COMPUTER-AIDED ORCHESTRATION BASED ON PROBABILISTIC INSTRUMENTS MODELS AND GENETIC EXPLORATION

Damien Tardieu, Grégoire Carpentier, Xavier Rodet
IRCAM-CNRS - STMS
1 place Igor Stravinsky - Paris, France F-75004
{carpenti,dtardieu,rodet}@ircam.fr

ABSTRACT

In this paper we introduce a tool aimed at assisting composers in orchestration tasks. Thanks to this tool, composers can specify a target sound and replicate it with a given orchestra. We discuss the problems raised by the realization of such a tool, concerning instrumental sound description and combinatorial optimization. Then we describe the solution adopted. We propose a machine learning method based on generative probabilistic modeling to represent and generalize instrument timbre possibilities from sample databases. This model allows to deduce the timbre of any mixture of instrument sounds. In a second part, we show that search of sound mixtures that match a given target is a combinatorial optimization problem that can be addressed with multicriteria genetic algorithms.

1. INTRODUCTION

In the last few decades contemporary music composers have widely experienced computer-aided composition (CAC) software in their works. Originally, those tools were designed to provide composers with the ability to easily manipulate musical symbolic objects, such as notes, chords, melodies or polyphonies, but the timbral aspect of the composition, orchestration, has stayed relatively unexplored. We define orchestration as the composition with the orchestral timbre. An orchestra is composed by many instruments, each of them being able to create a large variety of sounds. By the combination of those instruments, the composer have access to a huge set of timbres. In this paper we present a tool that helps composers to explore this set. In this tool, composers can specify a target sound and imitate it with a given orchestra. Some attempts have been made ([6], [12], [13]) to address this problem by combining instrument sound spectrums to match a target spectrum. But timbre is much more than spectrum. A sound is perceived through many different characteristics, like spectrum but also modulations, roughness and others. So our method is based on a description of sound that takes those characteristics into account. But extending the sound description raise the problem of timbre similarity. Comparing one characteristic of two sounds may be complicated but still achievable, but finding a global similarity measure that takes several characteristics into account is more difficult [1]. Does a clarinet with vibrato sounds closer to a clarinet without vibrato or to a flute with vibrato ? It is a problem of personnal preference depending on the aspect of the sound we focus on: is it this frequency modulation caused by the vibrato, or this lack of even harmonics so characteristic of the clarinet sound ? There is no unique way of comparing sounds globally. We thus chose to use multicriteria optimization methods that do not make any assumption on the descriptors relative saillance, and guess the user preferences by an interaction process. Another drawback of the previous approaches lies in the direct use of spectrum of instruments samples. Doing this, they ignore one of the main difficulty, then interest, of the problem: we deal with musical instruments and the databases samples are only examples of them. The system needs to be able to learn and generalize the timbre possibilities of the instruments from those examples. We use generative probabilistic models of the features allowing to extract a general knowledge of the instrument timbre from different samples. The last issue is the exploration of the solutions space. This is a hard combinatorial problem. Previous approaches use either greedy algorithms or decomposition of the target spectrum on a basis of instruments spectrum. Those methods are computationally efficient but do not allow a wide exploration of the solution space, and need the definition of a global similarity measure. We then adopt a genetic multicriteria algorithm.
The paper is organized as follow. In the first section we give an overview of our system, then we describe the methods used to learn the instrument timbre possibilities in section 3. Section 4 explains the exploration of the solution space and finally we discuss the evaluation of the proposed method.

2. ORCHESTRATION PROCEDURE

2.1. System overview

In our system the user specifies the sound to be produced (the target) and the instruments that can be used to produce it . Then, an orchestration engine uses an instruments samples database to suggest instruments notes combinations (mixtures) that sound close to the target. The engine can be divided into two parts, an instrument knowledge part and an exploration process. The instrument knowledge part performs the extraction and structuring of all the available information from the sounds databases. The ex-

ploration process is an algorithm dedicated to the efficient exploration of the possible mixtures.

2.2. Mathematical formulation

An instrument sound n, we call it an item, is defined by the instrument i, its articulation a (vibrato, tremolo, ...), its pitch p, its loudness l and its mute m. An item is represented by the probability density functions of its descriptors values, $(d_j^n)_{j\in[1,J]}$, $f(d_j|i,a,p,l,m)$. The sound of a mixture $\mathcal{K}$ of items is represented by the pdfs $f(d_j|\mathcal{K})$. The target t is defined by the set of its descriptors values $(\hat{d}_j)$. The orchestration problem can be formulated as follow: given a set of items E and a sound target t, the goal of the orchestration procedure is to find a subset of E that maximizes the similarity with t, *i.e.* that maximizes the probabilities $P(\hat{d}_j|\mathcal{K})$. Note that the timbre similarity is computed as a *vector* of probabilities (instead of a single value) along each descriptor in order to cope with the multidimensional mechanism of timbre perception.

