A musically motivated mid-level representation for pitch estimation and musical audio source separation

Jean-Louis Durrieu, Bertrand David, *Member, IEEE* and Gaël Richard, *Senior Member, IEEE*

Abstract

When designing an audio processing system, the target tasks often influence the choice of a data representation or transformation. Low-level time-frequency representations such as the short-time Fourier transform (STFT) are popular, because they offer a meaningful insight on sound properties for a low computational cost. Conversely, when higher level semantics, such as pitch, timbre or phoneme, are sought after, representations usually tend to enhance their discriminative characteristics, at the expense of their invertibility. They become so-called mid-level representations. In this paper, a source/filter signal model which provides a mid-level representation is proposed. This representation makes the pitch content of the signal as well as some timbre information available, hence keeping as much information from the raw data as possible. This model is successfully used within a main melody extraction system and a lead instrument/accompaniment separation system. Both frameworks obtained top results at several international evaluation campaigns.

Index Terms

Non-negative Matrix Factorization (NMF), audio signal representation, pitch estimation, audio melody extraction, musical audio source separation

J.-L. Durrieu is with the Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Laboratories (LTS5), CH-1015 Lausanne, Switzerland; e-mail: jean-louis.durrieu@epfl.ch.

B. David and G. Richard are with Institut Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France; e-mail: firstname.lastname@telecom-paristech.fr.

This work was partly funded by the European project IST-Kspace and by the French Oseo project Quaero.

The authors would like to thank the anonymous reviewers for their detailed comments which helped to improve the quality of this paper.

I. INTRODUCTION

The high diversity of music signals is particularly challenging when designing systems for audio analysis, indexing or modeling. This diversity is not only due to the multiple production mechanisms and the wide tessitura of the involved musical instruments but also to the large palette of possible instrument combinations. It is also worth emphasizing that music signals are, in most cases, polyphonic (*e.g.* produced as mixtures of individual musical sources). This polyphony undoubtedly is one of the major bottlenecks of musical signal processing since it considerably limits the analysis capabilities. For instance, compared with pitch estimation on monophonic signals, multiple pitch estimation adds several other challenges such as dealing with partial overlapping between concurrent sounds or the estimation of the number of notes.

When dealing with polyphonic signals, two strategies can be undertaken: either the whole signal is processed with a direct extraction of information, or it is split in several individual components ideally hypothesized as monophonic streams. Examples of the first case include multiple pitch analysis [1] or cover song detection [2]. For the second case, applications range from instrument recognition [3] to lyrics-to-audio alignment [4]. A third alternative strategy is emerging, following the latter strategy but without explicitly performing any separation. It consists in defining a mid-level representation that facilitates the subsequent processing (see for example [5] for tempo estimation, [6], [7] for instrument recognition and pitch estimation, or [8] for genre classification). Compared to traditional time domain and frequency domain representations (such as the short-time Fourier transform or STFT), mid-level representations are often viewed as a signal transformation from which a collection of indicators are extracted and indexed by their time instants. These indicators generally tend to emphasize higher semantics than the energy in the time-frequency plane. In a number of cases, designing such a mid-level representation consists in obtaining a salience function for the task at hand. For example, the representation proposed in [9] provides Instantaneous Frequency (IF) attractors, that can be compared to a reassigned STFT and are well adapted for audio processing systems based on sinusoidal models. It was in particular used in [10] for main melody extraction where the time domain signal is first mapped to its constituting sinusoids, and then transformed into the "pitch domain". As another example, the salience functions defined in [1] for multipitch extraction are directly obtained by a weighted alternative of the Harmonic Sum (HS).

Mid-level representations may also be built upon perceptually motivated time-frequency transforms such as gammatone filter-banks [11] or the constant-Q transform (CQT) [12], with logarithmically spaced frequency bands. Finally, numerous studies rely on a low-level feature extraction step which

can also be seen as a form of mid-level representation. Indeed, the widely used Mel-Frequency Cepstral Coefficients (MFCC) [13] globally provide information about the spectral envelope: more precisely, under mild assumptions, it carries information about the filter part of an underlying source/filter model. Other features such as those based on chroma vectors, also known as pitch class profiles (PCP), are often used as salience function for harmony related tasks, such as chord detection, tonality estimation or audio-to-score alignment tasks [14], [15].

A potential drawback of such representations is however a bias towards indexing tasks at the cost of information loss, thus limiting the potential for other applications such as sound transformation, compression or source separation. For example, the CQT neglects detailed information in high frequency bands and is not invertible. Similarly MFCCs and PCPs only describe particular aspects of sounds, which can be respectively interpreted as timbre and harmony, but only with rather strong assumptions, and are better defined when dealing with monophonic signals. It is therefore interesting to investigate models that allow to describe most, if not all, the contributions of the audio mixture: in this paper, we propose a novel "model-driven" mid-level representation. It combines some advantages of existing representations, especially invertibility allowing separation applications, with the access to semantically rich salience functions for pitch and timbre content analysis. Such a representation can be used in a number of applications including main melody extraction and separation, lyrics-to-audio alignment [4], music instrument recognition or singer identification [16], [17].

The proposed mid-level representation is an extension of previous works on the Instantaneous Mixture Model (IMM) presented in [18], [19]. The main contributions of this paper include:

- A more generic model for the IMM framework: the assumptions on the signal of [19] are relaxed. In particular, the part dedicated to the monophonic lead voice in [19] is here used to model several harmonic audio sources. This relaxation first allows to use the proposed model for a broader range of signals, and second provides new interpretations, leading to semantically rich mid-level representations;
- The extension of the initial model to multi-channel (*e.g.* stereo) signals;
- The incorporation of a specific dictionary element in the decomposition to allow the representation of unvoiced or noise components in the leading musical source;
- a detailed experimental validation for both main melody and lead instrument/accompaniment separation.

This paper is organized as follows. The signal model is presented in Section II, along with a brief

introduction to the estimation of the involved parameters and a discussion of the different facets of the proposed model. Then, in Section III, we discuss three applications of the proposed representation. At last, in Section IV, concluding remarks are followed by perspectives for future work.

II. Signal model

The proposed signal model is based on previous studies on main melody extraction [18], [19]. Here, it is presented in a generalized framework, with a specific focus on the interpretation of the parameters and their potential use.

A. Model description

1) Generic model: The input music signal X is the instantaneous sum of two signals: a signal of interest and a residual. For our applications, the signal of interest often refers to a leading instrument which is pitched (such as a singing voice) while the residual refers to the remaining background music. The signal of interest is therefore denoted V and the residual M, by reference to "Voice" and "Music". In this section, the single-channel case is introduced, while stereo-channel signal processing is addressed in Section II-A2, as a specific instance of the generic model presented here.

In the proposed framework, the $F \times N$ STFT of the single-channel mixture, $\mathbf{X} = [x_{fn}]_{fn}$, is modeled through its squared magnitude or power spectrum: $\mathbf{S}^X$, the short-time power spectrum (STPS) of X. The analysis window size for the STFT is $L = 46.44$ms (2048 samples at 44100Hz) and the hop size is fixed to 5.8ms (256 samples at 44100Hz), resulting in N analysis windows for the STFT. The discrete Fourier transforms are performed on L points, after applying a sinebell weighting window to the frame. The first $F = L/2 + 1$ coefficients (the bins of the positive frequencies) of frame n are stored as the n^{th} column of $\mathbf{X}$.

V and M are assumed independent one from the other, and the STPS of their sum is therefore assumed to be the sum of their STPS's:

$$\mathbf{S}^X = \mathbf{S}^V + \mathbf{S}^M \tag{1}$$

In this section, the signal of interest is defined in a broad sense, and will be specified for target applications in Section II-A2. This general definition however allows to better understand the principle of the proposed method and enhances the applicability of the model to a wider range of applications.

