a smaller scale, with harmonic analysis (Raphael and Stoddard, 2003) and chord estimation (Bello and Pickens, 2005; Lee and Slaney, 2008).

Problem 3 Learning is the first problem that needs to be solved in order to use a HMM model, unless the parameters of the model are externally specified. It means estimating the parameters of the models, usually iteratively done by the EM algorithm (Dempster et al., 1977).

For what concerns the task of AMT, the sequential structure that may be inferred from musical signals can be usefully integrated to systems with HMMs. Indeed, as approaches consisting of a Multi-Pitch Estimation stage, or even classification approach (Poliner and Ellis, 2007), display the obvious fault of processing each frame independently of its neighbors: the inherent temporal structure of music is not exploited. HMMs have then been used for three main tasks: note modeling, note segmentation and sequential post-processing. All these tasks fall within the second HMM problem defined above, and are now developed in the following sections. In this PhD thesis, we put a special focus on the last two ones, which will be used in Chapter 3 for KCMA incorporation.

2.4.3 Acoustic modeling of note events

HMMs have made use of the inherent temporal structure of audio and have shown to be particularly powerful in modeling sounds in which temporal structure is important, such as speech (Rabiner, 1989). Two different approaches can be distinguished here. A first one which integrates HMMs directly into MPE methods, such as NMF (Nakano, 2010) and PLCA (Mysore and Smaragdis, 2009; Benetos and Dixon, 2013). A second one which uses HMMs to model note events from the salience function of the system during a post-processing stage. Ryynanen and Klapuri (2005) and Ryynanen and Klapuri (2008) define a left-to-right HMM topology to model acoustically note events. Grosche et al. (2012) use two models for segmenting the semitone bands. An on model that captures properties of sounding notes and an off model which captures regions where no note is active. The on model exhibits three states with a left-right topology. This reflects the assumption that a note consists of an attack, decay, and sustain part.

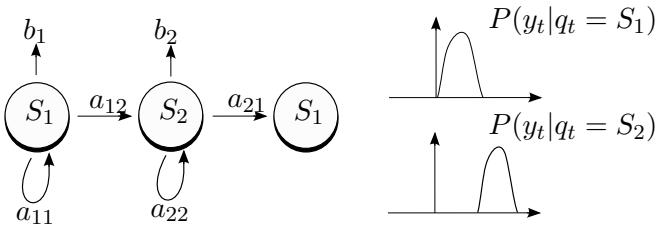


Figure 2.2 – On the left, graphical representation of transition probabilities between three hidden states. On the right, probability density functions for the observed data given the underlying state of the model y_t is the observation at time t, and q_t is the underlying state label at time t.

2.4.4 HMM-based note segmentation

As a replacement of the simple thresholding (see Sec. 2.3.3), HMM can be used to perform pitch-wise note segmentation from a salience matrix. This function of HMMs consists in a time filtering of note detection decision and generates smoothed note boundaries. We propose two HMM models for this task, the first order two-state on/off HMMs, developed by Poliner and Ellis (2007), and an original one which extends this model by including note duration modeling through a higher order. Here, the hidden stochastic process for HMMs in AMT is associated to the activation matrix $P(i, t)$ computed during the multi-pitch estimation stage.

First-order two-state on/off HMM

This model has been introduced in Poliner and Ellis (2007). Each pitch is modelled as a two-state on/off HMM, i.e. $S_i \in \{0, 1\}$, which denotes pitch activity/inactivity. The state dynamics, transition matrix, and state priors are estimated from our "directly observed" state sequences, i.e. the training MIDI transcripts, which are sampled at the precise times corresponding to the analysis frames of the activation matrix. The initial probability $P(q_1)$ for each note is supposed to be 1 for the off state, because all notes are inactive at the beginning of a recording. There are only four transition probabilities $P(q_t|q_{t-1})$, which correspond to the following state transitions: on/on, on/off, off/on, off/off. These probabilities are strongly dependent on the tempo of the considered composition. $P(y_t|q_t)$ is the observation probability, whose values are extracted frame-wisely from the PLCA activation matrix. Figure 2.2 provides a graphical representation of transition probabilities between two hidden states with this model, along with their observation distribution.

O_d-Order two-state on/off HMM

In this HMM model, as we only have two states, $S_i \in \{0, 1\}$, the left-to-right HMM topology allows for a simple modeling of their temporal dynamic through the duration that the last state has stayed. Through this model, we no longer consider consecutive states belonging to a same state as independent identically distributed, but aggregate them in groups according to the duration of staying in a state. The knowledge memorized in this modeling is then about note duration, and allows for an efficient reduction of the HMM parameters to be estimated. Mathematically, eq. 2.39 can be reduced to

$$P(q_1,...,q_T,y_1,...,y_T) = P(q_1)P(y_1|q_1,d_1(q_1)) \cdot P(q_t|q_{t-1},d_{t-1}(q_{t-1}))P(y_t|q_t,d_t(q_t)) \cdot P(q_T|q_{T-1},d_{T-1}(q_{T-1}))P(y_T|q_T,d_T(q_T)) \quad (2.40)$$

where $d_t(s_t) \in \{1,...,O_d\}$, represents the duration that state q_t has stayed up to time t, and is equal to O_d when it exceeds the maximum dependency order O_d. Figure 2.3 provides a graphical representation of transition probabilities between three hidden states with this model, along with their observation distribution. We now need to define the score of the candidate optimal partial, which will depend on state i and staying duration d, at a given time t. Such a score can be defined as, for $0 \leq$ d $\leq O_d$,

$$\delta(i,d,t) = \max_{q_1,...,q_{t-d-1}} ln(P(q_1,...,q_{t-d-1}, q_{t-d} = i-1, q_{t-d+1} = \cdots = q_t = i, y_1,...,y_T)) \quad (2.41)$$

and for d=O_d,

$$\delta(i,D,t) = \max_{q_1,...,q_{t-O_d-1}} ln(P(q_1,...,q_{t-O_d}, q_{t-D+1} = \cdots q_t = i, y_1,...,y_T)) \quad (2.42)$$

This HO-HMM model is actually equivalent to O_dN-state first order HMM and can then be recursively solved with a Viterbi algorithm. Algorithm 2 presents this resolution for two on/off states $S_i \in \{0,1\}$. In this algorithm, the probability of transition from state i to itself after staying for d frames is denoted by $\tilde{a}(i,d)$. Briefly, in the initialization step (step 1) of the algorithm, the first frame is constrained to be in state S_1=0 (i.e. a sequence begins with silence). For d=1 (step 6), the first entrance of state S_1 is from state S_2 with a stay time of 1 through O_d, and reciprocally. For d $>$ 2 (step 8), the previous frame must be at the same state with stay time equal to d-1. For d =O_d (step 10), the first entrance of state S_1 is from state S_2 with a stay time at state $O_d - 1$, and reciprocally. In the termination step (step 14), the last frame is constrained to be at state S_1=0, and it should be at the end boundary state after the last frame.

One major advantage of this algorithm is that it allows the complex topology of equivalent first order HMMs to be automatically learned from the training data. Initially, the training data are uniformly segmented according to the state number. The observation vectors belonging to the same state and staying time are grouped. From the uniformly segmented state sequence, we can accumulate the state transition counts and then estimate state transition probabilities. Let C(i,d), d $\leq O_d$, denote the times that state i has stayed d frames and let C(i,O_d+) denote the times that state i has stayed for more than O_d frames. Then, the state transition probabilities can be estimated by

$$\tilde{a}(i,d) = \frac{C(i,d+1)}{C(i,d)}, 1 \leq d < O_d \quad (2.43)$$

$$\tilde{a}(i,O_d) = \frac{C(i,O_d+)}{C(i,O_d)} \quad (2.44)$$

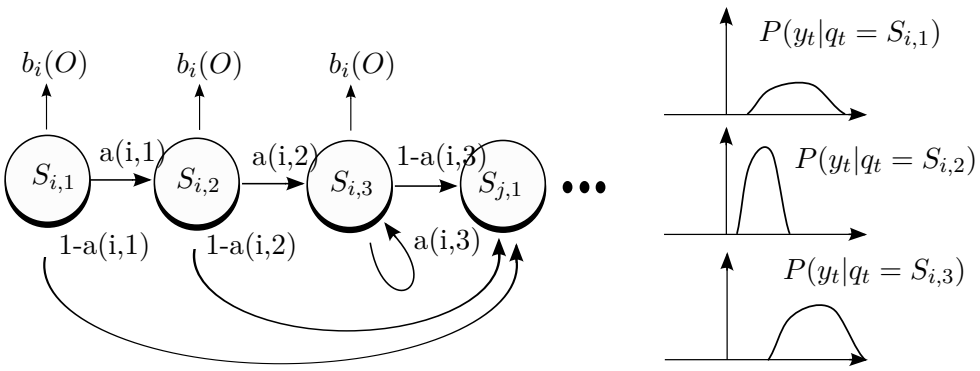


Figure 2.3 – On the left, graphical representation of transition probabilities between three hidden states. On the right, probability density functions for the observed data given the underlying state of the model y_t is the observation at time t, and q_t is the underlying state label at time t.

Algorithm 2 Viterbi Algorithm adapted for a O_d-order two-state on/off HMM

Initialization (t=1)
$\delta(1,1,1) = ln(b_{11}(y_1))$
$\delta(1,i,d) = -\inf, \quad \text{for} \quad i \in \{1,2\}, 1 < d < O_d$
for 1<t<T **do**
  for $i, j \in \{1,2\}$, with i $\neq$ j **do**
    $\delta(t,i,1) = \max_{1<\tau<O_d} \delta(t-1,i-1,\tau) + ln(\tilde{a}(j,\tau)) + ln(b_{i1}(y_t))$
    for $2 < d < O_d$ **do**
      $\delta(t,i,d) = \delta(t-1,i,d-1) + ln(\tilde{a}(i,d-1)) + ln(b_{id}(y_t))$
    end for
    $\delta(t,i,O_d) = \max_{O_d-1<\tau<O_d} \delta(t-1,j,\tau) + ln(\tilde{a}(i,\tau)) + ln(b_{iO_d}(y_t))$
  end for
end for
Termination (t=T) :
$\delta_{opt} = \max_{1<d<O_d} \delta(T,1,d) + ln(\tilde{a}(2,d))$

2.4.5 Note mixture-based HMM for sequential post-processing

In this section, we develop an original HMM-based post-processing method, which consists in rejecting unlikely successions of note mixtures based on prior knowledge of polyphonic harmonic transitions. Instead of keeping a list of binary notes as in $\hat{P}(i,t)$, we will define probabilistic note candidates and further weight their respective likelihoods using a note mixture-based HMM method.

Definition of the note mixture event

In the note segmentation stage (Subsec. 2.3), we have defined the time intervals $\boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})}$ on which a note is played, for each pitch $i \in \tilde{\boldsymbol{I}}$. Then, each appearance of a new note event, occurring at the different starting time frames $t_k \in \mathbf{T}$, will create a note mixture M_{t_k}, also indexed by t_k. Before defining these mixtures, in order to reduce the number of defined note mixtures, we aligned the intervals $\boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})}$ whose starting times are closer than 46 ms (i.e. 4 time frames, see Sec. 5.1.3 for details on numerical values).. This way, we aim to identify more conveniently symbolic music units by aggregating individual estimations, and only keep the most significant grouping of notes.

Over the duration of a musical sequence, and after note segmentation, we obtain a list $\boldsymbol{M} = \{M_{t_1}, \cdots, M_{t_k}, \cdots, M_{t_K}\}$, with K the number of starting points in a musical sequence, and where each note mixture is indexed in time by its starting time t_k. Each note mixture M_{t_k} is then parametrized by the vectors of size $N_{\tilde{I}}$, defined as

$$\boldsymbol{\Upsilon}_{\boldsymbol{t_k}} = [\Upsilon_{t_k}(1), \cdots, \Upsilon_{t_k}(N_{\tilde{I}})]^T \tag{2.45}$$

$$\boldsymbol{y}_{\boldsymbol{t_k}} = [y_{t_k}(1), \cdots, y_{t_k}(N_{\tilde{I}})]^T \tag{2.46}$$

$\Upsilon_{t_k}(\bullet)$ contains binary values $\Upsilon \in \{0,1\}$ indicating if each pitch is played in the note mixture or not. $\boldsymbol{y}_{\boldsymbol{t_k}}$ is the observation vector, where we defined the salience scores $\boldsymbol{s}_{\boldsymbol{t_k}}(i)$ as the median of the pitch activity contained in the activation interval $\boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})}$, i.e.

$$s_{t_k}(i) = \underset{t \in \boldsymbol{U}_{\boldsymbol{t_k}}^{(\boldsymbol{i})}}{median}(P(i,t)) \tag{2.47}$$

Then, for a given test musical sequence, the set of HMM states will be formed as follows $\mathbf{S} = \{\boldsymbol{S_{test}}, \boldsymbol{S_{train}}\}$, where $\boldsymbol{S_{test}} = [\boldsymbol{M_{t_1}}, \cdots, \boldsymbol{M_{t_{N_{S_{test}}}}}]_{\neq}^T$, where the suffix $\neq$ expresses the fact that the elements in this set are all distinct [1], with $N_{\boldsymbol{S_{test}}} \leq K$ the total number of states. A supplementary set of HMM states $\boldsymbol{S_{train}}$ from the training material is also added to this initial set. This operation can be performed as the observation probability for each note mixture is computed through their individual constitutive pitches. To extract these supplementary HMM states from training data, we adopt the strategy of selecting the mixtures that

1. are the most recurrent in the training material, identified through a simple frequency counting ;
2. share common pitches with the note mixtures from the test sequence. For example, a note mixture $\{D_3, F_3, A_3, B_3\}$ in the initial set can bring the supplementary states $\{D_3, F_3, A_3\}$ and $\{D_3, F_3\}$ from the training data

The numbers $N_{\boldsymbol{S_{test}}}$ and $N_{\boldsymbol{S_{train}}}$ are around 45, depending on the instrument repertoire and test sequences, and 20.

1. This distinctiveness between elements is done w.r.t the $\boldsymbol{\Upsilon}_{\boldsymbol{t_k}}$ vectors.

State-conditional observation likelihood

We then need to assign an observation probability likelihood to each M_{t_k} HMM state. The first step of this process consists in learning sample histograms of salience values respective to each pitch, and separating values from played notes and not-played notes, as illustrated in figure 2.4. These sample histograms are then smoothed using two analytical PDF models fitting their distributions, $H^{(i)}(\boldsymbol{s(i|i)})$ and $\bar{H}^{(i)}(\boldsymbol{s(i|\bar{i})})$, where $\boldsymbol{s(i|i)}$ and $\boldsymbol{s(i|\bar{i})}$ refer to the salience values of the played and not-played notes of pitch i, respectively. The first one describes the observed salience values given that note of pitch i was played according to the annotation of the training material. Similarly, the second PDF describes the observed salience values when note was not played according to the annotation. Normal and log-normal distributions are respectively used, as follows

$$H^{(i)}(\boldsymbol{s(i|i)}) = \frac{1}{\sigma_{\mathbf{s(i)}}\sqrt{2\pi}} e^{-\frac{(\mathbf{s(i)}-\mu_{\mathbf{s(i)}})^2}{2\sigma^2_{\mathbf{s(i)}}}} \tag{2.48}$$

$$\bar{H}^{(i)}(\boldsymbol{s(i|\bar{i})}) = \frac{1}{\mathbf{s(i)}\sigma_{\mathbf{s(i)}}\sqrt{2\pi}} e^{-\frac{(ln(\mathbf{s(i)})-\mu_{\mathbf{s(i)}})^2}{2\sigma^2_{\mathbf{s(i)}}}} \tag{2.49}$$

In figure 2.4, we represented three examples of learning of these sample histograms and their respective PDF models. The pitches C_4, G_4 and E_5 from the Velonjoro repertoire (see Sec. 4.4 for details on the datasounds of this PhD) have been used for this learning. Their salience values have been discretized into 80 observation scores.

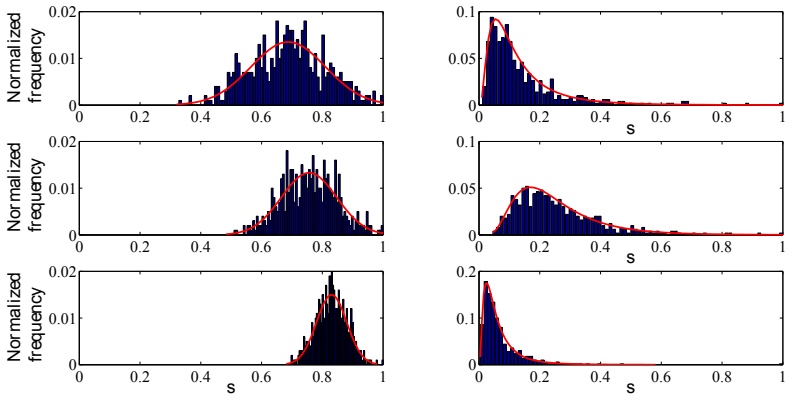


Figure 2.4 – On left graphs, sample histograms of the salience values $\boldsymbol{s(i|i)}$ in the training data when the notes with pitch i were played according to the annotation (bars), and the corresponding model PDFs $H^{(i)}(\boldsymbol{s(i|i)})$ (red curves). On right graphs, sample histograms of the salience values $\boldsymbol{s(i|\bar{i})}$ in the training data when the notes with pitch i were not played according to the annotation (bars), and the corresponding model PDFs $\bar{H}^{(i)}(\boldsymbol{s(i|\bar{i})})$ (red curves).

The state-conditional observation likelihood given a mixture M_{t_k} is then defined as

$$P(y_{t_k}|q_{t_k} = M_{t_k}) = \prod_{i=\tilde{\mathbf{I}}}^{N_{\tilde{\mathbf{I}}}} P(s_{t_k}(i)|\Upsilon_{M_{t_k}}(i)) \tag{2.50}$$

$$= \prod_{i=\tilde{\mathbf{I}}_1}^{N_{\tilde{\mathbf{I}}_1}} \prod_{j=\tilde{\mathbf{I}}_2}^{N_{\tilde{\mathbf{I}}_2}} \underbrace{P(s_{t_k}(i)|\Upsilon_{M_{t_k}}(i) = 1)}_{H^{(i)}(s_{t_k}(i))} \cdot \underbrace{P(s_{t_k}(j)|\Upsilon_{M_{t_k}}(j) = 0)}_{\bar{H}^{(j)}(s_{t_k}(j))} \tag{2.51}$$

where $\tilde{\mathbf{I}}_\mathbf{1}$ is the set of played pitches in the mixture, $\tilde{\mathbf{I}}_\mathbf{2}$ the complementary set of not-played pitches in the mixture (with $N_{\tilde{\mathbf{I}}_\mathbf{1}} + N_{\tilde{\mathbf{I}}_\mathbf{2}} = N_{\tilde{\mathbf{I}}}$). Eventually, at each starting time t_k, the different observation probabilities are normalized so as to satisfy condition given by eq. 2.32,

$$P(y_{t_k}|q_{t_k} = M_{t_k}) = \frac{P(y_{t_k}|q_{t_k} = M_{t_k})}{\sum_{j=1}^{N_\mathbf{S}} P(y_{t_k}|S_j)} \tag{2.52}$$

with $S_j \in \mathbf{S}$.

Guiding example

To complete this section, we propose a simple example to help understanding this method, as illustrated in figure 2.5. After learning sample histograms of salience values and their PDF models (step A), the note segmentation stage of the PLCA pitch activity matrix (step B) allows identifying the time frames during which a note is estimated to be played, filled with coloured rectangles, and defining the different time intervals $U_t^{(i)}$. The values of $P(i,t)$ are also noted in each pitch-time frame. Then, at step C, we define the different note mixtures M_{t_k}, indexed by the starting times t_k and the set of vectors $\boldsymbol{\Upsilon}$,

$$\boldsymbol{\Upsilon} = [\boldsymbol{\Upsilon}_{\boldsymbol{t_1}}, \cdots, \boldsymbol{\Upsilon}_{\boldsymbol{t_4}}]^T = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.53}$$

We can then compute the salience values $s_{t_k}(i)$ according to eq. 2.47. Eventually, the state-conditional observation likelihood is computed according to eq. 2.50, as detailed at step D.

2.5 Conclusion

This chapter has presented the baseline statistical methods we used to perform Automatic Music Transcription (AMT), which are mainly based on Probabilistic Latent Component Analysis (PLCA) and Hidden Markov Models (HMMs). PLCA is a probabilistic subspace analysis method which can be used for decomposing audio spectrograms. HMMs belong to dynamic Bayesian networks, and are especially known for their applications in modeling and recognition of time series. These two statistical methods use Bayes-based probabilistic frameworks, which allow powerful modelings with the incorporation of external musical knowledge. Put in combination, they cover both the time and frequency domains, a necessary condition to model music signals.

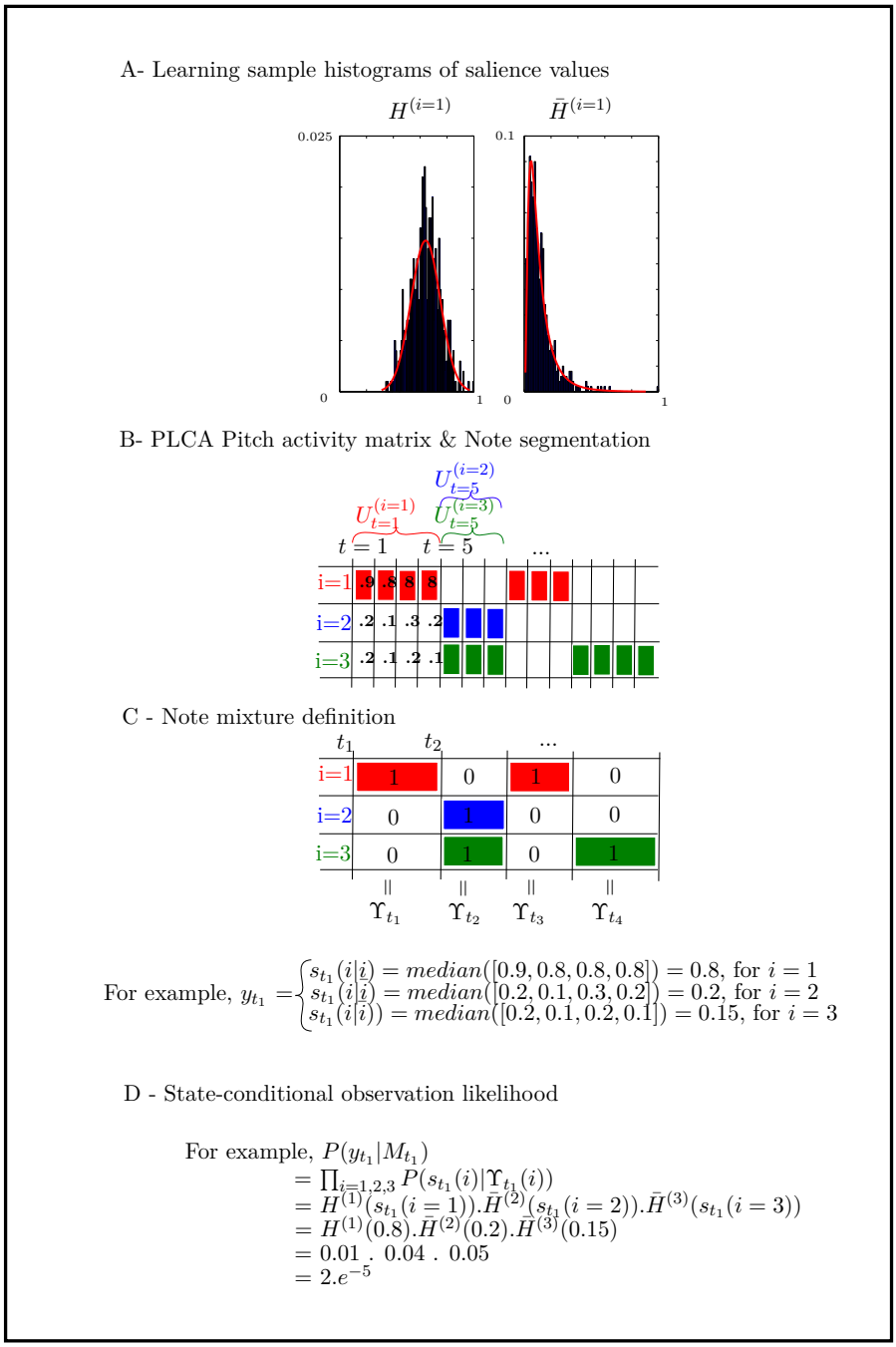


Figure 2.5 – Illustrative example explaining the computation of the state-conditional likelihood probability.

Chapter 3

Incorporating Knowledge Components from Musical Acoustics in AMT systems

Abstract

In this chapter, we develop methods to incorporate our Knowledge Components of Musical Acoustics (KCMA), as identified and extracted in chapter 1, in the baseline statistical methods described in chapter 2. In the PLCA framework, we incorporate frequency-domain knowledge, which acts as priors by constraining parameter estimation, and in the HMM framework, we incorporate time-domain knowledge in a post-processing stage. Figure 3.1 illustrates the relation of this chapter with the others.

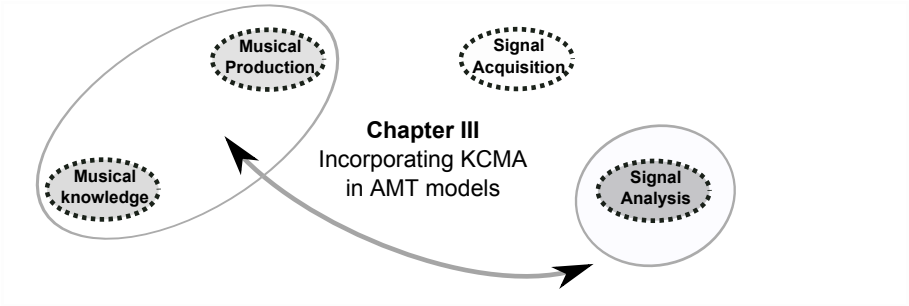


Figure 3.1 – Schematic diagram of the PhD organization for chapter 3.

3.1 Introduction

3.1.1 Background

A musical signal is highly structured, in both time and frequency domains. In time domain, tempo and beat specify the range of likely note transition times. In the frequency domain, as audio signals are both additive and oscillatory (musical objects in polyphonic music superimpose and not conceal each other), several notes played at the same time form chords, or polyphony [1], merging their respective spectral structures. When designing priors for an AMT system, one basically aims to help the system figuring out "which notes are present at time t" and "by which ones they will be followed". These two types of information belong respectively to frame-wise spectral priors (e.g. sparseness, spectrum modeling including inharmonicity (Rigaud et al., 2013)) and to frame-to-frame temporal priors (e.g. harmonic content transitions, smoothing of spectrum envelop).

Many previous works (Poliner and Ellis, 2007; Grindlay and Ellis, 2011; Benetos and Dixon, 2013) on AMT have used sequential priors to model each pitch activity/inactivity phases, which is done using two-state on/off HMMs for each of them during a post-processing stage. This operation performs a time filtering of note detection decision, which mainly avoids a lot of single miss errors and smooths note boundaries. But musically, the information is very restricted, as it consists only in knowing how long a given pitch note remains active, which can result from both playing techniques of the musician and vibratory properties of the instrument.

In principle any acoustic or score-related information that can facilitate the transcription process can act as prior information for the system. However, to be of use in a practical application, it is important that it does not require too much time and effort, and that the required information can be reliably extractable by the user, who might not be an expert musician. Depending on the expertise of the targeted users, information that is easy to provide could include key, tempo and time signature of the piece, structural information, information about the instrument types in the recording, or even asking the user to label a number of notes for each instrument. Although many proposed transcription systems -often silently- make assumptions about certain parameters, such as the number or types of instruments in the recording, not many published systems explicitly incorporate prior information from a human user.

In the context of source separation, Ozerov et al. (2012) proposed a framework that enables the incorporation of prior knowledge about the number and types of sources, and the mixing model. The authors showed that by using prior information, a better separation can be achieved than with completely blind systems. A system for user-assisted music transcription was proposed in Kirchhoff et al. (2012), where the user provides information about the instrument identities or labels a number of notes for each instrument. This knowledge enabled the authors to sidestep the error-prone task of source identification or timbre modelling, and to evaluate the proposed non-negative framework in isolation.

3.1.2 Two main families of knowledge incorporation

Knowledge is commonly incorporated in retrieval methods to better fit the decomposition to specific properties of input data. These properties can be either physics-based or signal-based. In the first case, knowledge can be explicitly modeled and incorporated in the statistical framework.

1. Here polyphonic music refers to a signal where several sounds occur simultaneously. Whereas in monophonic signals, at most one note is sounding at a time.

