Instrument identification in solo and ensemble music using independent subspace analysis

Emmanuel Vincent, Xavier Rodet

INSTRUMENT IDENTIFICATION IN SOLO AND ENSEMBLE MUSIC USING INDEPENDENT SUBSPACE ANALYSIS

Emmanuel Vincent and Xavier Rodet
IRCAM, Analysis-Synthesis Group
1, place Igor Stravinsky – F-75004 PARIS – FRANCE
{vincent,rod}@ircam.fr

ABSTRACT

We investigate the use of Independent Subspace Analysis (ISA) for instrument identification in musical recordings. We represent short-term log-power spectra of possibly polyphonic music as weighted non-linear combinations of typical note spectra plus background noise. These typical note spectra are learnt either on databases containing isolated notes or on solo recordings from different instruments. We show that this model has some theoretical advantages over methods based on Gaussian Mixture Models (GMM) or on linear ISA. Preliminary experiments with five instruments and test excerpts taken from commercial CDs give promising results. The performance on clean solo excerpts is comparable with existing methods and shows limited degradation under reverberant conditions. Applied to a difficult duo excerpt, the model is also able to identify the right pair of instruments and to provide an approximate transcription of the notes played by each instrument.

1. INTRODUCTION

The aim of instrument identification is to determine the number and the names of the instruments present in a given musical excerpt. In the case of ensemble music, instrument identification is often thought as a by-product of polyphonic transcription, which describes sound as a collection of note streams played by different instruments. Both problems are fundamental issues for automatic indexing of musical data.

Early methods for instrument identification have focused on isolated notes, for which features describing timbre are easily computed. Spectral features such as pitch, spectral centroid (as a function of pitch), energy ratios of the first harmonics and temporal features such as attack duration, *tremolo* and *vibrato* amplitude have proved to be useful for discrimination [1].

These methods have been extended to solo and ensemble music using the Computational Auditory Scene Analysis (CASA) framework [1, 2, 3, 4]. The principle of CASA is to generate inside a blackboard architecture note hypotheses based on harmonicity and common onset and stream hypotheses based on timbre, pitch proximity and spatial direction. Hypotheses are validated or rejected according to prior knowledge and complex precedence rules. The best hypothesis is selected for final explanation.

Feature matching methods [3, 4] use the same timbre features as in isolated notes. Features computed in zones where several notes overlap are modified or discarded before stream validation depending on their type. Template matching methods [2] compare the observed waveform locally with sums of template waveforms, that are phase-aligned, scaled and filtered adaptively.

A limitation of such methods is that often timbre features or templates are used only for stream validation and not for note validation (except in [3]). This may result in some badly estimated notes, and it is not clear how note errors affect instrument identification. For example a bass note and a melody note forming a two-octave interval may be described as a single bass note with a "strange" spectral envelope. This kind or error could be avoided using the features or templates of each instrument in the note estimation stage.

Timbre features for isolated notes have also been used on solo music with statistical models which do not require note transcription. For example in [5, 6] cepstral coefficients are computed and modeled by Gaussian Mixture Models (GMM) or Support Vector Machines (SVM).

In order for the cepstral coefficients to make sense, these methods suppose implicitly that a single note is present at each time (or that the chords in the test excerpt are also present in the learning excerpts). Thus they are not applicable to ensemble music or to reverberant recordings and not robust towards background noise changes. Moreover they do not model the relationship between pitch and spectral envelope, which is an important cue.

In this article we investigate the use for instrument identification of another well-known statistical model: Independent Subspace Analysis (ISA). Linear ISA transcribes the short-time spectrum of a musical excerpt as a weighted sum of typical spectra, either adapted from the data or learnt in a previous step. Thus it performs template matching in the spectrum domain. Linear ISA of power spectrum has been applied to polyphonic transcription of drum tracks [7, 8] and of synthesized solo harpsichord [9]. But its ability to discriminate musical instruments seems limited, even on artificial data [10]. Linear ISA of cepstrum and log-power spectrum has been used for instrument identification on isolated notes [11] and general sound classification in MPEG-7 [12]. But, as the GMM and SVM methods mentioned above, it is restricted to single class data and sensitive to background noise changes.

