HAL
open science

Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds

Jun Wu, Emmanuel Vincent, Stanislaw Andrzej Raczynski, Takuya Nishimoto, Nobutaka Ono, Shigeki Sagayama

To cite this version:

Jun Wu, Emmanuel Vincent, Stanislaw Andrzej Raczynski, Takuya Nishimoto, Nobutaka Ono, et al.. Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds. IEEE Journal of Selected Topics in Signal Processing, 2011. inria-00594965v1

HAL Id: inria-00594965

https://inria.hal.science/inria-00594965v1

Submitted on 23 May 2011 (v1), last revised 1 Jun 2011 (v2)

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.

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds

Jun Wu, Emmanuel Vincent, Stanisław Andrzej Raczyński, Takuya Nishimoto, Nobutaka Ono and Shigeki Sagayama

Abstract — **Polyphonic pitch estimation and musical instrument identification are some of the most challenging tasks in the field of Music Information Retrieval (MIR). While existing approaches have focused on the modeling of harmonic partials, we design a joint Gaussian mixture model of the harmonic partials and the inharmonic attack of each note. This model encodes the power of each partial over time as well as the spectral envelope of the attack part. We derive an Expectation-Maximization (EM) algorithm to estimate the pitch and the parameters of the notes. We then extract timbre features both from the harmonic and the attack part via Principal Component Analysis (PCA) over the estimated model parameters. Musical instrument recognition for each estimated note is finally carried out with a Support Vector Machine (SVM) classifier. Experiments conducted on mixtures of isolated notes as well as real-world polyphonic music show higher accuracy over state-of-the-art approaches based on the modeling of harmonic partials only.**

Index Terms—**Instrument identification, Harmonic model, attack model, EM algorithm, PCA, SVM**

I. INTRODUCTION

Polyphonic musical instrument identification consists of estimating the pitch, the onset time and the instrument associated with each note in a music recording involving several instruments at a time. This is often addressed by conducting multiple pitch estimation first, then classifying each note into an instrument class using suitable timbre features [1,2,3,4].

Multiple pitch estimation is the task of estimating the fundamental frequencies and the onset times of the musical notes simultaneously present in a given musical signal. It is considered to be a difficult problem mainly due to the overlap between the harmonics of different pitches, a phenomenon common in Western music, where combinations of sounds that share some partials are preferred. Several approaches have been proposed, including perceptually motivated [5,6,7,8], parametric signal model-based [9,10], classification-based [11] and parametric spectrum model-based [12,13,14,15,16] algorithms. Parametric spectrum model-based algorithms represent the power spectrum or the magnitude spectrum of the observed signal as the sum or the mixture of individual note spectra or harmonic partial spectra and perform parameter estimation in the Maximum Likelihood (ML) sense. These algorithms are particularly suitable in the context of polyphonic instrument identification since they do not only provide the pitch of each note but also additional parameters encoding part of its timbre.

Timbre features have been widely investigated for the classification of isolated notes or single-instrument recordings and gradually applied to polyphonic recordings. Typical features computed on the signal as a whole include power spectra [17], spectral or cepstral features [18,19] as well as temporal features [20]. These features are not directly computable from the parameters of a multiple pitch estimation model. By contrast, timbre features have been derived in an unsupervised fashion from the amplitudes of the harmonic partials in [21,22,23] either via Multidimensional Scaling (MDS) or Principal Component Analysis (PCA). Supervised timbre models involving a source-filter-decay model or a dynamic statistical model of the amplitudes of the partials trained over labeled training data were also considered in [1,2]. Classification is then performed either via the Euclidean distance between the feature vectors or via maximum likelihood (ML) under the above models. In addition to their ease of use in the context of multiple pitch estimation, these algorithms reduce the dimension of the timbre parameter set, resulting in increased robustness with respect to parameter estimation errors. Feature weighting techniques were proposed in [3,4,24] to further improve robustness by associating a smaller weight to the parameters of overlapping partials, which are likely to be less accurately estimated.

While the attack part of musical notes is essential for timbre perception [20], the above multiple pitch estimation and timbre feature models have focused on the representation of harmonic partials only. The attack part consists of an inharmonic sound and may be characterized in particular by its spectral envelope and its power, both of which depend on the instrument. Designing an instrument model able to deal both with harmonic and inharmonic features is essential for reflecting the timbre

Jun Wu, Stanisław A. Raczyński and Shigeki Sagayama are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: wu@hil.t.u-tokyo.ac.jp, raczynski@hil.t.u-tokyo.ac.jp, sagayama@hil.t.u-tokyo. ac.jp). Emmanuel Vincent is with INRIA, Centre de Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France (e-mail: emmanuel.vincent@inria.fr). Takuya Nishimoto is with Olarbee Japan, Akiku, Hiroshima 736-0088, Japan (e-mail: nishimotz@olarbee.com). Nobutaka Ono is with the Principles of Informatics Research Division, The National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: onono@nii.ac.jp). This research was performed while Takuya Nishimoto and Nobutaka Ono were with the Graduate School of Information Science and Technology, The University of Tokyo.

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

characteristics of any musical instrument. In [25], a joint parametric harmonic and non-parametric inharmonic model was proposed and used for source separation given the pitch and instrument of all notes. In [26], we defined a joint parametric model of harmonic and attack sounds but considered timbre features derived from the harmonic part only. Therefore, attack timbre features have not been exploited for polyphonic musical instrument identification to date.