Signal-based

For instance, when isolated note recordings are available, on the same instrument and with the same recording conditions, the spectra of the dictionary can be learned independently (Niedermayer, 2008; Dessein et al., 2010). In that case, since the dictionary is learned on monophonic data and fixed at the learning step, high transcription performances can be obtained.

Physics-based

For example, note templates can be parametrized by some prior information on harmonicity (Bertin et al., 2010; Hennequin et al., 2010), temporal evolution of spectral envelop (Hennequin et al., 2011), sparsity of simultaneously activated notes (Hoyer, 2004), beat structure (Ochiai et al., 2012), etc., which is used to model the dictionary or the time-activation matrices.

3.1.3 Objectives & Plan

In this chapter, we make the link between the music related knowledge detailed in chapter 1, and listed as KCMA in table 1.2, and the algorithms for AMT in which they can be incorporated. The resulting overall transcription system is then of probabilistic nature, with a cascade of note probability weighting, delaying the final decision made on the estimation of active notes. In the first section 3.2, we describe the different theoretical methods used for the incorporation of our KCMA into our baseline methods for AMT. The first two methods incorporate knowledge as priors, which apply a direct constraint on the PLCA parameter estimation through modifications of EM equations. The next three methods incorporate KCMA in a post-processing stage, re-weighting candidate note probabilities in the activation matrix before note segmentation. It is noteworthy that in recent literature Sigtia et al. (2014, 2015), the expression of "re-transcription" has been used, with the computation of priors from a first transcription step, and using it in a second transcription step by re-weighting model parameter estimation in EM equations. Our processing stage is different from this approach in the sense that we do not go through the EM algorithm again to re-weight pitch probabilities. Eventually, in section 3.4 we list all KCMA implemented in our AMT baseline system. They are categorized accordingly to their respective datasound sources, as already done in Sec. 1.

3.2 Mathematical prior incorporation

3.2.1 Sparse entropy-like prior in PLCA

As mentioned in Sec. 2.2, PLCA has the advantage that it enables to use a priori knowledge of domains. From the very first works on PLCA (Shashanka et al., 2008; Smaragdis and Mysore, 2009; Smaragdis et al., 2009), sparseness assumptions have been introduced in the PLCA framework as "entropic priors" to PLCA. The most generic mechanic to introduce priors in PLCA models is through the use of a Dirichlet distribution, which is a conjugate prior distribution to the PLCA multinomial distributions $P(f|z)$ and $P(t|z)$ (Smaragdis and Mysore, 2009). It can then be used to constrain the structure of the model distribution. The priors Pr for all the frequency distributions Λ_f, and temporal distributions Λ_t, are

$$Pr(\Lambda_f) \sim \prod_z \prod_f P(f|z)^{\kappa_z \alpha(f|z)} \tag{3.1}$$

$$Pr(\Lambda_t) \sim \prod_z \prod_t P(t|z)^{\mu_z \alpha(t|z)} \tag{3.2}$$

where $\alpha(\bullet)$ are the positive and real hyperparameters defining the Dirichlet distribution. The weight scalars κ_z and μ_z can be interpreted as the prior strengths. Using the above priors, our EM estimation equations for $P(f|z)$ and $P(t|z)$ (eq. 2.13 and 2.14, respectively) now change to:

$$P(f|m,i) = \frac{\sum\limits_{\delta_f,t} P(i,\delta_f,m|f+\delta_f,t)X_N(f+\delta_f,t) + \kappa_z\alpha(f|m,i)}{\sum\limits_{f,t,\delta_f} P(i,\delta_f,m|f+\delta_f,t)X_N(f+\delta_f,t) + \kappa_z\alpha(f|m,i)} \tag{3.3}$$

$$P(i|t) = \frac{\sum\limits_{f,\delta_f,m} P(i,\delta_f,m|f,t)X_N(f,t) + \mu_z\alpha(t|m,i)}{\sum\limits_{i,f,\delta_f,m} P(i,\delta_f,m|f,t)X_N(f,t) + \mu_z\alpha(t|m,i)} \tag{3.4}$$

The first prior type will impose a sparsity on $P(f|m,i)$ so that each spectral basis component consists of only a few bins on the spectrum, while the second prior will impose a sparsity on $P(i|t)$ so that only a few pitches i are active in the same time. A third type of entropy-like prior exists, imposing a sparsity between spectral basis components so that spectral basis components $P(f|m,i)$ for different pitches i are not similar to each other. This prior takes the form of a cross-entropy defined by (Smaragdis et al., 2009)

$$H(\{P_k\}_k, \{Q_k\}_k) = -\sum_k P_k log Q_k - \sum_k Q_k log P_k \tag{3.5}$$

which measures the similarity between the two distributions $\{P_k\}_k$ and $\{Q_k\}_k$ on their common dimension, and will be used as follows in our context

$$Pr = \beta \sum_{i,i'|i\neq i'} H(\{P(f|i)\}_f, \{P(f|i')\}_f) \tag{3.6}$$

where we defined the parameter β is the prior strength. These sparseness prior is added to the general likelihood eq. 2.9 as a cost function Pr:

$$L = \sum_{f\in\boldsymbol{F}} \sum_{t\in\boldsymbol{T}} n(f,t) log(P(f,t)) - Pr \tag{3.7}$$

which leads, following the classical EM resolving procedure, to this template update equation

$$P(f|i) = B(0, \sum_t P(i|f,t)X_N(f,t) - \sum_{i=i'} P(f|i')) \tag{3.8}$$

where

$$B(\beta,\gamma_i) = \frac{\varrho(i)}{\sum\limits_{i'=1,\neq i}^{I-1} \varrho(i')} \tag{3.9}$$

and

$$\varrho(i) = \beta\gamma_i + \frac{\alpha G(\beta,\gamma_i)^\alpha}{(\alpha-1)\sum_{i'=1,\neq i}^{I-1} G(\beta,\gamma_{i'})^\alpha} \tag{3.10}$$

The estimation of eq. 3.8 is computed iteratively with the two equations 3.9 and 3.10 until convergence. As already mentioned for the concept of conservative transcription, these sparseness assumptions also allow to avoid that one pitch is scattered into $P(i|t)$ and $P(f|i)$ for multiple bases, which consequently degrades the accuracy of $P(i|t)$ and that of $P(f|i)$, leading to an incorrect solution. For our AMT application, in particular this inter-basis sparseness prior is important to impose a discriminability between the different pitches.

3.2.2 Inter-pitch sparse prior in PLCA

The inter-pitch matrix $\boldsymbol{\iota}$ (eq. 1.1) is injected in the PLCA framework by constraining the impulse distribution coefficients $P(i,t,m)$ with the term

$$P(\theta) \propto \exp(-\theta'\boldsymbol{\iota}^*\theta)$$

where $\boldsymbol{\iota}^*$ is obtained by, $\forall i,j$,

$$\boldsymbol{\iota}^*(i,j) = \begin{cases} 0 & \text{for } i=j \\ \Pi(\iota_{i,j}) & \text{otherwise} \end{cases} \tag{3.11}$$

with the operator Π defined as

$$\Pi(x) = 1 - \frac{x}{max(x)} \tag{3.12}$$

which defines an information rejecting the hypothesis of certain pitch combinations, and where $\boldsymbol{\iota}^*$ is then the matrix M whose diagonal coefficients have been set to 0. It is noteworthy that simpler modeling of prior knowledge, through for example a pitch-dependent vector, can also take the form of a diagonal matrix of size I^2, with the vector values put into this diagonal (the zero-coefficients of $\boldsymbol{\iota}$ provide an unitary prior value which does not affect particle weights). When developing eq. 3.2.2, we get the following form

$$P(\theta) \propto \exp(-\sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\boldsymbol{\iota}^*(i,j)\theta_i\theta_j) \tag{3.13}$$

The usual maximization step is replaced by a maximum a posteriori (MAP) process, i.e. instead of maximizing the log expected likelihood Q_Λ, the log posterior probability $D(\theta) = Q_\Lambda + \ln(P(\theta))$ is maximized. This maximized function is defined by

$$D(\theta) : \Omega =]0,1[^{N_I} \rightarrow \mathbb{R} \tag{3.14}$$

$$D(\theta) = \sum_{j=1}^{N_I} w_j \ln(\theta_j) - \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\boldsymbol{\iota}^*(i,j)\theta_i\theta_j \tag{3.15}$$

with $w_j = \sum_{f,z} V_{ft} P(i,s,z,c=h|f,t)$. Working with probability distributions, the following system is obtained :

$$\max D(\theta) \tag{3.16}$$

$$u.c.g(\theta) = \sum_{j=1}^{N_I} \theta_j = 1 \tag{3.17}$$

A maximum $\hat{\theta}$ exists since D is continuous and upper bounded by 0 on Ω. To maximize that function under the constraint $g(\theta) = 1$, we introduce a Lagrangian system :

$$L(\theta, \lambda) : \Omega \times \mathbb{R} \to \mathbb{R} \tag{3.18}$$

$$(\theta, \lambda) \mapsto D(\theta) - \lambda(g(\theta) - 1) \tag{3.19}$$

Moreover, the maximum $\hat{\theta}$ verifies the first order necessary conditions proper to local maxima (Lagrange theorem) since D and g are both differentiable According to these two points, there exists a unique $\lambda \in \mathbb{R}$ such that :

$$\nabla L(\hat{\theta}, \lambda) = 0 \tag{3.20}$$

The previous equation (eq. 3.20) leads to, $\forall j \in \{1, \cdots, N_I\}$,

$$\hat{\theta}_j = \frac{w_j}{\sum\limits_{i=1}^{N_I} \iota^*_{ij} \hat{\theta}_i + \lambda} \tag{3.21}$$

There is no closed-form solution for each θ_j, but following Fuentes et al. (2013), a fixed point algorithm can be used to find a solution increasing the posterior probability at each iteration of the EM algorithm. Numerical simulations showed that the fixed point Algorithm 1 always converges to a solution that makes the posterior probability increase from one iteration of the EM algorithm to the next.

Algorithm 3 Fixed-point method for the mutual resonance prior

1: $\forall j \in \{1, \cdots, N_I\}, \hat{\theta}_j \leftarrow \frac{w_j}{\sum_k \theta_k}$;
2: **while** Convergence not reached **do**
3: $\quad \forall j \in \{1, \cdots, N_I\}, r_j \leftarrow \sum\limits_k \iota^*_{jk} \theta_k$
4: $\quad$ Find the unique λ such that $\sum\limits_j \frac{w_j}{r_j + \lambda} = 1$ and $\forall j, \frac{w_j}{r_j+\lambda} \geq 0$
5: $\quad \forall j \in [1, J], \hat{\theta}_j \leftarrow \frac{w_j}{r_j+\lambda}$
6: **end while**

3.2.3 A Bayesian estimation of frame-wise note number

At a given time t and playing mode m, conditionally to the data $y = P(i, t)$, let's define the posterior distribution $P(\theta = k|y)$ of the impulse distribution vector $P(i, t, m)$. Thanks to Bayes rule, and $\forall k \in \{1, \cdots, d_{max}\}$, we have

$$P(\theta = k|y) \propto P(y|\theta = k)P(\theta = k) \tag{3.22}$$

where d_{max} is the maximum degree of polyphony and $P(\theta = k) = (\alpha_1, \cdots, \alpha_{d_{max}})$ identifies to the distribution of the polyphonic degree associated to prior P_1. In a statistical

framework, θ follows a multinomial distribution in $\{1, \cdots, d_{max}\}$ associated to probabilities $\{\alpha_1, \cdots, \alpha_{d_{max}}\}$. Then, as $P(y|\theta = k)$ is composed of the k greater components $(y_1^*, \cdots, y_k^*)$ of y, we simply compute it as the probability $P(y|\theta = k) = P(y_1^*, \cdots, y_k^*)$, i.e. the activity scores $(s_1^*, \cdots, s_k^*)$ of the corresponding pitches, as $y = P(i,t)$. Considering the independence assumption between pitch activations, the posterior distribution 3.22 can be rewritten in

$$P(\theta|y) = \begin{pmatrix} \alpha_1 s_1^* \\ \vdots \\ \alpha_k \prod_{i=1}^{k} s_i^* \\ \vdots \\ \alpha_{d_{max}} \prod_{i=1}^{d_{max}} s_i^* \end{pmatrix}$$

Then, the degree of polyphony at a given time t is the k^{th} element of θ which maximizes $P(\theta|y)$.

3.2.4 Modifying activation temporal profiles

We propose here a method to incorporate the KCMA $\sim IsoNo1$ and $IsoNo2$, respectively on note duration and amplitude energy modulation. These descriptors have been computed on the sound template sets of each instrument, to get for each pitch i the prior parameters Δ_i and β_i (defined as the median of the descriptor values for the sound samples of pitch i, and normalized by the maximum of all values). The prior parameter Δ_i is used as a pitch-specific minimum duration for pruning, as defined in step 3 of the pseudo-algorithm 4. β_i is used in an acoustics-informed smoother function of pitch-wise activation intervals $U_{i,k}$, indexed by k. Following the pseudo-code 4, this smoother computes within each activation interval, the minimum/maximum filter defined as

$$P(i, U_{i,k}(n)) = \begin{cases} min(P(i, U_{i,k}(Int))) & \text{if median(P(i,}U_{i,k}\text{(Int))) } \leq \Theta \\ max(P(i, U_{i,k}(Int))) & \text{otherwise} \end{cases} \quad (3.23)$$

with $Int = [n - \eta/2 : n + \eta/2]$ and η defining the temporal window of activity, and Θ the threshold ordering the switch between minimum and maximum filters. These two parameters define the smoothing strength. The higher is η and the lower is Θ, the smoother are the resulting activation intervals. These two parameters are constrained to range from $[2 : min(10, length(U_{i,k}))]$ frames and $[0.1 : 0.9max(P(i, U_{i,k}))]$, respectively. Their positions in these ranges are linearly indexed by the value of β_i, so that the higher is β_i, the smoother are the resulting activation intervals (i.e. the higher is η and the lower is Θ). Also, the value of 250 ms is learnt from the training, defined as the minimal note duration so as a significant $AmpMod$ degree appears. Figure 3.2 illustrates the application of this intermodulation-informed smoother on the activation of an intermodulated note.

3.2.5 Note mixture-based HMM in post-processing

First-order N_{M_n}-state HMM for harmonic transitions

This model follows a classical first-order HMM with N_{M_n} states. Figure 3.3 provides a graphical representation of transition probabilities between three hidden states with this

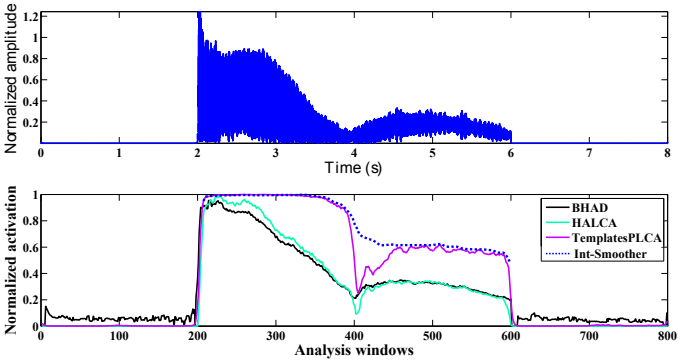


Figure 3.2 – Activation curve obtained with three different AMT systems, on an intermodulation note. Our intermodulation-informed smoother is also plotted, which results in flatting the activation dip.

Algorithm 4 Minimal Duration Pruning and Intermodulation-informed smoother

for Each pitch i, **do**
  Find pitch-wise intervals $U_{i,k}$, such that $median(P(i, U_{i,k})) > 0.4$;
  for Each interval $U_{i,k}$, **do** % *Minimal Duration Pruning*
    if length$(U_{i,k}) \leq \Delta_i$ **then**
      Pruning $U_{i,k}$
    end if
    if length$(U_{i,k}) \geq 250$ ms **then** % *AmpMod-informed smoother*
      Apply Minimum/Maximum filtering (eq. 3.23) to $U_{i,k}$
    end if
  end for
end for

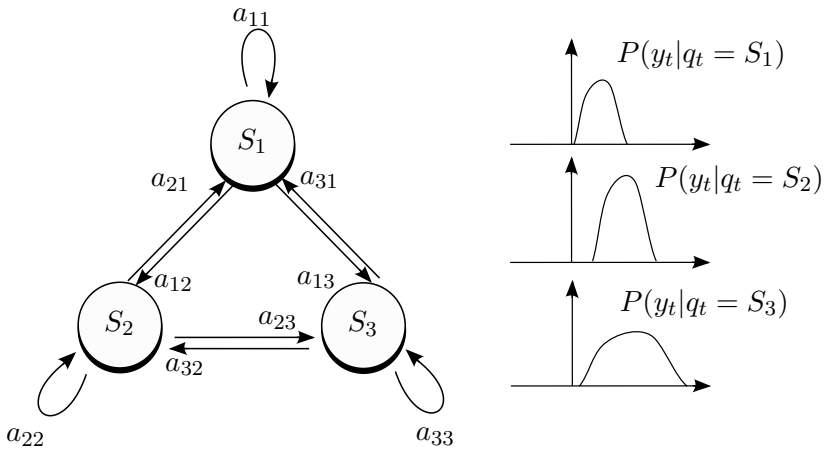


Figure 3.3 – On the left, graphical representation of transition probabilities between three hidden states. On the right, probability density functions for the observed data given the underlying state of the model y_t is the observation at time t, and q_t is the underlying state label at time t.

model, along with their observation distribution. In this model, transition probabilities are defined as the probability to switch between two successive mixtures of notes in a musical piece, simply computed by counting their occurrence frequency. These probabilities give us a global view of the usual and unusual harmonic transitions of an instrument repertoire, and are trained using MIDI scores. Despite the size of the learning database, novel note transitions may appear in test sequences without being trained, which will then not have estimated likelihoods. The distribution values must be smoothed to address tiny likelihood values for the note transitions that were not found in the learning database. Then, the Witten-Bell discounting algorithm (Witten and Bell, 1991) can be used to perform this smoothing. The algorithm is based on a principle of zero-frequency N-grams, i.e., the N-grams that have not yet occurred in the database. The zero-frequency N-gram likelihoods, however, may be estimated by the N-grams that have occurred once. Consequently, the zero-frequency likelihoods may be used to smooth the distribution.

Second-order N_{M_n}-state HMM for harmonic transitions

We used the extended second-order Viterbi algorithm already published in the literature (He, 1988).

3.3 Prior incorporation in the filtering particle framework

3.3.1 Background

In mathematical terms, priors are used to sharpen up estimation of model parameters by emphasizing the most likely values in their distributions. This prior integration is performed during state generation, by re-weighting each particle value with a corresponding prior gain. Adding prior knowledge on parameters p_t leads up to sample from the posterior distribution,

$$P(p_t|y_t) \sim P(y_t|p_t)P(p_t) \tag{3.24}$$

where $P(y_t|p_t)$ identifies to the observation density g and $P(p_t)$ the prior knowledge. When the prior and the likelihood are conjugate, sampling from the posterior distribution is rather straightforward. When the posterior does not have a well known form, as it is the case in most real-life applications, computational statistics methods can be introduced to sample from the posterior. The Metropolis-Hasting algorithm (Roberts et al., 1997; Newman and Barkema, 1999; Robert and Casella, 1999), based on Monte Carlo methods, brings a powerful framework to tackle that issue. This algorithm is a random walk that uses an acceptance/rejection rule to converge to the specified target distribution, and proceeds as described in the algorithm 5.

Algorithm 5 Metropolis-Hasting algorithm for prior integration.

1: Draw a starting point p_t^0, for which $P(p_t^0|y_t) > 0$, from a starting distribution $p_0(p_t)$;
2: **for** $q = 1, 2, \ldots$ **do**
3: Sample a proposal p_t^* from a jumping distribution at iteration q, $J_q(p_t^*|p_t^{q-1})$;
4: Calculate the ratio of densities,

$$r = \frac{P(p_t^*|y_t)/J_q(p_t^*|p_t^{q-1})}{P(p_t^{q-1}|y_t)/J_q(p_t^{q-1}|p_t^*)} \tag{3.25}$$

5: Set

$$p_t^q = \begin{cases} p_t^* & \text{with probability } \min(r,1) \\ p_t^{q-1} & \text{otherwise} \end{cases} \tag{3.26}$$

6: **end for**

Considering the filtering particle framework defined above, few remarks about the different steps can be highlighted. First, the initial draw is replaced by the PF draw we want to interfere in. Even if the posterior distribution is unknown, the ratio r can be computed as the ratio of the product of the likelihood and the prior since the normalization constant is removed in the ratio.

$$r = \frac{g(y_t|p_t^*)p(p_t^*)/J_q(p_t^*|p_t^{q-1})}{g(y_t|p_t^{q-1})p(p_t^{q-1})/J_q(p_t^{q-1}|p_t^*)} \tag{3.27}$$

The jumping distribution J_q is chosen as a normal distribution to simplify the ratio computation. Indeed, the symmetry property of the normal distribution involves that J_q can be removed in eq. 3.27, to get

$$r = \frac{g(y_t|p_t^*)p(p_t^*)}{g(y_t|p_t^{q-1})p(p_t^{q-1}))} \tag{3.28}$$

The PLCA-PF framework allows a general insertion of the matrix $\boldsymbol{\iota}$ through the term $P(p_t)$ of eq. 3.24, to which we can give the following form

$$P(p_t) \propto \exp(-p_t'\boldsymbol{\iota}K_{p_t}) \tag{3.29}$$

where $p_t \in \{A_t, B_t(s), Pr_t(s,\delta_f)\}$, $\forall(s,\delta_f) \in \{1,\ldots,S\} \times \{1,\ldots,\Delta_f\}$, and K_{p_t} is a vector of length N_I associated to p_t.

3.3.2 Sparse priors

During the multi-pitch estimation step of an AMT process, a too much large number of non-zero activation scores is often observed, making the operation of "finding the right notes" more difficult. In order to overcome this flaw, a sparseness prior can reduce the number of active notes per frame in selecting the most salient ones. Previous works mostly use pitch-independent sparse prior in PLCA-EM algorithms. Fuentes et al. (2013) compute a sparseness prior $P(A_t)$ to constrain the impulse distribution A_t, as follows

$$P(A_t) \propto \exp\left(-2\beta\sqrt{J}||A_t)||_{1/2}\right) \tag{3.30}$$

with $||A_t||_{1/2} = \sum_i \sqrt{A_t(i)}$ and β a positive hyperparameter indicating the strength of the prior. With this prior, a numerical fixed point algorithm is required to obtain a solution with the EM algorithm. Other works Grindlay and Ellis (2011); Benetos and Dixon (2011, 2013) impose sparsity on the pitch activity matrix and the pitch-wise source contribution matrix by modifying EM equations. The common point to all these EM-based sparse priors is that they are pitch-independent, and rely on hyper-parameters, which are either arbitrary set and/or optimized on a given sound dataset. In this PhD, we define sparse priors informed by explicit musical acoustics related knowledge.

Eventually, for a sparse type prior $\boldsymbol{\iota_{spa}}$, the set of prior parameters p_t in eq. 3.29 is equal to A_t and K_{p_t} is set to A_t , which becomes

$$P_{spa}(A_t) \propto \exp(-A_t'\boldsymbol{\iota_{spa}}A_t) \tag{3.31}$$

3.3.3 Sequential priors on harmonic transitions

Such a sequential prior P_{tra} is represented through a matrix $\boldsymbol{\iota_{tra}}$. Eventually, for this sequential type prior, p_t in eq. 3.29 is equal to A_t and K_{p_t} is set to A_{t-1}, which leads to

$$P_{tra}(A_t) \propto \exp(-A_t'\boldsymbol{\iota_{tra}}A_{t-1}) \tag{3.32}$$

3.3.4 Prior combination

The PLCA-PF framework offers an easy-to-implement unifying way of integrating priors from both time and frequency domains. In this framework, priors are injected during the filtering process through eq. 3.24, and modify the parameters without disturbing their generation. In the set of parameters p_t, the independence between each parameter p_t^n leads to

$$P(p_t) \propto \prod_n P(p_t^n) \tag{3.33}$$

Within a defined parameter p_t^n, the general prior $P(p_t^n)$ can be seen as the product of the different priors P_{prior}^n associated to p_t^n

$$P(p_t^n) \propto \prod_{prior} P_{prior}^n(p_t^n) \tag{3.34}$$

Using equations 3.29, 3.33 and 3.34, we combine the different priors, characterized by their respective matrices S_{prior}, as follows

$$P(p_t) \propto \exp(-\sum_n \sum_{prior} p_t^{n\prime} S_{prior}^n K_{prior}^n) \tag{3.35}$$

3.4 List of the KCMA implemented in our AMT systems

We complete this chapter by listing the different combinations of KCMA (Chap. 1) - baseline AMT methods (Chap. 2) - KCMA incorporation methods (Chap. 3), resuming the different families of KCMA presented in Chapter 1.

3.4.1 From theoretical concepts (KCMA source I)

In statistical modeling methods based on learning data, the problem is that most often typical learning music databases are very small compared to the complexity of the polyphonic music signal. The undesirable effect is that different instances of a given note mixture can be mapped to different available states, usually those that are initialized closely. If we had an infinite amount of data, we could simply represent each note mixture as a distinct observation. To compensate for this, theoretical concepts, or general model-based knowledge, can be incorporated in the statistical framework for recognition of note mixture progressions, in order to redistribute efficiently a certain amount of probability mass to unseen events during training. The KCMA from theoretical concepts (Source I, from Sec. 1.1.1) we implemented are detailed in the second column from the left of table 3.1. There are 7 such KCMA.

Prior index	KCMA Names	Baseline methods	KCMA incorporation methods
$H_{1:3}$	$Theo2, 4, 7$	PLCA	Sparse entropy prior (eq. 3.3, Sec. 3.2.1)
$H_{4:7}$	$Theo1, 3, 5, 6$	HMM	First-order N_{M_n}-state HMM for harmonic transitions (Sec. 3.2.5)

Table 3.1 – Table of the different KCMA from theoretical concepts (source I) implemented in our AMT system.

3.4.2 From isolated note samples (KCMA source II)

The KCMA from isolated note samples (Source II, from Sec. 1.1.1) we implemented are detailed in the second column from the left of table 3.2. There are 7 such KCMA in this PhD project. In the following, we make specific comments for a few.

Prior index	KCMA Names	Baseline methods	KCMA incorporation methods
T_1	Pitch range	PLCA	Spectral basis range
T_2	$IsoNo1$ and $IsoNo2$	Frame-to-note conversion	Re-weighting of activity matrix (Sec. 3.2.4)
T_3	$IsoNo0$	PLCA	Template learning (Sec. 2.2.8)
T_4	$IsoNo5$	PLCA	Inter-pitch prior matrix (Sec. 3.2.2)
T_5	$IsoNo0$ - Playing dynamic (local timbre alteration)	PLCA	Template learning (Sec. 2.2.8)
T_6	$IsoNo0$ - Excitation modes (local timbre alteration)	PLCA	Template learning (Sec. 2.2.8)
T_7	$IsoNo6$	SI-PLCA	Template shifting (Sec. 2.2.4)

Table 3.2 – Table of the different KCMA from isolated note samples (source II) implemented in our AMT system.

T_1 : Pitch range The pitch range of an instrument, given in a semi-tone spacing, is directly defined by the extrema of an instrument pitch range. For the *marovany*, taking the range extrema of our different models N_1 to N_4, we selected the following pitch range : from F_3 to D_6, with a semi-tone spacing, which gave us a range of 34 notes. By default, an AMT system would consider a much larger pitch range covering the 5 to 6 octaves of instruments such as the piano or the harp. Algorithmically, the knowledge of pitch range is set as the range of spectral basis in the PLCA algorithm.

Prior index	KCMA Names	Baseline methods	KCMA incorporation methods
M_1	$TraPla1$	PLCA	Bayesian note number estimation (Sec. 3.3)
M_2	Known played notes	PLCA	Re-weighting of Activity matrix
M_3	$TraPla2$	HMM	First-order N_{M_n}-state HMM for harmonic transitions (Sec. 3.2.5)
M_4	$TraPla3$	HMM	First-order Two-state HMM (Sec. 2.4.4)
M_5	$TraPla4$	HMM	Higher-order Two-state HMM (Sec. 2.4.4)
M_6	$TraPla5$	HMM	Second-order N_{M_n}-state HMM for harmonic transitions (Sec. 3.2.5)

Table 3.3 – Table of the different KCMA from transcripts from playing (source III) implemented in our AMT system.

T_4 **: Mutual resonances** This sparseness prior informed by sympathetic resonances is introduced to reduce the number of active notes in selecting the most salient ones, and applies a constraint on the EM algorithm update rules.