Here we show that linear ISA is not adapted for instrument identification in polyphonic music. We derive a new ISA model with fixed nonlinearities and we study its performance on real recordings taken from commercial CDs.

The structure of the article is as follows. In Section 2 we describe a generative model for polyphonic music based on ISA. In Section 3 we explain how to use it for instrument identification. In Section 4 we study the performance of this model on solo music and its robustness against noise and reverberation. In Section 5 we show a preliminary experiment with a difficult duo excerpt. We conclude by discussing possible improvements.

2. INDEPENDENT SUBSPACE ANALYSIS

2.1. Need for a nonlinear spectrum model

Linear ISA of power spectrum explains a series of observed polyphonic power spectra $(\mathbf{x}_t)$ by combining a set of normalized typical spectra $(\mathbf{\Phi}_h)$ with time-varying powers (e_{ht}). For simplicity, this combination is usually modeled as a sum. This gives the generative model $\mathbf{x}_t = \sum_{h=1}^{H} e_{ht}\mathbf{\Phi}_h + \boldsymbol{\epsilon}_t$ where each note from each instrument may correspond to several typical spectra $(\mathbf{\Phi}_h)$, and where $(\boldsymbol{\epsilon}_t)$ is a Gaussian noise [9]. As a general notation in the following we use bold letters for vectors, regular letters for scalars and parentheses for sequences.

This linear model suffers from two limitations.

A first limitation is that the modeling error is badly represented as an additive noise term $\boldsymbol{\epsilon}_t$. Experiments show that the absolute value of $\boldsymbol{\epsilon}_t$ is usually correlated with $\mathbf{x}_t$, and that the modeling error may rather be considered as multiplicative noise (or as additive noise in the log-power domain). This is confirmed by instrument identification experiments, which use cepstral coefficients (or equivalently log-power spectral envelopes) as features, instead of power spectral envelopes [5, 6, 11]. This limitation seems crucial regarding the instrument identification performance of the model.

A second limitation is that summing power spectra is not an efficient way of representing the variations of the spectrum of a given note between different time frames. Many typical spectra are needed to represent small f0 variations in *vibrato*, wide-band noise during attacks or power rise of higher harmonics in *forte*. Summation of log-power spectra is more efficient. For instance it is possible to represent small f0 variations by adding to a given spectrum its derivative versus frequency with appropriate weights. It can easily be observed that this first order linear approximation is valid for a larger f0 variation range considering log-power spectra instead of power spectra.

We propose to solve these limitations using nonlinear ISA with fixed $\log(.)$ and $\exp(.)$ nonlinearities that transform power spectra into log-power spectra and *vice-versa*. The rest of this Section defines this model precisely.

2.2. Definition of the model

Let $(\mathbf{x}_t)$ be the short-time log-power spectra of a given musical excerpt containing n instruments. As usual for western music instruments, we suppose that each instrument j, $1 \leq j \leq n$, can play a finite number of notes h, $1 \leq h \leq H_j$, lying on a semitone scale (however the model could also be used to describe percussions).

Denoting $\mathbf{m}_{jt}$ the power spectrum of instrument j at time t and $\mathbf{\Phi'}_{jht}$ the log-power spectrum of note h from instrument j at time t, we assume

$$\mathbf{x}_t = \log\left[\sum_{j=1}^{n} \mathbf{m}_{jt} + \mathbf{n}\right] + \boldsymbol{\epsilon}_t, \tag{1}$$

$$\mathbf{m}_{jt} = \sum_{h=1}^{H_j} \exp(\mathbf{\Phi'}_{jht}) \exp(e_{jht}), \tag{2}$$

$$\mathbf{\Phi'}_{jht} = \mathbf{\Phi}_{jh} + \sum_{k=1}^{K} v_{jht}^{k} \mathbf{U}_{jh}^{k}, \tag{3}$$

where $\exp(.)$ and $\log(.)$ are the exponential and logarithm functions applied to each coordinate. The vector $\mathbf{\Phi}_{jh}$ is the unit-power mean log-power spectrum of note h from instrument j and $(\mathbf{U}_{jh}^{k})$ are L_2-normalized "variation spectra" that model variations of the spectrum of this note around $\mathbf{\Phi}_{jh}$. The scalar e_{jht} is the log-power of note h from instrument j at time t and (v_{jht}^{k}) are "variation scalars" associated with the "variation spectra". The vector $\mathbf{n}$ is the power spectrum of the stationary background noise. The modeling error vector $\boldsymbol{\epsilon}_t$ is supposed to be a white Gaussian noise.