In this article, we propose an algorithm for polyphonic pitch estimation and instrument identification by joint modeling of harmonic and attack sounds. At first a flexible harmonic model is proposed to model the harmonic and attack parameters of musical notes via a mixture of time-frequency Gaussian distributions. These parameters are then estimated from a given recording together with the time-varying fundamental frequency using the Expectation-Maximization (EM) algorithm. Timbre features are subsequently derived by PCA from the model parameters after suitable logarithmic transformation and normalization. Finally, instrument classification is performed for each note via a Support Vector Machine (SVM)-based classifier instead of Euclidean distance or likelihood. We thereby extend our preliminary paper [26] by providing a more detailed treatment of the model, defining more efficient timbre features and separately evaluating the resulting performance in terms of pitch estimation and instrument identification. Experimental results show that the proposed features outperform the features in [26].

The overall flowchart of the proposed system is illustrated in Figure 1. The output of the proposed system is the estimated collection of pitches underlying the musical signal and the different colors represent different instruments.

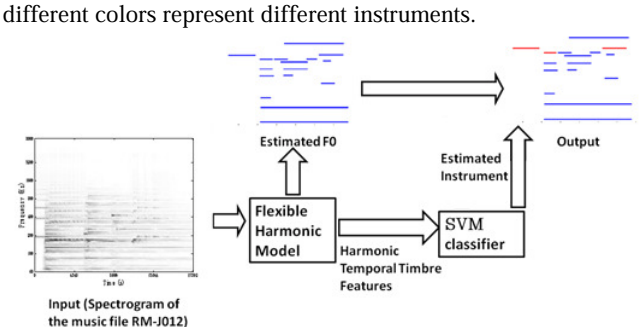

Figure 1. Flow chart of the proposed system.

The structure of the rest of this article is as follows. In Section II, the joint model of sustained and attack sounds is introduced. In Section III, parameter estimation and classification algorithms are presented. Experimental results on synthetic and real-world data are shown in Section IV. Finally, the conclusion is made in Section V.

II. JOINT MODELING OF SUSTAINED AND ATTACK SOUNDS

We adopt the same two-stage approach as a majority of algorithms [1,2,4,24]: a multipitch estimation stage provides the estimated pitch of all notes in the recording and an instrument identification stage classifies each note into a specific instrument category. However, while most algorithms rely on a different model for each stage, we use the same model for both stages. This model describes both the spectral and the temporal envelope by a mixture of Gaussian distributions as in [14] with significant improvements detailed hereafter. The main difficulty of polyphonic musical instrument identification is the overlapping of observed partials from different timbres. So an applicable model should also be able to associate the corresponding partials with specific timbres.

In the following, we assume that the input signal is sampled at 16 kHz and represented by its power constant-Q transform [14]. The transform is computed using Gabor-wavelet basis functions with a time resolution of 16 ms for the lowest subband. The time resolution is set to 16 ms for all subbands. The lower bound of the frequency range and the frequency resolution are 60 Hz and one semitone, respectively, as in [14].

Denoting by x and t the frequency bin and time frame indexes respectively, the proposed model approximates the observed nonnegative power spectrogram $W(x,t)$ by a mixture of K nonnegative parametric models, each of which represents a single musical note. Every note model is composed of a harmonic part, itself consisting of N harmonic partials, and an attack part. Figure 2 depicts the spectrogram of a piano note with the attack part being marked with a rectangle. The power spectrogram of the kth note is represented as

$$q_k(x,t) = w_k \sum_{n=1}^{N} H_{k,n}(x,t) + A_k(x,t). \quad (1)$$

where w_k is the total energy of the harmonic part, $H_{k,n}(x,t)$ represents the spectrogram of the nth harmonic partial and $A_k(x,t)$ the spectrogram of the attack part. The list of model parameters is shown in Table 1.

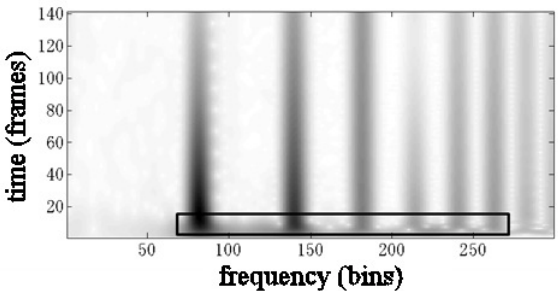

Figure 2. Spectrogram of a piano note signal. The rectangle marks the attack part of the note.

Parameter	Physical meaning
μ_k	Pitch of the kth note
w_k	Energy of the harmonic part of the kth note
$v_{k,n}$	Relative energy of the nth partial of the kth note
$u_{k,n,y}$	Coefficient of the spectro-temporal envelope of the kth note, nth partial, yth time instant
τ_k	Onset time of the kth note
$Y\phi_k$	Duration of the kth note (Y is constant)
σ_k	Bandwidth of the partials of the kth note
$\alpha_{k,j}$	Coefficient of the spectral envelope of the attack of the kth note, jth frequency band

Table 1. Free parameters of the proposed model.

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

A. Harmonic Model

The proposed model for the harmonic part is similar to [14]. However, in contrast to [14], the time-domain envelope is assumed to be different for each partial. This modification has significant impact on instrument identification since differences between the temporal evolution of the partials contribute to the characterization of timbre [2].

