AN INSTRUMENT TIMBRE MODEL FOR COMPUTER AIDED ORCHESTRATION

Damien Tardieu *Xavier Rodet*

IRCAM-STMS
1, place Igor Stravinsky, 75004 Paris
{Damien.Tardieu, Xavier.Rodet}@ircam.fr

ABSTRACT

In this paper we propose a generative probabilistic model for instrument timbre dedicated to computer aided orchestration. We define the orchestration problem as the search of instruments sounds combinations that sound close to a given target. A system that addresses this problem must know a large variety of instruments sounds in order to be able to explore the timbre space of an orchestra. The proposed method is based on gaussian mixture modeling of signal descriptors and on a division of the learning problems that allows to learn many different instrument sounds with few training data, and to deduce the models of sounds that are not in the training set but that are known to be possible.

1. INTRODUCTION

In the last few decades many attention has been drawn to symbolic aspects of computer aided composition but the timbral aspect of the composition, orchestration, has stayed relatively unexplored. Orchestration is the art of selecting and arranging instrument sounds to obtain a given timbre for a musical intention. The computer aided orchestration (CAO) task we focus on here is the following one : given a target timbre, specified by a recorded sound, find a mixture of instrument sounds that sounds close to the target. The instrument timbre knowledge must be extracted from sound samples databases. This can be seen as a transcription problem where the goal is to find an instrument combination that best reconstruct the target timbre and not the target signal and where the transcription must not only specify the instruments that are playing and their pitches but also the way the instrument is played. Indeed, a player can use a wide variety of techniques to alter the sound of its instrument, such as tremolo for strings or flatterzunge for winds, he can also add different mutes that produce different timbres. As a result, a CAO system must be able to learn and generalize all those possibilities from sounds samples databases to propose complex and interesting orchestration solutions. In this paper we will only address the aspects of computer aided orchestration related to instruments timbre description and learning. A promising solution, used in transcription and source separation ([1], [2]), is to model timbre descriptors with probability distributions. Standard methods rely on a modelization of the sound whose parameters are learned on a sample database. They achieve good results but to our knowledge such methods have never been used to learn and recognize a large set of playing techniques. On the other hand, a method dedicated to instrumental gesture extraction from the signal has been developped in [3], it can accuratly predict the plucking position on a guitar. The method is based on the knowledge of physical mechanisms taking place on the instrument. In our context, describing the physical properties of all the intruments would be a long and hard task, therefore we propose a more generic method based on probabilistic modeling of generic timbre descriptors. For probabilistic methods to be efficient, ie. to have good generalization abilties, they must be trained on large sets of sound examples. Obtaining many samples of all the possible sounds of each instrument is complicated. In this paper we propose a method that, first allows to learn many different instrument sounds with relatively few samples, and also enable the deduction of the model of playing techniques that are not contained in the training database. The method is based on gaussian mixture modeling of signal descriptors and on a factorization of the model that allows to separate the problem into smaller tractable problems. We decribe the model in section 2. The training method is explained in section 3. Finally an evaluation of the method based on classification tasks is proposed in section 4.

2. INTRUMENT TIMBRE MODEL

2.1. Description

We model the timbre of an instrument by the probability distribution $f((d_j)|f_0, e, S)$ of the sound descriptors (d_j) given the continuous variables f_0, the fundamental frequency, and e, the energy, and a set of discrete state variables $S = (s_k)$. S includes the instrument, the mute and the playing technique. The playing technique is described by several state variables. For instance, for a violin, there can be a variable specifying if it is bowed or plucked, another one defining the bow position (normal, on the fingerboard or on the bridge) and a last one telling if it is played with or without vibrato. The distribution is approximated by a Mixture Of Gaussians (GMM):

$$\begin{aligned} f((d_j)|f_0, e, S) &= \mathcal{GMM}((d_j); M, (\omega_m), (\mu_m), (\mathbf{\Sigma}_m)) \\ &= \sum_{m=1}^{M} \omega_m \mathcal{N}((d_j); \mu_m, \mathbf{\Sigma}_m) \end{aligned} \quad (1)$$

Where M is the number of components, ω_m, μ_m and $\mathbf{\Sigma}_m$ are, respectively, the weight, mean and covariance matrix of the mth component. To learn the distribution parameters, it is necessary to have many samples in the training set which is almost impossible with the above formulation. We propose two ways of transforming the model to make the training possible.