Since we are here interested in analyzing the polyphonic content of music signals, V is assumed to be generated by one or more harmonic instruments. Each frame n of V is characterized by its power

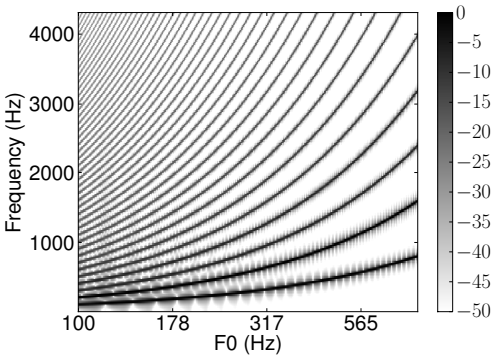


Fig. 1: Matrix $\mathbf{W}^{F_0}$, with KLGLOTT88 source model, $F_{\min}$ =100, $F_{\max}$ =800Hz, and U_{st} =20 (values used for the experiments, except for Fig. 5a, where $F_{\max}$ =2500Hz). Energy in dB.

spectrum for each frequency bin f, denoted s^V_{fn}. This time-frequency bin of the STPS is further modeled as follows: each frame is decomposed into an excitation spectrum ("source") $s^{F_0}_{fn}$ modulated by a spectral shaping envelope ("filter") s^{Φ}_{fn}, such that

$$s^V_{fn} = s^{\Phi}_{fn} s^{F_0}_{fn}. \tag{2}$$

F_0 recalls that the pitch information is included in the source part. Since it is aimed at building a pitch salience representation with possibly concurrent notes, the source part is further modeled as a combination of different hypothesized individual pitches:

$$s^{F_0}_{fn} = \sum_{u=1}^{U} h^{F_0}_{un} P_u(f) \tag{3}$$

where P_u, for all u, are fixed spectral shapes and $h^{F_0}_{un} \geq 0$. P_u can be any kind of spectral shape, for instance a spectral shape designed to correspond to a typical sound. Here, for each u, a fundamental frequency (or F0) $\mathcal{F}(u)$ is chosen, and P_u is then generated such that it is the power spectrum of a glottal signal with F0 equal to $\mathcal{F}(u)$, using the glottal source model KLGLOTT88 [20]. For convenience, the resulting harmonic spectral "combs" are then stored as the columns of the $F \times U$ dictionary matrix $\mathbf{W}^{F_0}$, such that $w^{F_0}_{fu} = P_u(f)$. $\mathcal{F}$ varies from $F_{\min}$ to $F_{\max}$, logarithmically every $1/U_{\text{st}}$ semitone:

$$\mathcal{F}(u) = 2^{\frac{u-1}{12U_{\text{st}}}} F_{\min}, \forall u = 1 \dots U \tag{4}$$

Note that when $U_{\text{st}} = 1$ and with $F_{\min}$ such that $\mathcal{F}(69) = 440$Hz, $\mathcal{F}$ is the mapping of a MIDI code number to its corresponding F0 in Hz. An example of a dictionary $\mathbf{W}^{F_0}$ is given on Fig. 1.

The "filter" part s^{Φ}_{fn} aims at providing more flexibility to the model, adapting it to a variety of possible instances (recording conditions, velocity of the played notes, intonations for a voice, etc.). It is

then decomposed into a linear combination of "smooth" filters $\Phi_k(f)$. The smoothness of these filters is controlled by generating them as a weighted sum of smooth spectral atoms $\Gamma_p(f)$:

$$s^{\Phi}_{fn} = \sum_{k=1}^{K} h^{\Phi}_{kn} \Phi_k(f) = \sum_{k=1}^{K} h^{\Phi}_{kn} \left(\sum_{p=1}^{P} h^{\Gamma}_{pk} \Gamma_p(f) \right), \tag{5}$$

where $h^{\Phi}_{kn} \geq 0$ and $h^{\Gamma}_{pk} \geq 0$. Contrary to P_u, Γ_p is constrained to be a smooth elementary envelope, describing broadband frequency behaviors. The decomposition onto the Γ_p function family therefore allows to catch a global spectral envelope. More precisely, the whole signal is described onto $K = 10$ spectral "envelopes" $\Phi_k(f)$, which are in turn decomposed onto the $P = 30$ smooth elementary envelopes Γ_p. A sensible choice for Γ_p is to use Hann functions overlapping at 75%, covering the whole frequency range, with centers linearly spaced in frequency. This can be seen as sub-sampling the spectral envelope, implicitly enforcing the smoothness of the estimated envelopes[1]. The choice of P fixes the frequency bands of these Hann functions. $P = 30$ allows to use functions that are narrow enough to describe a wide range of smooth envelopes, yet wide enough to avoid to capture spectra composed of isolated harmonics. Similarly to $\mathbf{W}^{F_0}$, we define $\mathbf{W}^{\Gamma}$ such that $w^{\Gamma}_{fp} = \Gamma_p(f)$ and $\mathbf{W}^{\Phi}$ such that $w^{\Phi}_{fk} = \Phi_k(f)$. The chosen $\mathbf{W}^{\Gamma}$ family and two examples of $\mathbf{w}^{\Phi}_k$ are illustrated on Fig. 2.

Finally, V is modeled such that:

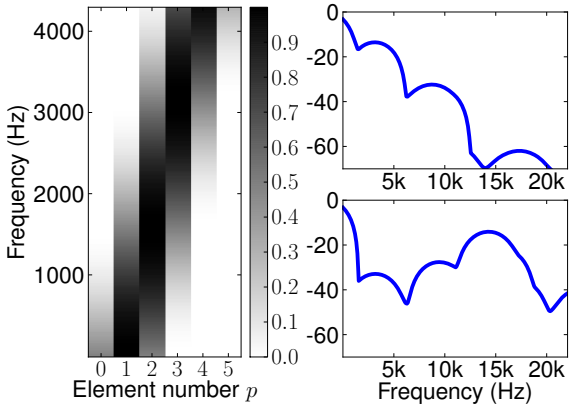
$$s^{V}_{fn} = \left(\sum_{k=1}^{K} h^{\Phi}_{kn} \sum_{p=1}^{P} h^{\Gamma}_{pk} \Gamma_p(f) \right) \left(\sum_{u=1}^{U} h^{F_0}_{un} P_u(f) \right) \tag{6}$$

In Eq. (6), the amplitude coefficients $h^{F_0}_{un}$, h^{Φ}_{kn} and h^{Γ}_{pk} give the decomposition of the signal onto the aforementioned dictionaries. They are estimated from the input signal, and respectively form the amplitude matrices $\mathbf{H}^{F_0}$ $(U \times N)$, $\mathbf{H}^{\Phi}$ $(K \times N)$ and $\mathbf{H}^{\Gamma}$ $(P \times K)$. These matrix conventions allow to write $\mathbf{S}^{V}$ in a compact way, underlining the link between the proposed framework and Non-negative Matrix Factorization (NMF) [21]:

$$\mathbf{S}^{V} = \underbrace{(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi})}_{\mathbf{S}^{\Phi}} \bullet \underbrace{(\mathbf{W}^{F_0}\mathbf{H}^{F_0})}_{\mathbf{S}^{F_0}}, \tag{7}$$

where the symbol $\bullet$ represents the Hadamard product. Matrices $\mathbf{S}^{\Phi}$ and $\mathbf{S}^{F_0}$ therefore capture different characteristics of the input signal of interest: they respectively catch the spectral envelope (related to timbre properties) and the pitch content, for each frame. Indeed, the purpose of this structure is to catch in $\mathbf{H}^{F_0}$ a pitch information which is independent from the timbre information, and conversely for $\mathbf{S}^{\Phi}$.

[1]Other bases were proposed as in [3], with logarithmically spaced centers, motivated by perceptual principles or by the physical properties of the sounds. Our choice, however, allows a broader variability in the spectral envelope.



(a) (b)

Fig. 2: Filter part: (a) $\mathbf{W}^{\Gamma}$, $P = 30$ Hann functions, overlap of 75%. The corresponding frequency band for each function is about 3000Hz. Only the first elements of the matrix, with non-null energy in the visible frequency bands are shown here. (b) 2 elements of $\mathbf{W}^{\Phi}$, in dB.

For instance, when a singer sings an A4 (440Hz) note, but sings different vowels, *e.g.* a `[a]` at a frame n_1 and a `[e]` at a frame n_2, then we would expect the columns $\mathbf{h}_{n_1}^{F_0}$ and $\mathbf{h}_{n_2}^{F_0}$ to roughly contain comparable values, while the spectral envelopes $\mathbf{s}_{n_1}^{\Phi}$ and $\mathbf{s}_{n_2}^{\Phi}$ should be rather different and characteristic of the pronounced vowel. This model is therefore remotely related to a **source/filter model** [22], which takes an almost literal meaning for the special case addressed in Section II-A2[2].