The harmonic model of each partial $H_{k,n}(x,t)$ is defined as the product of a spectral model $F_{k,n}(x)$ and a temporal model $U_{k,n}(t)$. Due to the use of a Gabor constant-Q transform, the spectral harmonic model follows a Gaussian distribution, as illustrated in Figure 3. The bandwidth is approximately equal for all partials on a log scale so a constant standard deviation σ_k can be used. Given the fundamental log-frequency μ_k of the *k*th note, the log-frequency of the *n*th partial is given by $\mu_k + \log n$. This results in

$$F_{k,n}(x) = v_{k,n}\mathcal{N}(x;\mu_k + \log n, \sigma_k^2) \tag{2}$$

where $v_{k,n}$ is the relative power of the *n*th partial satisfying

$$\sum_{n=1}^{N} v_{k,n} = 1 \ \forall k \tag{3}$$

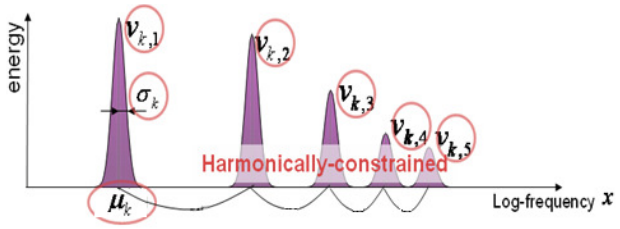

Figure 3. Representation of the spectral models $F_{k,n}(x)$ of all partials *n*.

The temporal model of each partial is designed as a Gaussian Mixture Model (GMM) with constrained means representing time sampling instants as shown in Figure 4. More precisely, the number of Gaussians is fixed to Y and the means are uniformly spaced over the duration of the note, resulting in

$$U_{k,n}(t) = \sum_{y=0}^{Y-1} u_{k,n,y}\mathcal{N}(t;\tau_k + y\phi_k, \phi_k^2) \tag{4}$$

where τ_k is the mean of the first Gaussian, which is considered as the onset time, $u_{k,n,y}$ is the weight parameter for each time instant, which allows the temporal envelope to have a variable shape for each harmonic partial, and ϕ_k is the spacing between successive sampling instants, which is proportional to the note duration $Y\phi_k$. The weight parameters are normalized as

$$\sum_{y=0}^{Y-1} u_{k,n,y} = 1 \ \forall k, \forall y. \tag{5}$$

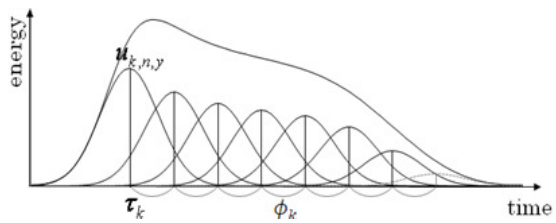

Figure 4. Representation of the temporal model $U_{k,n}(t)$ of one partial *n*.

The Dirichlet distribution is used as a prior distribution over $u_{k,n}$ and $u_{k,n,y}$

$$P(v_{k,1},\dots,v_{k,N}) = \frac{\Gamma(\sum_n (d_v\bar{v}_n+1))}{\prod_n \Gamma(d_v\bar{v}_n+1)} \prod_{n=1}^{N} v_{k,n}{}^{d_v\bar{v}_n} \tag{6}$$

$$P(u_{k,n,0},\dots,u_{k,n,Y-1}) = \frac{\Gamma(\sum_y (d_u\bar{u}_y+1))}{\prod_y \Gamma(d_u\bar{u}_y+1)} \prod_{n=1}^{N} u_{k,n,y}{}^{d_u\bar{u}_y} \tag{7}$$

where Γ is the gamma function, $\bar{v}_n$ and $\bar{u}_y$ denote the expected values of $v_{k,n}$ and $u_{k,n,y}$ and d_v and d_u regulate the strength of the priors.

B. Attack Model

We now define the attack model $A_k(x,t)$ as the product of a spectral model $F'_k(x)$ and a temporal model $U'_k(t)$. Our model differs from the nonparametric inharmonic model in [25] in two ways: it does not represent sustained inharmonic sounds but the attack part only and it involves much fewer parameters due to its parametric expression. These two differences make sense in our application context, where no prior information is available contrary to the informed source separation context in [25] where pitch, onset, duration and instrument are known.

The temporal attack model is expressed by a single Gaussian

$$U'_k(t) = \mathcal{N}(t;\tau_k,\phi_k^2) \tag{8}$$

Because the attack occurs at the same time as the onset of the harmonic partials, this distribution is equal to the first Gaussian component of the temporal harmonic model.

The spectral attack model is represented by a GMM with constrained means, where the number of Gaussians is fixed to J and the means are uniformly spaced over the whole log-frequency axis. This gives

$$F'_k(x) = \sum_{j=1}^{J} \alpha_{k,j}\mathcal{N}(x;\mu_j,\sigma^2) \tag{9}$$

where the means μ_j and standard deviation σ satisfy $\mu_{j+1} = \mu_j + \sigma$ and the weights $\alpha_{k,j}$ encode the spectral envelope.