2.2. Grouping pitches and dynamics

Since the different descriptors of an instrument do not vary with pitch and dynamic or are simply correlated with those variables,

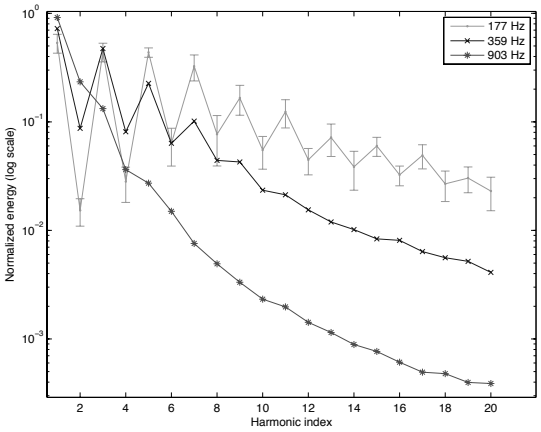

Figure 1: *Mean of the 3 gaussians of the Bb clarinet model for the descriptor "Normalized energy of the harmonics". The legend correspond to the mean of the fundamental frequency descriptor for each gaussian.*

we learn the parameters of the following distibution:

$$f((d_j), f_0, e|S) = \sum_{m=1}^{M} \omega_m \mathcal{N}((d_j); \bar{\mu}_m, \bar{\mathbf{\Sigma}}_m) \quad (2)$$

This means that all the different pitches and dynamics are described by only one model, and then that the training set of the model is much larger.
From this distribution, we can deduce the parameters of $f((d_j)|f_0, e, S)$:

$$\mu_m = \bar{\mu}_d + \bar{\mathbf{\Sigma}}_R \bar{\mathbf{\Sigma}}_{f_0,e}^{-1}(\mathbf{v} - \bar{\mu}_{f_0,e}); \quad (3)$$

$$\mathbf{\Sigma}_m = \bar{\mathbf{\Sigma}}_d + \bar{\mathbf{\Sigma}}_R \bar{\mathbf{\Sigma}}_{f_0,e}^{-1} \bar{\mathbf{\Sigma}}_R^T \quad (4)$$

with

$$\mathbf{v} = \begin{pmatrix} f_0 \\ e \end{pmatrix} \quad \bar{\mu}_m = \begin{pmatrix} \bar{\mu}_{f_0,e} \\ \bar{\mu}_d \end{pmatrix} \quad (5)$$

$$\bar{\mathbf{\Sigma}}_m = \begin{pmatrix} \bar{\mathbf{\Sigma}}_{f_0,e} & \bar{\mathbf{\Sigma}}_R \\ \bar{\mathbf{\Sigma}}_R^T & \bar{\mathbf{\Sigma}}_d \end{pmatrix} \quad (6)$$

where $\bar{\mu}_{f_0,e}$ and $\bar{\mathbf{\Sigma}}_{f_0,e}$ are the mean and covariance matrix of f_0 and e, $\bar{\mu}_d$ and $\bar{\mathbf{\Sigma}}_d$ are the mean and covariance matrtix of the descriptors (d_j), $\bar{\mathbf{\Sigma}}_R$ contains the covariances between the elements of (f_0, e) and the corresponding elements of (d_j) and T represent the matrix transposition operator.
Figure 1 shows a probability density function obtained with the above method with the descriptor "energy of the harmonics normalized by the global energy". It is learned on four of the five sample databases described in section 4.2. We see that each gaussian is centered on a different fundamental frequency (see legend), thus models a different register of the instrument. For clarity purpose we plot only the standard deviations of one component, the other components have similar standard deviations.