The STPS of the mixture X is therefore modeled as:

$$\mathbf{S}^X = (\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi}) \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0}) + \mathbf{S}^M \tag{8}$$

We refer to this model as the **Instantaneous Mixture Model (IMM)**: the signal X is indeed assumed to be the instantaneous mixture of the different contributions. This, in particular, means that the model does not explicitly take into account reverberations or echoes: these are rejected to the residual part. We also define our model in contrast with an alternative framework, the Gaussian Scaled Mixture Model (GSMM) [19], for which only one of the contributions is active at each frame.

2) Specific model for main instrument+accompaniment modeling: The IMM is particularly well suited for signals containing a main harmonic instrument, such as a singing voice V, played by a single

[2]According to [22], the filter should be a cascade of filters (product in frequency) and not filters in parallel (*i.e.* sum of filters). However, in [23], it is shown that parallel filters can also be used to successfully synthesize formants.

monophonic instrument backed by an accompaniment M, played by other instruments. In such a case, the main voice STPS is parameterized by the above source/filter framework, while an unconstrained NMF model of order R is assumed for $\mathbf{S}^M$, as in [19]:

$$\mathbf{S}^V = \mathbf{S}^\Phi \bullet \mathbf{S}^{F_0} \text{ and } \mathbf{S}^M = \mathbf{W}^M \mathbf{H}^M \tag{9}$$

This model is then called the **"Voiced"-IMM (VIMM)**. Strictly speaking, $\mathbf{S}^V$, and in particular $\mathbf{H}^{F_0}$, should reflect the fact that V is generated by a single monophonic instrument: for a given frame n, the vector $\mathbf{h}_n^{F_0}$ should contain only one non-null coefficient. This condition could for instance be controlled during the parameter estimation of Section II-B. However, as in [19], this constraint is applied in iterative steps, as shown on the general block diagram in Fig. 4 and further explained in Section III-B for melody estimation and Section III-C for lead instrument separation.

The aperiodic (or unvoiced for a singing voice) components of the main instrument are however not well modeled with the harmonic patterns of $\mathbf{W}^{F_0}$ and will mostly remain in the residual. To better capture these **unvoiced components**, it is proposed to extend the previous model by adding an element to the dictionary $\mathbf{W}^{F_0}$. This additional element, a vector of ones, is appended to the former $\mathbf{W}^{F_0}$ and corresponds to a white noise spectrum. The model including this element is called the **"Voiced+Unvoiced"-IMM (VUIMM)**. This extra element is only added in the source separation framework, once the melody of the lead instrument and all the other parameters have been accordingly estimated. This allows for an appropriate modeling of the unvoiced components of V while at the same time avoiding, to a certain extent, to capture the noise components of the other musical sources such as the drums.

At last, since many audio recordings are recorded on **several channels**, mostly with two channels, right $\mathcal{R}$ and left $\mathcal{L}$, we also propose an extension allowing to deal with such signals. The mixture is assumed to be an anechoic mixture of all the contributions: each source contributes to a channel only through the direct path between the corresponding microphone and the position of the source. We further neglect the delay of reception between microphones, hence reducing the spatial model to real amplitude gain differences between channels.

The main voice is placed at a single static position, and contributes to channel $\mathcal{C}$ with gain $\alpha_\mathcal{C} > 0$, while each of the R elements of M has its own position, with contribution gain $\beta_{\mathcal{C},r} > 0$. Let $\mathbf{B}_\mathcal{C} = \text{diag}(\left[\beta_{\mathcal{C},1}^2, \ldots, \beta_{\mathcal{C},R}^2\right])$, then, for each channel $\mathcal{C} \in \{\mathcal{R}, \mathcal{L}\}$:

$$\mathbf{S}^{V,\mathcal{C}} = \alpha_\mathcal{C}^2 \mathbf{S}^\Phi \bullet \mathbf{S}^{F_0} \text{ and } \mathbf{S}^{M,\mathcal{C}} = \mathbf{W}^M \mathbf{B}_\mathcal{C} \mathbf{H}^M \tag{10}$$

The constraints on the gains are given by:

$$\sum_{\mathcal{C}} \alpha_{\mathcal{C}} = 1 \text{ and } \sum_{\mathcal{C}} \beta_{\mathcal{C},r} = 1, \forall r = 1, \ldots, R \tag{11}$$

This model, although simple, still allows to consistently deal with multi-channel signals, even for more than 2 channels. It might however not be robust enough to discriminate the different contributions directly from their spatial positions. Our specific source/filter model for the main instrument compensates this simplicity by enforcing the signal of interest, V, to obey to the desired melody smoothness property [19].

B. Estimation of the parameters

In this section, we address the estimation of the parameter set $\mathbf{\Theta} = \{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{\Theta}^{M}\}$, such that $\mathbf{S}^{X}$, in the single-channel case, is the estimation of the observed STPS $\mathbf{S}^{X,o} = |\mathbf{X}|^2$:

$$\mathbf{S}^{X,o} \approx \mathbf{S}^{X} = \mathbf{S}^{\Phi} \bullet \mathbf{S}^{F_0} + \mathbf{S}^{M} \tag{12}$$

$\mathbf{S}^{X}$ is parameterized by $\mathbf{\Theta}$ and by the fixed dictionaries $\mathbf{W}^{\Gamma}$ and $\mathbf{W}^{F_0}$. $\mathbf{\Theta}^{M}$ is the set of variables which calibrates the residual STPS $\mathbf{S}^{M}$. The set $\mathbf{\Theta}$ is estimated as the $\widehat{\mathbf{\Theta}}$ that minimizes criterion $C(\mathbf{\Theta})$, defined as the divergence measure D between $\mathbf{S}^{X,o}$ and $\mathbf{S}^{X}$:

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} C(\mathbf{\Theta}) = \arg\min_{\mathbf{\Theta}} D(\mathbf{S}^{X,o}|\mathbf{S}^{X}) \tag{13}$$

We consider divergence measures of the following form:

$$D(\mathbf{A}|\mathbf{B}) = \sum_{ij} d(a_{ij}, b_{ij}) \tag{14}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two matrices with the same dimensions $I \times J$ and d is a scalar divergence measure. Typical measures for NMF methods are the Euclidean (EUC) distance d_{EUC}, the Kullback-Leibler (KL) divergence d_{KL} or the Itakura-Saito (IS) divergence d_{IS}. These divergences differ in properties but also in their interpretation on the signal model [24]. The β-divergence generalizes these divergences [25]:

$$d_{\beta}(a,b) = \begin{cases} a\frac{a^{\beta-1}-b^{\beta-1}}{\beta(\beta-1)} + b^{\beta-1}\frac{b-a}{\beta}, & \beta \in \mathbb{R}^{+}\backslash\{0,1\} \\ a\log\frac{a}{b} - a + b, & \beta = 1 \\ \frac{a}{b} - \log\frac{a}{b} - 1, & \beta = 0 \end{cases} \tag{15}$$

d_{β} therefore corresponds to d_{EUC}, d_{KL} and d_{IS} when β respectively equals 2, 1 and 0.

The chosen estimation algorithm relies on the multiplicative gradient principle developed in [21]. Some insights about its convergence properties can be found in [26]. The V(U)IMM estimation algorithm, where $\mathbf{\Theta}^{M} = \{\mathbf{W}^{M}, \mathbf{H}^{M}\}$, is an iterative algorithm for which the updating rules are given in Tab. I, where the Hadamard products are denoted by $\bullet$, the fractions are meant element by element, as well as the exponent

TABLE I: Single-channel IMM parameter estimation, with NMF accompaniment model for residual $\mathbf{S}^M$.