T_5 **: Dynamics of instrument playing (i.e. local timbre alteration)** To model dynamics of the instrument playing, we assigned multiple templates per pitch in the PLCA framework. These templates are extracted from pre-recorded isolated notes covering a complete range of playing dynamics.

T_6 **: Excitation modes (i.e. local timbre alteration)** To model excitation modes, multi-templates per pitch can also be used as in the prior T_5, only now we use sound samples related to specific excitation modes.

T_7 **: Non-tempered tuning** For what concerns the algorithmic implementation of T_7, we use the Shift-Invariant PLCA (Smaragdis et al., 2008) (Sec. 2.2.4) extension of the PLCA, exploiting the fact that in a CQT, a change of fundamental frequency is traduced by a simple frequency translation of its partials, resulting in a shift invariance over log-frequency. SI-PLCA then performs a multi-pitch detection with a frequency resolution higher than MIDI scale.

3.4.3 Transcripts from playing (KCMA source III)

The KCMA from transcripts from playing (Source III, from Sec. 1.1.1) we implemented are detailed in the second column from the left of table 3.3. There are 6 such KCMA in this PhD project. In the following, we make specific comments for a few.

M_1 **: Polyphonic degree** As already mentioned, polyphony is due either to an "acoustic polyphony", for instruments whose repertoires do not have a vertical writing properly speaking (e.g. the *marovany* repertoire with its fast arpeggios and strong mutual resonances), or a polyphonic writing (e.g. contrapuntal Bach pieces). The former has already been characterized by the prior T_4, and the prior M_1 deals with the second.

M_2 **: Knowing the played notes** When performing a musical piece, an instrument only plays pitches from a reduced part of its pitch range. This information of knowing which notes are actually played can be used as a simple prior in our AMT system, which is directly used to define the set of PLCA spectral basis which can be activate during parameter estimation, all others receiving a null activity score.

M_3 **: Probabilistic transition between note mixtures** This prior is based on the first-order HMM detailed in Sec. 3.2.5, whose states are defined with the different note mixtures M_k identified after note segmentation. The sequential post-processing we propose now consists in correcting the succession of M_k based on some knowledge of polyphonic harmonic transitions.

M_6 **: Second-order N_c-state HMM for harmonic transitions** We used the extended second-order Viterbi algorithm already published in the literature He (1988). Only now, training of state transition probabilities implies counting two-state sequences.

3.5 Conclusion

In this chapter, we were interested in developing a generic transcription system based on a PLCA post-processed with a HMM block, which makes possible to supply it with all kind of prior knowledge at different levels of abstraction.

Chapter 4

Multichannel capturing sensory systems for transcription ground-truth creation

Abstract

In this chapter we propose a method of automating the daunting task of providing the machine-learning models with labeled data for training and evaluation. This method is based on the technology of Multichannel Capturing Sensory Systems (MCSSs), which allow decomposing a multi-source audio signal into simple identifiable components, and simplifying more particularly the complex analysis of a polyphonic sequence by processing individually several monophonic sequences. A large amount of reliable ground truth can thus be provided thanks to this automatic generation process. This is a crucial step in the investigation of automatic transcription of orally transmitted music repertoires, including the *marovany* zither one. Figure 4.1 illustrates the relation of this chapter with the others.

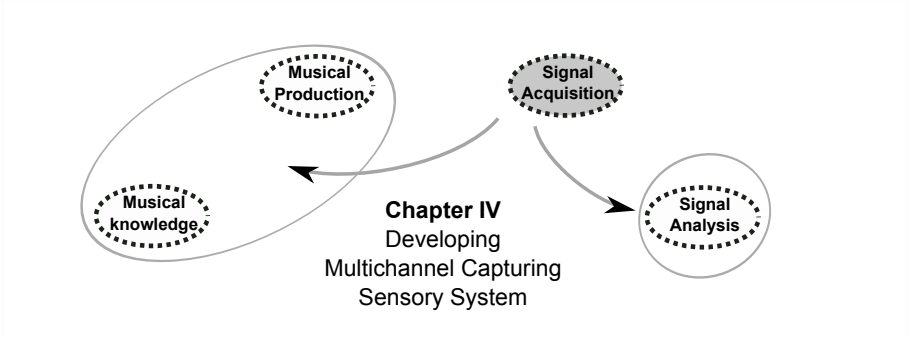


Figure 4.1 – Schematic diagram of the PhD organization for chapter 4.

4.1 Introduction

4.1.1 Background on MCSSs

In this chapter, we explore the use of Multichannel Capturing Sensory Systems (MCSSs) (also named "divided pickup" in specialized literature) for AMT of acoustic plucked string instruments. MCSSs are made of several pickup devices, one for each string. A pickup device is a transducer that captures mechanical vibrations from stringed instruments (e.g. the electric guitar or electric violin), and converts them to an electrical signal that is amplified and recorded. A MCSS is then capable of picking up signals from each individual string and outputting these signals individually. This information allows to detect accurate performance information for each individual string, and supports analysis of an instrument performance that would be challenging to extract and decouple from the audio signal. Efforts made in the development of capturing sensory retrieval systems aim to select the signal type which provides the best support to extract the basic note parameters necessary for music transcription. Additionally, since the output signal is analog -the same as traditional pickups -it faithfully conveys all guitar-playing techniques. For the AMT task, such a system has an obvious advantage in this application, as it allows to break down a polyphonic musical signal into the sum of monophonic signals respective to each string. Then, we come back to a monophonic transcription problem, which can be understood as a special simple case of polyphonic transcription, and is considered as practically solved (Klapuri, 2004b).

Among the pioneer works towards the development of MCSSs, we can mention the Gittler guitar, an experimental guitar with six pickups, one for each string. Some time after, Gibson created the HD-6X Pro guitar with The Hex Pickup that captures a separate signal for each individual string and sends it to the onboard analog/digital converter. The output can be routed as a single summed mono signal to an amplifier or recording console. It can also send the E, A, and D strings to one amp or recording channel and the G, B, and high E to a separate amp or channel. Or it can send the output of all six individual strings to six different amps or channels. These six individualized outputs can used to create various effects.

First projects developing MCSSs dedicated to AMT for acoustic instruments go back some time. In the 1980's, Trimpin designed a system to capture which fingers were pressing which key on a grand piano. Currently one of the most robust systems to capture this information is commercially available. It is the Piano Bar, designed by Don Buchla in 2002 and now sold by Moog Music. It captures the full range of expressive piano performance by using a scanner bar that lies above any 88-key piano, gathering note velocity as well as a pedal sensor which gathers a performer's foot movement. Researchers at Osaka University in Japan designed a system for real-time fingering detection using a camera-based image detection technique, by coloring the finger nails of the performer (Takegawa et al., 2006). In O'Grady and Rickard (2009), they propose an alternative paradigm to electric guitar transcription, where the signal generated by each string at the guitar pickup is captured separately using a hexaphonic guitar ; thus changing a polyphonic transcription problem into a monophonic one, which ameliorates subsequent analysis and allows finger position identification for tablature generation.

4.1.2 Ground truth for AMT evaluation

First, a caveat: any evaluation system inherently assumes some idea of ground truth against which the candidate is evaluated. Then, when interesting to the evaluation of AMT

performances, the question is: what to compare the output of an algorithm with ? It is common practice in MIR fields to collect reference set or ground-truth data by inquiring musicological experts or by crowd-sourcing. These data are considered the intended output of the computational model and the model is evaluated in terms of its ability to reproduce the ground-truth.

In the field of Automatic Music Transcription, annotated sound databases are then needed both to develop and to evaluate algorithms. However, such ground truth databases are rather scarce, mostly due to the fact that manual transcription is a very cumbersome and time-consuming task. Private databases like those used in contests like MIREX exists (MIREX, 2011), but are not made available to the entire community of researchers, hence the development of public databases. For example, the RWC music database (Goto et al., 2003) is a widely-used database in this field, which is composed of real-world music excerpts with several levels of polyphony and different instrumental sounds covering different music styles. But actually only a very few exist, as a number of issues are commonly encountered : little amount of sounds, either to copyright and distribution problems, or recording conditions, the ground truth is often generated *a posteriori*, with some inaccurate or erroneous values of pitch or onset and offset times, and a time-consuming process. Also, Hainsworth (2004) within MIR literature figured out that manual transcription strategies can be quite different resulting various degrees of divergence from the original performance. Similarly, the study of Cemgil (2004) shows that there is no unique ground truth for manual transcription even among well-trained musicians.

The question of ground truth datasets was then at the core of this PhD thesis, which raises many methodological questions when one is interested to original instrument repertoires.

4.1.3 MCSSs for ground truth generation

One of the main reasons explaining why the investigation of non-eurogenetic music has fallen behind the one dedicated to eurogenetic music is because of the current lack of ground truth sound datasets. This is related to the absence of written supports often associated to orally transmitted repertoires, and so an absence of extensive inventories of these repertoires. Technologically, also, traditional instruments are mostly unplugged, i.e. wholly acoustic, which do not make easier the process of recording and collecting musical pieces of these repertoires, whereas instruments such as piano and guitar have sensory systems to allow for efficient extensive census of their repertoires.

The use of MCSSs has direct application to the automatic generation of ground truth for AMT, by capturing *in-situ* the musical production of a musician through the resonating parts of its instrument. The classical piano is the most studied music instrument in MIR, which is to be related to the fact that technology similar to MCSSs, such as the Disklavier (used in the MAPS dataset (Emiya et al., 2010)), is already widely used for this instrument. Research on MCSSs should help transposing such technology to different music instruments.

The main advantage of MCSSs is that it captures original precise musical interpretations of musicians, and is not fixed by scores. Indeed, the problem is whether the original notation or the manual transcription can exactly match with performance due to personal interpretations of both performer and transcriber. However this point especially becomes a problem when automatic transcription is defined as obtaining original notation from performance as a kind of reverse-engineering (Klapuri, 2004a).

4.1.4 Other Applications of MCSSs

Musical analysis

Retrieval systems with multichannel outputs could help in performing string-based musical analysis, such as isolating the specific rhythmical characteristic within groups of strings, and better understanding the playing methods. In our database, low-level scores [1] are now available in machine-readable formats, it is possible to perform directly all kinds of musical analysis.

Human-machine interaction

Human-machine interaction is part of research, called Augmented Instruments or Multimodal Environment, aiming to obtain accurate perception about computer system which can analyze, perform and compose human action (Rowe, 2004). To do so, multimodal sensory retrieval systems are often used. The idea of extending traditional acoustic instruments with sensors to capture performance information has been explored in Machover and Chung (1989). Other systems response which have influenced the community in this domain are Dannenberg's score following system (Dannenberg, 1984), George Lewis's Voyager (Lewis, 2000), and Pachet's Continuator (Pachet, 2002). There are few systems that have closed the loop to create a real live human/robotic performance system: Mari Kimura's recital with the LEMUR GuitarBot (Singer et al., 2003), Gil Weinberg's robotic drummer Haile (Weinberg et al., 2005, 2006), Benning et al. (2007)'s interactive robot for Indian music. The ESitar developed by Kapur (2002) is an Indian sitar retrofitted with a variety of sensors for capturing gestures of the performer while still producing sound acoustically and being playable as a traditional sitar. An exponentially distributed set of resistors is used in order to detect what fret is played by the performer. A sensor is placed under the right hand thumb and is used to deduce the direction and patterns of plucking. A force sensing resistor captures the applied force, which varies based on the stroke direction.

MIDIfication of acoustic instruments

The process of MIDIfication involves to equip an acoustic instrument with an array of electronic features allowing MIDI connectivity that supports communication with computing devices and external MIDI instruments. Famous models of MIDIfication of acoustic instruments already exist. The typical Disklavier is a real acoustic piano outfitted with electronic sensors for recording and electromechanical solenoids for playback. Sensors record the movements of the keys, hammers, and pedals during a performance, and the system saves the performance data as a Standard MIDI File. On playback, the solenoids move the keys and pedals and thus reproduce the original performance.

4.2 Sensoring systems: models, set-ups, quality check

This section aims to study comparatively different sensory systems in order to choose the right sensor for the task of AMT. Indeed, as this task calls for very specific signal processing operations, qualifying any of these operations of 'optimal' is strictly tied to the given sensory system, and the performance of an optimal processing with a given sensory

1. Similar to physical pianoroll, where notes are represented in terms of their physical duration and onset location, but using discrete MIDI-scaled pitches and amplitudes.

system may be less than the performance of a non-optimal processing with another sensory system. We then perform in the following a quantitative study to evaluate the appropriateness of several sensory systems to the task of AMT of plucked-string instruments. We start by listing the three different sensor types studied, and then by characterizing them through a set of specific numerical descriptors corresponding to required criteria on signals for AMT. Specific requirements must be respected when installing MCSSs on acoustic instruments, and exploiting their output signals for AMT. Our choice of sensor types has been made by attempting at best to comply with all the constraints listed above.

4.2.1 Sensor type I: optical

Background

Optical pickup guitars were first shown at the 1969 NAMM in Chicago, by Ron Hoag. In 2000, Christopher Willcox, founder of LightWave Systems [2], unveiled a new beta technology for an optical pickup system using infrared light. In May 2001, LightWave Systems released their second generation pickup, dubbed the "S2". The S2 featured LightWave Systems' monolithic bridge, six-channel motherboard, and a host of other improvements, making the technology more practical for use in both live and recording studio settings. LightWave Systems began producing their own guitars in the late 2000s. Currently the company features the Saber bass and the Atlantis ElectroAcoustic guitar. These models are the only guitars that come with the LightWave Systems optical pickup installed.

Optical-based systems have already found various applications, such as metrological measures of string displacement (Seydoux, 2012; Chabassier, 2012) or a MIDIfication [3] of a piano through the Moog piano-bar technology (Mowat, 2005; Assayag and Bloch, May 2008).

Physical principle

The selected optical sensor are slotted optical switches consisting of an infrared emitting LED and an NPN silicon phototransistor, that work by sensing the interruption of a light beam by a vibrating string. It has a fork design, with the string placed between the two branches as illustrated in the close-up of the figure 4.2. On one side, the light-emitting diode (LED) emits a light beam. On the other side, the phototransistor has a peak of sensitivity at 850 nm. The emitter casts a shadow of the string onto the photodetectors. As the string vibrates, the size and shape of the shadow changes accordingly and modulates a current which passes through the photodetectors. This current is the analog electrical signal which represents an accurate depiction of the vibrating string. In order to maximize the dynamic of the optical signals and obtain sharp transient attacks, the narrowest possible diameter for the laser is used. Such sensor then acts as a digital switch with a robust sensitivity to string displacements.

Sensor models and set-up

We mainly tested the two following models of optical sensors:

— Photomicrosensor (Transmissive) EE-SX398/498 at OMRON (technical sheets at) ;

— Slotted Optical Switch OPB610 (technical sheets at http://optekinc.com/datasheets/opb610.pdf) ;

2. See http://lightwave-systems.com/

3. Acronym meaning the in-situ conversion of an acoustic instrument into its homologous MIDI.

These two sensors are represented in figure 4.2. The optical sensor OPB610 presents the most advantageous features for our application: reduced size, narrow beam light with a diameter of 0.5mm and a wavelength of 940 nm, reduced sensitivity to external lights. The power pack of the optical sensors needs a continue tension of 5 V. An enhanced low current roll-off is used to improve contrast ratio and immunity to background irradiance. The power pack of the sensors is thermally isolated, which makes it well aligned with field conditions.

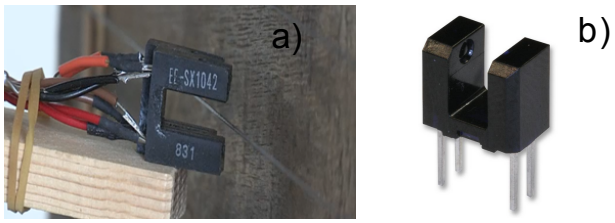


Figure 4.2 – Photos of the two models of optical sensors tested, with on the left the EE-SX398/498 model, and on the right the OPB610 model.

Figure 4.3 shows photos of the set-up installation, where all optical sensors are fixed on a vertical bar along the strings, and also this MCSS in condition of playing.

Figure 4.3 – Photos of the set-up installation, where all optical sensors are fixed on a vertical bar along the strings, and also this MCSS in condition of playing

4.2.2 Sensor type II: piezoelectric

Background

Piezoelectric belong to the family of contact pickup. Many semi-acoustic and acoustic guitars, and some electric guitars and basses, as well as other string instruments such as violins, have been fitted with piezoelectric pickups instead of, or in addition to, magnetic pickups. Solid bodied guitars with only a piezo pickup are known as silent guitars, which are usually used for practicing by acoustic guitarists. Piezo pickups can also be built into electric guitar bridges for conversion of existing instruments. Contact microphones are usually used to amplify acoustic instruments for live performance, or to record sounds.

Physical principle

The piezoelectric element is composed of a piezoelectric ceramic and a metal plate held together with adhesive. Both sides of the piezoelectric ceramic plate contain an electrode for electrical conduction. Piezo materials exhibit a specific phenomenon known as the

piezoelectric effect and the reverse piezoelectric effect. Exposure to mechanical strain will cause the material to develop an electric field, and vice versa.

Piezo-elements are made from two conductors separated by a layer of piezo crystals. Quartz crystals are used in most piezoelectric sensors to ensure stable, repeatable operation. The quartz crystals are usually preloaded in the housings to ensure good linearity. When the crystal is stressed with an external pressure, bending the metal conductor layers, a voltage is generated across the crystal layer. Piezoelectric pickups have a very high output impedance, appearing as a capacitance in series with a voltage source, and routed through a special low-noise cable to an impedance-converting amplifier.

When an alternating voltage is applied to the piezoceramic element, the element extends and shrinks diametrically. This characteristic of piezoelectric material is utilized to make the ceramic plate vibrate rapidly to generate voltage.

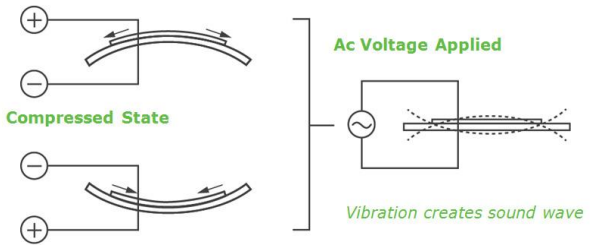


Figure 4.4 – Scheme of the functioning principle of piezolectric sensors (from http://www.cui.com/product-spotlight/piezo-and-magnetic-buzzers/).

Sensor models

In our study, we examined two different types of piezoelectric sensors:

— the acoustic gold RMC pickup sold by Roland http://www.rmcpickup.com/acousticgold.html;

— the Variax piezoelement sold by Line 6 http://fr.line6.com/store/item/267/

Both are sold by their respective providers as individual components. These two sensor models are represented in figure 4.5. These piezo pickups have been mounted under the central easels of the *marovany*, forming part of the bridge assembly itself, which capture the vibrations imparted on the bridge by the strings.

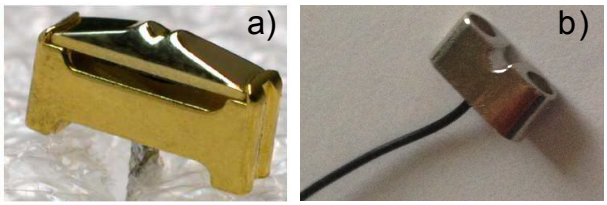


Figure 4.5 – Photos of the two models of piezolectric sensors tested, with on the left the Gold RMC pick-up, and on the right the Variax piezoelement.

4.2.3 Sensor type III: electromagnetic

Background

Electromagnetic sensors are commonly used as pick-up in electric guitars.

Physical principle

Magnetic pick-ups, as applied in electric guitars, register the vibrations of nickel or steel strings in a magnetic field. A magnetic pickup consists of a permanent magnet with a core of material such as alnico or ceramic, wrapped with a coil of several thousand turns of fine enameled copper wire. The permanent magnet creates a magnetic field; the motion of the nearby soft-magnetic vibrating steel strings modulates the magnetic flux linking the coil, which induces a voltage in the coil. This signal is then carried to amplification or recording equipment via a cable. More generally, the pickup operation can be described using the concept of a magnetic circuit. In this description, the motion of the string varies the magnetic reluctance in the circuit created by the permanent magnet.

Some high-output pickups achieve this by employing very strong magnets, thus creating more flux and thereby more output. This can be detrimental to the final sound because the magnet's pull on the strings can cause problems with intonation as well as damp the strings and reduce sustain. The turns of wire in proximity to each other have an equivalent self-capacitance that, when added to any cable capacitance present, resonates with the inductance of the winding. This resonance can accentuate certain frequencies, giving the pickup a characteristic tonal quality. The more turns of wire in the winding, the higher the output voltage but the lower this resonance frequency. The inductive source impedance inherent in this type of transducer makes it less linear than other forms of pickups, such as piezo-electric or optical.

Sensor models

— The GK-3 by Roland http://roland.com/V-Guitar/about.html, designed a strip of electromagnetic sensors dedicated to acoustic guitars ;
— Handcrafted sensor based on magnetic scanning heads of mini-K7

The electromagnetic sensors we used were handcrafted sensors built with magnetic scanning heads of mini-K7. The commercialized GK-3 Roland sensor was used to calibrate our own electromagnetic sensors, illustrated on the right graph of figure 4.6. These two sensors are represented in figure 4.6. The commercialized GK-3 Roland sensor was used to calibrated our own electromagnetic sensors.

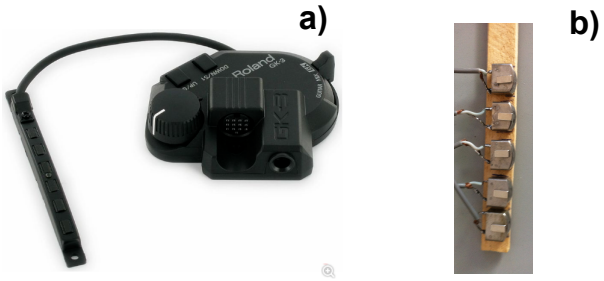


Figure 4.6 – Photos of the two models of electromagnetic sensors tested, with on the left the GK3 model, and on the right our handcrafted sensors.

Set-up installation

The pickup is most often mounted on the body of the instrument, but can be attached to the bridge, neck or pickguard, as on many electro-acoustic archtop jazz guitars and string basses. In our project, we used a portable bridge for our pickup sensors, and attached it to the body of the instrument. This MCSS was especially used for the *mvet* zither.

4.2.4 Quality check on Sensor Installation

Non-intrusiveness

The system must not be too cumbersome and disturbs the playability of the instrument. Specific playing techniques such as palm muting and excitation point displacements require a playing zone which must not be disturbed by the devices. The ideal sensory system would also not require a specific instrument-making.

Sensor placement in regards to string vibrations

The measuring point of string displacement may create a bias in the amplitude measure. Indeed, as this displacement consists of a superposition of vibratory modes, defined as a succession of nodes and anti-nodes, if a sensor is placed on a modal node the energy contribution of the corresponding mode is null. To answer these two constraints, the bar of sensors is positioned near the easel, in such a way that the playing zone is less disturbed and that the sensors are roughly placed on the ascending slope of the anti-node following directly the easel-related node common to all modes. When installing the sensors on instruments, one has to take care of avoiding placements on "nodal modes".

4.2.5 Quality check on Sensor Signal

We now quantify the quality of each sensor signal in view of AMT on the *marovany* instrument. Although pickup devices are most often evaluated for sensor transparency in the restitution of sound quality, for the task of AMT we rather focus on the ability of sensors to efficiently retrieve the four basic musical parameters in AMT: onset location, pitch, duration and amplitude. This quality check is performed with classical acoustic descriptors (detailed in Annex A).

A mechanical arm of calibration

A mechanical arm, illustrated in figure 4.7, is used to set precisely a reference strength to the excitation source. Three different heights h of the arm were used to define three different reference excitations (arbitrarily assigned to the musical nuances of pp, mf and ff), while keeping the height hs fixed and equal to the height of the string excited. The distance d is also kept fixed for the different excitations. Also, using this mechanical arm, we were able to calibrate the output level of each sensor, relatively to the reference excitation levels provided by the arm. Electronic gains respective to each sensor can then be deduced and apply in a post-processing stage to the signals.

Evaluation protocol

Eight different signal criteria C_c, with $c \in \{1, \cdots, 8\}$, will be evaluated on each sensor signal $x_s(t)$, with $s \in \{1, 2, 3\}$. Each signal criteria C_c employs a characteristic function

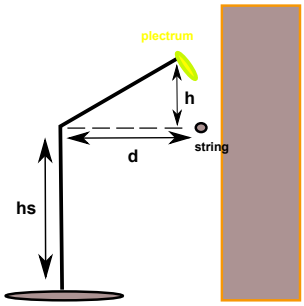


Figure 4.7 – Mechanical device used to provide reference excitation amplitudes.

χ_c emphasizing a specific feature of the signal. All measures are based on the same set of test notes I_{test}:

1. Three different amplitude levels (pp, mf and ff), provided by the mechanical arm ;
2. Six different strings with the following pitches: D_3#-A_3-C_4-E_4-B_4-D_5 ;

The three MCSSs have been installed on the *marovany* zither N_1, and their respectively signals are acquired simultaneously. The size of this set of notes is then $N_{I_{test}} = 3 \times 6 = 18$. Furthermore, microphone signals, on which audio processing is usually applied, are here used as reference to characterize physical properties of our sensor signals. Unless mentioned otherwise, we then define the normalized measure $\lfloor C_c \rfloor$ as follows

$$\lfloor C_c \rfloor = \underset{i \in I_{test}}{\text{median}}(\chi_c(x_s^i(t)) - \chi_c(x_0^i(t))) \tag{4.1}$$

The sign resulting from this operation does not have any physical meaning, but just indicates whether the value is higher or smaller than the respective$\chi_c(x_0(t))$ value.

Salience measures

We first compute the onset detection function F_{odf} described in Ellis (2007), followed by a peak finding function which identifies the location of each onset. Our first signal quality criteria C_1 is the degree of salience of the detected peaks. We simply compute the signal-to-noise-ratio associated to each test note p, defined as

$$\chi_1 = \frac{max(h(t))}{h(T_n)} \tag{4.2}$$

where $h(t) = F_{odf}(x(t))$, and T_n is defined as the first 10 windows. The sharpness of the peak (C_2), which informs on the accuracy of time location of the onset, is also computed. This measure of sharpness is made through the following dispersion measure

$$\chi_2 = arg_t(h(t) = \frac{max(h(t))}{4}) \tag{4.3}$$

These measures are illustrated in figure 4.8.

Spectral measures

Our third numerical criteria C_3 characterizes the harmonic content of the signal, and is computed using eq. 4.1, with χ_3 defined by the Harmonic Energy descriptor, i.e. HarmErg in Peeters et al. (2011) (detailed in Annex A). This descriptor ranges from 0 to 1. We

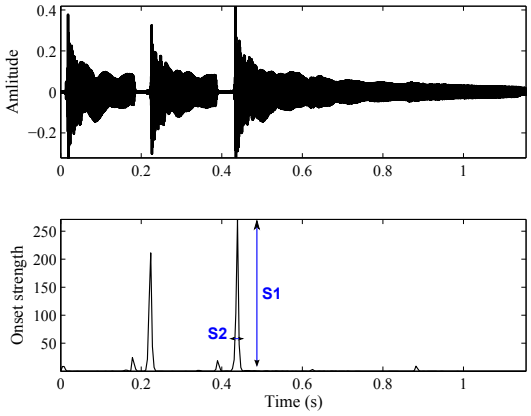


Figure 4.8 – Illustration of the salience measures $S1$ and $S2$, respectively the salience and the sharpness of the onset, through the onset detection function F_{odf}.

also compare the pitch estimations obtained with the YIN algorithm (de Cheveigné and Kawahara, 2002) on the different sensor signals, defining C_4, with χ_4 defined by the YIN pitch tracking function. We normalize the values of this descriptor.

Temporal measures

The signal criteria C_5 is on note duration, taking $\chi_5 \sim EffDur$, with EffDur the descriptor Effective Duration defined in Peeters et al. (2011). The Modulation Ratio of the temporal envelop (MOD) is also computed for criteria C_6, taking $\chi_6 \sim MOD$. This second descriptor ranges from 0 to 1.

Inter-note temporal interval measure

We must have a good separability of successive notes of a same string, with ideally inter-notes blanks resulting corresponding to the finger-string contacts. The descriptor C_7 measures the time interval between two successive notes. As illustrated in figure 4.9, we first computed the energy envelop in the location of the inter-note space, and then simply measured the time interval Ts between these notes at a same energy level, set at three times the noise level of the sensor signal.