2.3. Dividing the problem

2.3.1. A factorization of the distribution

As a second transformation, we propose the following factorization of the probability density function (pdf):

$$f((d_j), f_0, e|S) = \prod_k f((d_j), f_0, e|s_k) \quad (7)$$

Using this factorization allows to learn each terms independently and later perform the aggregation. In other words, we learn a model for each state variable value separately and deduce the model of a specific combination of those values by multiplying the pdfs. It can be seen as a supervised *Mixture of Experts* [4] where each state variable is an expert giving an opinion on the distribution of the descriptors.

2.3.2. Providing instrumentation knowledge

The above factorization allows to find the model of any state variable combination, even if it does not exist in the training set. But, in another hand, it can generate the model of impossible combinations. A flute cannot play "on the fingerboard", or a violin cannot use a wawa mute. It is therefore necessary to introduce external instrumental knowledge in the system. We propose a structuration of the possible values of the state variables (s_k). A state variable is defined by a set of values and a set of dependencies. The dependencies are the necessary values of other variables for this variable to make sense. For instance, the variable "bowPosition" can take the values "normal", "on the bridge" and "on the fingerboard" and this variable makes sense only if "family" is set to "string" and "rightHand" is set to "bow". This structure allows to generate only the possible combinations of playing techniques, instruments and mutes.

3. LEARNING BY CLASSIFICATION

3.1. Subproblem definition and training

Each state variable define a classification problem (we will call it a subproblem) that will be the basis for learning instrument models. A state variable s_k defines the subproblem P_k. Each value of s_k define a class whose model is

$$\mathcal{M}_{s_k} = \mathcal{GMM}(M^{(s_k)}, (\omega_m^{(s_k)}), (\mu_m^{(s_k)}), (\mathbf{\Sigma}_m^{(s_k)})) \quad (8)$$

For instance, the variable "vibrato" defines two classes "with vibato" and "without vibrato" and then two different models.
The learning of the model parameters is a two stage process. First, since the different state variables does not have an effect on all the descriptors we can reduce the dimension of the subproblem by selecting the descriptors that show significant variation between classes. we use a simple criterion based on mutual information between state variable values and descriptors. All the descriptors whose mutual information exceed a certain threshold are selected. We call D_k the descriptor set of P_k and $\mathcal{M}'_{s_k}$ the model based on the selected descriptors. This dimension reduction forces the learning algorithm to focus on the relevant descriptors and then significantly improve its perfomances.
Second, the parameters of $\mathcal{M}'_{s_k}$ are estimated by an EM algorithm. Several estimations are performed by successively increasing the number of gaussian components. The selected number is the one that gives the best recognition rate in a cross database classification task.

3.2. Subproblems models aggregation

For a given state $S = (s_k)$ we want to deduce the model $\mathcal{M}_S$ of S from the models $(\mathcal{M}_{s_k})$. Since all the pdfs are approximated by Mixture Of Gaussians, the aggregated pdf is itself a GMM [5]:

$$p((d_j)|S) = \mathcal{GMM}((d_j); \hat{M}, (\hat{\omega}_{\mathbf{m}}), (\hat{\mu}_{\mathbf{m}}), (\hat{\mathbf{\Sigma}}_{\mathbf{m}})) \quad (9)$$

$$\hat{\mu}_{\mathbf{m}} = \mathbf{\Sigma}_{\mathbf{m}} \left(\sum_{s_k \in S} \mathbf{\Sigma}_{m_{s_k}}^{(s_k)-1} \mu_{m_{s_k}}^{(s_k)} \right) \quad (10)$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{m}} = \left(\sum_{s=1}^{S} \mathbf{\Sigma}_{m_{s_k}}^{(s_k)-1} \right)^{-1} \quad (11)$$

$$\hat{\omega}_{\mathbf{m}} = \prod_{s=1}^{S} \omega_{m_{s_k}}^{(s_k)} \quad (12)$$