Initialize with random $\mathbf{\Theta}$ (*e.g.* modulus of values drawn from the standard normal distribution),

for $i = 1, \ldots, N_{\text{iter}}$ **do**

Update the matrices of $\mathbf{\Theta}$ in the following order, recomputing $\mathbf{S}^X$ after each of the following updates:

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{(\mathbf{W}^{F_0})^T(\mathbf{S}^\Phi \bullet (\mathbf{S}^X)^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,o})}{(\mathbf{W}^{F_0})^T(\mathbf{S}^\Phi \bullet (\mathbf{S}^X)^{\cdot(\beta-1)})}$$

$$\mathbf{H}^{\Phi} \leftarrow \mathbf{H}^{\Phi} \bullet \frac{(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma})^T(\mathbf{S}^{F_0} \bullet (\mathbf{S}^X)^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,o})}{(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma})^T(\mathbf{S}^{F_0} \bullet (\mathbf{S}^X)^{\cdot(\beta-1)})}$$

$$\mathbf{H}^{M} \leftarrow \mathbf{H}^{M} \bullet \frac{(\mathbf{W}^{M})^T((\mathbf{S}^X)^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,o})}{(\mathbf{W}^{M})^T(\mathbf{S}^X)^{\cdot(\beta-1)}}$$

$$\mathbf{H}^{\Gamma} \leftarrow \mathbf{H}^{\Gamma} \bullet \frac{(\mathbf{W}^{\Gamma})^T(\mathbf{S}^{F_0} \bullet (\mathbf{S}^X)^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,o})(\mathbf{H}^{\Phi})^T}{(\mathbf{W}^{\Gamma})^T(\mathbf{S}^{F_0} \bullet (\mathbf{S}^X)^{\cdot(\beta-1)})(\mathbf{H}^{\Phi})^T}$$

$$\mathbf{W}^{M} \leftarrow \mathbf{W}^{M} \bullet \frac{((\mathbf{S}^X)^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,o})(\mathbf{H}^{M})^T}{(\mathbf{S}^X)^{\cdot(\beta-1)}(\mathbf{H}^{M})^T}$$

end for

in $\mathbf{A}^{\cdot(\omega)}$. N_{iter} is the number of iterations for the gradient algorithm. The divergence value could be used to set a convergence condition, hence dynamically setting the optimal number of iterations. However, the link between this divergence value and the resulting score for the desired applications is sometimes not clear if not misleading [27]. In our experiments, several numbers of iterations were tried.

To derive the stereo-channel algorithm, the updating rules have to be modified. For instance, for $\mathbf{H}^{F_0}$, the updating formula should be:

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{(\mathbf{W}^{F_0})^T(\sum_{\mathcal{C}} \alpha_{\mathcal{C}}\mathbf{S}^\Phi \bullet (\mathbf{S}^{X,\mathcal{C}})^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,\mathcal{C}o})}{(\mathbf{W}^{F_0})^T(\sum_{\mathcal{C}} \alpha_{\mathcal{C}}\mathbf{S}^\Phi \bullet (\mathbf{S}^{X,\mathcal{C}})^{\cdot(\beta-1)})}$$

The updates for the channel gains are given by the following equations:

$$\alpha_{\mathcal{C}} \leftarrow \alpha_{\mathcal{C}} \frac{\text{sum}\left(\mathbf{S}^\Phi \bullet \mathbf{S}^{F_0} \bullet (\mathbf{S}^{X,\mathcal{C}})^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,\mathcal{C}o}\right)}{\text{sum}\left(\mathbf{S}^\Phi \bullet \mathbf{S}^{F_0} \bullet (\mathbf{S}^{X,\mathcal{C}})^{\cdot(\beta-1)}\right)}$$

$$\mathbf{B}_{\mathcal{C}} \leftarrow \mathbf{B}_{\mathcal{C}} \frac{(\mathbf{W}^{M})^T((\mathbf{S}^{X,\mathcal{C}})^{\cdot(\beta-2)} \bullet \mathbf{S}^{X,\mathcal{C}o})(\mathbf{H}^{M})^T}{(\mathbf{W}^{M})^T((\mathbf{S}^{X,\mathcal{C}})^{\cdot(\beta-1)})(\mathbf{H}^{M})^T}$$

where the operator sum(.) is the sum over all the elements of the input matrix.

We have more specifically focussed on the estimation for $\beta = 0$, that is when the β-divergence is the **Itakura-Saito (IS) divergence**. One can indeed show that the estimation of the parameters by minimizing the IS divergence is equivalent to assuming that the Fourier vector of each frame n follows a complex

Gaussian distribution, centered, with a diagonal covariance matrix whose diagonal is the vector $\mathbf{s}_n^X$ [24]. This view of the model is further studied in [19] and motivated by the applications to audio signals of Section III.

It is important to note that, because of the divisions in the formulas of Tab. I, the values of the parameters have to be controlled so as to avoid numerical errors such as divisions by zero. In addition, the indeterminacies related to the model and the chosen criterion can be avoided by normalizing the columns of $\mathbf{H}^\Gamma$, $\mathbf{H}^\Phi$ and $\mathbf{W}^M$. The columns of $\mathbf{W}^\Gamma$ and $\mathbf{W}^{F_0}$ should also be normalized, such that the values in $\mathbf{H}^{F_0}$ all have the same dynamics. At last, especially for the formulas used for stereo signal decompositions, the multiplicative gradient can be raised to some exponent between 0 and 2 as suggested in [26], [28]. A small value (0.1 for $\alpha_\mathcal{C}$ and $\mathbf{B}_\mathcal{C}$) of this exponent usually avoids an evolution of the parameters that would be, according to our tests, too chaotic and often converging towards the bounds of the search space, namely 0 or 1 for $\alpha_\mathcal{C}$ and $\beta_\mathcal{C}$. Although the convergence of the algorithm in Tab. I has not been proved yet, in practice, both the single- and stereo-channel algorithms decrease the criterion in Eq. (13) after each iteration. The resulting decompositions are also satisfying, as discussed in Section III.

C. Interpretation: three views of a model

In order to fully take advantage of the proposed signal model, it is important to understand what kind of representation it yields. The proposed model,

$$\mathbf{S}^X = \underbrace{(\mathbf{W}^\Gamma \mathbf{H}^\Gamma \mathbf{H}^\Phi)}_{\text{filter}} \bullet \underbrace{(\mathbf{W}^{F_0} \mathbf{H}^{F_0})}_{\text{source}} + \underbrace{(\mathbf{W}^M \mathbf{H}^M)}_{\text{residual}}, \tag{16}$$

exhibits several matrix multiplications. This fact makes our model very close to NMF models, in which, as we did previously, the decomposition of the signal onto a basis of spectral atoms is explicit. When using NMF-like methods for source separation [29], [30] or music analysis [31], most studies estimate both the spectral atoms and the activation coefficients, directly from the signal, in an "unsupervised" way.

Conversely, we develop in this section the reasons why supervised NMF methods such as ours are particularly interesting and why fixing some parameters is appropriate. An example of how the remaining parameters can be used is then sketched: the filter parameters indeed give an interesting insight on the timbre of the mixture.

1) (Non-)estimation of the pitch: As in [10], we propose a signal model that allows to decompose the signal onto several harmonic patterns: in this supervised case of NMF, we fix the spectral shapes $\mathbf{W}^{F_0}$ and only estimate the corresponding amplitudes $\mathbf{H}^{F_0}$. The estimation of the pitch with estimated spectral shapes may be unreliable, especially when these shapes are unconstrained, as shown in [32]. In

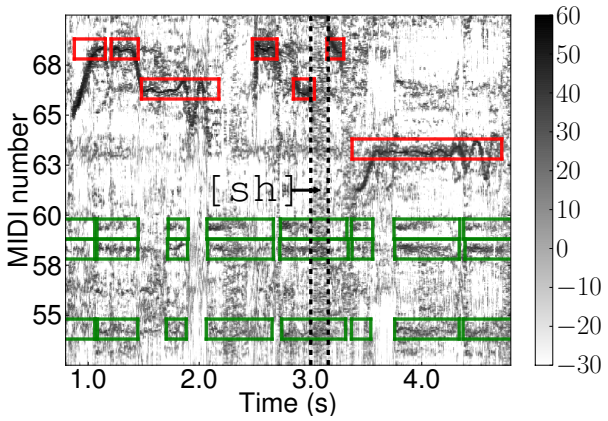


Fig. 3: Estimated $\mathbf{H}^{F_0}$ for the song by Tamy, from [33], the intended notes are drawn as red rectangles for the singing voice and green rectangles for the guitar.

the proposed framework, the pitch used to generate the basis $\mathbf{W}^{F_0}$ prevents from the need to determine the pitch of each estimated shape.

2) Pitch salience: The interpretation of the estimates we obtain are also similar to [10]: $h_{un}^{F_0}$ reflects the energy of the source component u at frame n, which is related to the fixed F0 value $\mathcal{F}(u)$. An example of an estimated $\mathbf{H}^{F_0}$ is given on Fig. 3. In [31], a similar pitch salience representation is proposed, but the reference pitch has to be estimated from the obtained spectral atom.