Note that explicit modeling of the background noise is needed in order to prevent it being considered as a feature of the instruments present in the excerpt.

This nonlinear model could be approximated by the simpler one $\mathbf{x}_t = \max_{jh}\left[\mathbf{\Phi'}_{jht} + (e_{jht}, \ldots, e_{jht})^T\right] +$

$\boldsymbol{\epsilon}_t$. Indeed the log-power spectrum can be considered as a "preferential feature" as defined in [3], meaning that the observed feature is close to the maximum of the underlying single instrument features.

Eq. (1-3) are completed with probabilistic priors for the scalar variables. We associate to each note at each time a discrete state $E_{jht} \in \{0, 1\}$ denoting absence or presence. We suppose that these state variables are independent and follow a Bernoulli law with constant sparsity factor $P_Z = P(E_{jht} = 0)$. Finally we assume that given $E_{jht} = 0$ e_{jht} is constrained to $-\infty$ and v^k_{jht} to 0, and that given $E_{jht} = 1$ e_{jht} and v^k_{jht} follow independent Gaussian laws.

2.3. Computation of acoustic features

The choice of the time-frequency distribution for $(\mathbf{x}_t)$ is not imposed by the model. However comparison of spectral envelopes on auditory-motivated frequency scales or logarithmic scales has usually lead to better performance than linear scales for instrument identification [5]. Thus precision in upper frequency bands is not needed and could lead to over-learning. The modeling of f0 variations with Eq. (3) also advocates for a logarithmic frequency scale at upper frequencies, since f0 variations have to induce small spectral variations for the linear approximation to be valid.

In the following we use a bank of filters linearly spaced on the ERB scale $f_{ERB} = 9.26 \log(0.00437 f_{Hz} + 1)$ between 30 Hz and 11 KHz. The width of the main lobes is set to four times the filter spacing. We compute log-powers on 11 ms frames (a lower threshold is set to avoid drop-down to $-\infty$ in silent zones).

3. APPLICATION TO INSTRUMENT IDENTIFICATION

For each instrument j, we define the instrument model $\mathcal{M}_j$ as the collection of the fixed ISA parameters describing instrument specific properties: the spectra $(\mathbf{\Phi}_{jh})$ and $(\mathbf{U}^k_{jh})$ and the means and variances of the Gaussian variables e_{jht} and (v^k_{jht}) when $E_{jht} = 1$. We call orchestra $\mathcal{O} = (\mathcal{M}_j)$ a list of instrument models.

The idea for instrument identification is now to learn instrument models for several instruments in a first step, and in a second step to select the orchestra that best explains a given test excerpt. These two steps called learning and inference are discussed in this Section.

3.1. Inference

The probability of an orchestra is given by the Bayes law $P(\mathcal{O}|(\mathbf{x}_t)) \propto P((\mathbf{x}_t)|\mathcal{O})P(\mathcal{O})$. The determination of $P((\mathbf{x}_t)|\mathcal{O})$ involves an integration over the state and scalar variables which is intractable. We use instead the joint posterior $P_{\rm trans} = P(\mathcal{O}, (E_{jht}), (\mathbf{p}_{jht})|(\mathbf{x}_t))$ with $\mathbf{p}_{jht} = (e_{jht}, v^1_{jht}, \ldots, v^K_{jht})$. Maximizing $P_{\rm trans}$ means finding the best orchestra $\mathcal{O}$ explaining $(\mathbf{x}_t)$, but also the best state variables (E_{jht}), which provide an approximate polyphonic transcription of $(\mathbf{x}_t)$. Here again instrument identification and polyphonic transcription are intimately related.