Where $\mathbf{m} = [m_{s_1} ... m_{s_k}]$ determines a particular component of $\mathcal{M}_S$ and m_{s_k} specifies the component from $(\mathcal{M}'_{s_k})$, which means that each resulting component is the product of subproblems components, and hence, that the effective number of components, $\hat{M}$, in the aggregated model is the number of possible combinations of components from each subproblem model,

$$\hat{M} = \prod_{s=1}^{S} M^{(s_k)} \quad (13)$$

Since we have selected the relevant descriptors for each subproblem, we have to aggregate models that do not rely on the same dimensions. To do this, we create a "virtual model" describing the union of all the descriptors. To simplify the notations, we explain the method by an example with two models. Let $\mathcal{M}'_1$ and $\mathcal{M}'_2$ be two classes models that we want to combine. Let $D'_1 = (D_c, D_1)$ and $D'_2 = (D_c, D_2)$ be the descriptor sets of the models. D_c is the subset of descriptors common to D'_1 and D'_2. We create two models $\mathcal{M}_1$ and $\mathcal{M}_2$ that describes $D_{12} = (D_c, D_1, D_2)$ with the following parameters :

$$\mu_1 = \begin{pmatrix} \mu'_1 \\ \mu'_{c,1} \\ \mathbf{0} \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} \mathbf{0} \\ \mu'_{c,2} \\ \mu'_2 \end{pmatrix} \quad (14)$$

$$\mathbf{\Sigma}_1 = \begin{pmatrix} \mathbf{\Sigma}'_1 & 0 & 0 \\ 0 & \mathbf{\Sigma}'_{c,1} & 0 \\ 0 & 0 & \sigma^2 I \end{pmatrix} \quad \mathbf{\Sigma}_2 = \begin{pmatrix} \sigma^2 I & 0 & 0 \\ 0 & \mathbf{\Sigma}'_{c,2} & 0 \\ 0 & 0 & \mathbf{\Sigma}'_2 \end{pmatrix} \quad (15)$$

Then applying aggregation equations to those parameters and making σ tends to ∞, we find the parameters of the aggregated model $\mathcal{M}_{12}$. Making σ tend to ∞ may be considered as making the underlying dimension non-informative. In other words we make explicit the information that the model does not know anything about some descriptors. One may notice that we cancelled the terms of the covariances between the shared and non shared descriptors, the only explanation is that it has shown better classification performances than keeping them.

4. EVALUATION ON CLASSIFICATION TASKS

We evaluate the proposed method on classification tasks. The goal of those tests is not to attain the highest possible recognition rates but to verify that the proposed method allows to learn more robust models. Then we will always compare results obtained with and without the factorization of the probabilty distribution.

4.1. Sound descriptors

We use signal descriptors coming either from psychoacoustic field [6] or from the automatic classification field [7]:

- 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,
- spectrum flatness,
- roughness.

Details on the descriptors computation can be found in [7]. We also add 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 [8]. 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 dynamics. Those values are used as standard values for the remaining of the samples.

4.2. Sound databases

We use five sound sample databases RWC, IOWA, Studio On Line, Vienna Symphonic Library and Virtual Orchestra. Some of them contain only the normal playing technique of each instrument, while others contain many techniques. We have, therefore, bowed strings with or without mute and two different mutes. They are played vibrato, non vibrato, tremolo, on the bridge and on the fingerboard. Woodwinds are played with and without vibrato, aeolian (breathy sound) and flatterzunge. Finally, brasses are played with 4 different mutes and flatterzunge.

