A roadmap towards versatile MIR
Emmanuel Vincent, Stanislaw A. Raczynski, Nobutaka Ono, Shigeki Sagayama

To cite this version:

HAL Id: inria-00544047
https://inria.hal.science/inria-00544047
Submitted on 7 Dec 2010

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.
A ROADMAP TOWARDS VERSATILE MIR

Emmanuel Vincent  Stanislaw A. Raczyński, Nobutaka Ono, Shigeki Sagayama
INRIA The University of Tokyo

ABSTRACT

Most MIR systems are specifically designed for one application and one cultural context and suffer from the semantic gap between the data and the application. Advances in the theory of Bayesian language and information processing enable the vision of a versatile, meaningful and accurate MIR system integrating all levels of information. We propose a roadmap to collectively achieve this vision.

1. INTRODUCTION

MIR has the vocation of covering all music and all music-related applications, e.g., transcription, structuration, alignment, tagging, personalization, composition and interaction. Yet, most systems to date are designed for one class of applications and one cultural context, namely Western popular music, which limits their reusability and their meaningfulness in the sense of [12]. In addition, most systems rely on general pattern recognition techniques applied onto a bag of low-level features, which bounds their accuracy to some glass ceiling. A system integrating all levels of information would make it possible to address virtually any application on any data in a versatile, meaningful and accurate fashion. For instance, it would enable much higher-level interaction, e.g. changing the music genre of a recording without affecting some of its other features. While many share the vision of this complete system [1], no fully satisfying approach has yet been proposed to achieve it.

One integration approach adopted e.g. by the NEMA project or by [5] is to queue several audio and symbolic feature extraction modules, so as to compute higher-level “features of features”. This bottom-up approach greatly improves accuracy but will eventually reach a glass ceiling too due to the propagation of errors from one processing stage to the next. Also, it is not fully versatile since each application requires the implementation of a specific workflow and the inputs and outputs of a module cannot be swapped otherwise than by developing a new module. Top-down approaches based on probabilistic graphical models address these issues by estimating the hidden features best accounting for the observed features [10]. All applications then amount to inferring and possibly manipulating some of the hidden features, without changing the model nor the general decoding algorithm. For instance, small graphical models such as Hidden Markov Models (HMMs) integrating harmony and pitch are routinely used to infer the most probable chord sequence given a set of MIDI notes [8] or conversely to generate the most probable melody given a chord sequence [2] using the general Viterbi algorithm.

It is common belief that the formalism of graphical models has the potential to yield a versatile and accurate MIR system by integrating more and more features into a hierarchical model. Yet, this formalism alone does not suffice to achieve this vision, as challenging issues pertaining to the definition of the model structure, the parameterization of conditional distributions, the design of unsupervised learning algorithms and the collection of development data have often been overlooked. In this paper, we explicitly state and clarify these issues and derive a feasible roadmap.

2. COMPLETE MODEL STRUCTURE

The first task is to collectively define a taxonomy of music features and their statistical dependencies that virtually covers all existing and future music. This involves the following steps: building an exhaustive list of music features and their definition, identifying features amenable to the same theoretical treatment, e.g. genre and mood, and organizing these features into a dependency network. This is not straightforward, since the current MPEG-7 standard mostly addresses low-level or application-specific features and agreed-upon definitions of features such as musical structure or rhythm remain to be found. Also, the dependency network is not unique and a sparse network is preferred.

We propose a draft model of a music piece as a dynamic Bayesian network in Figure 1. While it may be incomplete, we believe that it provides a useful basis for community discussion. In this graph, each node represents a sequence of uni- or multidimensional features considered as a vector random variable. Statistical dependencies are indicated by arrows such that the conditional distribution of a variable given its ancestors depends on its parents only. “Vertical” hierarchical dependencies are explicitly displayed, while “horizontal” temporal dependencies within and between nodes are implicitly accounted for. We adopt a generative modeling point of view [10] where lower-level features depend on higher-level features. The joint distribution of all variables then factors as the product of the distribution of each variable given its parents [10].
A wide application range is ensured by allowing the feature value sets $\mathcal{V}_{\text{sect}}$, $\mathcal{V}_{\text{meter}}$, $\mathcal{V}_{\text{tempo}}$, $\mathcal{V}_{\text{loud}}$, $\mathcal{V}_{\text{chord}}$, $\mathcal{V}_{\text{inst}}$, $\mathcal{V}_{\text{artic}}$, $\mathcal{V}_{\text{pitch}}$ to be either fixed or adaptive and to contain a $\emptyset$ symbol denoting the lack of structure, meter and so on. The variables $T_1$, $T_3$, $S_5$ and $S_7$ also implicitly depend on a set of structural, rhythmic, harmonic, melodic and bass patterns denoted $\mathcal{P}_{\text{sect}}$, $\mathcal{P}_{\text{rhythm}}$, $\mathcal{P}_{\text{chord}}$, $\mathcal{P}_{\text{melo}}$ and $\mathcal{P}_{\text{bass}}$. More generally, all model parameters can themselves be regarded as variables [10].