$P_{\rm trans}$ is developed as the weighted Bayes law

$$P_{\rm trans} \propto (P_{\rm spec})^{w_{\rm spec}} (P_{\rm desc})^{w_{\rm desc}} P_{\rm state}\, P_{\rm orch}, \qquad (4)$$

involving the four probability terms $P_{\rm spec} = \prod_t P(\boldsymbol{\epsilon}_t)$, $P_{\rm desc} = \prod_{jht} P(\mathbf{p}_{jht}|E_{jht}, \mathcal{M}_j)$, $P_{\rm state} = \prod_{jht} P(E_{jht})$ and $P_{\rm orch} = P(\mathcal{O})$ and correcting exponents $w_{\rm spec}$ and $w_{\rm desc}$. Experimentally the white noise model for $\boldsymbol{\epsilon}_t$ is not perfectly valid, since values of $\boldsymbol{\epsilon}_t$ at adjacent time-frequency points are a bit correlated. Weighting by $w_{\rm spec}$ with $0 < w_{\rm spec} < 1$ is a way of taking into account these correlations [13].

Maximization of $P_{\rm trans}$ with respect to the orchestra $\mathcal{O}$ is carried out by testing all possibilities and selecting the best one. For each $\mathcal{O}$, the note states (E_{jht}) are estimated iteratively with a jump procedure. At start all states are set to 0, then at each iteration at most one note is added or subtracted at each time t to improve $P_{\rm trans}$ value. The optimal number of simultaneous notes at each time is not fixed *a priori*. The scalar variables $(\mathbf{p}_{jht})$ are re-estimated at each iteration with an approximate second order Newton method.

The stationary background noise power spectrum $\mathbf{n}$ is also considered as a variable, initialized as $\min_t \mathbf{x}_t$ and re-estimated at each iteration in order to maximize $P_{\rm trans}$.

The variance of $\boldsymbol{\epsilon}_t$ and the sparsity factor P_Z are set by hand based on a few measures on test data. The correcting exponents $w_{\rm spec}$ and $w_{\rm desc}$ are also set by hand depending on the redundancy of the data (larger values are used for ensemble music than for solos).

Setting a relevant prior $P(\mathcal{O})$ on orchestras would need a very large database of musical recordings to determine the number of excerpts available for each instrumental ensemble and each excerpt duration. Here for simplicity we use $P(\mathcal{O}) = P_Z^{-T(H_1+\cdots+H_n)}$ where T is the number of time frames of $(\mathbf{x}_t)$. This gives the same posterior probability to all orchestras on silent excerpts (*i.e.* when all states (E_{jht}) are equal to 0).

Obviously this prior tends to favor explanations with a large number of instruments, and thus cannot be used to determine the number of instruments in a relevant way. Experiments in the following are made knowing the number of instruments *a priori*.

Note that even if the prior was more carefully designed, the model would not be able to discriminate a violin solo from a violin duo. Indeed the selection of the good orchestra would only be based on the value of $P(\mathcal{O})$, independently of the monophonic or polyphonic character of the excerpt. To avoid this, the Bernoulli prior for state variables should be replaced by a more complex prior con-

straining instruments to play one note at a time (plus reverberation of the previous notes).

3.2. About "missing data"

We mentioned above that log-power spectra are "preferential features" as defined in [3]. It is interesting to note that inference with ISA treats "missing data" in the same way that preferential features are treated in [3]. Indeed the gradients of P_{trans} versus e_{jht} and v^k_{jht} involve the quantity

$$\pi_{jhtf} = \frac{\exp(\Phi'_{jhtf})\exp(e_{jht})}{\sum_{h'=1}^{H_j}\exp(\Phi'_{jh'tf})\exp(e_{jh't})} \qquad (5)$$

which is the power proportion of note h from instrument j into the model spectrum at time-frequency point (t, f). When this note is masked by other notes, $\pi_{jhtf} \approx 0$ and the value of the observed spectrum x_{tf} is not taken into account to compute e_{jht}, (v^k_{jht}) and E_{jht}. On the contrary when this note is preponderant $\pi_{jhtf} \approx 1$ and the value of x_{tf} is taken into account.

This method for "missing data" inference may use available information more efficiently than the bounded marginalization procedure in [4]. When several notes overlap in a given time-frequency point, the observed log-power in this point is considered to be nearly equal to the log-power of the preponderant note, instead of being simply considered as an upper bound to the log-powers of all notes.