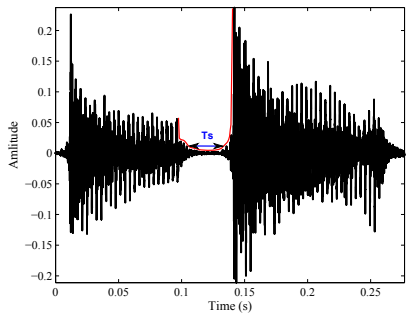


Figure 4.9 – Illustration of the inter-note temporal interval Ts measure.

Signal criteria	Measures	Unity	$MCSS_1$	$MCSS_2$	$MCSS_3$
Salience	$\lfloor C_1 \rfloor$	dB	- 5.1 ± 2.2	- 2.3 ± 2.9	- 8.7 ± 1.7
	$\lfloor C_2 \rfloor$	ms	- 23 ± 23	- 11.5 ± 11.5	-11.5 ± 23
Spectral	$\lfloor C_3 \rfloor$	X	- 0.23 ± 0.09	-0.11 ± 0.03	0.19 ± 0.05
	$\lfloor C_4 \rfloor$	Hz	23.4 ± 8.2	10.1 ± 4.5	55.9 ± 5.6
Temporal	$\lfloor C_5 \rfloor$	ms	- 320 ± 32	- 168 ± 21	- 563 ± 69
	$\lfloor C_6 \rfloor$	X	-0.1 ± 0.03	- 0.36 ± 0.07	- 0.26 ± 0.11
Segmentation	C_7	ms	19	16	6
Separability	C_8	X	2.7	3.4	1.1

Table 4.1 – Numerical values of the different signal criteria evaluated, for our three different MCSS, being respectively optical, piezoelectric and electromagnetic. Values of $\lfloor C_c \rfloor$ allows comparison between the different sensor signals, as well as with the microphone signal. Values of C_8 allow only relative comparisons between sensor signals. The sign X means dimensionless criteria.

Multichannel separability

This signal criteria C_8 is a measure of the degree of "repisse" in our multichannel output, using all pitches of our instrument pitch range. Each of these pitches is successively played, and we measured the residual RMS-energy level occurring at the same time in all other sensor signals. A single value is obtained by summing all elements of this matrix, providing a general estimation of sensor energy residual due to inter-sensor repisse.

Results and Discussion

Table 4.1 provides the numerical results obtained for each signal criteria with our different sensor types. We also propose in figure 4.10 a more concise representation of our results, allowing a direct quantitative comparison between our different MCSSs. We normalized the values of table 4.1 using simple mathematical transformations following the rules of quality for the task of AMT:

— $\lfloor C_1 \rfloor$ The higher the onset salience, the easier the identification of onset locations ;
— $\lfloor C_2 \rfloor$ The higher the onset sharpness, the easier the identification of onset locations ;
— $\lfloor C_3 \rfloor$ The richer the harmonic content of a sensor signal, the harder the pitch estimation is. In other words, sensors filtering out higher frequency components while emphasizing the string fundamental vibratory mode tends to greatly make the pitch estimation easier ;
— $\lfloor C_4 \rfloor$ The higher the accuracy of F_0 measure, the better the transcription ;
— $\lfloor C_5 \rfloor$ The higher the accuracy of note duration, the better the transcription ;
— $\lfloor C_6 \rfloor$ The smaller the temporal envelop modulation, the easier it is to segment note boundaries ;
— C_7 The larger the inter-note space, the easier it is to segment successive notes ;
— C_8 To perform monophonic transcriptions, the independence of each sensor signal is of first importance, each sensor having to detect solely the vibration associated to its string. Then, the higher the independence of each sensor signal, the easier the monophonic transcription is. Furthermore, although the multichannel "repisse" can result from sympathetic resonances between strings, it is still seen as a parasite signal from the point of view of AMT, and sensor signals which flatten out residual signals are consequently more advantageous.

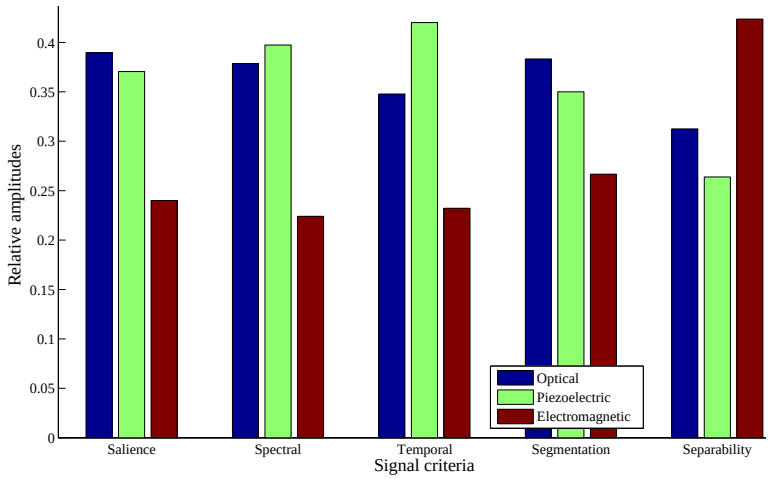


Figure 4.10 – Numerical scores on our signal criteria for the different sensor types.

Based on table 4.1 and figure 4.10, we can first observe that globally all sensor signals have sharper onsets than those in the microphone signal, which can be partly explained by the fact that a sharp signal pattern is obtained at the beginning of plucking during the period of finger-string contact. However salience values are slightly below microphone ones. Piezoelectric and optical sensors offer the highest salience and sharpness of onsets. The magnetic pickups appear to extraneous noise to the sound, while optical has virtually no inherent noise. Even when putting an amp at high volume, piezoelectric and optical sensors experience an increased dynamic range and sensitivity, i.e. loud notes are loud and clear, while soft notes and subtle nuances are not masked by background noise.

Second, our sensor signals tend to have a poorer harmonic energy than microphone signals, but a stronger harmonicity, which can be explained by the fact that a direct measure of string displacement privileges its own vibratory behavior, emphasizing the string fundamental frequency and minimizing the effects of coupling with the more complex modes of the soundboard. Furthermore, piezoelectric and electromagnetic sensors tend to exert more their own influence or color on the strings' vibratory properties.

Third, all sensors have a detrimental impact on temporal profiles of *marovany* sounds, and none are able to read string vibration for its full duration. Piezoelectric and electromagnetic sensors interfere with string vibration. Piezoelectrics as they are in direct contact with strings, and electromagnetic sensors with the magnetic influence which "pulls" on the metal strings. As a non-contact sensor, optics do not interact with the movement of the string, and even if they restore the most faithfully the string vibration temporal waveform, they also distort it due to their intrinsic response.

Fourth, one of the most conspicuous advantages of sensors over the microphone is the formation of a space between successive notes, both in time and amplitude, resulting from the mechanical contact between string and finger. Piezoelectric sensors appear to provide the largest inter-note space.

Fifth, we observed that globally most leakage signal are negligible in our sensor signals, expect for certain couples of pitches resulting from the sympathetic resonances of the instrument. Electromagnetic sensors offer the best channel separability, while optical and piezoelectric sensors appear to be a little more affected by the repisse between each string.

4.3 Methods for Automatic transcription of Monophonic sensor signals

We will restrict ourselves to retrieve the musical information necessary for low-level transcription, i.e. measuring the activity and basic parameters (onset location, amplitude, pitch, duration) of played notes. Figure 4.11 shows the block diagram of successive processing steps in the MCSS-based AMT. We tested two basic transcription algorithms: a feature based algorithm, which can be implemented in real-time applications, and a spectrogram factorization based algorithm, based on a PLCA-learning of a template dictionary for both harmonic and noise signals.

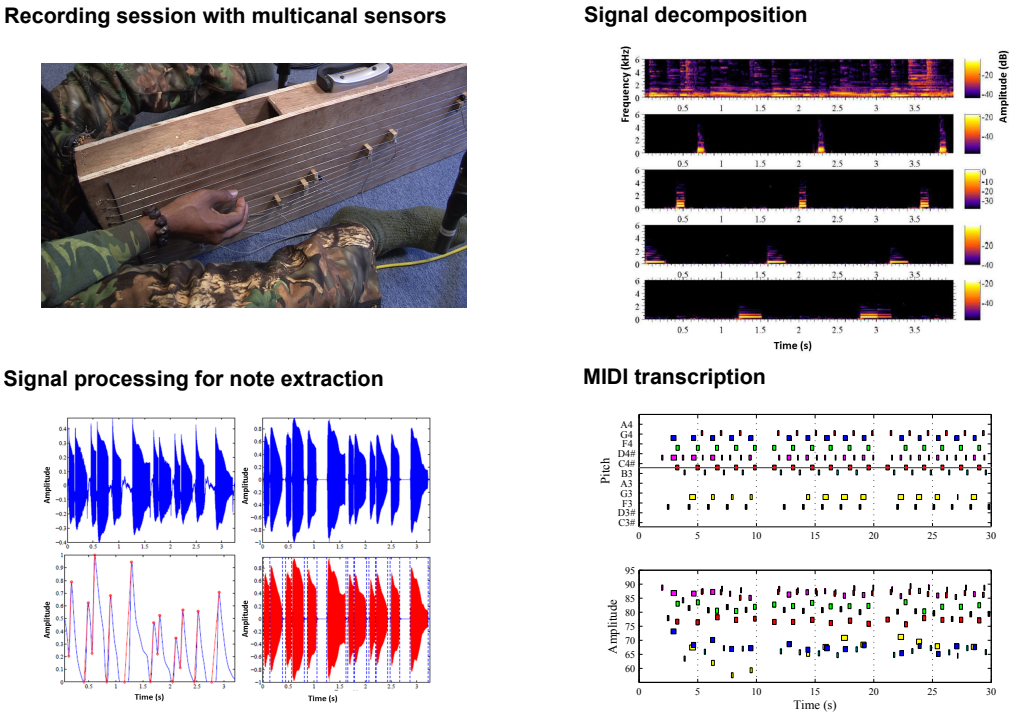


Figure 4.11 – Block diagram.

4.3.1 Feature-based

Left bottom graph in figure 4.11 represents the evolution of a waveform signal processed through this algorithm. From left top to right bottom: original optical signal, denoised signal, residual r with location of onsets, and segmented signals.

Blind adaptive denoiser

A blind denoising method (Ephraim and Malah, 1985) is applied to optimize the signal to noise ratio, mainly deteriorated by parasite noise coming indiscriminately from electronics and mutual resonances of strings. This algorithm of denoising takes as inputs segments of noises (defined as the first second of acquisition signals), and allows their subtraction to the signal by minimizing a prediction error with a least-mean square optimization.

Onset detector

Then, a 0.049-s hamming window with a 0.005-s overlapping (that is 11.6 ms, providing a temporal resolution whose order of magnitude is similar to the time attack) scans the entire sequence. Each onset of notes is detected using a spectral difference which takes into account the phase increment, as introduced by (Bello et al., 2004):

$$\hat{X}_{k,n} = |X_{k,n-1}|e^{j(2\phi_{k,n-1}-\phi_{k,n-2})} \tag{4.4}$$

with n the index of each window. As *marovany* sounds consist roughly of a superposition of short stationary sinusoids, the occurrence of an onset generates a peak in the prediction error defined by

$$r(n) = \sum_{n=1}^{N} |\hat{X}_{k,n} - X_{k,n}| \tag{4.5}$$

Windows for which this residual exceeds a fixed threshold are validated as onsets.

Note segmentation

From this detected onset, the descriptor E (eq. A.7) is computed for the neighbouring windows to search the local maximum $E_{max}(i)$ associated to pitch i, assuming this maximum is located near the onset, as expected for notes played by plucked string instruments. Then, E is computed on all the windows following the onset until the energetic value decreases below 5 % of $E_{max}(i)$, which may then be read as an adaptive note-specific energy threshold, or until another peak in the residual r is found. This estimation allows us to deduce note duration (eq. A.2), and its amplitude by averaging the energy over all windows within the note. We are not interested in the absolute amplitude of the notes, but only in their relative values within an air, in reference to a value determined by the MIDI gain.

4.3.2 Matrix factorization based method

The frequency marginals $P(f|z)$ of a PLCA model can be used as a model for certain kinds of sounds, which will now allow us to tackle the tasks of detection using separate class modeling. In a given MCSS sequence, two sound classes are assumed to be present, namely the background noise from the electronic device, and the instrumental notes. In the following, we then define two different template dictionaries, whose elements are defined by the frequency marginals distribution $P_i(f|z)$ and $P_n(f|z)$, with $z \in \boldsymbol{Z_i}$ and $\boldsymbol{Z_n}$, respectively the sets of latent variables for the note and noise classes. The following two-class PLCA for sound/noise detection will be used,

$$\begin{aligned} P(f,t) = P(t)(&\sum_{z\in Z_n} P(f|z)P(z|t) \\ &+ \sum_{z\in Z_i} P(f|z)P(z|t)) \end{aligned} \tag{4.6}$$

To impose a better discriminability between the spectral basis of our two latent families, i.e. enforcing a certain dissimilarity between them, we use the sparseness prior eq. 3.8 (Sec. 3.2.1). In particular, as we do not know exactly the optimal number of latent variables to best explain vocal sounds, this prior then allows to adapt our basis decomposition to the specific data multi-dimensionality.

A crucial step before performing a PLCA-based analysis is to properly initialize template dictionaries, which allows EM-based parameter estimation to be much more precise. We perform a first initialization of these templates using generic sounds from each class. Figure 4.12 provides some examples of these template dictionaries. The frequency marginals extracted from the instrument notes have a clear and spaced harmonics structure with a distinct pitch, while the marginals of noise display a more broadband and inharmonic spectrum. Once the frequency marginals are known for a certain sound in a mixture, they can be used to extract this kind of sound from the mixture in a supervised way Smaragdis et al. (2007).

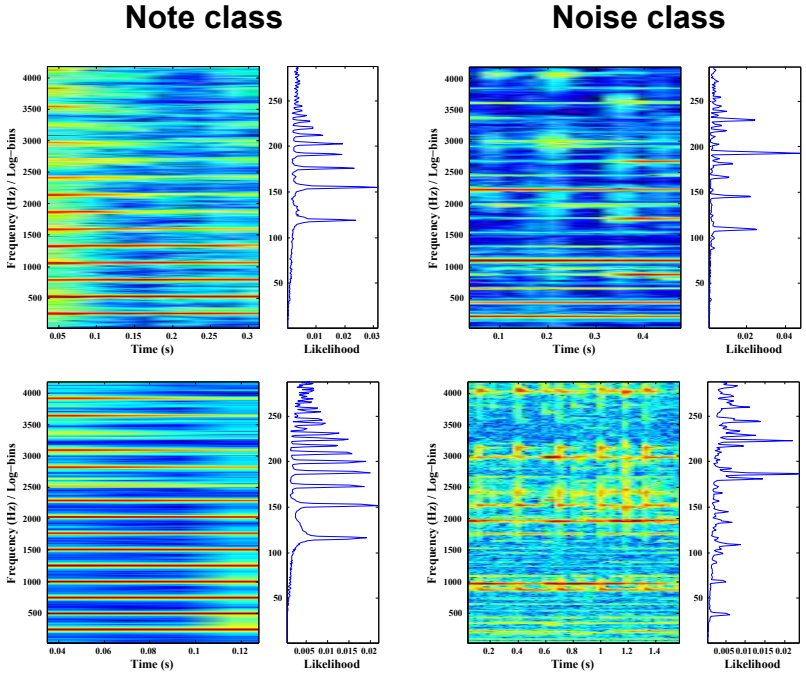


Figure 4.12 – Examples of spectral templates for the two PLCA latent classes of instrument notes and noise.

We then perform a data-based learning of template dictionaries using the conservative transcription method (see Sec. 2.2.8). After adapting our template dictionaries to input data, we extract the instrumental ntotes from the mixture by setting noise activations $P(z)$ for $z \in \boldsymbol{Z_n}$ to zero, and therefore, a denoised reconstructed PLCA model of sound spectra $P_i(f,t)$ can be obtained by:

$$P_i(f,t) \approx P(t) \sum_{z \in \boldsymbol{Z_i}} P(f|z)P(z|t) \tag{4.7}$$

To make this PLCA-based denoising process even more efficient, we also implemented in the preceding system the spectral re-weighting operation developed by Ye (2014), based on a frequency band-wise reconstruction error. It is grounded on the fact in human auditory that noise corrupted sound is recognized via processing the audio in local high SNR frequency bands Cooke (2006). It computes the following frequency band-wise reconstruction error possibility distribution

$$P(f|e) = \frac{\sum_t (X(f,t) - P_i(f,t))^2}{\sum_t \sum_f (X(f,t) - P_i(f,t))^2} \tag{4.8}$$

where X is input noise spectrogram and $\tilde{X}$ is the denoised reconstructed spectrogram by PLCA model. $P(f|e)$ presents spectral error possibility distribution.

$$\omega_f = 1 - \frac{P(f|e)}{\max(P(f|e))} \tag{4.9}$$

From this model error distribution ω_f, we can know which frequency bands of the the denoised reconstructed spectrogram by PLCA model present the strongest similarity with the input spectrogram. After this denoising process, we use a simple threshold-based detection of the note activations from the activity matrix $P(z,t)$, followed by a minimum duration pruning. The threshold for minimum duration for pruning was set to 200 ms.

4.3.3 Results and discussion

We created ground truth by semi-automatically hand-labeling several musical sequences captured with different MCSSs, extracting precise information on note onset location, duration, amplitude and pitch. To do so, we use a custom user-friendly interface, developed in Matlab and illustrated in figure 4.13, which implements an energy-based semi-automatic system optimized for the transcription of our MCSS signals.

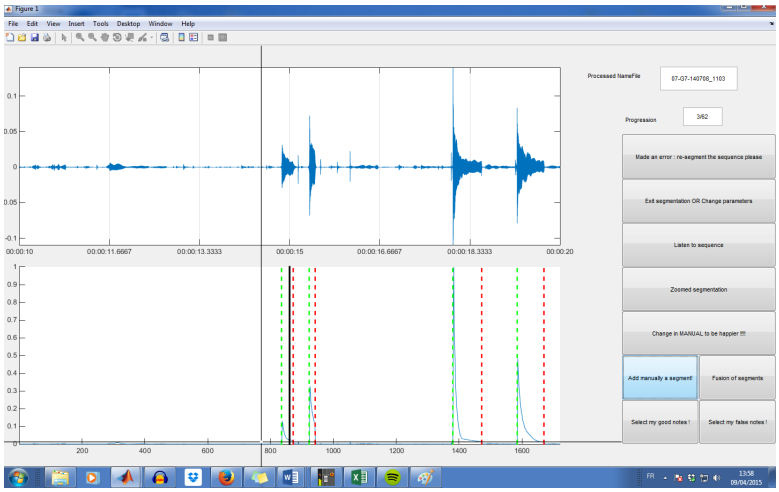


Figure 4.13 – Screen shot of our semi-automatic transcription system to extracted instrumental notes from our MCSS signals.

Figure 4.14 provides the ROC (Receiver Operating Characteristic, see Kay (1998) for details) curves with the error metrics FPR and TPR on our monophonic transcriptions of sensor signals, for our two feature-based and PLCA-based methods. Solid lines show transcription performance with an evaluation based only on pitch and onset location, and dashed lines show performance with a full evaluation based on the four parameters of a note, i.e. adding duration and amplitude. Our PLCA-based method provides the best performance for this transcription task, with accuracy (average F-measure) higher than 96 % over a large interval of thresholds (i.e. 0.11 -0.87). When considering all note parameters, transcription performance decreases.

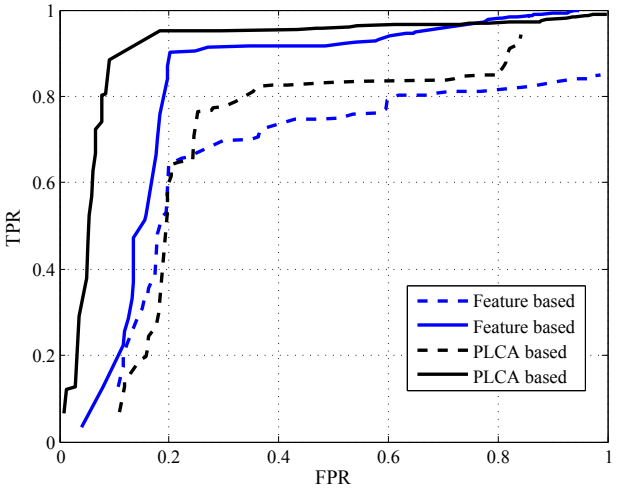


Figure 4.14 – ROC curves with the error metrics FPR and TPR on our monophonic transcriptions of sensor signals, for our two feature-based and PLCA-based methods. Solid lines show transcription performance with an evaluation based only on pitch and onset location, and dashed lines show performance with a full evaluation based on the four parameters of a note, i.e. adding duration and amplitude.

4.4 Development of the PSIFAMT (Plucked-String Instruments For Automatic Music Transcription) sound database

4.4.1 Context

Large and very diverse solo database are needed for training and evaluation of the AMT process. In the international MIR community, euro-genetic music, especially through the classical piano, has received the most attention from researchers, while non-eurogentic music has been rather left outside. Also, sound database for AMT may include a Musical Instrument Sample Database of Isolated Notes (MISDIN), which is a collection of sound samples of one or more musical instruments where each sample contains a recording of a single note played by one instrument. MISDINs are commonly used by electronic musical instruments, such as synthesizers and samplers, to reproduce sounds of other instruments. MISDINs are also utilized by the majority of music information retrieval (MIR) algorithms, including pitch estimation (Li and Wang, 2007), music representation (Leveau et al., 2008) and others, as evaluation data for experiments and for modeling sounds of different musical instruments. One major issue with creating a spectrogram factorization based transcription system for non-eurogenetic music is the lack of isolated note recordings for creating a training set. An isolated sounds database tends to considerably improve the performance (Benetos and Holzapfel, 2013).

4.4.2 Motivations

In this context, we have launched this research project in June 2013 of developing a Big Music Database of non-commercial recordings dedicated to plucked-string instruments. Our main long-term motivation is to provide original sound material for AMT evaluation, so as to bring the development of AMT systems closer to the actual world-wide music diversity and complexity. In response to this concern, our sound database first consists exclusively of the family of plucked-string instruments. These instruments are particularly

interesting as they present a great diversity in terms of timbre (e.g. inharmonicity, envelop spectrum variability, temporal profile modulation), playing styles and effects (e.g. different string excitation modes, palm muting, glissandi, vibrato, tremolo, harmonic notes). Second, our sound database wishes to include non-eurogenetic instrument repertoires, emphasizing even more the diversity related to timbre and playing techniques mentioned above. Indeed, acoustic measurements on isolated tone datasets of the *marovany* and the *mvet* reveal a more complex timbre (noticeably in terms of spectral envelop variability) than a standard instrument like the classical piano. Also, non-eurogentic musical repertoires encompass a great diversity in musicological codes (e.g. micro-rhythm, tuning based on non-equal temperament, modal scale), especially for orally transmitted musical repertoires, which also raise complex questions on transcription representation.

4.4.3 Technical specifications for recording

Place

Recordings mainly took place in a semi-anechoic chamber of the LAM, designed to perform recordings with absorbent panels on the walls (SNR $\approx$ 50 dBA). Recordings in Madagascar with insular musicians have also been performed, where we took care of finding rooms as quiet as possible. Furthermore, close proximity of the microphone to the instrument (around 40 cm) also favours to remove surrounding sound ambient. This homogenization of recording conditions allows for inter-recording comparisons which process only features related to the instrument repertoire, such as timbre, musicology and playing technique. It is noteworthy that nowadays, diversity in recording conditions of real-life audio recordings can be simulated subsequently, with computational tools such as the Degradation Toolbox (Mauch and Ewert, 2013).

Equipment

Figure 4.15 details the different elements of our recording equipment. It is composed of a sound card RME Fireface UFX, connected to RME pre-amps. Sensor signals are eventually stored synchronously onto the hard drive of a PC laptop with the music production/recording software Cubase LE. Each string's signal then starts out analog, and stays that way throughout up to the digital sound card.

All our recordings respect the following technical details:

— Schoeps stereo microphone, CMC 6 (capsule MK4), arranged following the ORTF setup (18 cm d'Ãľcart inter-microphones) ;
— Microphone is held at the instrument height, in front of it, with an average distance microphone/instrument soundboard of 50 cm ;
— The recordings are sampled at 44.1 kHz and stored as 24 bit WAV files ;
— To remove infrasonic disturbances the signals were highpass filtered with a fourth-order Butterworth filter with a cutoff frequency of 52 Hz ;

4.4.4 Recording types

Musical pieces

All the musical pieces recorded in this database are fully human-played. All the musicians are recognized professional, specialist of their instrument, and are asked to play representative pieces of their repertoire. Each piece can last from 2 to 6 minutes. The musician is also asked to define a pulsation track on which he will play all his musical

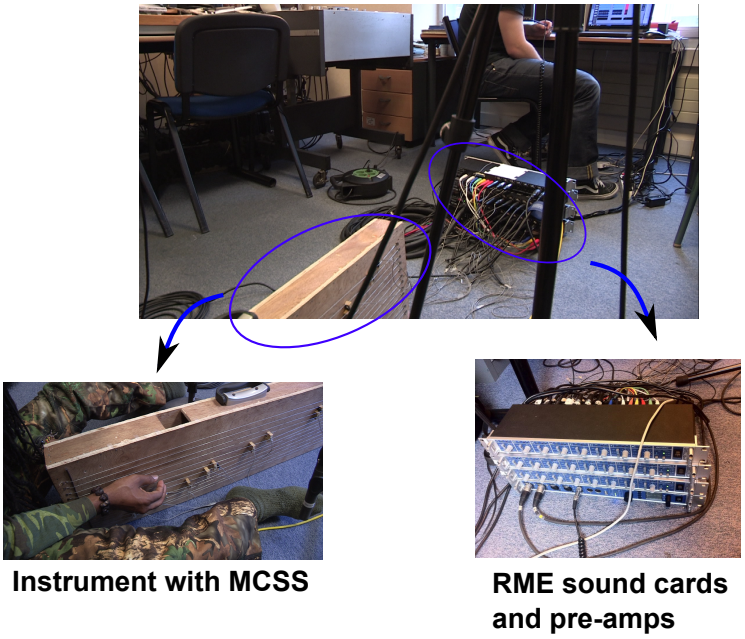


Figure 4.15 – Photos of the different elements of our recording set-up.

sequences. Certain musicians preferred to play their instruments without predefined pulsation, and reproduced it afterwards. We did not consider it as a bias in our experiment, as such musicians feel an interior pulsation. The tuning of each musical piece is identified trough precise F_0 measures on isolated note samples for each pitch.

Ground truth

Our ground truth has been built from transcriptions not based on examinations of musicological experts but rather on signal features, making it more objective. Our ground truth for each musical piece results from the use of Multichannel Capturing Sensory Systems, able to capture independent string-specific signals. When fulfilling the specific signal criteria previously exposed in Sec. 4.2.5, their transcription does not pose great difficulties, in comparison to the processing of the audio microphone signal. Although quite invasive as needing a complex experimental set-up during recording sessions, such a system allows for fast and very reliable polyphonic transcriptions.

Isolated tones

Each single instrument model recorded in our database has its own set of isolated notes, recorded in a same period (never exceeding 1 week). Pitch tones are tuned in the MIDI scale, so as to normalize our different datasets on the same pitch scale. 10 different occurrences of notes are recorded for each pitch, varying playing dynamic over three levels (pp, mf and ff), as well as more variable factors due to playing techniques such as distance to the easel, string pluck intensity, attack mode (nail or finger pulp) and finger inclination. The complete pitch range of the instrument is recorded. Also, we minded all spurious noises (e.g. finger scraping on the string) during recording. Also, each musician was asked to play all notes of the instrument once, from which we extracted the precise tuning ($\pm$ 20 cents) respective to each musical piece.

For each pitch of the instrument pitch range, we have 12 different tones per instrument

model, which are equally divided into two categories. A first category in which note samples are obtained by varying playing dynamic only, with a fixed position of the excitation point, centred in the playing area. A second category in which only excitation modes are varied, i.e. distance from the easel and plucking mode (e.g. fingertip or nail), with an almost-constant playing dynamic. All the isolated note samples have been recorded using a semi-tone scale, so as to guarantee generality regardless specific repertoires presenting micro-deviations in tunings.

4.4.5 Current state of database (ongoing database)

Details of the databases are given in table 4.2. All in all, this database currently holds X solo performances, by X different musicians. The overall size of the database is about X GB, i.e. about X hours of audio recordings. This sound database contains mainly recordings of the *maraovany*, as it was the research instrument of this PhD. For this instrument, it supplies a good generalization of the different sound possibilities of each instrument in various playing techniques and musical repertoires. It thus provides a good generalization of the sounds each instrument is producing in different recordings. We wish to make this sound database available under a Creative Commons license in a near feature.