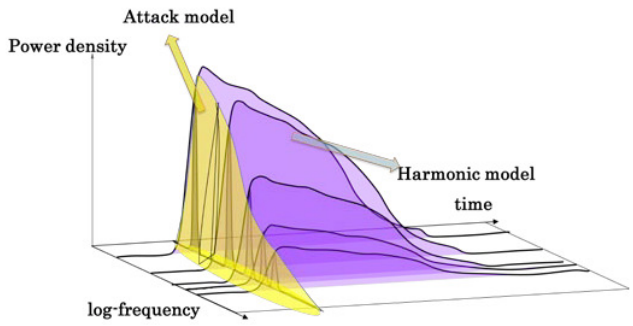

Figure 5. Overall representation of the proposed model.

C. Overall model

The whole proposed model including the harmonic part and attack part is illustrated in Figure 5. The harmonic model part is a GMM in the time and log-frequency direction while the attack model part is a GMM in the log-frequency direction. Overall, this can be expressed as

$$q_k(x,t;\theta) = \sum_z S_{k,z}(x,t;\theta) \tag{10}$$

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

where z indexes $N \times Y + J$ Gaussians representing either the harmonic part (one Gaussian per partial n and per time sampling instant y) or the attack part (one Gaussian per subband j) and θ denotes the full set of parameters of all notes. Therefore the whole signal is also represented as a mixture of spectro-temporal Gaussian distributions $S_{k,z}(x,t;\theta)$.

III. PARAMETER ESTIMATION AND CLASSIFICATION ALGORITHMS

A. Inference with the EM algorithm

We subsequently employ the EM algorithm [27] to estimate the parameters of our model. We assume that the observed power density $W(x,t)$ has an unknown fuzzy membership to the kth note, represented by a spectro-temporal mask $m_k(x,t)$. To minimize the difference between the observed spectrogram $W(x,t)$ and the note models, we use the Kullback–Leibler (KL) divergence as the global cost function

$$J = \sum_k \iint_D m_k(x,t) W(x,t) \log \frac{m_k(x,t)W(x,t)}{q_k(x,t;\theta)} dxdt \tag{11}$$

where D denotes the whole time-frequency plane. Therefore the problem is regarded as the minimization of (11) under the constraints

$$\sum_k m_k(x,t) = 1 \forall x, \forall t \tag{12}$$
$$0 \le m_k(x,t) \le 1 \forall k, \forall x, \forall t. \tag{13}$$

The parameters θ of the note models $q_k(x,t;\theta)$ and the corresponding masks $m_k(x,t)$ are both unknown and must be estimated. These quantities are initialized as described in Section IV.B and iteratively optimized using the EM algorithm, where the E-step updates $m_k(x,t)$ with θ fixed and the M-step updates θ with $m_k(x,t)$ fixed. The number of notes K is also estimated as explained in Section IV.B.

Since each note model is composed of several Gaussians $S_{k,z}(x,t;\theta)$, we use a complementary set of masks $m_{k,z}(x,t)$ to represent the fuzzy membership of $m_k(x,t)W(x,t)$ to the zth Gaussian. By apply Jensen's inequality, we get

$$\iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{\sum_{k,z} S_{k,z}(x,t;\theta)} dxdt \leqq \sum_z \iint_D m_k(x,t)m_{k,z}(x,t)W(x,t) \log\frac{m_k(x,t)m_{k,z}(x,t)W(x,t)}{S_{k,z}(x,t;\theta)} dxdt \tag{14}$$

Equality holds when

$$m_{k,z}(x,t) = \frac{S_{k,z}(x,t;\theta)}{\sum_z S_{k,z}(x,t;\theta)} \tag{15}$$

satisfying the following conditions:

$$\sum_z m_{k,z}(x,t) = 1, \forall k, \forall x, \forall t \tag{16}$$
$$0 \le m_{k,z}(x,t) \le 1, \forall k, \forall z, \forall x, \forall t. \tag{17}$$

The E-step is achieved by setting

$$m_k(x,t)m_{k,z}(x,t) = \frac{S_{k,z}(x,t;\theta)}{\sum_k \sum_z S_{k,z}(x,t;\theta)} \tag{18}$$

The M-step consists of updating each parameter in turn, where the updates can be obtained analytically using Lagrange multipliers. The update equations are given in Appendix. The computation time of the proposed approach is about 1.1 times that of the original HTC algorithm [14].

B. Feature extraction

Assuming that the model parameters have been estimated, we now exploit these parameters to derive relevant features for instrument identification. By contrast with previous approaches, we extract features jointly from harmonic and attack parameters. Also, contrary to [26], we do not consider the parameters themselves but apply a logarithmic transformation which increases correlation with subjective timbre perception [2] and makes their distribution closer to Gaussian [1], as needed by PCA. The impact of these choices is analyzed in Section IV.

For each note k, we extract a large feature vector consisting of the following six categories of features:

1. note energy feature $\log(w_k)$,
2. relative partial energy features $\log(v_{k,n})$ for all n,
3. partial bandwidth feature $\log(\sigma_k)$,
4. harmonic temporal envelope features $\log(u_{k,n,y})$ for all n and y,
5. note duration feature $\log(\phi_k)$,
6. attack spectral envelope features $\log(\alpha_{k,j})$ for all j.

Note that the choice of a GMM as the temporal model for the harmonic part enables the extraction of a fixed number of harmonic temporal envelope features from all notes, regardless of their duration.