3.3. Learning

Instrument models can be learnt from a large variety of learning excerpts, ranging from isolated notes to ensemble music. The learning procedure finds in an iterative way the model parameters that maximize P_{trans} on these excerpts. Each iteration consists in transcribing the learning excerpts as discussed above and then updating the instrument models in accordance.

The size of the model and the initial parameters are fixed by hand. In our experiments we set $K = 2$ for all instruments. The mean spectra $(\mathbf{\Phi}_{jh})$ were initialized as harmonic spectra with a -12 dB per octave shape. The "variation spectra" $(\mathbf{U}^1_{jh})$ and $(\mathbf{U}^2_{jh})$ initially represented wide-band noise and frequency variations respectively.

Experiments showed that learning on isolated notes is more robust since the whole playing range of each instrument is available and the state sequences are known *a priori*. We obtained lower recognition rates with instrument models learnt on solo excerpts only than with models learnt on isolated notes only (and the learning duration was also considerably longer).

The learning set used in the rest of the article consists in isolated notes from the RWC Database [14]. To make comparisons with existing methods easier, we consider the same five instruments as in [4]: flute, clarinet, oboe, bowed violin and bowed cello, abbreviated as Fl, Cl, Ob, Vn and Vc. All instruments are recorded in the same room, and for each one we select only the first performer and the most usual playing styles. Thus the learning set is quite small.

4. PERFORMANCE ON SOLO MUSIC

4.1. Clean conditions

The performance of the proposed method was first tested on clean solo music. For each instrument, we collected 10 solo recordings from 10 different commercial CDs. Then we constructed the test set by extracting 2 excerpts of 5 seconds out of each recording, avoiding silent zones and repeated excerpts.

Results are shown in Table 1. The average recognition rate is 90% for instruments and 97% for instrument families (woodwinds or bowed strings). This is similar to the 88% rate obtained in [4]. The main source of error is due to cello phrases containing only high pitch notes being easily confused with violin. However cello phrases containing both high pitch notes and low pitch notes are correctly classified. Ambiguous features of some notes inside a phrase are compensated by non ambiguous features of other notes.

To assess the relative importance of pitch cues and spectral shape cues, the same experiment was done with the default instrument models used for learning initialization, which all have -12 dB per octave spectra. The average instrument and family recognition rates dropped to 32% and 56% respectively, which is close to random guess (20% and 50%). Only cello had a good recognition rate (80%). This proves that the ISA model actually captures the spectral shape characteristics of the instruments and uses them in a relevant way for instrument discrimination.

		Identified instrument				
		Fl	Cl	Ob	Vn	Vc
Test excerpt	Fl	**100%**				
	Cl	5%	**85%**	5%	5%	
	Ob			**95%**	5%	
	Vn		5%		**95%**	
	Vc				25%	**75%**

Table 1. Confusion matrix for instrument recognition of clean five second solo excerpts from commercial CDs

4.2. Noisy or reverberant conditions

We also tested the robustness of the method against noisy or reverberant conditions.

We simulated reverberation by convolving the clean recordings with a room impulse response recorded at IRCAM (1 s reverberation time) having a non flat frequential response. The average instrument recognition rate decreased to 85%. Confusion was mainly augmented be-

tween close instruments (such as high pitch cello and low pitch violin).

Then we added white Gaussian noise to the clean recordings with various Signal to Noise Ratios (SNR). The average instrument recognition rate decreased to 83% at 20 dB SNR and 71% at 0 dB SNR when the noise spectrum $\mathbf{n}$ was provided *a priori*, and to 85% and 59% when it was estimated without constraints. Thus useful spectral information for instrument identification is still present in low SNR recordings and can be used efficiently. However the noise spectrum estimation procedure we proposed works at medium SNR but fails at low SNR. A first reason for this is that the hyper-parameters (variance of $\boldsymbol{\epsilon}_t$, P_Z, w_{spec} and w_{desc}) were given the same values for all test conditions, whereas the optimal values should depend on the data (for example the variance of $\boldsymbol{\epsilon}_t$ should be smaller at low SNR). A second reason is that the shape of the posterior is quite complex and that the simple jump procedure we proposed to estimate the note states becomes sensitive to noise initialization at low SNR. Small improvements (+2% at 20 and 0 dB SNR) were observed when initializing $\mathbf{n}$ *a priori*. Other Bayesian inference procedures such as Gibbs Sampling may help solve this problem.