3. SOUND DESCRIPTION AND LEARNING

3.1. Sound description

The set of sound descriptors must be reasonably small and understandable to a composer, in order to facilitate interaction during the exploration process. Thus we chose relatively high level signal descriptors coming either from psychoacoustic field [10] or from the automatic classification field [11]. The orchestration procedure we are introducing relies on a set of descriptors concerning different aspects of the sound. For the moment we use the following ones:

- energy of the harmonics normalized by the global energy,
- global noisiness,
- frequency and amplitude of the fundamental frequency modulation,
- frequency and amplitude of the energy modulation,
- attack time

We also need two information that do not relate to timbre, the fundamental frequency (f_0) and the energy of the signal. The f_0 is extracted with [3]. The signal energy is not extracted from the signal but guessed from the sample name. Indeed, the real energy of the sound is not available in the database samples. However, all the samples we have are named with a dynamic indication, such as *pp* or *mf*. From a subset of sounds for which we know that the relative dynamics are realistic, we extracted the mean energy of each dynamic. Those values are used as standard values for the remaining of the samples.

3.2. Mixture of gaussians

The descriptors pdfs are approximated by mixtures of gaussians. The distribution of a descriptor d_j is defined by the number of gaussian components M_j, the weights ω_{jm}, the means μ_{jm} and the covariance matrices $\mathbf{\Sigma}_{jm}$ by the following equation:

$$f(d_j^n) = \sum_{m=1}^{M} \omega_{jm}^n \mathcal{N}(d_j^n; \mu_{jm}^n, \mathbf{\Sigma}_{jm}^n) \qquad (1)$$

The parameters estimation is performed by an iterative EM algorithm, where the number of gaussian components is increased at each step. The selected model is the one that gives the best recognition rate in a cross database classification task.

3.3. Learning strategies

Learning one model for each item present a major drawback, it requires many sound samples for each item. We use two different strategies to avoid this problem.

3.3.1. Fundamental frequency and energy

The first strategy consists of learning $f(d,p,l|i,a,m) \sim f(d,f_0,e|i,a,m)$, where f_0 is the fundamental frequency and e is the energy, instead of learning $f(d|i,a,p,l,m)$. In other words, we learn the joint distribution of the descriptor, the fundamental frequency and the energy, next by conditionalization we can obtain $f(d|i,f_0,e,a,m)$ [8]. Now there is only one model for each instrument and articulation and the learning set of one model contains all the pitches and dynamics. Note that we transformed the discrete variables p and l into continuous variables f_0 and e, which allows to find the model for any pitch and dynamic even those that are not in the sample databases.

3.3.2. Dividing the problem

The second strategy is based on a reorganization of the problem. Instead of learning one model for each combination of instrument, mute and articulation available in the database, which lead to many high dimensional models, we learn a model for each articulation, mute and instrument individually, and we aggregate them. Note that, doing this, the articulation becomes a set (a_k) of articulations. Indeed, in the first approach a complex articulation available in the database, like aeolian+vibrato, was considered as one articulation whereas in the second approach it is considered as two. The aggregation is done in the following way:

$$f(d_j,f_0,e|i,(a_k),m) = f_i f_m \prod_k f_{a_k} \qquad (2)$$

$$where \quad f_i = f(d_j,f_0,e|i) \qquad (3)$$

$$f_m = f(d_j,f_0,e|m) \qquad (4)$$

$$f_{a_k} = f(d_j,f_0,e|a_k) \qquad (5)$$

This aggregation method is called Logarithmic Opinion Pool [5] and relates in our context to Product Of Mixtures of Gaussians [4]. Given that f_i, f_m and f_{a_k} are Mixtures of Gaussians, $f(d,f_0,e|i,(a_k),m)$ is itself a Mixture Of

Gaussians whose parameters can be calculated from the above pdfs parameters [4].
An interesting point about this reorganization is that we can reduce the dimension of each model by selecting the relevant descriptors for a problem. For example, the models of vibrato and non vibaro sounds, will only describe the modulation of the fundamental frequency and of the energy. If a descriptor is not selected, its distribution is assumed to be uniform, hence will have no influence in equation 2.
Another advantage of the method is that it allows to deduce the model of sounds that are not in the database. For instance, suppose we do not have any sound of a clarinet with vibrato, but we have sounds of a clarinet without vibrato and sounds of other instruments with vibrato. We can find the model of the vibrato clarinet by aggregating the clarinet model, learned on clarinet sounds without vibrato, with the vibrato model, learned on all the vibrato sounds of the other instruments.