C. PCA for dimension reduction

While this feature vector encodes relevant timbre information, it cannot be directly used as the input to an instrument classifier. Indeed, its large dimension makes it sensitive to overfitting and to outliers, due to e.g. possible misestimation of the parameters of overlapping partials. These issues are classically addressed by dimension reduction techniques [21,22,23].

We here use PCA to transform the above feature vector into a low-dimension vector. This transformation is carried over the whole feature vector, so as to account for possible redundancies between harmonic and attack features. Because centering and normalization play a crucial role in PCA (features with low variance are discarded even when they are discriminative), we subtract the mean of each feature and normalize it by its largest absolute value over the training data beforehand so that it ranges from -1 to 1.

In order to illustrate the result, we computed the proposed features for five instruments among the training data of Section IV and plot the first three principal components of the feature set without attack features in Figure 6 and of the full feature set with attack features in Figure 7. These figures show that harmonic features allow some discrimination of the instruments to a certain extent, but that attack features contribute to increasing the margin between certain pairs of instruments, e.g. alto sax and piano or piano and violin.

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

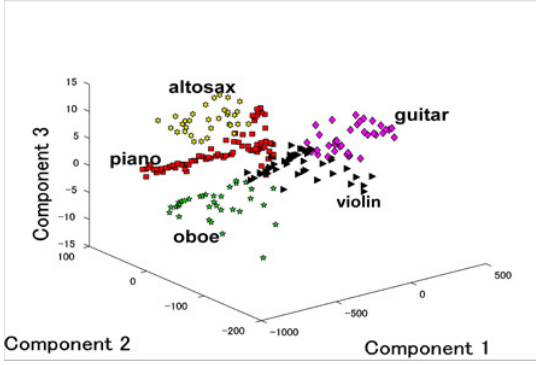

Figure 6. First three principal components of the proposed feature set without attack features.

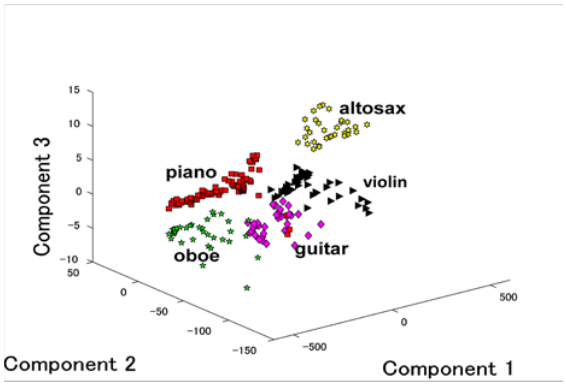

Figure 7. First three principal components of the proposed feature set with attack features.

In order to increase discrimination, a larger number of components is used in our experiments. We attempted a qualitative interpretation of these components. However, due to the normalization step, most features were active in some component, so that there was no obvious interpretation.

D. SVM for instrument classification

For each note k, instrument identification is achieved by classifying the corresponding low-dimension feature vector into one instrument class. To this aim, we use a set of SVM classifiers with radial basis function (RBF) kernel $K(x, x') = \exp(-\gamma \parallel x - x' \parallel^2)$ [28] where x is the feature vector composed of the values in Section III-B. SVMs are state-of-the art classifiers which maximize the margin between two classes of feature vectors in a high-dimensional space associated with the kernel. In order to solve the multi-class classification problem at hand, we use the one-versus-all approach: we train a SVM to classify each instrument versus all others and select the class which yields the greatest margin.

Training is performed on feature vectors extracted from isolated notes of each instrument. In order to account for the dependency of timbre features on pitch, a separate set of SVMs is trained for each pitch on the semitone scale. Since the accuracy of an SVM largely depends on the selection of the kernel parameters, we use 10-fold cross-validation to optimize the parameter γ of the RBF kernel on the training database.

IV. EXPERIMENTS

Since the proposed system aims to address both pitch estimation and instrument identification, we evaluate it according to three complementary tasks, namely multiple pitch estimation, instrument identification given the true pitches, and joint pitch estimation and instrument identification.

A. Training and test data

Training is performed on isolated notes from 9 instruments taken from three databases: the RWC database [29], McGill University Master Samples CD library [30] and the UIowa database [31]. The number of notes from each database is listed in Table 2.

Testing is performed on both synthetic mixtures of isolated notes and on real-world data. For each instrument of each database, we randomly generate 60 signals of 6 s duration. Each signal contains more than two notes and consists of both notes with similar onset times and notes in a sequence. We then randomly sum with each other the signals of different instruments within the same database so as to obtain 45 synthetic polyphonic test mixtures with the same duration. In addition, we use the real-world development data of the Multiple Fundamental Frequency Estimation & Tracking track of the 2007 Music Information Retrieval Exchange (MIREX) [32]. These data consist of five synchronized woodwind tracks, which we randomly cut to 6 s and sum together in order to obtain 30 real-world polyphonic test mixtures.

Since the timbre features of each instrument depend on the recording conditions, it is essential to use different databases for training and testing. In the following, we evaluate multiple pitch estimation and instrument identification performance on each of the three above databases (RWC, McGill or UIowa), while using the remaining two for learning. The results are then averaged over the three databases.