5. PERFORMANCE ON ENSEMBLE MUSIC

Finally the performance of the method was tested on ensemble music. Since we encountered difficulties in collecting a significant amount of test recordings, we show here only the preliminary results obtained on an excerpt from Pachelbel's canon in D arranged for flute and cello. This is a difficult example because 10 flute notes out of 12 are harmonics of simultaneous cello notes, and melody (flute) notes belong to the playing range of both instruments, as can be seen in Fig 1.

The results of instrument identification are shown in Fig 2. Using the number of instruments as *a priori* knowledge, the model is able to identify the right orchestra. Note that there is a large likelihood gap between orchestras containing cello and others. Orchestras containing only high-pitched instruments cannot model the presence of low-pitch notes, which is a coarse error. Orchestras containing cello but not flute can model all the notes, but not with the right spectral envelope, which is a more subtle kind of error.

The note states $E_{1,ht}$ and $E_{2,ht}$ inferred with the right orchestra are shown in Fig 1. All the notes are correctly identified and attributed to the right instrument, even when cello and flute play harmonic intervals such as two octaves or one octave and a fifth. There are some false alert notes, mostly with with short duration. If a precise polyphonic transcription is needed, these errors could be removed using time integration inside the model to promote long duration notes. For example the Bernoulli prior for state variables could be replaced with a Hidden Markov Model (HMM) [15], or even with a more complex model involving rhythm, forcing instruments to play monophonic phrases or taking into account musical knowledge [2].

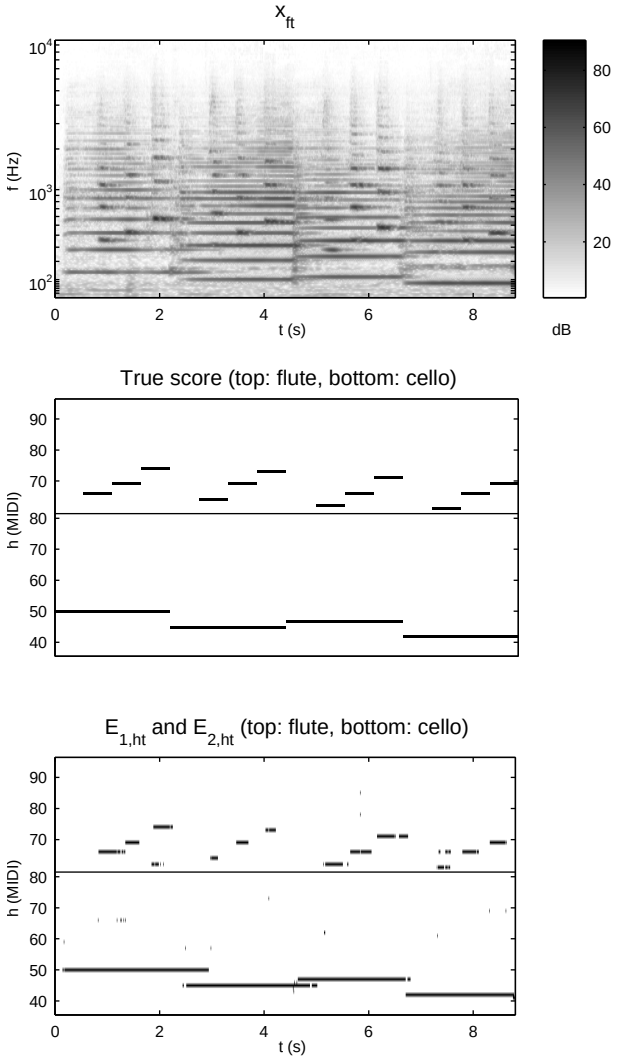

Figure 1. Spectrogram of a flute and cello excerpt and approximate transcription (with the right orchestra) compared with the true score.