Instrument type	Recording sessions	Musician name	Place / Date of recordings	Instrument model	MCSS Technology	Number of Pieces / Total duration (in h)	Total number of played notes
Marovany	D_1	Velonjoro	Madagascar / June 2012	N_3	Optical	4 / 0.29	11569
	D_2	Velonjoro	Madagascar / June 2013	N_1	Optical	5 / 0.4	16548
	D_3	Kilema	LAM (Paris) / May 2014	N_1	Piezoelectric	6 / 0.45	15496
	D_4	Charles Kely	LAM (Paris) / April 2014	N_1	Piezoelectric	6 / 0.43	18962
	D_5	Velonjoro	Madagascar / June 2014	N_1	Piezoelectric	8 / 0.77	31324
	D_6	Charles Kely	LAM (Paris) / May 2015	N_1	Piezoelectric	6 / 0.41	13265
	D_7	Kilema	LAM (Paris) / June 2015	N_1	Piezoelectric	6 / 0.4	12267
Mvet	D_8	Francois Essindi	LAM (Paris) / February 2014	Mvet1	Electromagnetic	4 / 0.44	6598
	D_9	Francois Essindi	LAM (Paris) / April 2014	Mvet1	Electromagnetic	5 / 0.48	7123
Folk Steel Gutiar	D_{10}	Adrien	LAM (Paris) / March 2014	Taylor 114-CE	Electromagnetic	9 / 0.72	17469
	D_{11}	Charles Kely	LAM (Paris) / September 2014	Taylor 114-CE	Electromagnetic	4 / 0.39	8756
	D_{12}	Damily	LAM (Paris) / May 2015	Taylor 114-CE	Electromagnetic	5 / 0.55	12009
N'Goni	D_{13}	Joseph	LAM (Paris) / February 2015	N'Goni1	Piezoelectric	2 / 0.16	2641

Table 4.2 – Catalogue of musical pieces in the PSIFAMT database.

4.4.6 Biographies of the *marovany* players

In the following, we provide short biographies of the three *marovany* players who participated to this PhD project.

Velonjoro Born in Ambovombe, Velonjoro has never left Madagascar, and has lived in Tuléar in isolation from western culture. One of his occupations is *marovany* player in trance rituals of *tromba*, making him an emblematic player of a vernacular and traditional repertoire of the *marovany*.

Kilema Born in Tuléar, Cément Randrianantoandro KILEMA is a malagasy profesionnal musician of marovany, but also of Kabosy and Katsa (native instruments), as well as a lead vocal singer. With the "Justin Vali Trio", a malagasy band, he performed in Japan, Australia, New Zealand, and Woodstock 1994 among others. In 1999, Kilema presented his first solo album *Ka Malisa*, an introspective into the music of southern Madagascar, his own birthplace. During the past three years, Kilema has been touring the main Ethnic and world Music Festivals in Spain. Through the recordings of four CDs, Kilema's continued investigation and diffusion of traditionnal Malagasy music.

Website: http://www.aido.fr/kilema.html

Charles Kely Born in Tananarive, the capital of Madagascar, Charles Kely is a professional player of guitar, *marovany*, *valiha*. He begun his career with his brother by covering traditional malagasy songs that they adapt to folk music. In 1997, Charles Kely is chosen by Rajery, the great *valiha* player, to tour around the world with his band. Together, they play in Paris, Chicago, Seattle, New Orleans and at the International festival of Louisiane. Since 2008, as a guitarist, Charles has also toured with prestigious world music artists as Tony Rabeson, Mounira Mitchala, Razia Said... He is now focusing on his solo career with the release of the CD Zoma Zoma in his own and unique manner : the open gasy style, an acoustic music from Madagascar, with a touch of bossa, jazz, blues, funk and subtle African influences.

Website: http://www.charleskely.com/

Figure 4.16 – Photos of the *marovany* musicians recorded in the PSIFAMT database.

4.5 Conclusion

In this chapter we propose to use the technology of Multichannel Capturing Sensory Systems (MCSSs) to automatically generate a large amount of reliable ground truth, in particular for the repertoires of the *marovany* zither. Among the most conspicuous technological advantages of these systems, we can mention the high signal to noise ratio, the multichannel output with independent signals corresponding to the played strings and the automatic demarcation between successive notes of a same string.

Chapter 5

Results and Discussions

Abstract

In this chapter, we present our main results obtained in this PhD project, and propose several discussions around the use of knowledge from musical acoustics in automatic transcription systems of music. Our different components of musical knowledge have been successively injected in our baseline transcription systems, consisting mainly of a Probabilistic Latent Component Analysis (PLCA) algorithm post-processed with a Hidden Markov Model (HMM), and their impacts on transcription results have been comparatively evaluated. Figure 5.1 illustrates the relation of this chapter with the others.

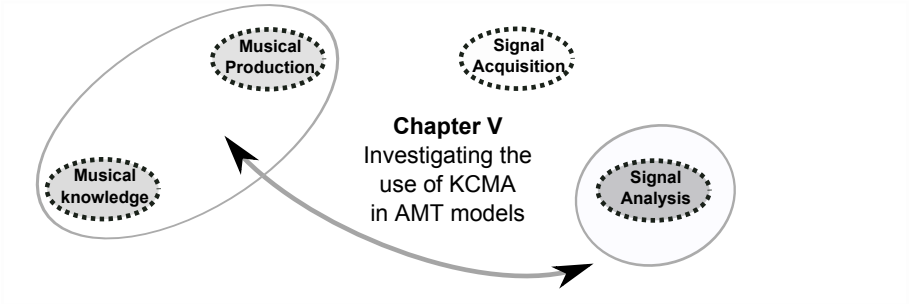


Figure 5.1 – Schematic diagram of the PhD organization for chapter 5.

5.1 Evaluation Methods

5.1.1 Evaluation procedure

Background

There are various evaluation metrics for AMT based on comparison of transcription and reference notation. Fonseca and Ferreira (2009) classify evaluation metrics as frame-based and note-based approaches. Frame-based approach is based on the comparison of two notations for every analysis temporal windows. The note-oriented approach is based on numerical rules identifying transcribed notes as corrects, depending on whether their onsets, pitches and/or durations are within a certain neighbourhood of the respective reference data. Another note-oriented evaluation metric is edit distance (ED), where the transcription is compared with reference notation on the basis of number of correct, inserted and deleted notes (Unal et al., 2008).

Note-based evaluation

For assessing the performance of our proposed transcription system, we perform a note-level evaluation stating that a note event is assumed to be correct if it fills four conditions on:

1. **Onset location**. Its onset must be within a Δ_o (in ms) range of a ground-truth onset. A value of 50 ms is commonly used for Δ_o (Bello et al., 2005; MIREX, 2011). Such tolerance level is considered to be "a fair margin for an accurate transcription" according to Seeger (1958), although it is far more tolerant than human ears would, as we remind that those are able to distinguish between two onsets as close as 10 ms apart (sounds arriving to the ear with lower time intervals are perceptually merged) (Moore, 1997). A stricter tolerance error of 40 ms has been used by Collins (2005) for percussive sounds. Our error tolerance was set based on figure 5.2, which provides the distribution of inter-onset intervals in the studied corpus. We put our tolerance error far below the smallest value of these intervals in order to be sure that our transcription systems are able to distinguish between all successive notes of our corpus. A value of 40 ms was retained for Δ_o ;
2. **Pitch**. Its pitch needs to be within a $\pm\ \Delta_i$ (in cents) tolerance around the ground-truth pitch. A value of 20 cents is commonly used for Δ_i in eurogenetic music (Fonseca and Ferreira, 2009). This error resolution is equal to one fifth of a semitone ;
3. **Duration**. Its duration needs to be within a $\pm\ \Delta_d$ (in s) tolerance around the ground-truth duration. In most AMT studies, duration is not evaluated, stating that especially for "decay instruments", the offset time is only important in a limited period of time, because after some point the sound energy will be below the threshold of hearing (Zwicker and Fastl, 1999), even if from the musician point of view the note is still playing. But Δ_d can be set relatively to 20% of the ground truth note duration (MIREX, 2011). Actually, the definition of a note offset has been a long-time ill-posed problem, and its effect on the success rate has actually never been clear in past studies. Much more than any other musical parameters, the offset suffers from the arbitrary of the ground truth, that is why we set a relatively higher error tolerance on this parameter ;
4. **Amplitude**. Its amplitude needs to be within a $\pm\ \Delta_a$ (normalized dimensionless value) tolerance around the ground-truth normalized amplitude. All amplitudes

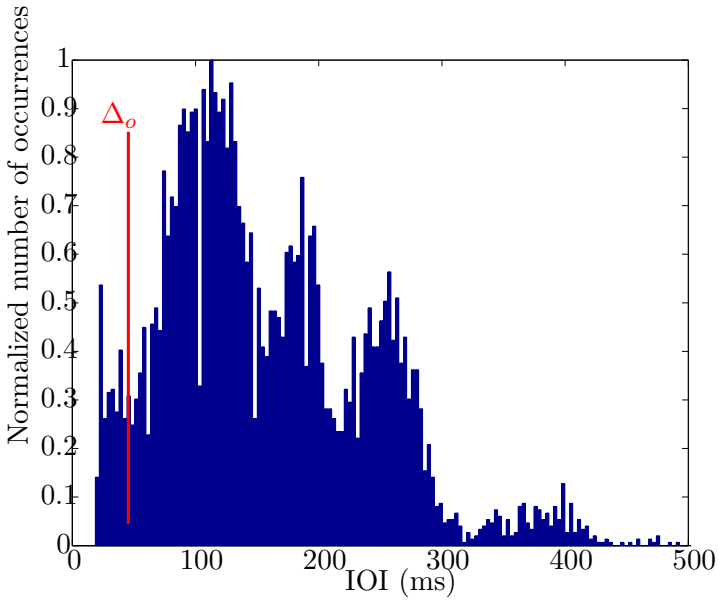


Figure 5.2 – Distribution of Inter-Onset intervals on the complete repertoire of the *marovany.*, including in red the tolerance threshold Δ_{Onset} for onset evaluation. This distribution present two broad beaks around 120 ms and 250 ms.

to be evaluated are first normalized by the maximum value, and only relative amplitude rapports to this value are then evaluated. Only a few AMT studies have tackled the problem of estimating note amplitudes, among which we can mention Marolt (2004), who computed this note amplitude based on the energy of its first harmonic, and Ewert and Muller (2011), who used a parametric model of the spectrogram to estimate this amplitude.

Table 5.1 details our Δ values for our different note parameters. As mentioned above, our thresholds of error tolerances on note duration and amplitude are pretty high, both because these parameters have never been properly defined in literature, and also because we know from section 4.3.3, that our ground truth transcriptions do not have a sufficient precision on their estimation (i.e. TPR and $FPR \leq 80$ % from the ROC curves).

Note parameters	Δ values
Onset location	$\Delta_o = 25$ ms
Pitch	$\Delta_i = 20$ cents
Duration	$\Delta_d = 40\% \ . \ R_{GroundTruth}$
Velocity	$\Delta_a = 40\% \ . \ A_{GroundTruth}$

Table 5.1 – Tolerance thresholds for the evaluation of each note parameter. We defined $R_{GroundTruth}$ and $A_{GroundTruth}$ the note duration and amplitude from the ground truth.

Furthermore, we used a binary approach to transcription (i.e. the original note is perfectly transcribed or it is not transcribed at all) with a one-to-one mapping when comparing transcription results, i.e., if a match is detected between an original note and a transcribed one, none of those can be used again on other note matches (Ryynanen and Klapuri, 2005; MIREX, 2011). The main idea is to avoid situations on which 2 different original notes could be matched to a single transcribed one (for instance, two quick notes

been transcribed as one), or vice-versa, allowing a perfect score even in situations that have different number of notes.

Evaluation metrics

Evaluation metrics are defined by equations 5.1-5.4 MIREX (2011), resulting in the note-based recall (TPR), precision (PPV), fall-out (FPR) and the harmonic mean of precision and recall (F-measure):

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$PPV = \frac{TP}{TP + FP} \tag{5.2}$$

$$FPR = \frac{FP}{FP + TN} \tag{5.3}$$

$$F - measure = \frac{2.PPV.TPR}{PPV + TPR} \tag{5.4}$$

where TP, FP and FN scores stand for the well-known True Positive, False Positive and False Negative detections. The recall is the ratio between the number of relevant and original items; the precision is the ratio between the number of relevant and detected items; and the F-measure is the harmonic mean between precision and recall. For all these evaluation metrics, a value of 1 represents a perfect match between the estimated transcription and the reference one.

Evaluation technique

From our different sound databases, we extracted different sets of training and test data. Within each dataset, the musical sequences were randomly split into training and testing sequences, using by default 30 % of sequences for testing, and the 70% remaining ones for training. Sequences from a same musical piece were constrained to appear either in training or in test data. Also, the learning of note templates (see T_3, T_5 and T_6) was made on instrument models different from the one used in test sequences. This procedure allows preventing any overfitting of our data in our simulation experiments, and is repeated five times, with an average computed on the resulting transcription scores. It is noteworthy that usually, results on AMT evaluation (Dessein et al., 2010; Benetos and Dixon, 2011; Grindlay and Ellis, 2011) are presented by selecting the parameter values (e.g. the sparsity coefficient in Benetos and Dixon (2013)) that maximizes the average accuracy in a dataset.

5.1.2 Evaluation AMT algorithms

Tolonen2000's algorithm

This algorithm (Tolonen and Karjalainen, 2000) is an efficient model for multipitch and periodicity analysis of complex audio signals. The model essentially divides the signal into two channels, below and above 1000 Hz, computes a "generalized" autocorrelation of the low-channel signal and of the envelope of the high-channel signal, and sums the autocorrelation functions.

Emiya2010's algorithm

This algorithm (Emiya et al., 2010) models the spectral envelope of the overtones of each note with a smooth autoregressive model. For the background noise, a moving-average model is used and the combination of both tends to eliminate harmonic and sub-harmonic erroneous pitch estimations. This leads to a complete generative spectral model for simultaneous piano notes, which also explicitly includes the typical deviation from exact harmonicity in a piano overtone series. The pitch set which maximizes an approximate likelihood is selected from among a restricted number of possible pitch combinations as the one.

Fuentes2013's HALCA algorithm

The Harmonic Adaptive Latent Component Analysis (HALCA) algorithm [1] (Fuentes et al., 2013) is used as a state-of-the-art reference. This algorithm was recently evaluated by MIREX and obtained the 2^{nd} best score in the MPE task, 2009-2012 (MIREX, 2011). It mainly differentiates from our baseline algorithm B_0 by the EM estimation of kernel weights, which are not kept fixed but update with data. This model also includes a noise model and different priors on sparsity, monomodality and temporal continuity of spectral envelops.

Benetos2013's algorithm

This PLCA-based AMT system [2] (Benetos et al., 2013a) uses pre-fixed templates, and has been ranked first in the MIREX transcription tasks MIREX (2011).

Proposed system 1 (BaseCaz1(C_i))

This first system corresponds to our baseline PLCA plus HMM system, with the incorporation of a combination of KCMA given by C_i. When C_i is empty, this system then uses the simple PLCA system presented in 2.2.8.

Proposed system 2 (BaseCaz2(C_i))

This first system corresponds to our baseline PLCA plus PF system, with the incorporation of a combination of KCMA given by C_i.

5.1.3 Numerical parameters & Default configuration

For all methods, as a time-frequency representation, the constant-Q transform (CQT) with 60 bins/octave was used, with a half-overlapped Hamming window of 23 ms (i.e. 1024 coefficients at 44.1-kHz sampling rate). The number of EM iterations is experimentally set to 25. By default, and unless said otherwise, we use the simple thresholding method (see section 2.3.3), which should allow one to better highlight the differences brought by musical knowledge. Also, unless said otherwise, we will only evaluate Onset location and Pitch, as judging that we did not have sufficiently reliable ground truth data for this evaluation, and also because those are not evaluated in most AMT studies. Also, after numerical experimentations, an optimal value of $\varepsilon = 0.08$ in eq. 2.28 has been set.

1. Codes are available at http://www.benoit-fuentes.fr/.
2. Codes are available at https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast.

5.2 Results of AMT performance I : on the *marovany* repertoires

In this section, we wish to investigate the impact of KCMA incorporation in AMT systems on our different *marovany* repertoires. Our research publication Article 2 resumes a large proportion of the content of this chapter, although with a smaller sound dataset.

5.2.1 Evaluation sound dataset

For this first set of numerical experimentations, three different repertoires R_i of the *marovany* zither were composed from our PSIFAMT database, corresponding respectively to $R_1 = \{D_1, D_5\}$ (Velonjoro's repertoire), $R_2 = \{D_4, D_6\}$ (Charles Kely's repertoire) and $R_3 = \{D_3, D_7\}$ (Kilema's repertoire). 12 different musical pieces were taken from each musical repertoire. We then selected 8 non-overlapping sequences of 30 seconds in each piece, for a total of 144 sequences (i.e. 48 minutes) per repertoire. All musicians played the major musical pieces of their repertoire on the same *marovany* model (model N_1), equipped with an original multi-sensor retrieval system Conference 2. Datasets R_1 to R_3 in combination with the note sample dataset N_1 compose our oracle systems [3]. A fourth "universal" sound database R_4 was also constituted by mixing all MIDI repertoires and template sets. Figure 5.3 provides an illustrative example of a test sequence (cut to 15 seconds) from the *marovany* piece Folera played by the musician Kilema, with different transcription outputs.

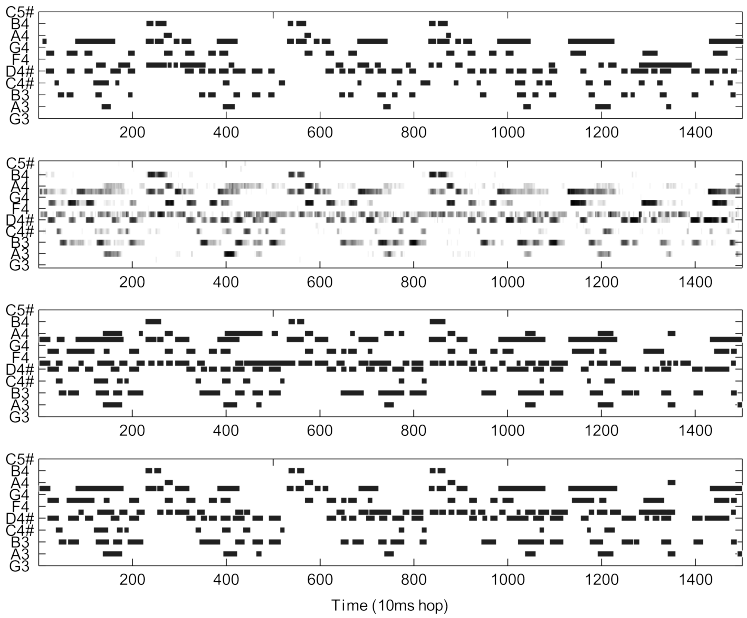


Figure 5.3 – Illustration of different stages of our AMT system on a test musical sequence, with from top to bottom: ground truth, pitch activity matrix $P(i,t)$, piano-roll transcription output using a simple thresholding and piano-roll transcription output using the combination of priors $\{T2, T4, T5, M4\}$ (taken from our Article 2).

3. An oracle system is defined as a transcription system using note samples from the same instrument source, which is meant to demonstrate the upper performance limit of the transcription system.

5.2.2 Performance of Baseline systems

In this first section, we test the performance of the different algorithms on the test bench, which allows placing our baseline transcription algorithm performance in the context of state-of-the-art algorithms.

Evaluation on Onset location and Pitch

As displayed in table 5.2, our baseline system algorithm BaseCaz1 provided transcription results of 57.5 %, 60.3 % and 62.5 % for the different R_1 to R_3, respectively, with slightly worst results for our algorithm BaseCaz2. Globally, the state-of-the-art AMT system of Benetos2013 provides the best transcription results. However, it must be reminded that, on the contrary to other the other systems, Benetos2013 uses pre-fixed instrument templates per pitch, i.e. musical knowledge specific to the instrument repertoire. Whereas the worst performing system is Tolonen2000, which is a feature-based algorithm without any prior information on the signals to be retrieved. This last approach appears to suffer the most from ambiguities in multi-pitch estimation.

Baseline systems	R_1	R_2	R_3	R_{tot}
Tolonen2000	52.2 %	57.4 %	57.1 %	55.5 %
Emiya2010	55.9 %	56.1 %	55.7 %	56.8 %
Fuentes2013	58.3 %	61.6 %	**61.2** %	59.1 %
Benetos2013	**60.5** %	**64.1** %	60.1 %	61.3 %
BaseCaz1	**57.5** %	**60.3** %	**62.5** %	58.6 %
BaseCaz2	56.8 %	60.6 %	61.4 %	58.2 %

Table 5.2 – Average F-measures obtained with our baseline systems using the two note parameters of Onset location and Pitch for evaluation, for our different evaluation datasets.

Evaluation on all note parameters

As displayed in table 5.3, our baseline system algorithm BaseCaz1 provided transcription results of 52.7 %, 48.5 % and 50.2 % for the different R_1 to R_3, respectively. As expected, by taking into account all note parameters in the AMT evaluation (with the error tolerance thresholds detailed in table 5.1), our transcription performance is drastically deteriorated.

Baseline systems	R_1	R_2	R_3	R_{tot}
Tolonen2000	49.1 %	51.9 %	45.8 %	48.7 %
Emiya2010	49.7 %	53.1 %	52.3 %	51.2 %
Fuentes2013	49.8 %	51.2 %	54.6 %	52.9 %
Benetos2013	51 %	54.2 %	49.5 %	50.6 %
BaseCaz1	52.7 %	48.5 %	50.2 %	51.4 %
BaseCaz2	48.7 %	49.5 %	50.1 %	49.9 %

Table 5.3 – Average F-measures obtained with our baseline systems using all note parameters for evaluation, for our different evaluation datasets.

Evaluation of a monophonic mode transcription

This lower level AMT task of monophonic transcriptions may be useful in many different application such as melody extraction, and identification of motive structures on larger time-scales. As displayed in table 5.4, our baseline system algorithm BaseCaz1 (using the note segmentation mode described in section 2.3.2) provided transcription results of 67.1 %, 67.9 % and 65.1 % for the different R_1 to R_3, respectively. In this evaluation set-up, our different AMT systems are rebalanced, while still outperforming more feature-based algorithms such as Tolonen2000.

Baseline systems	R_1	R_2	R_3	R_{tot}
Tolonen2000	59.4 %	62.3 %	59.1 %	59.6 %
Emiya2010	59.9 %	62.5 %	64.2 %	61.3 %
Fuentes2013	68.4 %	69.5 %	66.7 %	68.3 %
Benetos2013	67.5 %	68.9 %	66.2 %	68.7 %
BaseCaz1	67.1 %	67.9 %	65.1 %	66.4 %
BaseCaz2	62.1 %	63.4%	64.8 %	61.2 %

Table 5.4 – Average F-measures obtained with our baseline systems in the monophonic mode transcription, for our different evaluation datasets.

Sensitivity to threshold-based note segmentation methods

We now study dependency of our results on our threshold-based note segmentation methods, being either the simple fixed threshold (see section 2.3.3) or the adaptive one (see section 2.3.4). Figure 5.4 displays the evolution of the $F-measure$ (in %) against the value of the fixed threshold $Thres_{fix}$ (in dB) for three different baseline methods. A same tendency can be observed for the different curves, with a clearly identified optimal threshold for transcription, i.e. best compromise between true transcribed notes and false alarms. The adaptive-threshold segmentation method brings more enhancements for the AMT systems HALCA and BaseCaz1, supposedly because more variations occur in their activity matrices, in comparison to methods based on pre-fixed templates such as Benetos2013.

Sensitivity to template learning methods

Figure 5.5 proposes a comparison of transcription performance based on ROC curves using our different template learning methods, presented in section 2.2.8, in our baseline PLCA model BaseCaz1. The method of pre-fixing the templates before model parameter estimation proves to provide the highest performance in our simulations, while the method of data-based template adaptation using a conservative transcription presents similar performance to the HALCA template learning process.

5.2.3 Testing each KCMA individually

Each single KCMA incorporated in our AMT system BaseCaz1, described in tables 3.1-3.3, was successively incorporated in our transcription framework, composed of the three blocks (PLCA-Postprocessing-HMM). To quantify the impact of KCMA on transcription performance, we computed the measure G_p, called Gain prior, and defined as $F_{Prior_k} - F_{B_0}$, where F_{B_0} is our reference F-measure obtained with our baseline algorithm BaseCaz1, and

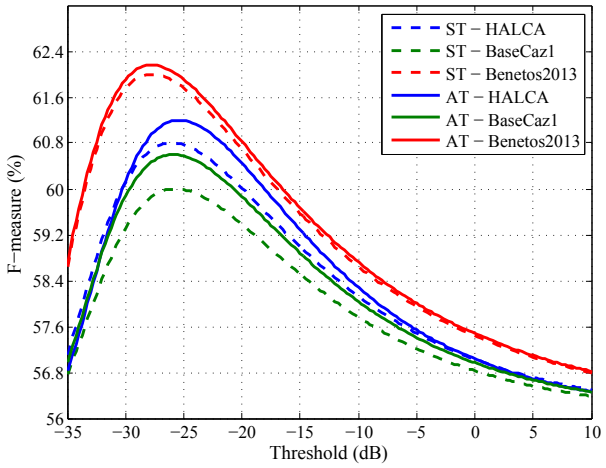


Figure 5.4 – Evolution of the $F-measure$ (in %) against the value of the fixed threshold $Thres_{fix}$ (in dB) for three different baseline methods. ST is short for the Simple Threshold method from section 2.3.3 and AT for the Adaptive Threshold method from section 2.3.4.

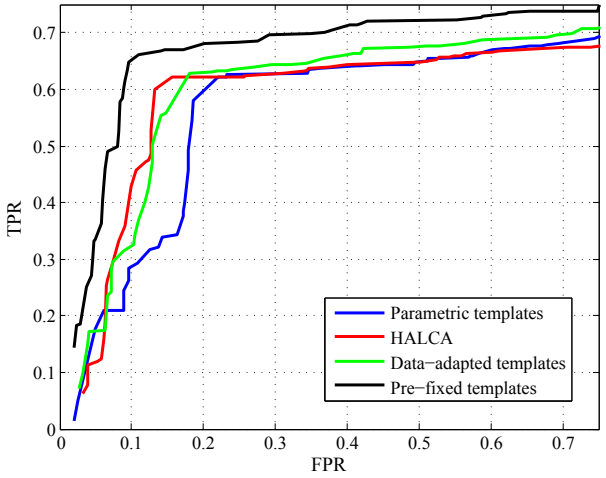


Figure 5.5 – Comparison of transcription performance based on ROC curves using our different template learning methods, presented in section 2.2.8, in our baseline PLCA model BaseCaz1.

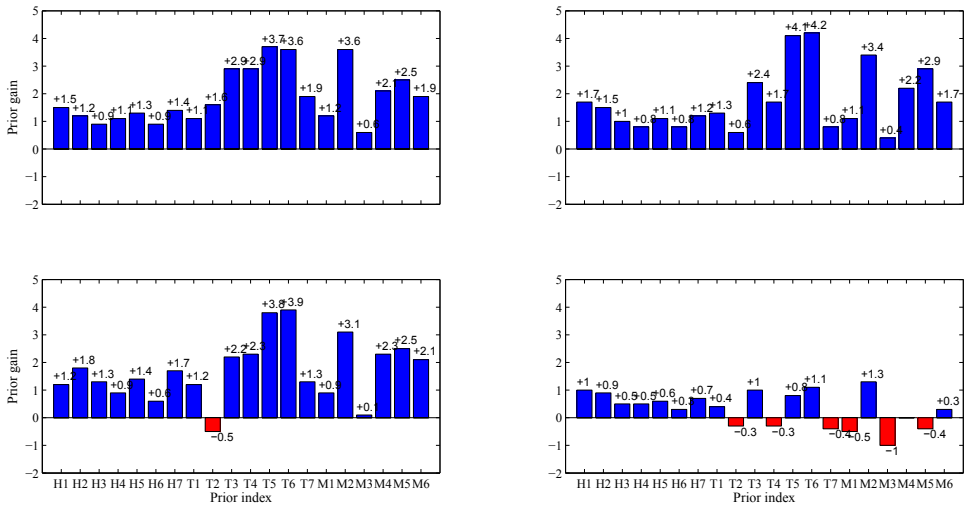


Figure 5.6 – Prior gains G_p for each individual prior, detailed for the different datasets.