	McGill	RWC	UIowa	Total
bassoon	16	112	113	241
cello	40	430	337	807
clarinet	47	120	423	590
flute	90	36	226	352
oboe	27	34	104	165
piano	67	88	88	243
tuba	16	90	111	217
viola	32	467	271	770
violin	93	45	283	421
Total	428	1422	1956	3806

Table 2. Number of isolated notes from databases.

B. Model settings

The proposed model includes a number of hyper-parameters, which are either fixed or estimated from the data as follows. The number of harmonic partials N and the number of time sampling instants Y are fixed to 20 and 10, respectively. The number of coefficients J of the attack model is set to 20, since we found it to provide the best accuracy experimentally. Fol-

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

lowing [14], the parameters of the prior distributions $\bar{v}_n$, $\bar{u}_y$, d_v and d_u are set to $0.6547n^{-2}$, $0.2096e^{-0.2y}$, 0.04 and 0.04, respectively.

The other model parameters are initialized as in [14]. In particular, the number of note models K is initialized as 60 and the fundamental log-frequency μ_k and the onset time τ_k of each note are initialized to the log-frequency and time frame of the K largest peaks in the observed spectrogram. σ_k is initialized as 2.0, ϕ_k is initialized as 5.0. After the EM algorithm has converged, the notes k whose energy per unit time $w_k/Y\phi_k$ is smaller than the average energy per unit time over all notes are discarded. This procedure allows automatic determination of the number of notes K.

Finally, we then extract the first 20 principal components of the feature vector. This number of components accounts for 99.3% of the variance of the training data and was found to provide good results experimentally.

Figure 8.b illustrates the result of the proposed algorithm with the above setting on an excerpt from the song *RM-J012* in the RWC database [29].

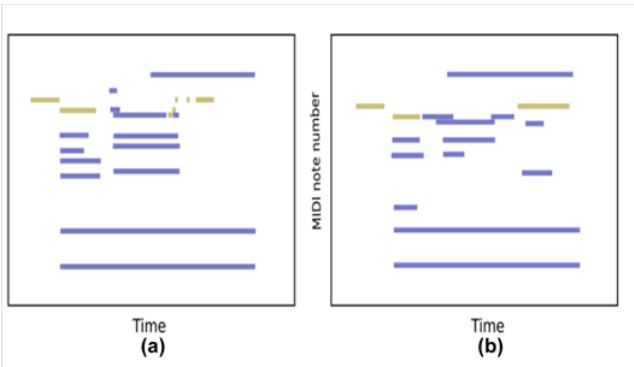

Figure 8. Comparison of the ground truth pitches (a) and the estimated pitches (b) for song *RM-J012* of the RWC database. Piano notes are represented in blue and flute notes in yellow.

C. Evaluation of multiple pitch estimation

In a first experiment, we assess multiple pitch estimation performance alone using the MIREX note tracking criteria [32]. A returned pitch-onset pair is considered as correct if it is within 1/4 tone and 50ms of a ground-truth note. The proportion of deleted and inserted notes is measured in terms of recall R and precision P. The F-measure is calculated from these two values as F = 2RP/(R + P).

We compare the proposed model with the NMF algorithm in [13] and the original HTC algorithm in [14]. The parameters of NMF are set as in [13] and those of HTC as in Section IV.B. To detect notes in the coefficient matrix of NMF, we use the procedure in [13] based on median filtering, thresholding and discarding of notes with short duration.

The results are shown in Table 3. Our algorithm outperforms NMF and HTC both in terms of recall and precision. The resulting improvement in terms of F-measure is equal to 13% and 6% on synthetic data and 15% and 6% on real-world data, respectively. This improvement is due in particular to the introduction of the attack model, which avoids errors due to fitting of inharmonic sounds by harmonic partials.

	Synthetic data			real-world data		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
NMF	72.5	74.4	73.4	44.1	46.6	45.3
HTC	82.0	78.7	80.3	57.4	51.3	54.2
Proposed	85.3	86.5	**85.9**	59.7	61.4	**60.5**

Table 3. Multiple pitch estimation performance

D. Evaluation of instrument identification given the true pitches

In a second experiment, we assume that the pitch and onset time of each note are known. We use the proposed multiple pitch estimation algorithm to estimate the remaining unknown parameters of each note and assess the subsequent instrument identification performance alone. The estimated instrument is considered as correct if it is the ground truth instrument. The resulting accuracy is the percentage of notes associated with the correct instrument.

The proposed algorithm is compared with conventional 12-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) [33], with the source-filter model of harmonic partials in [2] and with the harmonic features proposed in our previous work [26]. MFCCs are extracted from the power spectrum of each note $m_k(x,t)W(x,t)$ and classified by SVM. Source-filter features are classified by ML using the likelihood function defined in [2]. Finally, in order to directly compare SVM and ML, we also classify the proposed features by ML, where the likelihood function stems from the Gaussian model underlying PCA. We calculated the Euclidean distance between the training data and testing data, for every testing note the smallest Euclidean distance is obtained when the note is projected into the correct category.

Number of Instruments	1	2	3	4	Average
MFCC + SVM	66.5	60.8	53.4	44.1	56.2
Source-filter + ML	78.7	74.3	69.2	67.5	72.4
Harmonic features [26]	77.5	72.8	66.4	66.3	70.8
Proposed features (without attack) + ML	79.5	73.8	70.4	68.7	73.1
Proposed features (without attack) + SVM	82.7	78.5	72.0	70.2	75.9
Proposed features (with attack) + ML	82.3	77.4	71.9	70.4	75.5
Proposed features (with attack) + SVM	**84.5**	**80.7**	**73.8**	**72.7**	**77.9**

Table 4. Accuracy (%) for instrument identification given the true pitches (synthetic data).