4.3. results

First we evaluate the overall performance of the GMM modeling by performing a classification task on the family of the sounds (string, woodwind and brass). The model is trained on four databases and tested on the fifth one, which results in 5 tests. The average recall (recognition rate) is 77.4%. Second we test the ability to recognize the instruments. In that case we select, for each of the 5 tests, the instruments that belong to, at least, two databases. The recall is 57% for an average of 9 instruments. Those results are good considering the size of the descriptor set and the very large variety of sounds for each instruments. Now to evaluate the proposed method we give 2 examples:
First we try to classify the sounds of three brass instruments (trumpet, horn and trombone) that are played with and without flatterzunge. Without the factorization, the average recall is 27%, but

	M1	M2	random
All Instrument	27/28	59/64	16
Flatterzunge	54/59	59/64	33
	49/49	95/97	50
All	-	35/31	4,2
Instrument/ Bow Position	31/24	40/33	8,3
Instrument	55/53	70/75	25
Bow Position	58/44	57/42	33
Tremolo	-	87/90	50

Figure 2: *Results of the two tests. M1 is the method without factorization, M2 is the method with factorization. "Random" is the mean recall of a random guess. Results are presented in percentage in the following way : mean recall/mean precision*

the interesting fact is that very few flatterzunge sounds are recognized leading to the same recall as a random guess. This can be explained by the relatively high dimension of the problem (35) and by the few available samples for flatterzunge sounds, resulting in difficulty for the learning algorithm to focus on the distinctive characteristics of flatterzunge sounds. Now, using the factorization, the average recall becomes 59%, with 95% of the flatterzunge sounds that are recognized as flatterzunge. By dividing the problem, we helped the system to focus on the relevant characteristics of a problem and then to be more robust.

The second set of tests deals with strings. we use a test set containing four bowed strings (violin, viola, cello and contrabass) played with three bow positions (on the bridge, normal, on the fingerboard) and with or without tremolo. This results in a total of 24 classes. The interesting point here is that the training set contains only samples of tremolo sounds with normal bow position, but not any tremolo sounds for the other positions. Therefore, We test the ability of the method to model and recognize sound classes that are missing in the training set. The three classification subproblems give the following results: 70% of average recall for the instruments classification, 57% for the bow position, and 87% for the tremolo. From those results, it seems that none of the descriptors is really appropriate to explain the different timbres resulting from the various bow positions whereas the tremolo problem is fully describe by amplitude modulation. When performing classification on both instrument and bow position, the factorization method give slightly better results.Finally the results for the three variables together is 35% that has to be compared to 4.2% for a random guess. But here, the most important fact is that we succeeded in creating the model of sounds that were not in the training set with for instance 73% of the contrabass *on the fingerboard tremolo* sounds recognized.

5. CONCLUSION

We presented a generative instrument timbre model based on Gaussian Mixture Model dedicated to computer aided orchestration. We underlined the problem raised by the learning of a large set of instrument sounds and propose a factorization of the model probability density function. This factorization allowed to learn instrument playing techniques from few samples and to deduce the model of techniques that are not in the training database but are known to be possible by the introduction of external instrumental knowledge. The evaluation of the method with classification tasks gave encouraging results. Future research will focus on the introduction of other dimension reduction techniques such as principal component analysis and on the extension of the descriptor set to better describe some playing techniques.

6. ACKNOWLEDGEMENTS

The authors wish to deeply thank the composer Yan Maresz for his involvement in this project.

7. REFERENCES

[1] E. Vincent and X. Rodet, "Music transcription with isa and hmm," in *ICA*, Granada, Espana, 2004, pp. 1197–1204.

[2] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *Proc. ICASSP*, Hong-Kong, 2003, pp. 613–616.

[3] C. Traube, P. Depalle, and M. M. Wanderley, "Indirect acquisition of instrumental gesture based on signal, physical and perceptual information." in *Proc. NIME*, Montreal, Canada, 2003, pp. 42–48.

[4] G. Hinton, "Products of experts," in *Proc. ICANN*, Edinburgh, Scotland, 2003, pp. 1–6.

[5] M. J. F. Gales and S. S. Airey, "Product of gaussians for speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 22–40, january 2006.

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

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

[8] B. Doval and X. Rodet, "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs," in *Proc. ICASSP*, Minneapolis, USA, 1993, pp. 221–224.