F_{Prior_k} the F-measure obtained by integrating the prior k to the transcription system. This gain can be either positive or negative, depending on whether it enhances or degrades transcription results of the PLCA baseline system algorithm. Figure 5.6 shows their impacts on transcription performance in terms of the measure G_p. From this figure, we see that the G_p values respective to each individual prior ranged from + 4.2 to -1.3. Also, the dataset R_1 is the only one to benefit from a positive G_p for all priors, whereas the dataset R_4 is the one on which priors have the strongest negative impact, with only 4 positive contributions.

5.2.4 Testing combinations of KCMA

Best transcription gains

Table 5.5 shows the best transcription gains obtained from our baseline algorithms BaseCaz1 and BaseCaz2. By incorporating KCMA in AMT systems, we see that the prior gains brought in our system BaseCaz1($P1_c$), which is as high as +6.6% for Velonjoro's repertoire, outperforms the state-of-the-art AMT algorithms reported in table 5.2. This tendency is confirmed through our different *marovany* repertoires, although the prior gains significantly differ from one repertoire to another.

Methods	R_1	R_2	R_3	R_{tot}
BaseCaz1($P1_c$)	64.1 %	66 %	67.3 %	60.7 %
$G_p(P1_c)$	+ 6.6 %	+ 5.7 %	+ 4.8 %	+ 2.1 %
$P1_c$	$\{T1, T3, T4, T7, M5\}$	$\{T1, T5, M2, M5\}$	$\{H1, T1, T5, M2, M4\}$	$\{H2, H6, T3, M2\}$
BaseCaz2($P2_c$)	63.2 %	65.3 %	65 %	61.1 %
$G_p(P2_c)$	+ 6.4 %	+ 4.7 %	+3.6 %	+ 2.9 %
$P2_c$	$\{T3, T4, M4\}$	$\{H5, T1, T5, M2, M3\}$	$\{H1, T5, M4\}$	$\{H2, H7, M2\}$

Table 5.5 – Average F-measures obtained with the systems BaseCaz1($P1_c$) and BaseCaz2($P2_c$) (with $P1_c$ and $P2_c$ the combination of priors with the highest G_p value, for each dataset, respectively for BaseCaz1 and BaseCaz2), and the $G_p(P1_c)$ and the $G_p(P2_c)$ values, for our different evaluation datasets.

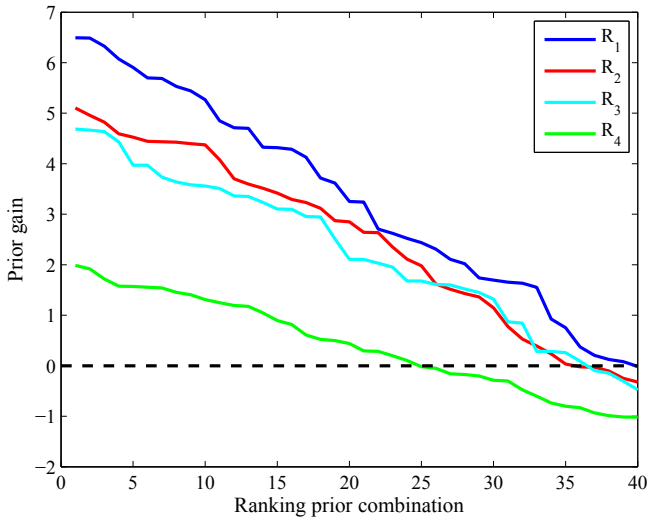


Figure 5.7 – Prior gains G_p of the first 40 best ranking prior combinations, detailed for the different datasets.

In details

We will now evaluate the impact of prior knowledge on AMT performance by considering different combinations of individual priors. To do so, we first formed all possible combinations of priors, and successively evaluated their impact on transcription results using each time the same test and learning sequences. In the same way that for figure 5.6, figure 5.7 plots the prior gain G_p for each prior combinations and each dataset. It can be noted that the dataset R_1 has a positive prior gain G_p over the 40 first best performing prior combinations, while the dataset R_2, R_3 and R_4 cross the zero gain score at the ranks 36, 37 and 24, respectively. The higher this rank, and the higher the usefulness of knowledge prior in enhancing the transcription accuracy of a repertoire. It can then be mentioned that the repertoire R_1 benefits the most from prior knowledge. On the contrary, dataset R_4 has a negative prior gain G_p from the 24^{th} prior combination.

To provide a more detailed insight in these prior combinations, figure 5.8 represents the relative proportions according to which each individual prior is present over the first 40 prior combinations presenting the highest G_p values, for the four repertoires. For example, we can see that in most repertoires, high values are obtained for priors T_5 and T_6, which mean that these priors are present in many of the best performing prior combinations.

5.2.5 Discussion

Our baseline performance remains not satisfactory enough to provide robust musical analysis supports, but they remain close to the average performance level of current AMT systems on polyphonic solo-instruments, which is around 60 % in the average $F-measure$ with note-based evaluation metrics (Benetos et al., 2013b). By incorporating KCMA in our system BaseCaz1 (see table 5.5), we achieve transcription enhancements as high as +6.6% in the average $F-measure$ for Velonjoro's repertoire, and outperforms significantly the state-of-the-art AMT algorithms reported in table 5.2 for all repertoires. This confirms well that model-driven transcription systems (Ellis, 1996), which make use of prior knowledge specific to their data under analysis, outperform globally data-based systems, whose more

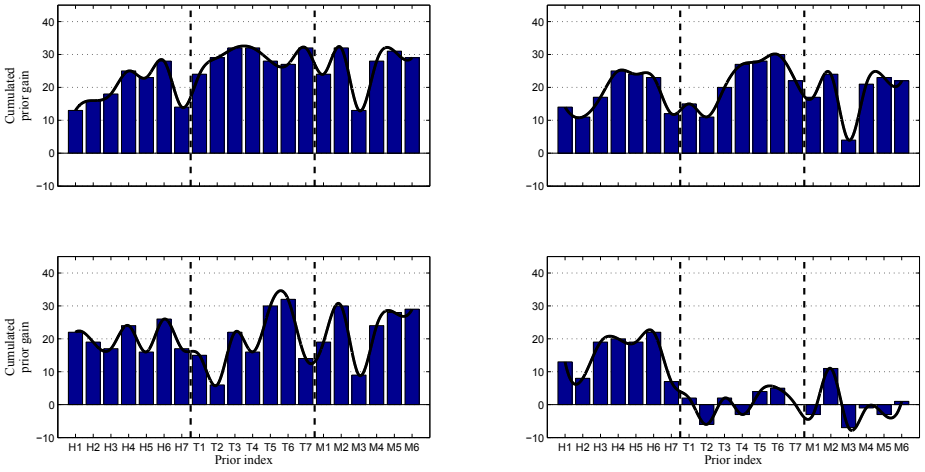


Figure 5.8 – G_p values cumulated over the first 40 best ranking prior combinations, respective to each prior, and for the four different datasets R_1 to R_4.

general application often costs a decrease in transcription performance.

We highlight the major contribution of pre-recorded templates per pitch in transcription tasks, which was particularly noticeable in contrast to the individual parametric spectral models, as in the HALCA algorithm. Templates have been used in pattern-matching operations and are very efficient in multi-pitch tracking tasks (Mysore, 2010; Benetos and Dixon, 2013). As far as the choice of templates is concerned, Kirchhoff et al. (2012) stated that the highest transcription accuracies are obtained when spectral templates are learned directly from the recording under analysis or from isolated notes of the same instrument model, whereas Benetos and Dixon (2013) seemed to suggest that having a large set of templates that might include instruments not present in the recording does in fact improve transcription accuracy. Also, Benetos and Dixon (2013) suggested the use of isolated sounds database to build spectral templates, instead of extracting them from solo performances. Our simulation experiments seem to confirm these last two statements, especially when adding a variety of template models able to encompass both the intrinsic timbre properties of an instrument and the acoustic modifications which can be induced to different playing style, as proposed by the RWC database (Goto et al., 2003). Such a combination of different instrument templates might better approximate the spectra of the produced notes, as observed by Benetos and Dixon (2013). Prior gains associated to the use of templates show great promises for all *marovany* repertoires, and even higher transcription performance can be expected with a more advanced characterization of their instrument acoustics. One can speculate on the fact that more complex and variable acoustic signatures exist for these instruments due to their non-standardization and "precarity" of their making, in contrast to commercialized eurogenetic instruments. A higher acoustic complexity which has been indeed observed in Conference 4, particularly in regards to inharmonicity and spectral envelope variability. Further studies will be needed to validate a correlation between acoustic timbre complexity and the usefulness of timbre-related prior knowledge for non-eurogenetic repertoires. However, the current rarity of datasounds formatted for AMT evaluation of more traditional instrument repertoires is the first problem to be solved. In that direction, the authors have undertaken the development of a sound

database dedicated to plucked string instruments of traditional African repertoires.

Also, our results provide interesting basis to discuss inter-repertoire differences, which will be further explored in the following by studying how prior knowledge impact each repertoire transcription. It is the musical repertoire R_1 of Velonjoro which gets both the lowest baseline and HALCA transcription results and the highest positive gain from priors. In other words, this repertoire leaves much room for prior-based transcription improvements, and benefits the most from specific prior knowledge. The repertoire of Velonjoro consists of a succession of melodic phrases most often played in arpeggios. There is no vertical writing properly speaking in this music, excepting a few punctual chords. However, the high tempo at which notes are played, combined with the facts that strings are barely muted within a musical phrase and that they often resonate with each other, confers to this music a complexity of analysis comparable to that provided by polyphonic music. The high-level of polyphony and virtuosity style of this musical corpus constitutes classical difficulties for automatic transcription. In addition to that, Velonjoro tuning is quite deviating from the typical well-tempered scale, as well as "non-octaving", i.e. scale is not the same depending on the octave (e.g., in 1^{st} variation of Sojerina, the flat C in octave 3 is a natural C in octave 4 and 5). All these musical features contribute in decreasing transcription performance obtained with the HALCA algorithm, while their proper identification and characterization allowed the development of specific priors quite efficient on this repertoire. Mostly, this repertoire benefits the most from the KCMA category of timbre (see relations between KCMA and their categories in table 1.1), which fits well the fact that in the traditional musical playing of the *marovany*, players like Velonjoro interfere at the minimum on its intrinsic timbre, and even excite it as louder as possible (an effect called *mafo be*). In particular, prior $T3$ (related to timbre) is at the same level as the other template-based priors $T5$ and $T6$ (related to playing style), which are much more predominant in the other *marovany* repertoires.

Concerning sequential prior knowledge, injected in the HMM framework, we can mention that the knowledge-based priors H globally outperforms the data-based prior M_3 for the modeling of harmonic transitions, resulting simply from the fact that taking into account musical knowledge such as tonality provides more accurate transition probabilities, which are robust to the different datasets, whereas the probabilistic note transition contains a superposition of the different tonalities, with a too local information to be representative of different musical repertoires. In repertoires with simpler harmonic structures, as in Velonjoro, the prior gains of these two types of priors (knowledge-based vs data-based) are rebalanced.

Priors integrating information on note duration and temporal envelop appears to be beneficial, especially for Velonjoro's repertoire. In particular, a previous in-depth analysis of model errors on an annotated *marovany* dataset showed that most of the false negatives are actually produced by a strong intermodulation on certain notes. The valleys appearing in temporal envelops of intermodulated notes are a strong cause of false alarm detection, as they can be seen as quick re-occurrences of a same note, which should actually be caused by particular musical figures (e.g. fast arpeggio). These two quite simple priors work surprisingly well against these errors, especially in Velonjoro's repertoire where notes are most often let in free-resonance.

Eventually, when considering the dataset R_4, in comparison to the oracle datasets R_1 to R_3, we observe that globally all prior contributions decrease, with more noticeable negative impacts on priors T_5 and T_6, the two multi-template based priors. Then, we can state that even within a same instrument repertoire, the integration of prior knowledge should also take into account specificities from different musicians and instrument models,

as is the case here for the *marovany*.

5.3 Results of AMT performance II : on inter- instrument repertoires

In this section, we wish to investigate the impact of KCMA incorporation in AMT systems on different instrument repertoires, especially between classical piano, folk guitar and the *marovany*. To do so, we will proceed using the same steps as in section 5.2. Our research publication Article 1 resumes a large proportion of the content of this chapter.

5.3.1 Evaluation sound dataset

Three different decay instruments were selected, namely classical piano, steel-string acoustic guitar and the *marovany* zither from Madagascar, labelled R_1 to R_3. For the first two instruments, isolated note samples were extracted both from the RWC database (Goto et al., 2003). For the *marovany*, all note samples were recorded by the authors in a semi-anechoic chamber of our laboratory, as no publically available sound databases exist for this instrument. For each pitch of an instrument pitch range, we have three different note samples from three different instrument models, for a total of 9 note samples per pitch per instrument type. These note samples are obtained by varying playing dynamic only (covering approximately the nuances of piano, mezzo forte, and forte), and with a fixed position of the excitation point, centred in the playing area, for the two plucked-string instruments.

For what concerns musical pieces, as in section 5.2, 12 different musical pieces were taken from each musical repertoire. We then selected 8 non-overlapping sequences of 30 seconds in each piece, for a total of 144 sequences (i.e. 48 minutes) per repertoire. For classical piano, musical pieces were taken randomly from the MAPS database (Emiya et al., 2010). These pieces belong to the euro-genetic classical music from the 18^{th} and 19^{th} centuries. For our two plucked-string instruments, our musical pieces were randomly taken from our PSIFAMT database, corresponding respectively to $R_2 = \{D_{10} : D_{12}\}$ and $R_3 = \{D_1, D_4, D_5\}$. Figure 5.3 provides an illustrative example of a test sequence (cut to 15 seconds) from a classical piano piece (MAPS database), with different transcription outputs.

5.3.2 Performance of Baseline systems

In this first section, we test the performance of the different algorithms on the test bench, which allows placing our baseline transcription algorithm performance in the context of state-of-the-art algorithms. As displayed in table 5.6, our baseline system algorithm BaseCaz1 provided transcription results of 62.2 %, 59.7 % and 59.2 % for the different repertoires R_1 to R_3, respectively. The best performing system is the one of Benetos2013, with a score of 64.8 % for the classical piano repertoire, which is of the order of magnitude found in the literature (Benetos et al., 2013a), and 61.5 % for the *marovany* repertoire.

5.3.3 Testing combinations of KCMA

Best transcription gains

Table 5.7 shows the best transcription gains obtained from our baseline algorithms BaseCaz1 and BaseCaz2. With the prior gain, we see that our AMT system BaseCaz1($P1_c$)

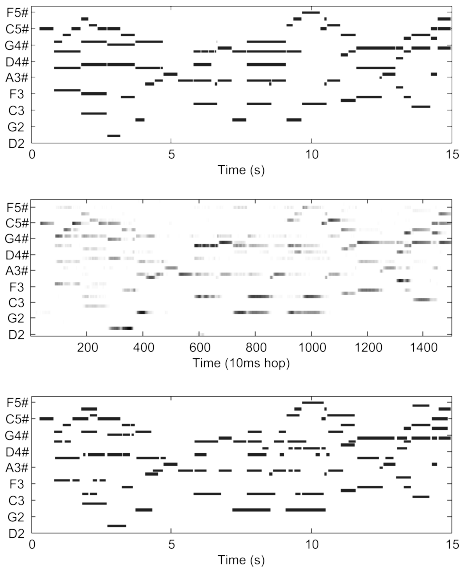


Figure 5.9 – Illustration of different stages of our BaseCaz2 system on a test musical sequence, with from top to bottom: ground truth, pitch activity matrix $P(i,t)$ and piano-roll transcription output.

Baseline systems	R_1	R_2	R_3	R_{tot}
Tolonen2000	55.4 %	57.6 %	53.4 %	56.6 %
Emiya2010	61.5 %	58.7 %	56.4 %	57.9 %
Fuentes2013	63.2 %	61.4 %	59.5 %	60.6 %
Benetos2013	**64.8** %	**62.3** %	**61.5** %	62.9 %
BaseCaz1	**62.2** %	**59.7** %	**59.2** %	61.4 %
BaseCaz2	61.8 %	59.9 %	58.1 %	59.7 %

Table 5.6 – Average F-measures obtained with our baseline systems, for our different evaluation datasets.

outperforms the state-of-the-art AMT algorithms reported in table 5.6. Once again, the incorporation of optimal KCMA in our systems BaseCaz1 and BaseCaz2 enhances significantly our transcription performance, which brings it to the same order of magnitude as Benetos2013 for the classical piano, and outperforms this one for the *marovany* repertoires. Also, we can observe strong inter-repertoire differences in the impact of KCMA on the AMT system BaseCaz1($P1_c$) , with a difference in $G_p(P1_c)$ as high as 3.1 % between the repertoires of classical piano R_1 and *marovany* R_3.

In details

We will now evaluate the impact of prior knowledge on AMT performance by considering different combinations of individual priors. To do so, we first formed all possible combinations of priors, and successively evaluated their impact on transcription results using each time the same test and learning sequences. Figure 5.10 plots the prior gain G_p for each prior combinations and each dataset. It can be noted that the dataset R_1 has a positive prior gain G_p over the 40 first best performing prior combinations, while the dataset R_1, R_2 and R_4 cross the zero gain score at the ranks 31, 33 and 15, respectively. The higher this rank, and the higher the usefulness of knowledge prior in enhancing the

Methods	R_1	R_2	R_3	R_{tot}
BaseCaz1($P1_c$)	65.2 %	64.1 %	65.3 %	59 %
$G_p(P1_c)$	+ 3 %	+ 4.4 %	+ 6.1 %	+ 1.1 %
$P1_c$	$\{H1, M1, M2, M4\}$	$\{T1, T5, M2, M4\}$	$\{T1, T3, T4, T7, M5\}$	$\{H3, H7, M2\}$
BaseCaz2($P2_c$)	64.5 %	64 %	63.4 %	61.3 %
$G_p(P2_c)$	+ 2.7 %	+ 4.1 %	+ 5.3 %	+ 1.6 %
$P2_c$	$\{H1, M2, M4\}$	$\{H5, T6, M2, M4\}$	$\{T1, T4, T5, M5\}$	$\{H2, H6, M2\}$

Table 5.7 – Average F-measures obtained with the systems BaseCaz1 + $P1_c$ and BaseCaz2 + $P2_c$ (with $P1_c$ and $P2_c$ the combination of priors with the highest G_p value, for each dataset, respectively for BaseCaz1 and BaseCaz2), and the $G_p(P1_c)$ and the $G_p(P2_c)$ values, for our different evaluation datasets.

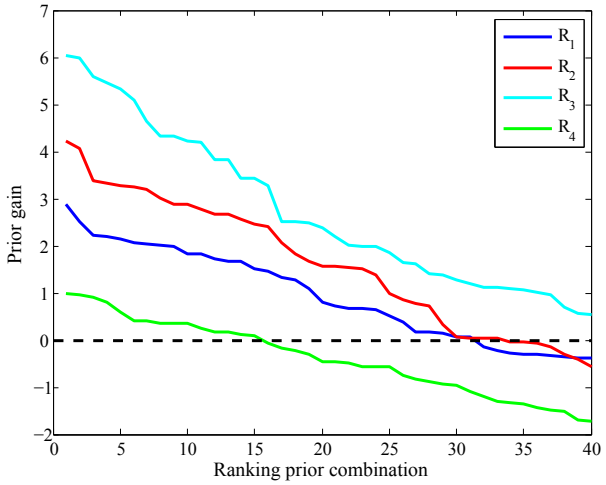


Figure 5.10 – Prior gains G_p of the first 40 best ranking prior combinations, detailed for the different datasets.

transcription accuracy of a repertoire. It can then be mentioned that the repertoire R_3 benefits the most from prior knowledge. On the contrary, dataset R_4 has a negative prior gain G_p from the 15^{th} prior combination.

To provide a more detailed insight in these prior combinations, figure 5.11 represents the relative proportions according to which each individual prior is present over the first 40 prior combinations presenting the highest G_p values, for the four repertoires.

5.3.4 Discussion

To begin with, it can be first noticed that all priors considered in our analysis induce enhancements in the transcription results of state-of-the-art algorithms, on at least one of the instrument repertoires. With the results from this section, some inter-repertoire differences related to KCMA incorporation are clearly highlighted. Numerically, prior gains range are of 3 %, 4.4 % to 6.1 %, respectively for the classical piano, folk guitar and *marovany* repertoires. This tendency can be first explained by the fact that transcription performance of baseline systems are higher for classical piano (see table 5.6), and so there is less room for improvement with the incorporation of musical knowledge. Also, transcription enhancements are more important for our two plucked-string instruments, in comparison to classical piano, but are more similar for knowledge-based KCMA (source I). This supports the idea that musical features related to both timbre and playing style may be more complicated in plucked-string instrument repertoires. This makes sense indeed

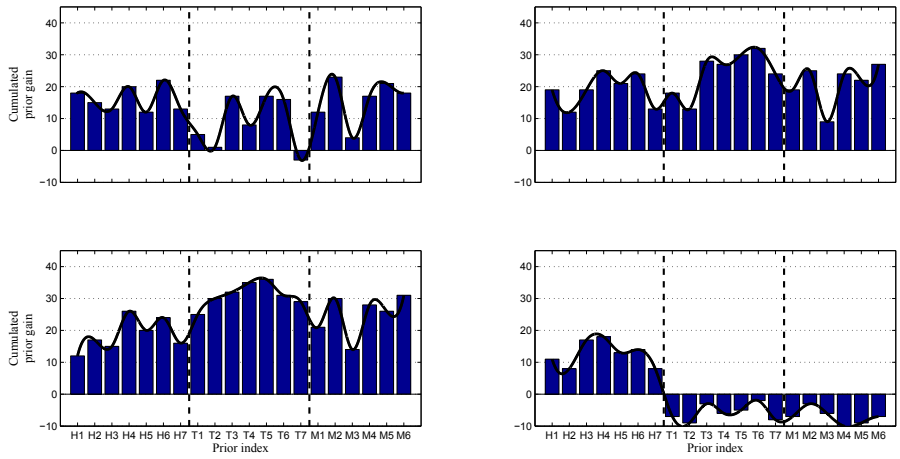


Figure 5.11 – G_p values cumulated over the first 40 best ranking prior combinations, respective to each prior, and for the four different datasets R_1 to R_4.

when comparing the different playing techniques feasible on plucked-string instruments, such as plucking position and excitation modes.

Then, the results of this section 5.3 learns us that the *marovany* repertoire benefits the most from musical knowledge, whatever musical knowledge is considered, whereas our "universal" dataset R_4 calls for more carefully selected and well-dosed external information (in our simulation experiments, most knowledge components result unnecessary and often detrimental), risking to decrease drastically transcription results. A tendency supported by the important predominance of very specific musical information which do not depend on data, such as the known played notes and the tonality-based probabilistic transitions between note mixtures. Such a result suggests the non-universality of prior knowledge, and the consequent need to adapt information depending on repertoires. Also, when comparing this universal dataset from section 5.3 to the one from section 5.2, we see that the transcription performance gains from KCMA are less degraded within the *marovany* repertoires.

Like for results from section 5.2, the most stable KCMAs are, as expected, the ones from the Source I, whose contributions are not really affected by the mixing of the different repertoires in the dataset R_4. Such KCMAs are naturally in generic AMT systems (see e.g. Benetos et al. (2014b)), although our results reveal that their performance gains are quite limited in comparison to KCMA specific to an instrument repertoire.

5.4 Conclusion

In this chapter, we were interested in performing an in-depth investigation of their impact on music transcription. In order to identify prior contributions in transcription results, we used only knowledge components built on explicit information from various music-related domains of knowledge, namely the timbre, musicological and playing style classes. Our numerical experiments lead us to several interesting observations on the use of knowledge in a state-of-the art AMT system, such as the complementarity of the different knowledge classes and the non-universality of prior knowledge in regards to the diversity of instrument repertoires across the world.

Conclusion

Summary content

This PhD thesis deals with the development of an automatic transcription system dedicated to the repertoires of the traditional Marovany zither from Madagascar. The common denominator of the different approaches developed to the task of AMT lays in the use of explicit music-related prior knowledge in computational systems. The AMT framework was build upon two main methods, namely Probabilistic Latent Component Analysis (PLCA) for multi-pitch estimation and Hidden Markov Models (HMMs) for note segmentation and sequential post-processing, and different configurations of these methods were developed to provide a powerful probabilistic framework, covering the time-frequency domain on different time-scales, in which original integration methods of prior knowledge were developed. We were interested in developing a generic transcription system based on a PLCA post-processed with a HMM block, which makes possible to supply it with all kind of prior knowledge at different levels of abstraction, and performing an in-depth investigation of their impact on music transcription. In order to identify prior contributions in transcription results, we used only knowledge components built on explicit information from various music-related domains of knowledge, namely the timbre, musicological and playing style classes. Our numerical experiments lead us to several interesting observations on the use of knowledge in a state-of-the art AMT system, such as the complementarity of the different knowledge classes and the non-universality of prior knowledge in regards to the diversity of instrument repertoires across the world.

Also, an important part of this PhD project dealt with the constitution of a database of ground-truth transcriptions of a not yet studied instrument repertoires. This involved the development of specific acquisition and processing tools to automatize this process, and takes part to a larger ambition of making available to the MIR community a sound database dedicated to various plucked-string instrument repertoires, especially including traditional ones.

Review of Objectives

In terms of our initial objectives, this PhD has achieved the three main following points

— First, the development of Multichannel Capturing Sensory Systems dedicated to plucked-string instruments, and more specifically to the *marovany*, has achieved to provide numerous reliable transcriptions of instrument repertoires, in an automatic and time-saving way. Such transcriptions have been exploited as ground truth transcriptions in the task of Automatic Music Transcription (see below), and also in computational musicological studies ;

— Second, the development of an automatic music system with original knowledge components from musical acoustics incorporated in it, specially conceived and op-

timized for the *marovany* instrument repertoire. A first major result was the greater transcription performance of templates learned from pre-recorded samples, including "interpreted" note samples which result from specific playing techniques, in comparison to more generic parametric note models. The timbre of *marovany*, which can vary a lot in playing condition, on the contrary to notes of classical piano for example, then really benefit from a more instrument-specific modeling of note spectral envelops. Other knowledge components specific to this instrument repertoire include sympathetic resonances, intermodulated notes, shift-invariant templates, pitch-wise note duration. Optimal incorporations of this knowledge in music transcription systems have allowed significant enhancements in transcription performance, outperforming state-of-the-art systems ;

— Third, the development of a vast set of knowledge components from musical acoustics, covering both timbre and music language features of an instrument repertoire, have allowed investigating the impact of knowledge components on different instrument repertoires. We then highlighted significant inter-repertoire differences through the incorporation of these components in our baseline music transcription system.

We really wish this work to help researchers in the community of computational ethnomusicology, who would like to apply tools of automatic music transcription to their repertoires. The analytical framework developed in this PhD should help such researchers in choosing the appropriate knowledge to incorporate in computational systems, as we showed that it is inappropriate to use all music knowledge, in terms of both the quantity and nature, regardless the instrument repertoire under study. The methodology and results of this PhD should also help in better understanding how current music transcription algorithms could be adapted to new instrument repertoires. Through the explicit links we made between musical acoustics knowledge and automatic transcription accuracy, one could directly predict the best knowledge components to use for transcription after having identified and characterized the most characteristic musical features in its repertoires (e.g. highly variable spectral envelops of played notes → learning of pre-recorded templates; non-tempered tuning → shifting templates in frequency)

Prospects

Dealing with musicological issues, and ethnomusicological issues in particular, is not a simple subject for computer scientists, as explained by Lartillot et al. (2008): "We hope expert ethnomusicologists will understand the experimental aspect of such a cross-disciplinary undertaking, and will pardon the potential imperfection in this computational attempt toward cross-cultural understanding. The argumentation of the paper might betray for some readers some residual of Eurocentrism. We would like to emphasize that the authors of this work are aware of the limitation of this opposition between Western and non-Western music, which is proposed as an experimental state before the development of further research aimed at the modeling of "stylistic features proper to different stylistic areas"."This PhD thesis tried at best to follow these recommendations, while bringing clear evidences that the study of a non-eurogenetic repertoire such as the *marovany* repertoire raises specific issues, and a consequent need of adapting and developing specific tools to answer them. Also, the authors are aware of focusing on a very specific music culture, which make the selected priors and results not easily generalizable. However, there are clear analogies and similar phenomenon occurring in many music cultures. Therefore we believe this document can be a good reference to many researchers working in the AMT

tasks, and also a useful starting point in tailoring AMT systems as an aid to musicological studies involving melody.

Our research interests for future studies will be to test the validity of our results on sound databases with real recordings, to quantify the precise representativeness and amount of musical pieces in regards to a complete instrument repertoire, and to undertake the development of a semi-automatic transcription tool, integrating a large and flexible number of prior knowledge which can all be directly understood and modified by an external user.