3.4. Mixture models

3.4.1. How do descriptors add ?

To evaluate the possibility for a mixture $\mathcal{K}$ to imitate the target, we have to determine the model of this mixture from the models of its components. Thus, we have to know how the descriptors of two or more sounds add. This is a specific problem, related to sound perception, that we will not detail here. For instance, we hypothesize that the harmonics energies add, or that the attack time of a mixture of two sounds equals the shortest of the two attack times. All those hypothesis are being tested either by psychoacoustic experiments or by less formal tests.

3.4.2. Finding the mixture model

The distributions of $\mathcal{K}$ depend on the addition method of the underlying descriptors. Hence, a specific operator is needed for each descriptor. We will not detail here this operator for all the descriptors but, we just show an example where the addition method is a sum weighed by the energy, which can be used for $d_j =$ *normalized harmonics energy*.

$$d_j^{\mathcal{K}} = \sum_n e^n d_j^n \tag{6}$$

Since the d_j^n are described by mixtures of gaussians, the distribution of $d_j^{\mathcal{K}}$ cannot be calculated. We go back to a gaussian case by selecting, for each item n, the gaussian component that gives the highest probability for the f_{0n} and the energy e_n of the item.
Therefore, assuming that d_j^n follows a gaussian distribution $\mathcal{N}(\mu_j^n, \mathbf{\Sigma}_j^n)$ and that the e^n are known, the pdf of the mixture descriptor is:

$$f(d_i|\mathcal{K}) = \mathcal{N}(d_i, \sum_n e^n \mu_j^n, \sum_n (e^n)^2 \mathbf{\Sigma}_j^n) \tag{7}$$

This last equation will be used in the following section to compute a similarity measure between the target and the mixture.

4. EXPLORING THE SOLUTION SPACE

As explained in [2] searching efficient sound combinations with capacity constraints (due to the limited instrumental ressource of an orchestra) may be seen as a multi-objective, multidimensional 0/1 knapsack problem (MOKP-0/1). Briefly speaking, a knapsack problem consists in finding a set of items to be inserted in a knapsack in order to maximize some profit function without exceeding the knapsack capacity. Formally, the orchestration task is defined as follows:

$$\begin{cases} \max z_j = P(d_j|\mathcal{K} = \{x_1, ..., x_n\}) \\ j = 1, ..., J \\ \text{s.t.} \quad x_k \in \{0; 1\} \\ \qquad R.x \leq C \end{cases} \tag{8}$$

where J is the number of descriptors, C the orchestra's capacity vector (the number of instruments of each type) and R a ressource allocation matrix handling constraints due to the restricted number of instruments. The orchestration problem only differs from MOKP-0/1 by the use of complex profit functions $P(d_j|\mathcal{K})$ along each objective rather than a simple sum of item profits.
Knapsack problems are $\mathcal{NP}$-hard. They have been widely studied in operational research and combinatorial optimization and many ad-hoc search methods have been designed (see [9] for a review of mono-objective problems). The orchestration problem is particularly complex as capacities and profits are multi-dimensional and profits are correlated. In [2] we have proposed an efficient, evolutionary, multi-objective method inspired of Jaszkiewicz's algorithm [7]. Our algorithm uses a population of solutions and iteratively alternates between genetic recombination and local search optimization, with ad-hoc operators for each phase. The multi-objective approach is handled thanks to a weighted Tchebycheff function which aggregates the probabilities $P(d_j|\mathcal{K})$ into a single fitness value:

$$F(D, \mathcal{K}) = \max \lambda_j P(d_j|\mathcal{K} = \{x_1, ..., x_n\}) \tag{9}$$

where the weights $(\lambda_j)_{1 \leq i \leq J}$ define the current direction of optimization. The weights are randomly drawn at each generation by a pseudo-random number sequence, insuring an uniform sampling of the search directions. In other words, the relative importance of the objectives randomly change over the iterations, letting the population approximate the set of efficient solutions, also called Pareto set.
The use of weighted Tchebycheff functions was also motivated by the opportunity to easily design a user interaction process: When a first stopping criterion is met, the algorithm stops and the user is asked to choose *one best solution* $\mathcal{K}_{best}$ among the current Pareto set. $\mathcal{K}_{best}$'s coordinates in the criteria space are used to infer the user's preferences, i.e. the relative weights (λ_i^{best}) that rank $\mathcal{K}_{best}$ first when injected in eq. 9. The algorithm then restarts with fixed weights (λ_i^{best}), favorizing the search direction of the best solution found so far.