Any variable may be either fully observed, partially observed or hidden, leading to a huge range of scenarios. For instance, automatic accompaniment consists of inferring $A_2$ given part of $A_1$, while symbolic genre classification consists of inferring part of $O$ given $S_7$. Playlist generation or cover detection may also be addressed by comparing all features $O$, $T_*$, $S_*$ and $E_*$ inferred from $A_2$ in each piece of a database according to some criterion.

3. SCALABLE CONDITIONAL DISTRIBUTIONS

Once the variables have been defined, the next step consists of designing conditional distributions between these variables. While recent studies in the field of audio source separation have already led to complete family of acoustic models $P(A_1|E_4, E_5)$ [9], many other dependencies have either been studied in a deterministic setting or not investigated yet. More crucially, current probabilistic symbolic models, e.g. [4, 8, 11], only account for short-term dependencies between two or three variables taking few possible values and rely on hand tying of probabilities based on Western musicology. This design method does not scale with the long-term high-dimensional dependencies found in Figure 1. For instance, the virtually infinite set of overall features $O$ affects most other features and the probability of quantized notes $P(S_7|O, T_1, T_2, T_3, S_2, S_3, S_4, S_5, S_6)$ depends on as many as 9 other features. Scalable methods must hence be found to parameterize each conditional distribution so as to avoid overfitting.

A promising approach consists of modeling the conditional distribution of a variable given its parents by interpolation of its conditional distributions given each parent individually. This approach is widely used in the language processing community [3] but does not account for possible interactions between parents. This issue may be tackled by reparameterizing the space of parent variables in terms of a smaller number of factors using e.g. Latent Semantic Indexing (LSI) techniques [7] developed for text retrieval and collaborative filtering. We believe that the extension of these approaches to all symbolic music data will lead to a similar breakthrough as in the above domains.

4. UNSUPERVISED LEARNING ALGORITHMS

The design of conditional distributions is closely related to that of learning and decoding algorithms. Indeed, due to the above dimensionality issues and to the variety of music and individual music experiences, most distributions cannot be fixed a priori but must be learned from possibly user-specific training data or from the test data. Similarly, the
feature value sets $V_c$ and the patterns $P_c$, must be learned to identify e.g. the most relevant set of chord symbols and patterns for a given song, in line with human listening that picks up regularities based on prior exposure without actually naming them [12]. These learning tasks are always unsupervised since training data annotated with all features of Figure 1 will most probably never exist.

The estimation of some hidden variables consists of marginalizing i.e. integrating the likelihood over the values of the other hidden variables [10]. This can be achieved using the modular sum-product and max-product junction tree algorithms [10] that generalize the classical Baum-Welch and Viterbi algorithms for HMMs. The considered objective is often the maximization of the posterior distribution of the inferred variables given the data. Although this Maximum A Posteriori (MAP) objective may be used for unsupervised learning of the model parameters (i.e. conditional probabilities) [8], it cannot infer the model order (i.e. the dimension, the value set and the parents of each feature). Unsupervised pruning of feature dependencies would also considerably accelerate the speed of the junction tree algorithm, that is exponential in the number of dependencies, and make it possible to match the available computational power in an optimal fashion. Suitable model selection criteria and algorithms are hence of utmost importance.

Automatic Relevance Determination (ARD) and several other popular model selection techniques employ prior distributions over the model parameters favoring small model orders [7]. The alternative variational Bayesian inference technique [10] selects the model with highest marginal probability by integrating the posterior distribution of all hidden variables. This technique also provides an approximation of the posterior distribution of all hidden variables as a by-product. This increases the meaningfulness and interpretability of the results compared to the estimation of the MAP variable values only, at the cost of higher computational complexity. Again, we believe that advances will eventually be achieved by combining these approaches.

The choice of algorithms will guide that of feature formats. Efficient graph formats exist for the representation of posterior distributions of symbolic feature sequences [6], but they must yet be extended to e.g. polyphonic note sequences. More generally, the high dimensionality of all features will necessitate compressed feature formats.

5. MULTI-FEATURE ANNOTATED DATABASE

Finally, although the development of a database annotated with all features of Figure 1 appears infeasible, some multi-feature annotated data will nevertheless be needed to initialize and evaluate the unsupervised learning process. This implies strong community coordination to push current annotation efforts towards the same data and bridge the cultural gap between experts of different music styles. Also, this advocates for the evaluation of conditional feature estimation tasks within MIREX, where some other features would be known, as opposed to the current full-fledged estimation tasks, where only audio or MIDI are given.

6. SUMMARY AND IMPLICATIONS

We provided a feasible roadmap towards a complete MIR system, emphasizing challenges such as scalable parameterization and unsupervised model selection. As recommended in [1], this implies that most efforts in the MIR community now focus on symbolic data. Yet, other strong implications also arise regarding the need for a coordinated effort and the definition of MIREX tasks and data.

7. ACKNOWLEDGMENT

This work is supported by INRIA under the Associate Team Program VERSAMUS (http://versamus.inria.fr/).

8. REFERENCES