Following Goto [10], we believe that representations such as his or ours are very powerful: they provide salience functions that are specially designed to avoid classical octave estimation errors. Indeed, assuming that an ideal decomposition of the signal onto the given basis $\mathbf{W}^{F_0}$ exists and is sparse, the only non-zero coefficients correspond to sounds that are present in the mixture. However, due to the flexibility of the approaches, several spurious non-zero coefficients may occur, as seen on Fig. 3. For instance, at around 3s, the unvoiced component `[sh]` leads to an amplitude vector $\mathbf{h}_n^{F_0}$ with values relatively uniform: a wide band noise can indeed be roughly approximated by a weighted sum of harmonic combs. A post-processing is therefore needed, depending on the target application, as will be seen in Section III.

Such a representation should be compared with other pitch salience functions such as those proposed in [34], [35]. These approaches rely more on a perceptual basis for analysis, while decomposition approaches as the proposed one are analysis-by-synthesis approaches. The former approaches tend to focus on what is assumedly important in the signal, while the latter ones first model the signal, and then interpret the desired parameters. Indeed, our method first estimates the parameters, and then analyzes

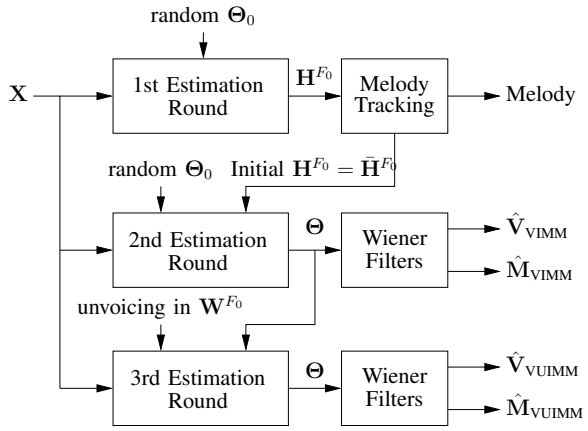


Fig. 4: Block diagram of the complete system: a melody estimation block followed by a first parameter re-estimation block for the VIMM separation, then a second re-estimation, for the VUIMM separation.

$\mathbf{H}^{F_0}$. This aspect is discussed in detail in Section III-A.

3) Timbre analysis: At last, using the filter parameters, we can extract a "spectral envelope" for the mixture. A first estimation round does not usually provide a meaningful envelope, since it corresponds, in our framework, to the global envelope of the mixture. We could estimate one envelope per frame and per component, as in [10]. The choice of a limited number of filters however comes from the desire to limit the range of possible timbres, hence limiting the number of instruments that are caught. Since the proposed model is originally designed to focus on only one instrument, a second round of estimation of the parameters, with an explicit restriction on the instrument to catch, provides a more useful estimation of the spectral envelope.

One can thereafter process these envelopes for instance to infer the lyrics, without an explicit separation of the signals as done in [4].

III. Applications

The block diagram of the complete system is given in Fig. 4. It illustrates each target application, namely the main melody estimation and the separation of the corresponding instrument (with both variants VIMM and VUIMM), respectively detailed in Section III-B and III-C. The use of the IMM as a mid-level representation is first discussed in Section III-A, along with its underlying interpretations, in order to illustrate the characteristics of the proposed model. In addition, some supporting material, sound examples, annotation files and source code can be found at http://www.durrieu.ch/research/jstsp2010.html.

A. IMM as a mid-level representation

It is possible to build, using the IMM, a new mid-level representation, interesting for pitch content analysis. Indeed, the matrix $\mathbf{H}^{F_0}$ gathers the necessary information for inspecting the pitched content, as discussed in Section II-C2.

In this section, the resulting F0 salience representation for a polyphonic music excerpt is first described in detail, highlighting both the advantages and drawbacks of the proposed model. These characteristics are then discussed against other representations based on harmonic summations (HS).

1) Detailed example: A polyphonic excerpt from "Three views of a secret" (J. Pastorious) is analyzed by the single-channel version of the IMM. The chosen excerpt is interesting because it exhibits a rather dense polyphony, with a leading theme played by several instruments, on two octaves. The parameters of the model for the analysis were fixed as follows, using the NMF model for $\mathbf{S}^M = \mathbf{W}^M\mathbf{H}^M$:

Parameter	$F_{\min}$	$F_{\max}$	U_{st}	K	P	R	N_{iter}
Value	100Hz	2.5kHz	20	20	35	1	60

Fig. 5a shows a detail of the resulting matrix $\mathbf{H}^{F_0}$. The presence of several concurrent tones, at t=2s, is evidenced by the strong contours marking the corresponding midi notes at numbers 81 (A5) and 93 (A6).

A first property arising from our decomposition-based model is the **sparsity of the representation**: in the best cases, non-null coefficients correspond to active F0s, with very localized peaks in the representation. It therefore provides a comfortable pitch representation, with a good **readability**. It is worth noting how the different F0 lines are distinguishable on Fig. 5a: in the 1s-2s interval, the evolutions of two different F0 lines at MIDI number 81 can be observed, and the F0 line at 93 is clearly different from those at 81, which tends to confirm that it is another distinct F0 line, and not an artefact of the algorithm which could be caused by the active F0s one octave below.

Second, the lead instruments are usually well represented in $\mathbf{H}^{F_0}$, and the main F0 lines on Fig. 5a indeed correspond to lead instruments. This makes the model particularly suitable for the applications explored in Section III-B and III-C, respectively main melody estimation and lead instrument separation.

For real world signals, as can be seen on Fig. 3 and Fig. 5a, the obtained $\mathbf{H}^{F_0}$ is however much noisier than wished for. Several reasons explain this result: first, since $\mathbf{W}^{F_0}$ is not a basis, the decomposition is not unique. The algorithm in Tab. I may therefore lead to local minima of the criterion in Eq. (13). For instance, coefficients of overtones of active F0s are very likely to be non-null. Second, since that criterion does not explicitly include sparsity constraints, the result itself is not guaranteed to be sparse.

Such constraints have been proposed, for instance in [36] or [37] and could easily be included in the proposed framework. However, for the applications presented in the following sections, the $\mathbf{H}^{F_0}$ matrices, estimated without such constraints, lead to satisfying results in terms of melody estimation and separation.

Other potential limitations lie in the possible discrepancies between the assumed properties of the model and the characteristics of real world signal. The chosen matrix $\mathbf{W}^{F_0}$ may be unadapted for inharmonic instruments, for which the correction brought by $\mathbf{S}^{\Phi}$ may be too restricted. Another limitation is the quantification of the F0 scale. If the real F0 does not belong to the scale, the method will likely use the neighboring quantified values to represent the sound. This can be seen on Fig. 3, where the fast pitch variations of the singer lead to a blurrier graph.

2) Qualitative comparison: In this section some of the advantages and drawbacks of the proposed representation are discussed and qualitatively compared to other representations.

Two other existing methods are discussed: the weighted harmonic sum (HS) of [1] and its improved version [35]. The HS method, implemented using the same parameters as in [1], is one of many methods using sub-harmonic summation as pitch salience function. For a given F0, it consists in adding the amplitude of the frequency bins that lie within a certain range from the expected harmonics of F0, weighted by a function of the harmonic number. The method described in [35] performs a further processing improving the salience of fundamental frequencies, while reducing the salience of "spurious" peaks which inherently appear in the representation[3]. This section mainly focusses on HS because many pitch salience functions rely on a similar principle, including [38], [39].

Fig. 5b and 5c respectively show the representation obtained with the weighted HS (WHS) [1] and the pitch salience of [35]. For all the figures, the colors have been scaled such that the result is visually satisfying. The F0 granularity (y-axis) of each representation was chosen or re-mapped to fit the above IMM choices.

First, the HS-based representations are inherently more dense than the previously presented $\mathbf{H}^{F_0}$ matrix: a single sinusoid indeed leads to a rather complex pattern [35], with many non-zero coefficients which do not correspond to any "true" F0. Harmonic sounds therefore correspond to sophisticated patterns, as illustrated on Fig. 5b and post-processing steps from HS results need to take into account these patterns. For instance, the problem of finding the optimal salience function in [35] can be seen as an inverse problem of finding the sources (and their F0s) which have generated these patterns. To a certain extent, as discussed earlier, the proposed method provides such a solution directly from the power spectrum.