For what our PSIFAMT database, the motivations of developing this new sound database are multifold:

— Making available a new sound database for AMT evaluation in the MIR community, which is now indispensable to test the robustness of current algorithms to music diversity, and face them to more complex musical features ;

— Plucked-string instruments present a great diversity in terms of timbre (e.g. inharmonicity, envelop spectrum variability, temporal profile modulation), playing styles and effects (e.g. different string excitation modes, palm muting, glissandi, vibrato, tremolo, harmonic notes) ;

— Put the attention on non-eurogentic traditional repertoires, encompassing the greatest diversity of musicological codes (e.g. micro-rhythm, tuning based on non-equal temperament, modal scale) and of musical practice, and complexity in music encoding ;

We wish to complete the formatting of this sound database so it can be permanently hosted on the Web, and keep on expanding it with other recordings from new instruments.

Publication list

This thesis consists of several publications, listed in the following, along with their section locations of the thesis in which they appear, or partially appear.

Peer-reviewed journal article

Article 1 **Cazau, D.**, Revillon, G., Krywyk, J. and Adam, O. (*Accepted in JASA after minor revisions*). An investigation on Musical Acoustics knowledge in Automatic Music Transcription systems

Article 2 **Cazau, D.**, Wang, Y., Chemillier, M. and Adam, O. (In reviewing). An automatic music transcription system dedicated to the repertoires of the *marovany* zither from Madagascar, *Submitted to J. of New Music Research*

Article 3 **Cazau, D.**, Revillon, G., Wang, Y. and Adam, O. (In reviewing). Particle filtering for PLCA model, with an application to Automatic Music Transcription, *Submitted to IEEE, Trans. on Signal Processing*

Article 4 **Cazau, D.**, Revillon, G. and Adam, O. (2015). Deep scattering transform applied to onset detection and instrument recognition of decay instruments, *Submitted to Acta Acustica, Signal Processing*

Peer-reviewed conferences

Conference 1 **Cazau**, D.; Adam, O. and Chemillier, M. "Système de captation optique pour la transcription automatique de la musique de cithare malgache Marovany", Proc. JIM (JournÃľes d'Informatique Musicale), p. 51-58, Saint-Denis, France, May 2013.

Conference 2 **Cazau**, D.; Adam, O. and Chemillier, M. "Information retrieval of marovany zither music with an original optical-based system", Proc. DAFx, p. 1-6, Maynooth, Ireland, September 2013.

Conference 3 **Cazau**, D.; Adam, O. and Chemillier, M. "An original optical-based retrieval system applied to music automatic transcription of the marovany zither", Proc. FMA (Folk Music Analysis workshop), p. 44-50, Amsterdam, Netherlands, June 2013.

Conference 4 **Cazau**, D.; Adam, O. and Chemillier, M. "Etude comparative des timbres d'instruments Ãă cordes pincées occidentaux et traditionnels d'Afrique", Proc. Congress of the Acoustical Society of France, p. 1608, Poitiers, France, April 2014.

Conference 5 **Cazau**, D.; Adam, O. and Chemillier, M. "A study of contrametricity in traditional musical repertoire of Africa", Proc. FMA (Folk Music Analysis workshop), p. 100-101, Istanbul, Turkey, June 2014.

Conference 6 **Cazau, D.**, Adam, O. and Chemillier, M. "Automatic Music Transcription of the *marovany* repertoires, based on musical acoustics knowledge", Proc.

FMA (Folk Music Analysis workshop), Paris, France, June 10th-12th, 2015

Seminar

Seminar 1 **Cazau**, D. "Multichannel capturing sensory system for the *marovany* zither of Madagascar", Seminar at EHESS, Paris, 2014 May 14^{th}, video available on the web [4].

Seminar 2 **Cazau**, D. "Vers une automatisation de l'analyse ethnomusicologique ? Un bilan", Journées Doctorales d'Ethnomusicologie at Sorbonne, Paris, France, 2014 October 18^{th}.

Book chapter

Book 1 **Cazau**, D.; Chemillier, M. and Adam, O. (Accepted) "Computational Music Analysis approaches for automatic transcription of orally transmitted repertoires, with application to human-machine interaction environment", in book : Trends in Music Information Seeking, Behavior, and Retrieval for Creativity, Editors: Kostagiolas, P., Martzoukou, K. and Lavranos, C.

Technical note

Note 1 **Cazau**, D. and Nuel, G. "Understanding the Probabilistic Latent Component Analysis Framework"

Note 2 **Cazau**, D. and Nuel, G. "Relating sigmoid parameters to explicit musical features in view of HMM-based note segmentation"

Other research works

— Adam, O.; **Cazau, D.**; Gandilhon, N.; Fabre, B.; Laitman, J. T. and Reidenberg, J. S. (2013). New acoustic model for Humpback whale sound production, *Applied Acoustics*, 74, 1182-1190

— **Cazau, D.**; Adam, O.; Laitman, J. T. and Reidenberg, J. S. (2013). Understanding the intentional acoustic behavior of humpback whales: a production-based approach, *J. Acoust. Soc. Am.*, 134, 2268-2273

— Gandilhon, N.; Adam, O.; **Cazau, D.**; Laitman, J. T. and Reidenberg, J. S. (2014). Two new theoretical roles of the laryngeal sac of humpback whales, *Marine Mammal Science, DOI: 10.1111/mms.12187*

4. http://www.dailymotion.com/video/x1zkcla_kilema-ehess-5-7_music

Complementary research studies

The *marovany* in ImproteK

Some research (e.g. Dubnov (2003)) seek to capture some of the regularity apparent in the composition process by using statistical and information-theoretic tools to analyze musical pieces. The resulting models can be used for inference and prediction and, to a certain extent, to generate new works that imitate the style of the great masters. A generative theory of music can be constructed by explicitly coding music rules in some logic or formal grammar. HMMs have been used in this context of style modeling, allowing to build a computational representation of the musical surface that captures important stylistic features. Statistical analysis of a corpus reveals some of the possible recombinations that comply with the constraints or redundancies typically found in a particular style. Interesting applications include style characterization tools for the musicologist, generation of stylistic metadata for intelligent retrieval in musical databases, music generation for Web and game applications, machine improvisation with or without interaction with human performers, and computer-assisted composition. One of machine learning's main purposes is to create the capability to sensibly generalize.

While the real-time conversion of an acoustic instrument into a multichannel numerical signal provides a robust optimal method for audio acquisition and post-processing, it also opens original artistic prospects. The capacity of integrating real-time audio transformations (filtering, dynamic and timbre changing âĂę) in live performance has already found some interests in renowned cithara players like Rajery. Also, a MIDI conversion method of the cithara audio signal could also bring new applications in the field of musician-machine interaction systems. Actually this study is connected to the IMPROTECH project entitled "Technologies and Musical Improvisation" (grant ANR-09-SSOC-068 by the French National Research Agency) that includes a partnership with IRCAM. A computer environment called OMax, which learns in real time from human performers, has been developed in this context. The improvisation kernel is based on sequence modeling and statistical learning. It allows capturing stylistic musical surface rules in a manner that allows musically meaningful interaction between humans and computers. Recently an international workshop was held in New York **?** dedicated to the exploration of the links between musical improvisation and digital technologies that gathered researchers and artists from both research & creation scenes. Through this current study we aim to connect our optical-based MIDI conversion system developed for the *marovany* cithara to improvisation environments such as OMax, in order to explore the musical interest of such an interaction and help folk music to be represented in such events.

Study of the *marovany* repertoire in trance *tromba*, founded on musical criteria, and complementing other behavioral indices observed with an audio-visual device and audio data, should bring original elements of investigation to the fascinating relationships between music and trance. Another application theme of such a system would be the Human-

Machine musical interaction, through the OMAX improvisation IT environment Nika and Chemillier (2012) (developed from the OMax environment Assayag et al. (2010-2014) in collaboration with IRCAM). Future musical projects could involve malagasy musicians in this environment, using MIDI data from our retrieval system. Questions of a more aesthetic character (acceptability of musical formulas derived from a known repertoire, oral transmission of this skill, musical interest in the amplification, virtual re-orchestration of a musical environment and real-time modifications of musical parameters) will be considered in future investigations following this direction.

Research publications This thematic has been the subject of the following research publications:

Annex

Appendix A

Acoustic & Musical Descriptors

The features are calculated using a short time analysis window with duration 10-40 milliseconds. In addition, the means and variances of the features over a larger texture window (0.2-1.0 seconds) are computed resulting in a feature set with eight dimensions.

A system for the extraction of audio descriptors is usually organized according to the properties of the descriptors. We can distinguish three main properties of an audio descriptor: (1) the temporal extent over which the descriptor is computed (a specific region in time, such as the sustain, or the whole duration of a sound file), (2) the signal representation used to compute it (e.g., the waveform, the energy envelope or the short-term Fourier transform), and (3) the descriptor concept described by it (e.g., the description of the spectral envelope or the energy envelope over time) (Peeters et al., 2011). We first used classical descriptors covering both physical domains, temporal and spectral, allowing to quantify temporal variations of acoustic properties within a sound unit. Most of them have already been used in the acoustic characterization of mysticetes sound units (Fristrup and Watkins, 1992; Au et al., 2006; Mercado et al., 2010). They were selected for their physical meaning easily interpreted.

The temporal extent denotes the segment duration over which the descriptor is derived. A descriptor can either directly represent the whole sound event (e.g., the Attack Time descriptor, because there is only one attack in a sound sample) or represent a short-duration segment inside the event (e.g., the time-varying spectral centroid, which is derived from a spectral analysis of consecutive short-duration segments of a sound, usually of 60 ms duration).

Descriptors of the first group are called "global descriptors," and those of the second group are called "time-varying descriptors." Time-varying descriptors are extracted within each time frame of the sound and therefore form a sequence of values. In order to summarize the sequence in terms of a single value, we use descriptive statistics, such as minimum or maximum values, the mean or median, and the standard deviation or interquartile range (i.e., the difference between the 75th and 25th percentiles of the sequence of values). As such, the structure of an audio descriptor system usually separates the extraction of global descriptors (which are directly considered as the final results) from the extraction of time-varying descriptors (which are subsequently processed to derive the descriptive statistics).

A.1 Global descriptors

AT , the attack time (in s) is defined by the necessary time for the signal to reach 95 % of its maximal energy E_{max}

$$s(n = AT) = 0.95E_{max} \tag{A.1}$$

EffDur , the physical duration of the signal (in s) will be defined as the time during which the signal energy remains between 5 % and 95 % of its maximal energy E_{max}

$$EffDur = \{n/s(n) > 0.05E_{max} \ \& \ s(n) < 0.95E_{max}\} \tag{A.2}$$

DT , the decreasing time (in s) is defined by the necessary time for the signal to reach 5 % of its maximal energy E_{max} from this value

$$s(n = DT/n > AT) = 0.05E_{max} \tag{A.3}$$

TGC , the Temporal Gravity Center (in s) is the average value of the time energy distribution, which divides a temporal profile in two parts of equal energy :

$$TGC = \frac{\sum_{k=0}^{K} kEnv(k)}{\sum_{k=0}^{K} Env(k)} \tag{A.4}$$

with *Env(k)* the temporal envelop of the signal ;

AmpMod, the Amplitude Modulation quantifies energetic variations within a temporal profile :

$$AmpMod = \frac{\max_{0 \leq k \leq N} Env(k)}{\frac{1}{N} \sum_{k=1}^{N} Env(k)} \tag{A.5}$$

A value close to one traduces a stationary behavior of the signal, whereas strong values traduce transitory variations ;

FM, the Frequency Modulation quantifies energetic variations within a temporal profile :

$$FM = \frac{\max_{0 \leq k \leq N} F0(k)}{\frac{1}{N} \sum_{k=1}^{N} F0(k)} \tag{A.6}$$

with F_0 the temporal frequency curve. A value close to one traduces a stationary behavior of the signal, whereas strong values traduce transitory variations ;

Enveloppe , the temporal envelope e(tn) of the audio signal s(tn) is derived from the amplitude of the analytic signal sa(tn) given by the Hilbert transform of s(tn). This amplitude signal is then low-pass filtered using a third-order Butterworth filter with a cutoff frequency of 5 Hz. e(tn) has the same sampling rate and duration as that of s(tn).

A.2 Time-varying descriptors

It should be noted that, in our system, all window durations and hop sizes are defined in seconds and then converted to samples according to the sampling rate of the input audio signal. This guarantees that the same spectral resolution will be obtained whatever the sampling rate of the signal. However, the content of the representation itself will differ according to the sampling rate. This is because the upper frequency of the STFT depends on the sampling rate (it is equal to $f_{max}.s_r/2$).

All descriptors will be computed computed in a 0.049-s (1024 samples) hamming window at 0.01-s intervals with a 0.005-s overlapping. Classical speech analysis is usually conducted over such a duration.

Energy , the energetic level rms (in Pa) of a signal is defined by

$$E(k) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}|(x+kN)|^2} \tag{A.7}$$

computed for K successive frames of N samples. Given the difficulty of characterizing amplitude values of sounds recorded from free-ranging animals in an open acoustic environment (Au et al., 2006), as the amplitude of any field-recorded sound is necessarily affected by a number of sources of measurement error, including variations in the distance between source and microphone, in the relative orientations of source and microphone, in the effects of intervening sound barriers, and in ambient background noise levels, these figures on the energy must be only seen as relative indicators allowing to compare very roughly the intensity levels respective to each type.

Also, since microphone sensitivity (or clipping level) is often unavailable, sound levels could only be found as values relative to some arbitrary level, chosen such that the weakest click level corresponded to 0 dB. Hence, we report only "relative vocalization levels", by which we mean the difference between the current click level and the minimum click level (over all clicks) ;

HD, the Harmonicity Detector (adimensioned) is an indicator of harmonicity. The principle Youngmoo and Whitman (2002) is to automatically scan the spectral density of a signal with a comb filter whose fundamental frequency F_0 and varies within a given range of interest. When the valleys of this filter coincides with the peaks of an harmonic sequence for a particular F_0, their product will result in a very weak value which traduces the presence of an important harmonicity. Mathematically, we define it as

$$HD = min(\frac{E_{pond}}{E_{init}}) \tag{A.8}$$

with $E_{init} = \sum|Y(k)|^2$ and $E_{pond} = Filt(k,k_o)E_{init}$, where Filt is a comb filter defined as Filt $= 2(1-|cos(\frac{\pi F}{F_0})|)$.

SGC, the Spectral Gravity Center is the average value of the spectral energy distribution, which divides a spectral profile in two parts of equal energy :

$$SGC = \frac{\sum_{k=0}^{K} f_k a_k}{\sum_{k=0}^{K} a_k} \tag{A.9}$$

where a_k represents the spectral amplitude of the sample k, f_k its bin and K the total number of bins ;

ROF, the roll-off point estimates the amount of high frequency in the signal consists in finding the frequency such that a certain fraction (α_{ROF}) of the total energy is contained below that frequency. This ratio is fixed by default to .85 Tzanetakis and Cook (2002). We can define as

$$\sum_{k=0}^{ROF} a_k^2 = \alpha_{ROF}\sum_{k=0}^{Fe/2} a_k^2 \tag{A.10}$$

THD, the Total Harmonic Distorsion measures the energetic weigth of harmonics relatively to the fundamental one, defined by :

$$THD = \frac{\sqrt{(\sum_i^N H_i)^2)}}{H_0} \tag{A.11}$$

with N the number of harmonics with non-negligible amplitudes ;

Df, the formant dispersion (in Hz), which is the average distance between each adjacent pair of formants, was calculated using the following formula

$$D_f = \frac{\sum_i^N (F_{i+1} - F_i)}{N - 1} \tag{A.12}$$

where N is the total number of formants measured, and F_i is the frequency (in Hz) of the formant i ;

ZCR, the Zero Crossing Rate is a simple indicator of noisiness consisting in counting the number of times the signal crosses the X-axis (or, in other words, changes sign).

$Delta_F0$, this descriptor computes the difference between successive frames of the F_0 curve, i.e.

$$Delta_F0(i) = (F_0(i+1) - F_0(i)) \tag{A.13}$$

The sign of this descriptor gives the sense (upwards / backwards) of the sound modulation.

A.3 Musical descriptors and transforms

A.3.1 Key detection

Prior knowledge of the key results from an automatic detection based on a chroma-based frequency analysis Shenoy et al. (2004), performed on each training and test sequence.

A.3.2 Constant-Q transform

Appendix B

Musical sound representation

For humans the perceptual distance between 220 and 440 Hz is the same as between 440 and 880 Hz. A pitch representation that takes this logarithmic relation into account is more practical for some purposes. Luckily there are a few:

MIDI note number The midi standard defines note numbers from 0 to 127, inclusive. Normally only integers are used but any frequency f in Hz can be represented with a fractional note number n using this equation

$$n = 69 + 12.log_2(\frac{f}{440}) \tag{B.1}$$

$$n = 12.log_2(\frac{f}{r}) \tag{B.2}$$

with

$$r = \frac{440}{2^{69/12}} = 8.176Hz \tag{B.3}$$

Rewriting Equation (A1) to (A2) shows that midi note number 0 corresponds with a reference frequency of 8.176 Hz which is C_{-1} on a keyboard with A4 tuned to 440 Hz. It also shows that the midi standard divides the octave into 12 equal parts. To convert a midi note number n to a frequency f in Hz one of the following equations can be used.

$$f = 440.2^(n - 69/12) \tag{B.4}$$

$$f = r.2^(n/12) \tag{B.5}$$

Using pitch represented as fractional midi note numbers makes sense when working with midi instruments and midi data. Although the midi note numbering scheme seems oriented towards Western pitch organization (12 semitones) it is conceptually equal to the cent unit which is more widely used in ethnomusicology.

Cent Von Helmholtz and Ellis (1912) introduced the nowadays widely accepted cent unit. To convert a frequency f in Hz to a cent value c relative to a reference frequency r also in Hz:

$$c = 1200.log_2(\frac{f}{r}) \tag{B.6}$$

With the same reference frequency r Equations (A5) and (A2) differ only by a constant factor of exactly 100. In an environment with pitch representations in midi note numbers and cent values it is practical to use the standardized reference frequency of 8.176 Hz.

$$r = r.2^{c/2000} \tag{B.7}$$

Savart & Millioctaves Divide the octave in 301.5 and 1000 parts respectively, which is the only difference with cents.

Pitch ratio representation Pitch ratios are essentially pitch intervals, an interval of one octave, 1200 cents equal to a frequency ratio of 2/1. To convert a ratio t to a value in cent c:

$$c = 1200.\frac{ln(t)}{ln(2)} \tag{B.8}$$

The natural logarithm, the logarithm base e with e being Euler's number, is noted as ln. To convert a value in cent c to a ratio t:

$$t = \exp\frac{cln(2)}{1200} \tag{B.9}$$

Discussion of the different representations Further discussion on cents as pitch ratios can be be found in appendix B of Sethares (2005). There it is noted that: There are two reasons to prefer cents to ratios: Where cents are added, ratios are multiplied; and it is always obvious which of two intervals is larger when both are expressed in cents. For instance, an interval of a just fifth, followed by a just third is $(3/2)(5/4) = 15/8$, a just seventh. In cents, this is $702 + 386 = 1088$. Is this larger or smaller than the Pythagorean seventh 243/128? Knowing that the latter is 1110 cents makes the comparison obvious.

The cent unit is mostly used for pitch interval representation while the midi key and Hz units are used mainly to represent absolute pitch. The main difference between cent and fractional midi note numbers is the standardized reference frequency. In our software platform Tarsos we use the exact same standardized reference frequency of 8.176 Hz which enables us to use cents to represent absolute pitch and it makes conversion to midi note numbers trivial. Tarsos also uses cents to represent pitch intervals and ratios.

B.1 Table of main harmonic relations

Interval name	Size (semitones)	F_0 relation
Octave	12	2:1
Perfect fifth	7	3:2
Perfect fourth	5	4:3
Major third	4	5:4
Minor third	3	6:5
Major second	2	9:8

Appendix C

Filtering Particle

In the framework of Bayesian variable selection, Markov Chain Monte Carlo, or Particle filtering (PF), type approaches have been proposed Févotte and Godsill (2006a); Févotte et al. (2008). These methods consist in scanning the whole posterior distribution, making them more demanding than their EM-like counterparts, but which also, in return, offer increased robustness in convergence (i.e. reduced problems of convergence to local minima) and a complete Monte Carlo description of this parameter posterior density Févotte and Godsill (2006b,a); Févotte et al. (2008).

C.1 General Overview

Many problems in statistical signal processing Fong et al. (2002); Andrieu et al. (2003); Vermaak et al. (2000) can be stated in a state space form as follows,

$$x_{t+1} \sim f(x_{t+1}|x_t) \tag{C.1}$$

$$y_{t+1} \sim g(y_{t+1}|x_{t+1}) \tag{C.2}$$

where $\{x_t\}$ are unobserved states of the system and $\{y_t\}$ are observations made over some time, t. $f(.|.)$ and $g(.|.)$ are pre-specified state evolution and observation densities. A primary concern in many state-space inference problems is the sequential estimation of the filtering distribution $p(x_t|y_{1:t})$, and the simulation of the entire smoothing distribution $p(x_{1:t}|y_{1:t})$, where $y_{1:t} = (y_1, y_2, \cdots, y_t)$ and $x_{1:t} = (x_1, x_2, \cdots, x_t)$. Updating of the filtering distribution can be achieved, in principle, using the standard filtering recursions Robert and Casella (1999)

$$p(x_{t+1}|y_{1:t}) = \int p(x_t|y_{1:t}) f(x_{t+1}|x_t) dx_t \tag{C.3}$$

$$p(x_{t+1}|y_{1:t+1}) = \frac{g(y_{t+1}|x_{t+1}) p(x_{t+1}|y_{1:t})}{p(y_{t+1}|y_{1:t})} \tag{C.4}$$

Smoothing can also be performed recursively backwards in time using the smoothing formula Robert and Casella (1999)

$$p(x_t|y_{1:T}) = \int p(x_{t+1}|y_{1:T}) \frac{p(x_t|y_{1:t}) f(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})} dx_t \tag{C.5}$$

In practice, these filtering (eq. C.3) and smoothing (eq. C.5) computations can only be performed in closed form for linear Gaussian models using the Kalman filter / smoother, and for finite state-space hidden Markov models. In the case of non-linear non-Gaussian

models, there is no general analytic expression for the computations of these density functions. As a consequence, an approximation strategy is required to estimate the filtering and smoothing densities, which is commonly performed with the PF method, also known as sequential Monte Carlo methods. Within the PF framework, the filtering distribution is approximated with an empirical distribution formed from point masses also called particles,

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(x_t - x_t^{(i)}) \tag{C.6}$$

$$\sum_{i=1}^{N} w_t^{(i)} = 1, w_t^{(i)} \geq 0 \tag{C.7}$$

where $\delta(.)$ is the Dirac delta function and $w_t^{(i)}$ is a weight attached to particle $x_t^{(i)}$. Given this particle approximation to the posterior distribution, we can estimate the expected value of any function f $w.r.t$ the distribution $I(f,t)$, defined as $I(f_t) = \int f(x_t)p(x_t|y_{1:t})dx_t$, using the following Monte Carlo approximation

$$I(f_t) \approx \sum_{i=1}^{N} f(x_t^{(i)}) w_t(i) \tag{C.8}$$

Particle smoothers generate batched realisations of $p(x_{1:T}|y_{1:T})$ based on the forward PF results. In other words, the particle smoothers are an efficient method for generating realisations from the entire smoothing density $p(x_{1:T}|y_{1:T})$ using filtering approximation.

C.1.1 Filtering

We consider the filtering distribution $p(x_t|y_{1:t})$. Using the Bayes' rule, this distribution can be rewritten as follows,

$$\begin{aligned} p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:t-1}) && \text{(C.9)} \\ &\propto p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1}) && \text{(C.10)} \\ &\propto g(y_t|x_t)p(x_t|y_{1:t-1}) && \text{(C.11)} \\ &\propto \int g(y_t|x_t)f(x_t|x_{t-1})p(x_{1:t-1}|y_{1:t-1})dx_{1:t-1} && \text{(C.12)} \end{aligned}$$

Assuming that a particle approximation to $p(x_{1:t-1}|y_{1:t-1})$ has already been generated,

$$p(x_{1:t-1}|y_{1:t-1}) \approx \sum_{i=1}^{N} \delta(x_{1:t-1} - x_{1:t-1}^{(i)}) \tag{C.13}$$

Then, assuming that $f(x_t|x_{t-1})$ and $g(y_t|x_t)$ can be evaluated pointwise, we generate, for each state trajectory $x_{1:t-1}^{(i)}$, a random sample from a proposal distribution $q(x_t|x_{1:t-1}^{(i)}, y_{1:t})$. Then, the weights w_t of the filtering distribution (eq. C.6) can be approximated by

$$w_t^{(i)} \approx \frac{g(y_t|x_t^{(i)})f(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{1:t-1}^{(i)}, y_{1:t})} \tag{C.14}$$

Finally, we perform a multinomial resampling step, such that the probability that $x_t^{(i)}$ is selected is proportional to $w_t^{(i)}$, to obtain an unweighted approximate random draw

from the filtering distribution $p(x_t|y_{1:t})$. It is noteworthy that if the resampling step is forgotten, a degeneracy phenomenon can occur. Indeed, after a few iterations, all but one particle will have negligible weight. Doucet et al. (2000) has shown that the variance of the importance weights can only increase over time, and thus, it is impossible to avoid the degeneracy phenomenon. This degeneracy implies that a large computational effort is devoted to updating particles whose contribution is almost zero. As a result, a resampling step is needed to eliminate particles with small weights and generate a new set $\{x_t^{(i)}\}_i$, which is an i.i.d. (independent and identically distributed) sample from the approximate density $p(x_t|y_{1:t})$, with a resetting of the weights $\{w_t^{(i)}\}_i$ to $1/N$.

C.1.2 Smoothing

The entire smoothing density $p(x_{1:T}|y_{1:T})$ can be factorized as :

$$p(x_{1:T}|y_{1:T}) = p(x_T|y_{1:T}) \prod_{t=1}^{T-1} p(x_t|x_{t+1:T}, y_{1:T}) \tag{C.15}$$

Using the filter approximation (eq. C.6) to $p(x_t|y_{1:t})$ and the Markovian assumptions of the model, we can write,

$$\begin{aligned} p(x_t|x_{t+1:T}, y_{1:T}) &\propto p(x_t|y_{1:t}) f(x_{t+1}|x_t) \\ &\approx \sum_{i=1}^{N} w_{t|t+1}^{(i)} \delta(x_t - x_t^{(i)}) \end{aligned} \tag{C.16}$$

with the modified weights

$$w_{t|t+1}^{(i)} = \frac{w_t^{(i)} f(x_{t+1}|x_t^{(i)})}{\sum_{j=1}^{N} w_t^{(j)} f(x_{t+1}|x_t^{(j)})} \tag{C.17}$$

This revised particle distribution can be used to generate states successively in the reverse-time direction, conditioning upon future states.

C.2 Our contributing work

The main objective of this paper is to propose an alternative formulation of current PLCA models applied to audio signals, replacing the EM algorithm by a more generic parameter estimation algorithm based on a PF method. We call this new algorithm PLCA-PF in the following. The main advantage expected from this new algorithm is to be able to scan the whole parameter space so as to take into account any features of the parameters, and thus overcoming the limitations underlined in our introduction specific to current PLCA models. In regards to prior integration particularly, this new framework allows releasing the constraints on prior mathematical forms and number. This paves the way towards more complete modelings of the multi-faceted information carried by musical signals, covering both time (e.g. tempo and rhythm) and frequency (e.g. note spectra and chords) domains, and the different prior knowledge classes related to musicology, timbre and playing style.

List of Figures

List of Tables

Bibliography

Andrieu, C., Davy, M., and Doucet, A. (**2003**). "Efficient particle filtering for jump markov systems. application to time-varying autoregressions." IEEE Trans. on Signal Proc., **51**, 1762–1770.

Assayag, G. and Bloch, G. (**May 2008**). "Omax. the software improviser." In *Documentation version 2, IRCAM.*

Assayag, G., Bloch, G., Chemillier, M., Cont, A., and Dubnov, S. (**2010-2014**). "The omax project page." URL omax.ircam.fr.

Au, W.W.L., Pack, A.A., Lammers, M.O., Herman, L.M., Deakos, M.H., and Andrews, K. (**2006**). "Acoustic properties of humpback whale songs." J. Acoust. Soc. Am., **120**, 1103–1110.

Aucouturier, J.J. and Sandler, M. (**2001**). "Segmentation of musical signals using hidden markov models." In *Audio Engineering Society Convention 110th, Amsterdam, Netherlands.*

Baruch, C. and Drake, C. (**1997**). "Tempo discrimination in infants." Infant Behavior and Development, **20**, 573–577.

Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M.B. (**2005**). "A tutorial on onset detection in music signals." IEEE Trans. on Speech and Audio Proc., **13**, 1035–1047.

Bello, J.P., Daudet, L., and Sandler, M.B. (**2006**). "Automatic piano transcription using frequency and time-domain information." IEEE Trans. Audio Speech Lang Proc., **14**, 2242–2251.

Bello, J.P., Duxbury, C., Davies, M., and Sanders, M.B. (**2004**). "On the use of phase and energy for musical onset detection in the complex domain." IEEE Signal Proc. Letters, **11**, 553–556.

Bello, J.P. and Pickens, J. (**2005**). "A robust mid-level representation for harmonic content in music signal." In *6th International Conference on Music Information Retrieval, London, UK.* pp. 304–311.

Benetos, E., Badeau, R., Weyde, T., and Richard, G. (**2014**a). "Template adaptation for improving automatic music transcription." In *15th International Society for Music Information Retrieval Conference, Taipei, Taiwan.* pp. 175–180.

Benetos, E., Cherla, S., and Weyde, T. (**2013**a). "An efficient shift-invariant model for polyphonic music transcription." In *6th Int. Workshop on Machine Learning and Music, Prague, Czech Republic.*

Benetos, E. and Dixon, S. (**2011**). "Multiple-instrument polyphonic music trancription using a convolutive probabilistic model." In *Proc. 8th Sound and Music Computing Conf.* pp. 19–24.

——— (**2013**). "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model." J. Acoust. Soc. Am., **133**, 1727–1741.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A. (**2013**b). "Automatic music transcription: Challenges and future directions." J. of Intelligent Information Systems, **41**, 407–434.

Benetos, E. and Holzapfel, A. (**2013**). "Automatic transcription of turkish makam music." In *14th International Society for Music Information Retrieval Conference, Curitiba, PR, Brazil*. pp. 355–360.

Benetos, E., Jansson, A., and Weyde, T. (**2014**b). "Improving automatic music transcription through key detection." In *AES 53RD INTERNATIONAL CONFERENCE, London, UK*.

Benning, M.S., Kapur, A., Till, B.C., and Tzanetakis, G. (**2007**). "Multimodal sensor analysis of sitar performance: Where is the beat?" In *IEEE 9th Workshop on Multimedia Signal Processing*. pp. 74–77.

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

Blacking, J. (**1979**). *The Performing Arts: Music and Dance* (The Hague: Mouton De Gruyter.), chap. The Study of Man as Music-Maker, pp. 3–15.

Canadas, F., Vera, P., Ruiz, N., Mata, R., and Carabias, J. (**2008**). "Note-event detection in polyphonic musical signals based on harmonic matching pursuit and spectral smoothness." Journal of New Music Research, **37**, 167–183.

Carterette, E. and Kendall, R. (**1999**). *The Psychology of Music* (San Diego, CA: Academic Press.), chap. Comparative Music Perception and Cognition, pp. 725–791.

Cemgil, A.T. (**2004**). "Bayesian music transcription." Ph.D. thesis, Radboud University of Nijmegen.

Chabassier, J. (**2012**). "Modélisation et simulation numérique d'un piano par modéles physiques." Ph.D. thesis, Ecole Polytechnique.

Cheng, T., Dixon, S., and Mauch, M. (**2013**). "A deterministic annealing algorithm for automatic music transcription." In *14th International Society for Music Information Retrieval Conference, Curitiba, PR, Brazil*.

Ching, W.K., Ng, M.N., and Fung, E.S. (**2008**). "Higher-order multivariate markov chains and their applications." Linear Algebra and its Applications, **428**, 492–507.

Ching, W.K.e.a. (**2004**). "Higher-order markov chain models for categorical data sequences." Naval Research Logistics (Wiley Periodicals, Inc.), **51**, 557–574.

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.*

Collins, N. (**2005**). "Using a pitch detector for onset detection." In *6th International Conference on Music Information Retrieval, London, UK*. pp. 100–106.

CompMusic (**2011-2017**). "Computational models for the discovery of the world's music." URL http://compmusic.upf.edu/.

Cooke, M. (**2006**). "A glimpsing model of speech recognition in noise." J. Acoust. Soc. Am., **119**, 1562–1573.

Dannenberg, R. (**1984**). "An online algorithm for real-time accompaniment." In *ICMC, p. 193-198.*

de Cheveigné, A. and Kawahara, H. (**2002**). "Yin, a fundamental frequency estimator for speech and music." J. Acoust. Soc. Am., **111**, 1917–1930.

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, Series B, **39**, 1–38.

Dessein, A., Cont, A., and Lemaitre, G. (**2010**). "Real-time polyphonic music transcription with nonnegative matrix factorization and beta-divergence." In *11th International Society for Music Information Retrieval Conference, Utretcht, Netherlands*. pp. 489–494.

DIADEMS (**2012-2015**). "Description, indexation, access to sound and ethnomusicological documents." URL http://www.irit.fr/recherches/SAMOVA/DIADEMS/fr/welcome/&cultureKey=en.

Domenichini, M. (**1984**). *Stanley Dani* (London:Macmillan), chap. Valiha, pp. 705–706, vol. 3.

Doucet, A., Godsill, S., and Andrieu, C. (**2000**). "On sequential monte carlo sampling methods for bayesian filtering." Statistics and Computing, **10**, 197–208.

Downie, J.S., Byrd, D., and Crawford, T. (**2009**). "Ten years of ismir: reflections on challenges and opportunities." In *ISMIR*.

Drake, C. (**1998**). "Psychological processes involved in the temporal organization of complex auditory sequences: Universal and acquired processes." Music Perception, **16**, 11–26.

Dubnov, S., e.a. (**2003**). "Using machine-learning methods for musical style modeling." Computer, **36 (10)**, 73–80.

Ellis, D.P.W. (**1996**). "Prediction-driven computational auditory scene analysis." Ph.D. thesis, Massachusetts Institute of Technology.

——— (**2007**). "Beat tracking by dynamic programming." Tech. rep., LabROSA, Columbia University, New York.

Emiya, V., Badeau, R., and Richard, G. (**2010**). "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle." IEEE Trans. on Audio, Speech, Lang. Proc., **18**, 1643–1654.

Engelmore, R. and Morgan, A., eds. (**1988**). *Blackboard Systems* (Addison-Wesley Longman Publishing (Boston, MA, USA)), 602 pp.

Ephraim, Y. and Malah, D. (**1985**). "Speech enhancement using a minmum-mean square error log-spectral amplitude estimator." IEEE Trans. on Acoustics, Speech and Signal Proc., **33**, 443–445.

Ewert, S. and Muller, M. (**2011**). "Estimating note intensities in music recordings." In *ICASSP*. pp. 385–388.

Févotte, C. and Godsill, S. (**2006**a). "A bayesian approach for blind separation of sparse sources." IEEE Trans. Audio Speech Language Processing, **14**, 2174–2188.

——— (**2006**b). "Sparse linear regression in unions of bases via bayesian variable selection." IEEE Signal Processing Letters, **13**, 441–444.

Févotte, C., Torrésani, B., Daudet, L., and Godsill, S. (**2008**). "Sparse linear regression with structured priors and application to denoising of musical audio." IEEE Trans. Audio Speech Language Processing, **16**, 174–185.

Fletcher, N.F. and Rossing, T.D. (**1998**). *The Physics of Musical Instruments* (New York: Springer), 613 pp.

Fong, W., Godsill, S., Doucet, A., and West, M. (**2002**). "Monte carlo smoothing with application to audio signal enhancement." IEEE Transactions on Signal Processing, **50**, 438–449.

Fonseca, N. and Ferreira, A. (**2009**). "Measuring music transcription results based on a hybrid decay/sustain evaluation." In *7th Triennial Conference of European Society for the Cognitive Sciences of Music, Jyvaskyla, Finland*. pp. 119–124.

Fristrup, K.M. and Watkins, W.A. (**1992**). "Characterizing acoustic features of marine animal sounds." Technical Report.

Fuentes, B., Badeau, R., and Richard, G. (**2013**). "Harmonic adaptive latent component analysis of audio and application to music transcription." IEEE Trans. on Audio Speech Lang. Processing, **21**, 1854–1866.

Gainza, M., Lawlor, R., Coyle, E., and Kelleher, A. (**2004**). "Onset detection and music transcription for the irish tin whistle." In *ISSC, Belfast, June 30 - July 2*.

Gedik, C. (**2012**). "Automatic transcription of tradional turkish art music recordings : a computational ethnomusicology approach." Ph.D. thesis, Engineering and Sciences of Izmir Institute of Technology.

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.).

Godsmark, D. and Brown, G.J. (**1999**). "A blackboard architecture for computational auditory scene analysis." Speech Communication, **27**, 351–366.

Gomez, E. (**2006**). "Tonal description of music audio signals." Ph.D. thesis, Universitat Pompeu Fabra for the program in computer science and digital communication.

Gomez, E. and Herrera, P. (**2008**). "Comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction." Empirical Musicology Review, **3**, 140–156.

Goto, M. (**2004**). "A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-word audio signals." Speech Communications, **43**, 311–329.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (**2003**). "Rwc music database: Popular, classical, and jazz music databases." In *3rd International Conference on Music Information Retrieval, Baltimore,MD.* pp. 287–288.

Grindlay, G. and Ellis, D.P.W. (**2010**). "A probabilistic subspace model for multi-instrument polyphonic transcripton." In *11th International Society for Music Information Retrieval Conference, Utrecht, Netherlands.*

——— (**2011**). "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments." IEEE J. Sel. Topics Signal Proc., **5**, 1159–1169.

Grosche, P., Schuller, B., MÃijller, M., and Rigoll, G. (**2012**). "Automatic transcription of recorded music." Acta Acustica, **98**, 199–215.

Hainsworth, S.W. (**2004**). "Techniques for the automated analysis of musical audio." Ph.D. thesis, University of Cambridge.

Handel, S. (**1995**). *Hearing* (Oxford Academic Press), chap. Timbre Perception and Auditory Object Identification, pp. 425–459.

He, Y. (**1988**). "Extended viterbi algorithm for second order hidden markov process." In *9th International Conference on Pattern Recognition, Roma, Italia.* pp. 718–720.

Heittola, T., Klapuri, A., and Virtanen, T. (**2009**). "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." In *ISMIR, Kobe, Japan.* pp. 327–332.

Hennequin, R., Badeau, R., and David, B. (**2010**). "Time-dependent parametric and harmonic templates in non-negative matrix factorization." In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria , September 6-10, 2010.*

Hennequin, R., David, B., and Badeau, R. (**2011**). "Score informed audio source separation using a parametric model of non-negative spectrogram." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic.* pp. 45–48.

Hoffman, M.D., Blei, D.M., and Cook, P.R. (**2009**). "Finding latent sources in recorded music with a shift-invariant hdp." In *12th Int. Conference on Digital Audio Effects (DAFx-09), Como, Italy.*

Hofmann, T. (**1999**). "Probabilistic latent semantic indexing." In *22th Annual International SIGIR Conference on Research and Development in Information Retrieval.*

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

Hu, D. and Saul, L.K. (**2009**). "A probabilistic topic model for unsupervised learning of musical key-profiles." In *10th International Society for Music Information Retrieval Conference*. pp. 441–446.

Itoyama, K.e.a. (**2008**). "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models." In *ISMIR*.

Iverson, P. and Krumhansl, C.L. (**1993**). "Isolating the dynamic attributes of musical timbre." J. Acoust. Soc. Am., **94**, 2595–2603.

Kameoka, H., Nishimoto, S., and Sagayama, S. (**2007**). "A multipitch analyzer based on harmonic temporal structured clustering." IEEE Trans. on Audio, Speech Lang. Proc., **15**, 982–994.

Kapur, A. (**2002**). "Digitizing north indian music: Preservation and extension using multimodal sensor systems, machine learning and robotics." Ph.D. thesis, B.S.E., Princeton University.

Kay, S.M. (**1998**). *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory* (Prentice Hall).

Kennedy, M. and Bourne, J. (**1996**). *The Concise Oxford Dictionary of Music* (Encyclopedia.com), 1011 pp.

Kirchhoff, H., Dixon, S., and Klapuri, A. (**2012**). "Shift-variant nonnegative matrix deconvolution for music transcription." In *IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan*. pp. 125–128.

Klapuri, A. (**2003**). "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness." IEEE Trans. Speech and Audio Proc., **11**, 804–816.

——— (**2004**a). "Automatic music transcription as we know it today." J. of New Music Research, **33**, 269–282.

——— (**2004**b). "Signal processing methods for the automatic transcription of music." Ph.D. thesis, Tampere University of Technology.

——— (**2008**). "Multipitch analysis of polyphonic music and speech signals using an auditory model." IEEE Trans. on Audio, Speech Lang. Proc., **16**, 255–266.

Klapuri, A.P., Eronen, A.J., and Astola, J.T. (**2004**). "Analysis of the meter of acoustic musical signals." IEEE Trans. on Acoustics Speech and Signal Proc., **14**, 342–355.

Kranenburg, v.P. (**2010**). "A computational approach to content-based retrieval of folk song melodies." Ph.D. thesis, University of Utretcht.

Krumhansl, C.L. (**1990**). *Cognitive foundations of musical pitch* (Oxford University Press), chap. A key-finding algorithm based on tonal hierarchies, pp. 77–110.

Krumhansl, C.L. and Shepard, R. (**1979**). "Quantification of the hierarchy of tonal functions within a diatonic context." Journal of Experimental Psychology: Human Perception and Performance, **5**, 579–594.

Lartillot, O., Toiviainen, P., and Eerola, T. (**2008**). "Commentary on "comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction" by emilia gomez, and perfecto herrera." Empirical Musicology Review, **3**, 157–160.

Lee, K. and Slaney, M. (**2008**). "Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio." IEEE Trans. Audio Speech Lang Proc., **16**, 291–301.

Lerdahl, F. and Jackendoff, G. (**1983**). *A generative theory of tonal music* (MIT Press Cambridge).

Leveau, P., Vincent, E., Richard, G., and Daudet, L. (**2008**). "Instrument-specific harmonic atoms for mid-level music representation." IEEE Trans. Audio, Speech, Lang. Process.,, **16**, 116–128.

Lew, M.S., Sebe, N., Djeraba, C., and Jain, R. (**2006**). "Content-based multimedia information retrieval: state of the art and challenges." ACM Transactions on Multimedia Computing, Communications and Applications, **2 (1)**, 1–19.

Lewis, G. (**2000**). "Too many notes : Computers, complexity and culture in voyager." Leonardo Music J., **10**, 33–39.

Li, Y. and Wang, D. (**2007**). "Pitch detection in polyphonic music using instrument tone models." In *Proc. IEEE Conf. Acoust., Speech, Signal Process, Honolulu, Hawaii*. pp. 481–484.

Lidy, T., Silla, C.N., Cornelis, O., Gouyon, F., Rauber, A., Kaestner, C.A.A., and Koerich, A.L. (**2010**). "On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-western and ethnic music collections." Signal Proc., **90**, 1032–1048.

Logan, B. and Chu, S. (**2000**). "Music summarization using key phrases." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. pp. 749–752.

Machover, T. and Chung, J. (**1989**). "Hyperinstruments : musically intelligent and interactive performance and creativity systems." In *ICMC*.

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

Mari, J.F., Haton, J.P., and Kriouile, A. (**1997**). "Automatic word recognition based on second-order hidden markov models." IEEE Transactions on Speech and Audio Processing, **5**, 22–25.

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

Marolt, M., Kavcic, A., and Privosnik, M. (**2002**). "Neural networks for note onset detection in piano music." In *International Computer Music Conference, Gothenberg, Sweden*.

Martin, K.D. (**1999**). "Sound-source recognition: A theory and computational model." Ph.D. thesis, MIT, USA.

Mauch, M. and Ewert, S. (**2013**). "The audio degradation toolbox and its application to robustness evaluation." In *14th International Society for Music Information Retrieval Conference, Curitiba, PR, Brazil*. pp. 83–88.

Mayor, O., Bonada, J., and Loscos, A. (**2009**). "Performance analysis and scoring of the singing voice." In *AES 35th International Conference: Audio for Games.*

McLachlan, G. and Basford, K.E. (**1988**). *Mixture Models* (Marcel Dekker, INC, New York Basel).

Meddis, R. and Hewitt, M.J. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification." J. Acoust. Soc. Am., **89**, 2866–2882.

Mercado, E., Schneider, J., Pack, A.A., and Herman, L.M. (**2010**). "Sound production by singing humpback whales." J. Acoust. Soc. Am., **127**, 2678–2691.

MIDIwebPage (**2015**). "Midi tuning standard." URL http://www.microtonal-synthesis.com/MIDItuning.html.

MIREX (**2011**). "Music information retrieval evaluation exchange (mirex)." Available at http://music-ir.org/mirexwiki/ (date last viewed January 9, 2015).

Moelants, D., Cornelis, O., Leman, M., Gansemans, J., Caluwe, R.T., Tré, G.D., Matthé, T., and Hallez, A. (**2006**). "Problems and opportunities of applying data- & audio-mining techniques to ethnic music." In *7th International Conference on Music Information Retrieval, Victoria, Canada*. pp. 334–336.

——— (**2007**). "The problems and opportunities of content-based analysis and description of ethnic music." International J. of Intangible Heritage, **2**, 59–67.

Moore, B.C.J. (**1997**). *An Introduction to the Psychology of Hearing* (New York: Academic), 441 pp.

Moorer, J.A. (**1977**). "On the transcription of musical sound by computer." Computer Music Journal, **1**, 32–38.

Mowat, W. (**2005**). "Bob moog piano bar : Midi output device for acoustic pianos." In *Sound On Sound, available at www.soundonsound.com/sos/mar05/\articles/moogpianobar.html*.

Mysore, G.J. (**2010**). "A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures." Ph.D. thesis, Stanford University, USA.

Mysore, G.J. and Smaragdis, P. (**2009**). "Relative pitch estimation of multiple instruments." In *International Conference on Acoustical Speech and Signal Processing, Taipei, Taiwan*. pp. 313–316.

Nakano, M.e.a. (**2010**). "Nonnegative matrix factorization with markov-chained bases for modeling time-varying patterns in music spectrograms." In *LVA/ICA 2010, LNCS 6365, V. Vigneron et al. (Eds.)*. pp. 149–156.

Newman, M.E. and Barkema, G.T. (**1999**). *Monte Carlo Methods in Statistical Physics* (USA: Oxford University Press).

Newton, M.J. and Smith, L.S. (**2012**). "A neurally inspired musical instrument classification system based upon the sound onset." J. Acoust. Soc. Am., **131**, 4785–4798.

Niedermayer, B. (**2008**). "Non-negative matrix division for the automatic transcription of polyphonic music." In *Proc. of the 9th Int. Soc. for Music Information Retrieval conference (ISMIR)*. pp. 544–549.

Nika, J. and Chemillier, M. (**2012**). "Improtek, integrating harmonic controls into improvisation in the filiation of omax." In *International Computer Music Conference (ICMC), Ljubljana, Slovenia*. pp. 180–187.

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

Ockelford, A. (**2005**). *Repetition in Music: Theoretical and Metatheoretical Perspectives* (Farnham, U.K.: Ashgate).

O'Grady, P.D. and Rickard, S.T. (**2009**). "Automatic hexaphonic guitar transcription using non-negative constraints." In *Signals and Systems Conference (ISSC), IET Irish*. pp. 1–6.

Ozerov, A., Vincent, E., and Bimbot, F. (**2012**). "A general flexible framework for the handling of prior information in audio source separation." IEEE Trans. Audio Speech Lang Proc., **20**, 1118–1132.

Pachet, F. (**2002**). "The continuator : musical interaction with style." In *ICMA, Gotheberg, Sweden*.

Papadopoulos, H. and Peeters, G. (**2007**). "Large-scale study of chord estimation algorithms based on chroma representation and hmm." In *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI) (Bordeaux, France)*. pp. 53–60.

Peeters, G., Giordano, B.L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The timbre toolbox: Extracting audio descriptors from musical signals." J. Acoust. Soc. Am., **130**, 2902–2916.

Piszczalski, M. and Galler, B. (**1977**). "Automatic music transcription." Computer Music Journal, **1**, 24–31.

Poliner, G. and Ellis, D. (**2007**). "A discriminative model for polyphonic piano transcription." J. on Advances in Signal Proc., **8**, 1–9.

Rabiner, L.R. (**1989**). "A tutorial on hidden markov models and selected applications in speech recognition." Proc. of the IEEE, **77**, 257–286.

Raphael, C. (**1999**). "Automatic segmentation of acoustic musical signals using hidden markov models." IEEE Transactions on Pattern Analysis and Machine Intelligence, **21**, 360–370.

Raphael, C. and Stoddard, J. (**2003**). "Harmonic analysis with probabilistic graphical models."

Razafindrakoto, J. (**1999**). "Le timbre dans le répertoire de la valiha, cithare tubulaire de madagascar." Cahiers d'ethnomusicologie, **2**, 2–16.

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

Robert, C.P. and Casella, G. (**1999**). *Monte Carlo Statistical Methods* (Springer Science+Business Media New York).

Roberts, G.O., Gelman, A., and Gilks, W.R. (**1997**). "Weak convergence and optimal scaling of random walk metropolis algorithms." Ann. Appl. Probab., **7 (1)**, 110–120.

Robinson, K. and Patterson, R.D. (**1995**). "The duration required to identify the instrument, the octave, or the pitch chroma of a musical note." Music Perception, **13**, 1–15.

Rossing, T.D. (**2010**). *The Science of String Instruments* (Springer-Verlag New-York), 470 pp.

Rouiller, E. (**1997**). *The Central Auditory System* (Oxford University Press, Oxford, UK), chap. Chap. 1 Functional organization of the auditory pathways.

Rowe, R. (**2004**). *Machine Musicianship* (Cambridge, MA : MIT Press).

Ruwet, N. and Everist, M. (**1987**). "Methods of analysis in musicology." Music Anal., **6**, 9–36.

Ryynanen, M. and Klapuri, A. (**2005**). "Polyphonic music trancription usng note event modeling." In *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics, New Paltz, NY*. pp. 319–322.

Ryynanen, M.P. and Klapuri, A.P. (**2008**). "Automatic transcription of melody, bass line, and chords in polyphonic music." Computer Music J., **32**, 72–86.

Saito, S., Kameoka, H., Takahashi, K., Nishimoto, T., and Sagayama, S. (**2008**). "Specmurt analysis of polyphonic music signals." IEEE Transactions on Audio, Speech and Language Processing, **16**, 639–650.

Schenker, H. (**1954**). *Harmony* (Chicago, IL: Univ. of Chicago Press).

Seeger, C. (**1958**). "Prescriptive and descriptive music writing." The Musical Quaterly, **44**, 184–195.

Sethares, W. (**2005**). *Tuning timbre spectrum scale* ((2nd ed.). Berlin: Springer.).

Seydoux, L. (**2012**). "Mesure de déplacement de cordes avec des fourches optiques." Master's thesis, Mémoire de fin d'études, LAM (Laboratoire Acoustique Musicale).

Shashanka, M., Raj, B., and Smaragdis, P. (**2008**). "Sparse overcomplete latent variable decomposition of counts data." In *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Inc.), pp. 1313–1320. URL http://papers.nips.cc/paper/3235-sparse-overcomplete-latent-variable-decomposition-of-counts-data.pdf.

Shenoy, A., Mohaptra, R., and Wang, Y. (**2004**). "Key determination of acoustic musical signals." In *IEEE International Conference on Multimedia and Expo, Taipei, Taiwan*. pp. 1771–1774.

Sigtia, S., Benetos, E., Boulanger, N., Garcez, A., Weyde, T., , and Dixon, S. (**2015**). "A hybrid recurrent neural network for music trancription." In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Sigtia, S., Benetos, E., Cherla, S., Weyde, T., Garcez, A., and Dixon, S. (**2014**). "An rnn-based music language model for improving automatic music trancription." In *15th International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan*.

Singer, E., Larke, K., and Bianciardi, D. (**2003**). "Lemur guitarbot: Midi robotic string instrument." In *NIME '03 Proceedings of the 2003 conference on New interfaces for musical expression*. pp. 188–191.

Smaragdis, P. and Mysore, G.J. (**2009**). "Separation by "humming": user-guided sound extraction from monophonic mixtures." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA*. pp. 69–72.

Smaragdis, P., Raj, B., and Shanshanka, M. (**2006**). "A probabilistic latent variable model for acoustic modeling." In *Neural Information Proc. Systems Workshop, Whistler, BC, Canada*.

Smaragdis, P., Raj, B., and Shashanka, M. (**2007**). "Supervised and semi-supervised separation of sounds from single-channel mixtures." Independent Component Analysis and Signal Separation, Lecture Notes in Computer Science, **4666**, 414–421.

——— (**2008**). "Sparse and shift-invariant feature extraction from non-negative data." In *International Conference Acoustical Speech and Signal Processing, Las Vegas, NV*. pp. 2069–2072.

Smaragdis, P., Shashanka, M., Raj, B., and Mysore, G. (**2009**). "Probabilistic factorization of non-negative data with entropic cooccurrence constraints." In *ICA*.

Takegawa, Y., Terada, T., and Nishio, S. (**2006**). "Design and implementation of a real-time fingering detection system for piano performance." In *International Computer Music Conference*. pp. 67–74.

Temperley, D. (**2002**). *Music and Artificial Intelligence* (Berlin, Germany: Springer), chap. A Bayesian approach to key finding, pp. 195–206.

Tidhar, D., Mauch, M., and Dixon, S. (**2010**). "High precision frequency estimation for harpsichord tuning classification." In *IEEE Int. Conf. Audio, Speech and Signal Processing, Dallas, USA*.

Tolonen, T. and Karjalainen, M. (**2000**). "A computationally efficient multipitch analysis model." IEEE Trans. on speech and audio processing, **8**, 708–716.

Typke, R., Wiering, F., and Veltkamp, R.C. (**2005**). "A survey of music information retrieval systems." In *ISMIR, Queen Mary, University of London*. pp. 153–160.

Tzanetakis, G. and Cook, P. (**2002**). "Musical genre classification of audio signals." IEEE Trans. Audio Speech Lang Proc., **10**, 293–302.

Tzanetakis, G., Kapur, A., Schloss, W.A., and Wright, M. (**2007**). "Computational ethnomusicology." J. of interdisciplinary music studies, **1**, 1–24.

Ueda, N. and Nakano, R. (**1998**). "Deterministic annealing em algorithm." Neural Networks, **11**, 271–282.

Unal, E., Chew, E., Georgiou, P.G., and Narayanan, S.S. (**2008**). "Challenging uncertainty in query by humming systems: A fingerprinting approach." IEEE Trans. Audio Speech Lang Proc., **16**, 359–371.

Vassilakis, P. (**1999**). "Chords as spectra, harmony as timbre." In *138th meeting of the Acoustical Society of America*.

Vermaak, J., Andrieu, C., and Doucet, A. (**2000**). "Particle filtering for non-stationary speech modelling and enhancement." In *6th International Conference on Spoken Language Processing*. pp. 594–597.

Von Helmholtz, H. and Ellis, A.J. (**1912**). *On the sensations of tone as a physiological basis for the theory of music* ((translated and expanded by Alexander J. Ellis, 2nd. (English ed.). London: Longmans Green.).

Wang, C.K., Lyu, R.Y., and Chiang, Y.C. (**2003**). "A robust singing melody tracker using adaptive round semitones (ars)." In *Proceedings of 3rd International Symposium on Image and Signal Processing and Analysis (ISPA03)*. pp. 18–20.

Weinberg, G., Driscoll, S., and Parry, M. (**2005**). "Haile - a perceptual robotic percussionist." In *ICMC*.

——— (**2006**). "Jam'aa - a middle eastern percussion ensemble for human and robotic players." In *ICMC*.

Wikipedia (**2015**). "Wikipedia page on midi standard." URL https://en.wikipedia.org/wiki/MIDI#cite_note-1.

Witten, I.H. and Bell, T.C. (**1991**). "The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression." IEEE Transactions on Information Theory, **37**, 1085–1094.

Ye, J.a. (**2014**). "Robust acoustic feature extraction for sound classification based on noise reduction." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Florence, Italy*.

Yeh, C., Roebel, A., and Rodet, X. (**2005**). "Multiple fundamental frequency estimation of polyphonic music signals." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, USA*.

Youngmoo, E.K. and Whitman, B. (**2002**). "Singer identification in popular music recordings using voice coding features." In *3rd International Conference on Music Information Retrieval, Paris, France, October 13-17*.

Zwicker, E. and Fastl, H. (**1999**). *Psychoacoustics: facts and models* (Springer series in information sciences, 22. Springer, Berlin ; New York, 2nd updated edition.).