5. RESULTS AND CONCLUSION

We have presented an approach to computer assisted orchestration that allows to find instrumental sound combinations that imitate a target sound. The approach is based on a generative probabilistic model of the instruments sounds and on a genetic algorithm for the exploration of the solution space. The evaluation of the overall system is a hard problem since the final rating of the goodness of a solution can only be achieved by listening. A first step is to evaluate the two parts separately. As an indication of the performance of the instruments models we provide the results of two classification tasks realized on five different sounds databases. For both tasks, the models are trained on four databases and tested on the fifth one. We performed a 8 classes task with 4 woodwind instruments: clarinet, bassoon, flute and oboe. Each instrument is played with and without flatterzungue. Learning one model per classe leads to an average recall of 51%, but almost none of the flatterzungue sounds are recognized. With our method, we obtain a score of 62% with 95% of the flatterzungue sounds that are recognized as flatterzungue. This shows that this method gives much more robust models. As a second test, we did another 8 classes experiment with 4 instruments: trumpet, trombone, horn and tuba. In the test set, each instrument plays with and without vibrato, but there is no vibrato brass sound in the training set. The vibrato and non vibrato models are trained on woodwinds and strings. The mean recall is 35%, to be compared with a random of 12%, and 71% of the vibrato sounds are recognized as vibrato. This experiment shows that the method allows to recognize sounds that are not in the training set which is interesting in an orchestration context since we want to know as many instrument sounds as possible.
Concerning the genetic algorithm, performances are especially hard to evaluate. Traditionally performances of multi-objective optimization methods are based on the size, shape, density, or homogeneity of the Pareto set, or on the distance between the theoretic Pareto set and its approximation, when the former is known, which is not the case here. Therefore, Pareto sets obtained by our algorithm are difficult to score. However, early experiments with our system gave encouraging results, allowing to find interesting mixtures with large orchestras in a very short time. Future research will focus on the design of objective and subjective evaluation procedures for the genetic method alone and for the overall system.

6. ACKNOWLEDGEMENTS

The authors wish to deeply thank the composer Yan Maresz for his involvement in this project. The authors also want to thank Geoffroy Peeters for his critical review of this paper.

7. REFERENCES

[1] The role of similarity in categorization: Providing a groundwork. *Cognition*, 52(2):125–157, aug 1994.

[2] Gregoire Carpentier, Damien Tardieu, Gérard Assayag, Xavier Rodet, and Emmanuel Saint-James. An evolutionary approach to computer-aided orchestration. In *EvoMUSART*, volume LNCS4448, pages 488–497, Valence, Espagne, Avril 2007.

[3] B. Doval and Xavier Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and hmms. In *Proc. IEEE-ICASSP*, pages 221–224, 1993.

[4] Gales M. J. F. and Airey S. S. Product of gaussians for speech recognition. *Computer Speech and Language*, 20(1):22–40, jan 2006.

[5] Genest, Christian and Zidek, James V. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, feb 1986.

[6] Thomas A. Hummel. Simulation of human voice timbre by orchestration of acoustic music instruments. In *Proceedings of International Computer Music Conference 2005*, 2005.

[7] A. Jaszkiewicz. Comparison of local search-based metaheuristics on the multiple objective knapsack problem. *Foundations of Computing and Design Sciences*, 26:99–120, 2001.

[8] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 285–288, 1998.

[9] S. Martello and P. Toth. *Knapsack problems: Algorithms and computer implementations*. John Wiley & Sons, Chichester, 1990.

[10] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192, 1995.

[11] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification). in the CUIDADO project. Paris, IRCAM, 2004.

[12] David Psenicka. Sporch: An algorithm for orchestration based on spectral analyses of recorded sounds. In *Proceedings of International Computer Music Conference 2003*, 2003.

[13] François Rose and James Hetrick. Spectral analysis as a ressource for contemporary orchestration technique. In *Proceedings of Conference on Interdisciplinary Musicology 2005*, 2005.