[3]The authors are thankful to Prof. A. Klapuri for providing a Matlab implementation of his algorithm [35].

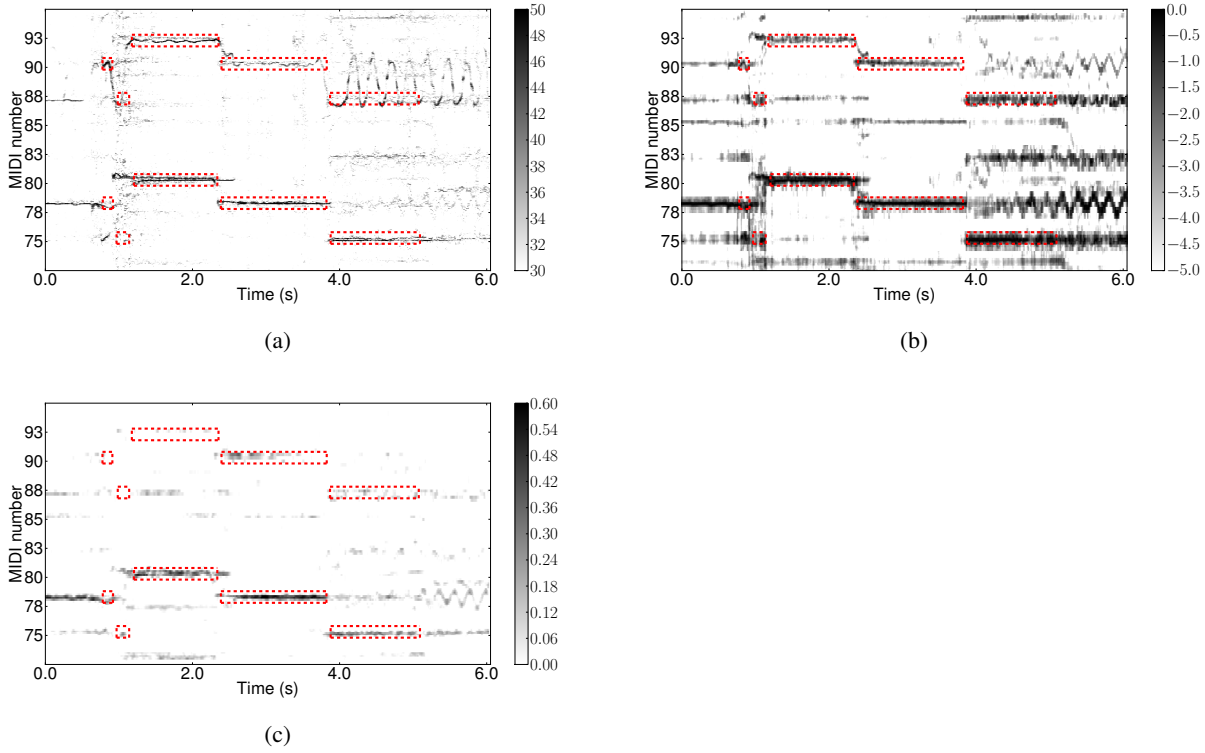


Fig. 5: Pitch representations for an excerpt from "Three Views Of A Secret" (J. Pastorius). The rectangles show some of the notes played by the predominant instruments (mostly trumpets): (a) Estimated $\mathbf{H}^{F_0}$ (60 iterations), each frame is normalized by its maximum, displayed in dB. ; (b) Salience function from [1] (HS with parametric weights), normalized frames, displayed in dB; (c) Salience function from [35] (HS with learnt weights), normalized frames.

Furthermore, on Fig. 5b or 5c, the F0 lines around note 81, in the 1s-2s interval, are hidden within one unique lobe. Although with each representation and appropriate algorithms, it would probably be possible to retrieve these lines, it is worth noting that with the IMM, such a task may be greatly simplified when the decomposition is "ideal". On that particular excerpt, the octave line, note 93, seems easier to distinguish from its lower octave on Fig. 5a than on the others: this is not surprising, since the sparsity induced by our decomposition approach very likely helps in obtaining F0 lines which are more localized than HS-based methods.

As mentioned above, the purpose of this section is not to formally compare or classify the above representations but rather to illustrate some properties of our framework. In practice, a mid-level representation has no intrinsic performance outside the framework within which it is developed: the proposed

representation could not replace that of the multiple F0 estimation system in [1] without proper changes in the whole processing chain. It however provides detailed and meaningful information about the pitch content (through $\mathbf{H}^{F_0}$), the timbre ($\mathbf{S}^{\Phi}$) and the residual ($\mathbf{S}^{M}$), tending towards invertibility, since the whole spectrogram is modeled (except for its phase). It then allows further processing steps such as main melody estimation or source separation based on pitch, as shown in the subsequent sections, in a unified framework. In comparison, the signal model used to estimate the melody in [40], also based on an HS salience, is different from the model used to remove the singing voice therein, which is based on an NMF decomposition approach.

At last, the main drawback of the proposed method, compared to HS-based ones, is its complexity. It requires quite a lot of memory to store all the matrices of the model. Furthermore, its use in a real-time system is not straightforward, since the estimation algorithm also requires many computationally heavy iterative steps.

B. Main melody extraction

The input song is here assumed to be composed of two main contributions: a leading voice, produced by a given instrument, playing a predominant melody line, and the complementary music, or accompaniment. Such a framework was already studied in [19], and is only briefly described here.

Ideally with the lead instrument/accompaniment model V(U)IMM, the estimated $\mathbf{S}^{V} = \mathbf{S}^{\Phi} \bullet \mathbf{S}^{F_0}$ represents the lead voice signal, while the estimated $\mathbf{S}^{M}$ represents the accompaniment. As seen previously, the estimation of $\mathbf{H}^{F_0}$ actually describes a pitch salience for the processed mixture signal: the desired main melody but also occasionally some accompaniment notes. For this reason, a post-processing step is necessary in order to extract the main melody from these other sources. We have proposed in [19] a simple yet satisfying system, shown in Fig. 4.

The matrix $\mathbf{H}^{F_0}$ is first estimated without any constraint on the number of active sources, such that more than one non-null pitch per frame can be active. As previously seen, the estimation is usually effective in keeping the most energetic F0s in $\mathbf{H}^{F_0}$. The melody F0s can therefore be detected using a Viterbi algorithm. The underlying assumption is that the melody line is smooth and that it can therefore be modeled thanks to a hidden Markov model (HMM), balancing the smoothness of the melody line with its energy dominance over the other active F0s. This results in the sequence of melody index $\{\xi_n\}_{n=1\ldots N} \in [1, U]^N$, such that the sequence of melody F0s is $\{\mathcal{F}(\xi_n)\}_{n=1\ldots N}$. In addition, as described in [19], the lead instrument present or absence was decided thanks to the energy of each frame.

In addition to the state-of-the-art results obtained at international campaigns [41], [42], we have tested

TABLE II: Precision (Pr), Recall (Rc) and F-measure (F) for melody estimation. The best result among all tests is reported, along with the corresponding parameters R and N_{iter}.

Song	R	N_{iter}	Pr (%)	Rc (%)	F (%)
Bearlin	100	50	55.0	76.1	64.0
Tamy	100	100	79.9	95.1	86.9
Another Dreamer	10	70	47.7	69.4	56.5
Fort Minor	1	100	36.1	45.5	40.2
"Ultimate"	20	100	52.6	68.2	59.4

our system on the development dataset from the SiSEC evaluation campaign for Professionally Produced Music Recordings [43] and [33]. We have annotated the 5 available stereo audio excerpts: using the proposed system, the melody was first estimated on the vocals, then manually corrected. Each annotation is therefore a sequence of melody F0s, in Hz, evaluated on windows of 46.44ms, every 5.8ms.

The proposed IMM model requires several manually set parameters for the leading instrument ($F_{\min}$, $F_{\max}$, U_{st}, P, the overlapping rate for $\mathbf{W}^{\Gamma}$, K), for the accompaniment (R) and N_{iter}. Several combinations of these parameters were tested in [19]. In this section, we propose to analyze the effect of the choice of R and N_{iter}. We set the other parameters such that $F_{\min} = 100$, $F_{\max} = 800$, $U_{\text{st}} = 20$, $P = 30$, $K = 10$ and an overlapping rate for $\mathbf{W}^{\Gamma}$ of 75%. The tested values of R were 1, 5, 10, 20, 32, 40, 50, 70, 100 and 200. N_{iter} was taken from 15, 30, 50, 70 and 100.