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

Number of instruments	1	2	3	4	Average
MFCC	57.5	52.4	43.3	38.7	48.0
Source-filter + ML	72.6	69.1	62.9	59.6	66.1
Harmonic features[26]	70.4	65.3	61.4	56.2	63.3
Proposed features (without attack) + ML	72.3	69.8	63.5	60.4	66.5
Proposed features (without attack) + SVM	75.8	72.4	65.7	62.5	69.1
Proposed features (with attack) + ML	74.9	73.4	64.7	62.8	69.0
Proposed features (with attack) + SVM	**76.3**	**74.2**	**67.5**	**64.7**	**70.7**

Table 5. Accuracy (%) for instrument identification given the true pitches (real-world data).

The results over synthetic data and real-world data are shown in Tables 4 and 5 as a function of the number of instruments in the test signals. The proposed algorithm based on joint harmonic and attack features and SVM outperforms all other algorithms on all tasks. The resulting improvement is equal to 22%, 5% and 7% compared to MFCCs, source-filter features and our previous features on average. Including the attack features or using a SVM classifier improves the accuracy compared to considering harmonic features only or using ML classification, but only using both attack features and the SVM classifier provides the best performance for all test data.

Number of instruments	1	2	3	4	Average
MFCC + SVM	58.5	52.1	46.0	34.3	47.7
Source-filter + ML	70.4	63.8	58.4	54.0	61.7
Proposed features (without attack) + ML	71.4	65.1	60.5	56.4	63.3
Proposed features (without attack) + SVM	73.5	69.5	63.6	59.8	66.6
Proposed features (with attack) + ML	72.6	68.4	63.4	59.0	65.9
Proposed features (with attack) + SVM	**75.4**	**70.2**	**65.7**	**61.9**	**68.3**

Table 6. F-measure (%) for joint pitch estimation and instrument identification (synthetic data).

E. Evaluation of joint pitch estimation and instrument identification

Finally, as a third experiment, we use the proposed multiple pitch estimation algorithm to estimate all note parameters and jointly evaluate multiple pitch estimation and instrument identification. An estimated note is considered as correct when its pitch, onset and instrument are all correct. The proposed algorithm features are compared with the same alternative features and classifiers as in the second experiment.

The results over synthetic data and real-world data are shown in Tables 6 and 7. Again, the proposed algorithm outperforms all other algorithms on all tasks. The resulting improvement is equal to 20% and 6% compared to MFCCs and source-filter features on average.

Number of instruments	1	2	3	4	Average
MFCC + SVM	35.8	32.0	27.7	20.5	29.0
Source-filter + ML	47.5	45.4	40.1	36.0	42.3
Proposed features (without attack) + ML	48.0	44.3	39.4	36.5	42.1
Proposed features (without attack) + SVM	51.4	48.5	44.2	40.9	46.3
Proposed features (with attack) + ML	50.5	49.2	43.2	39.0	45.5
Proposed features (with attack) + SVM	**52.7**	**50.5**	**47.0**	**42.4**	**48.2**

Table 7. F-measure (%) for joint pitch estimation and instrument identification (real-world data).

V. CONCLUSION

In this article, we proposed an algorithm for polyphonic pitch estimation and instrument identification based on joint modeling of sustained and attack sounds. The proposed algorithm is based on a spectro-temporal GMM model of each note, whose parameters are estimated by the EM algorithm. These parameters are then subject to a logarithmic transformation and to PCA so as to obtain a low-dimension timbre feature vector. Finally, SVM classifiers are trained from the extracted features and used for musical instrument recognition. The proposed algorithm was shown to outperform certain state-of-the-art algorithms based on harmonic modeling alone both for multiple pitch estimation and instrument identification. Future work will focus on explicitly accounting for overlapping partials so as to further improve the robustness of the proposed timbre features.

VI. ACKNOWLEDGMENT

The authors would like to thank Anssi Klapuri for providing the code of his source-filter model [2]. This work was supported by INRIA under the Associate Team Program VERSAMUS (http://versamus.inria.fr/).

APPENDIX

The update equations of the parameters are as follows.

Joint harmonic and attack parameters:

$$l_{k,z}(x,t) = m_k\,(x,t)m_{k,z}(x,t)W(x,t) \tag{19}$$

Copyright IEEE 2011. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

$$c_k = \sum_z \iint_D l_{k,z}(x,t)dxdt \tag{20}$$

$$\tau_k = \frac{1}{c_k}\sum_z \iint_D (t - y\phi_k)\, l_{k,z}(x,t)dxdt \tag{21}$$

$$\begin{cases} a = \sum_z \iint_D y(t-\tau_k) l_{k,z}(x,t)dxdt \\ b = \sum_z \iint_D (t-\tau_k)^2 l_{k,z}(x,t)dxdt \end{cases} \tag{22}$$

$$\phi_k = \frac{-a + \left(a^2 + 4bc_k\right)^{1/2}}{2c_k} \tag{23}$$

Harmonic parameters:

$$w_{k,n} = \sum_y \iint_D l_{k,n,y}(x,t)dxdt \tag{24}$$

$$w_k = \sum_n w_{k,n} \tag{25}$$

$$\mu_k = \frac{1}{w_k}\sum_{n,y} \iint_D (x - \log n)^2\, l_{k,n,y}(x,t)dxdt \tag{26}$$

$$\sigma_k^2 = \frac{1}{w_k}\sum_{n,y} \iint_D \left(x - \mu_k - \log n\right)^2 l_{k,n,y}(x,t)dxdt \tag{27}$$

$$v_{k,n} = \frac{1}{d_v + w_k}\left(d_v \bar{v}_n + \sum_y \iint_D l_{k,n,y}(x,t)dxdt\right) \tag{28}$$

$$u_{k,n,y} = \frac{1}{d_u + w_{k,n}}\left(d_u \bar{u}_y + \iint_D l_{k,n,y}(x,t)dxdt\right) \tag{29}$$

Attack parameters:

$$\alpha_{k,j} = \iint_D l_{k,j}(x,t)dxdt \tag{30}$$

In these equations, $l_{k,n,y}(x,t)$ and $l_{k,j}(x,t)$ denote $l_{k,z}(x,t)$ when the zth Gaussian encodes the nth harmonic partial at instant y or the jth frequency subband of the attack, respectively. Furthermore, the value of y in (21) and (22) is assumed to be 0 for those Gaussians associated with the attack.

REFERENCES

[1] J. Burred, A. Röbel, and T. Sikora "Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds," IEEE Trans. on Audio, Speech, and Language Processing, 18(3):663-674, 2010.

[2] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 53-56, 2007.

[3] T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in Proc. IJCAI Workshop on Computational Auditory Scene Analysis, pp. 18–24, 1999.

[4] J. Eggink and G. J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," in Proc. Int. Symp. on Music Information Retrieval (ISMIR), 2003.

[5] W. M. Hartmann, "Pitch, periodicity, and auditory organization," Journal of the Acoustical Society of America, 100(6):3491-3502, 1996.

[6] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness", IEEE Trans. on Audio, Speech and Language Processing, 11(6):804-816, 2003.

[7] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," IEEE Trans. on Speech and Audio Processing, 11(3):229–241, 2003.

[8] T. Tolonen, M. Karjalainen, "A computationally efficient multipitch analysis model," IEEE Trans. on Speech and Audio Processing 8(6):708–716, 2000.

[9] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of western tonal music," Journal of the Acoustical Society of America, 119(4):2498–2517, 2006.

[10] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 728–731, 1993.

[11] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," EURASIP Journal on Advances in Signal Processing, vol. 2007, article ID 48317, 2007.

[12] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," Speech Communication, 43(4):311–329, 2004.

[13] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in Proc. Int. Conf. on Music Information Retrieval (ISMIR), pp.381-386, 2007.

[14] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," IEEE Trans. on Audio, Speech and Language Processing, 15(3):982–994, 2007.

[15] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," IEEE Trans. on Audio, Speech and Language Processing, 18(3):528-537, 2010.

[16] C. Yeh, A. Röbel and X. Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," IEEE Trans. on Audio, Speech and Language Processing, 18(6):1116-1126, 2010.

[17] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in Proc. Int. Conf. on Music Information Retrieval (ISMIR), pp.576-581, 2004.

[18] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," Journal of the Acoustical Society of America, 105(3):1933–1941, 1999.

[19] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," EURASIP Journal on Applied Signal Processing, 2003(1):5–14, 2003.

[20] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 753–756, 2000.

[21] M. A. Loureiro, H. B. De Paula, and H. C. Yehia, "Timbre classification of a single musical instrument," In Proc. Int. Conf. on Music Information Retrieval (ISMIR), 2004.

[22] C. Hourdin, G. Charbonneau, and T. Moussa, "A multidimensional scaling analysis of musical instruments' time-varying spectra," Computer Music Journal, 21(2):40–55, 1997.

[23] G. Sandell and W. Martens, "Perceptual evaluation of principal-component-based synthesis of musical timbres," Journal of the Audio Engineering Society, 43(12):1013–1028, 1995.

[24] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," EURASIP Journal on Advances in Signal Processing, vol.2007, Article ID 51979, 2007.

[25] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 57-60, 2007.

[26] J. Wu, Y. Kitano, S. Raczynski, S. Miyabe, T. Nishimoto, N. Ono, and S. Sagayama, "Musical instrument identification based on harmonic temporal timbre features," in Proc. Workshop on Statistical and Perceptual Audition (SAPA), pp. 7-12, 2010.

[27] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society B, 39(1):1–38, 1977.

[28] J. A. K. Suykens, "Nonlinear modeling and support vector machines," IEEE Instrumentation and Measurement Technology Conf., pp. 287-294, 2001.

[29] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in Proc. Int. Symp. on Music Information Retrieval (ISMIR), pp. 287–288, 2002.

[30] http://www.music.mcgill.ca/resources/mums/html/MUMS_audio.htm

[31] http://theremin.music.uiowa.edu/MIS.html.

[32] http://www.music-ir.org/mirex/wiki/2007:Multiple_Fundamental_Frequency_Estimation_%26_Tracking

[33] F. Zheng, G. Zhang and Z. Song, "Comparison of different implementations of MFCC," Journal of Computer Science & Technology, 16(6): 582–589, 2001.