6. CONCLUSION

In this article we proposed a method for instrument identification based on ISA. We showed that the linear ISA framework is not suited for this task and we proposed a new ISA model containing fixed nonlinearities. This model provided good recognition rates on solo excerpts and was shown to be robust to reverberation. It was also able to determine the right pair of instruments in a difficult duo excerpt and to transcribe it approximatively.

Compared to other statistical models such as GMM and SVM, ISA has the advantage of being directly applicable to polyphonic music without needing a prior note tran-

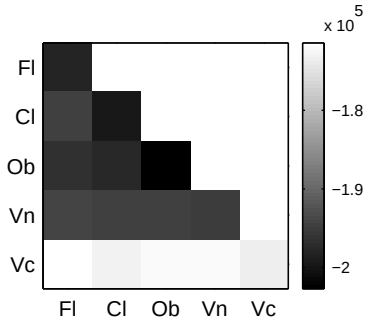

Figure 2. Log-likelihoods of the duo orchestras on the duo excerpt of Fig. 1

scription step. Instrument identification and polyphonic transcription are embedded in a single optimization procedure. This procedure uses learnt note spectra for each instrument, which makes it successful for both tasks even in difficult cases involving harmonic notes.

However a few problems still have to be fixed, for instance better estimating the background noise by selecting automatically the values of the hyper-parameters depending on the data, determining the number of instruments with a better orchestra prior, and separating streams using musical knowledge when one instrument plays several streams. The computational load may also be a problem for large orchestras, and could be reduced using prior information from a conventional multiple f0 tracker. We are currently studying some of these questions.

An interesting way to improve the recognition performance would be to add a prior on the time evolution of the state variables E_{jht} or of the scalar variables e_{jht} and v^k_{jht}. For example in [8] time-continuity of the scalar variables is exploited. In [11] a HMM is used to segment isolated notes into attack/sustain/decay portions and different statistical models are used to evaluate the features on each portion. This uses the fact that many cues for instrument identification are present in the attack portion [1]. This single note HMM could be extended to multiple notes and instruments supposing that all notes evolve independently or introducing a coupling between notes and instruments.

Besides its use for instrument identification and polyphonic transcription, the ISA model could also be used as a structured source prior for source separation in difficult cases. For example in [15] we couple instrument models and spatial cues for the separation of underdetermined instantaneous mixtures.

7. REFERENCES

[1] K.D. Martin, *Sound-source recognition : A theory and computationnal model*, Ph.D. thesis, MIT, 1999.

[2] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*, vol. 27, pp. 337–349, 1999.

[3] T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in *Proc. IJCAI Worshop on CASA*, 1999, pp. 18–24.

[4] J. Eggink and G.J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," in *Proc. ISMIR*, 2003.

[5] J. Marques and P.J. Moreno, "A study of musical instrument classification using Gaussian Mixture Models and Support Vector Machines," Tech. Rep., Compaq Cambridge Research Lab, june 1999.

[6] J.C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *Journal of the ASA*, vol. 109, no. 3, pp. 1064–1072, 2001.

[7] D. Fitzgerald, B. Lawlor, and E. Coyle, "Prior subspace analysis for drum transcription," in *Proc. AES 114th Convention*, 2003.

[8] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. ICMC*, 2003.

[9] S.A. Abdallah and M.D. Plumbley, "An ICA approach to automatic music transcription," in *Proc. AES 114th Convention*, 2003.

[10] J. Klingseisen and M.D. Plumbley, "Towards musical instrument separation using multiple-cause neural networks," in *Proc. ICA*, 2000, pp. 447–452.

[11] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Proc. ISSPA*, 2003.

[12] M.A. Casey, "Generalized sound classification and similarity in MPEG-7," *Organized Sound*, vol. 6, no. 2, 2002.

[13] D.J. Hand and K. Yu, "Idiot's bayes - not so stupid after all ?," *Int. Statist. Rev.*, vol. 69, no. 3, pp. 385–398, 2001.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: database of copyright-cleared musical pieces and instrument sounds for research purposes," *Trans. of Information Processing Society of Japan*, vol. 45, no. 3, pp. 728–738, 2004.

[15] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Proc. ICA*, 2004.