Each system outputs a sequence of F0, one per frame. The returned value is assumed to be a "True Positive" (TP) if it is within a semitone around the ground-truth value. "False Positives" (FP) are non-null returned values of F0 for silent frames, "True Negatives" (TN) count the number of times the system correctly detected silent frames, while "False Negatives" (FN) are the frames incorrectly detected as silent. The precision ($Pr = \frac{\#\text{TP}}{\#\text{TP}+\#\text{FP}}$), recall ($Rc = \frac{\#\text{TP}}{\#\text{TP}+\#\text{FN}}$) and F-measure ($F = 2\frac{PrRc}{Pr+Rc}$) were computed and analyzed.

Tab. II provides the best results for each of the SiSEC songs, obtained by the proposed single-channel VIMM system performed on the mean of the 2 channels. Fig. 6 also shows Pr with respect to N_{iter}.

We performed an analysis of variance (ANOVA) on the results. It appears that all the parameters (R, N_{iter} and the so-called "song" factor) have a non-negligible effect on the performances. Not surprisingly, the **differences between the songs** imply most of the differences in the results. As seen in Tab. II, the best overall result is obtained for the song by *Tamy* (with, in average, $P = 77\%$, $R = 92\%$ and

$F = 84\%$), while the worst result is obtained for the rap song by *Fort Minor* ($P = 31\%$, $R = 38\%$ and $F = 34\%$). The performance of the system clearly depends on the actual content of the excerpt: in the best case, there is only one singer, plus a guitar, with repetitive and soft chords, easily fitted by our accompaniment model. On the contrary, in the rap song, the accompaniment is dense and very present, and the "singer" recites, with voiced components that are less sustained, hence breaking the assumption of smoothness of the melody line, hence disturbing the algorithm.

A better discrimination between vocal and instrumental notes is also crucial to improve the system. This is particularly true for the song by *Tamy*, with a very good recall score, *i.e.* our algorithm finds many of the notes that were sung, but a relatively low precision, *i.e.* when the singer is silent, our algorithm tends to catch another instrument.

Furthermore, the ANOVA reveals that **the more elements there are in $\mathbf{W}^M$**, and the better the results are. Similarly, the performances **grow with the number of iterations**. It however seems that there exists some interaction between these two factors, such that, depending on the song, there are different optimal combinations, as shown in Tab. II. From our experiments, as a trade-off, the combination of $R = 40$ and $N_{\text{iter}} = 50$ can lead to satisfying results.

It is also interesting to note that, at low values of R, even with $R = 1$, and for a sufficient number of iterations, the melody estimation does not break down. This tends to confirm that a polyphonic signal can be modeled in $\mathbf{S}^V$, as presented in Section II-A1.

At last, a detailed analysis of the estimated pitch deviations to the ground-truth shows that, with a small number of iterations (under 10), octave errors seem to happen more often than with more iterations.

C. Lead instrument/Accompaniment separation

The proposed representation is also useful for audio source separation. We have demonstrated in [44] and [18] that the previously described melody estimation can be successfully used to separate the lead instrument from the accompaniment. The systems in these references, one single-channel and one stereo-channel system, are designed such that both the melody estimation and the separate signal estimations are done within a unified framework. Indeed, contrary to common representations in melody pitch estimation [10], [14], the proposed model provides a representation of the signal which does not miss important information, in particular the envelope of each note in the signal. Furthermore, as shown in [24], the statistical model underlying the choice of the IS divergence makes the computation of the Wiener filters straightforward, and the time-domain signals can be retrieved, by an overlap-add procedure, from

the estimated STFTs, for channel $\mathcal{C}$:

$$\widehat{\mathbf{V}}^{\mathcal{C}} = \frac{\widehat{\mathbf{S}}^{V,\mathcal{C}}}{\widehat{\mathbf{S}}^{X,\mathcal{C}}}\mathbf{X}^{\mathcal{C}} \text{ and } \widehat{\mathbf{M}}^{\mathcal{C}} = \frac{\widehat{\mathbf{S}}^{M,\mathcal{C}}}{\widehat{\mathbf{S}}^{X,\mathcal{C}}}\mathbf{X}^{\mathcal{C}} \tag{17}$$

We propose two main systems, based on the VIMM and the VUIMM models. As shown on Fig. 4, we first estimate the main melody $\{\xi_n\}$ (first block line), with the single-channel VIMM algorithm, performed on the mean of the 2 available channels. Then, $\overline{\mathbf{H}}^{F_0}$ is created from $\mathbf{H}^{F_0}$ and the estimated melody to simulate an "ideal" source coefficient matrix where only the coefficients on the path of the melody and around a quarter-tone thereof are non-null:

$$\begin{aligned}\overline{h}_{un}^{F_0} &= h_{un}^{F_0}, \text{ if } |u - \xi_n| < U_{\text{st}}/2 \\ &= 0, \text{ otherwise.}\end{aligned}$$

Such a $\overline{\mathbf{H}}^{F_0}$ matrix as initial $\mathbf{H}^{F_0}$ for the stereo-channel VIMM parameter estimation therefore limits the number of active F0s in V to 1 per frame, fulfilling the monophonic assumption for the lead instrument. This leads to the VIMM separation result (second block line). At last, we add the unvoiced basis element in $\mathbf{W}^{F_0}$ before a last parameter estimation round, leading to the VUIMM results (third block line).

One drawback of the proposed methods is the sub-optimal solution which consists in first estimating the melody, and then re-estimating the parameters to compute the Wiener masks. The joint estimation of the melody and of the separation parameters is however a difficult problem. The proposed solution, although sub-optimal, still provides good results, as discussed in this section. Improvements may rather come from a revised signal model, with more constraints narrowing the potential lead instrument, for instance, or directly integrating the HMM for the states of the source part during the estimation of the parameters.

The separation results for the above experiences, testing the number of iterations and the number of elements in the accompaniment part, were also computed, in terms of Signal-to-Distortion Ratio (SDR) [45]. The ANOVA suggests that the separation performance also depends on all the factors: R, N_{iter} as well as the differences between the songs. In particular, the SDR grows with R, up to about $R = 40$, then decreases as R grows. This effect can be explained by the imperfections of the model for the lead instrument V: when the harmonic combs in $\mathbf{W}^{F_0}$ or when the filter smoothness constraint is too rigid, some elements in $\mathbf{W}^{M}$ may be fitting the lead instrument part, replacing the estimation in $\mathbf{S}^{V}$. This will more likely occur when there are more elements in $\mathbf{W}^{M}$ than necessary. The parameter R should therefore be adapted to the processed song, in accordance with its actual content. Furthermore, a close inspection of the results suggests, as for Precision (Pr), that an optimal combination of R and

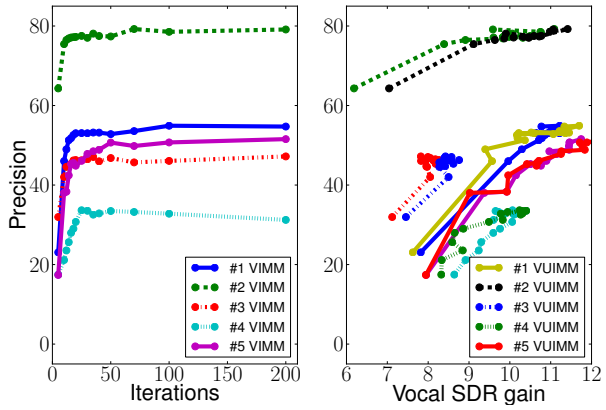


Fig. 6: Melody estimation Pr values against N_{iter}, and against the corresponding SDR gains.

N_{iter} exists for each song. Again, the values of $R = 40$ and $N_{\text{iter}} = 50$ seem to lead to a good trade-off for a general use of the systems for separation purposes.

In addition, Fig. 6 shows the melody estimation Precision Pr against N_{iter}, with $R = 40$, and the corresponding SDR gains with respect to Pr, for the 5 songs from SiSEC 2010. The SDR gains are computed as the SDR of the estimated vocals, to which the original Signal-to-Interference Ratio (SIR), here the vocal to accompaniment energy ratio, has been subtracted. As the melody Pr grows, the SDR gains also grow, but the absolute gain itself does not seem to be related in a linear way to Pr.

Tab. III provides the resulting vocal SDRs for several systems. First, the original SIR for each song is provided, in dB. The *Wiener* system corresponds to the estimation of the V and M using a Wiener mask computed thanks to the original separated tracks: in Eq. (17), $\widehat{\mathbf{S}}^{V,\mathcal{C}} = |\mathbf{V}^{\mathcal{C}}|^2$, $\widehat{\mathbf{S}}^{M,\mathcal{C}} = |\mathbf{M}^{\mathcal{C}}|^2$ and $\widehat{\mathbf{S}}^{X,\mathcal{C}} = \widehat{\mathbf{S}}^{V,\mathcal{C}} + \widehat{\mathbf{S}}^{M,\mathcal{C}}$. The *Oracle* systems compute the Wiener masks with V(U)IMM parameters directly estimated on the original individual tracks. Then, we ran the V(U)IMM systems with different initial conditions, providing the annotated melody (*Melody*), the voiced-vocal presence for each frame (*Pres.*, based on the annotated melody) and at last without prior knowledge. For the oracle systems, as well as for the *Pres.* systems, N_{iter} was fixed to 50. For the other V(U)IMM systems, the best results were chosen, among the different experiments with R and N_{iter}.

From Tab. III, several remarks can be made:

- All the SDR gains are positive, which shows that our systems always **improve the separation**, and rarely miss the desired lead instrument.
- The Wiener estimations are satisfying with regards to many criteria, although the underlying as-

TABLE III: Vocal SDR results for several systems: see text for details. The song labels are: #1 Bearlin, #2 Tamy, #3 Another Dreamer, #4 Fort Minor and #5 "Ultimate nz tour"

Song	#1	#2	#3	#4	#5
Original SIR	-5.3	0.2	-3.0	-7.2	-7.5
Wiener	10.9	13.7	11.5	11.4	10.1
VIMM - Oracle	8.5	12.4	8.4	7.4	6.4
VUIMM - Oracle	8.6	12.4	8.8	6.8	6.2
VIMM - Melody	6.9	11.3	5.7	4.2	4.9
VUIMM - Melody	7.5	11.8	6.6	4.5	5.5
VIMM - Pres.	5.9	10.5	5.3	2.6	4.4
VUIMM - Pres.	6.7	10.7	6.1	2.7	4.8
VIMM	6.0	11.3	5.3	3.3	4.3
VUIMM	6.5	11.6	5.8	3.3	4.9
Best SiSEC2010	x	x	3.1	3.9	2.6

sumption of independence between the audio sources is not always true.

- The performance differences between the **oracle systems** and V(U)IMM systems with annotated melody show that the melody alone is not a sufficient cue to achieve an optimal separation. A better estimation of the accompaniment part might explain the difference and future work should probably aim at better exploiting the accompaniment characteristics, such as repetitions or steadiness in comparison with vocal signals.
- The automatic systems sometimes perform better than the *Pres.* systems: the provided voiced-presence might be too restrictive for the systems. A fuzzier knowledge, provided as probabilities of presence, may improve the results, especially at the boundaries of presence/non-presence and for frames with lead instrument unvoiced parts.
- **VUIMM improves the result of VIMM**, thanks to the addition of the unvoiced element in $\mathbf{W}^{F_0}$. However, many spurious unvoiced sounds, especially drum elements, are caught in VUIMM. A pre-processing step reducing these effects could be held, for instance using [46]. Note however that for some signals, such as the rap song (by Fort Minor), the vocal signal seems easier to understand with VUIMM than with VIMM.
- The lowest SDR is obtained for the **rap song** by Fort Minor. In this case, the singing part is closer to speech than singing. For rap songs, a possible work-around is probably to explicitly take into account the repetitive aspect of the accompaniment: the lead vocal is predominant mostly because it

varies more than the rest, not because of its energy. Our systems are essentially based on the energy cue to detect the melody and are therefore less suitable for these specific signals.

- Our V(U)IMM systems achieve better results than the algorithms that participated to the **SiSEC 2010 evaluation campaign**, on the vocal SDR basis for songs #3 and #5 [43]. The interested reader is also invited to compare our separated vocals to those at http://www.irisa.fr/metiss/SiSEC10/professional/dev_eval.html. These encouraging results show that, when the singing style exhibits sufficiently smooth melody lines, a lead instrument separation system based on melody estimation achieves state-of-the-art results.

IV. Conclusion

The proposed method models an audio mixture power spectrum as a decomposition onto a dictionary of pre-defined spectral shapes. The algorithm to estimate the decomposition parameters, both for single and multiple channel cases, is also described. The model and its parameters can be successfully used as a mid-level representation of the mixture, displaying in our case its polyphonic pitch content, but also for applications such as melody extraction and lead instrument separation from its background accompaniment.

These applications obtain state-of-the-art results, as shown by international evaluation campaigns. However, there is room for improvement, especially in modeling specific singing styles, such as rap or spoken texts. Deciding whether an unvoiced sound belongs to the lead instrument, e.g. a singer, rather than to some other instrument, such as the drums, might actually be an ill-posed problem. Indeed, some people can rather genuinely imitate percussive sounds, which also means that some unvoiced parts of speech signals might be harder to discriminate from the accompaniment, without an explicit learning stage. Using several channels to take advantage of spatial information may help in the decision process.

At last, it is believed that the proposed model could be advantageously used in other scenarios than those explored in this article, such as lyrics recognition, chroma computation or multiple pitch extraction.

References

[1] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. of the International Conference on Music Information Retrieval*, Victoria, Canada, October 2006.

[2] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, pp. 1138–1151, August 2008.

[3] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. of the International Society for Music Information Retrieval Conference*, Kobe, Japan, October 26 - 30 2009, pp. 327 – 332.

[4] H. Fujihara and M. Goto, "Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, March 31-April 4 2008, pp. 69–72.

[5] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram – a mid-level tempo representation for music signals," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, March 14-19 2010, pp. 5522 – 5525.

[6] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 116–128, January 2008.

[7] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Musical instrument recognizer "instrogram" and its application to music retrieval based on instrument similarity," in *Proc. of the IEEE international symposium on multimedia*, San Diego, California, 2006, pp. 265–272.

[8] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 576–588, March 2010.

[9] T. Abe, T. Kobayashi, and S. Imai, "The IF spectrogram: a new spectral representation," *Proc. of the International Symposium on Simulation, Visualization and Auralization for Acoustic Research and Education*, vol. 97, pp. 423–430, 1997.

[10] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *ISCA Speech Communication*, vol. 43, no. 5, pp. 311 – 329, September 2004.

[11] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer, Perception Group - Advanced Technology Group, Tech. Rep., 1993.

[12] J. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[13] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357 – 366, August 1980.

[14] M. P. Ryynänen and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, March 2008.

[15] C. Joder, S. Essid, and G. Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, USA, March 14-19 2010, pp. 409 – 412.

[16] W. Tsai, H. Wang, and D. Rodgers, "Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal," in *Proc. of Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 1-4 2003, pp. 2993–2996.

[17] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, March 2010.

[18] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proc. of European Signal Processing Conference*, Glasgow, Scotland, August 24-28 2009.

[19] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from

polyphonic audio signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564 –575, March 2010.

[20] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, 1990.

[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems*, 2000, pp. 556–562.

[22] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1970.

[23] J. N. Holmes, "Formant synthesizers: Cascade or parallel?" *Speech Communication*, vol. 2, no. 4, pp. 251 – 273, December 1983.

[24] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, March 2009.

[25] A. Cichocki, R. Zdunek, and S. Amari, "Csiszars divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. of the International Conference on Independent Component Analysis and Blind Source Separation*, Charleston, South Carolina, USA, March 5-8 2006, pp. 32–39.

[26] R. Badeau, N. Bertin, and E. Vincent, "Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization," *IEEE Trans. on Neural Networks*, vol. 21, no. 12, pp. 1869–1881, December 2010.

[27] N. Bertin, C. Févotte, and R. Badeau, "A tempering approach for itakura-saito non-negative matrix factorization. with application to music transcription," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, R.O.C., April 2009, pp. 1545–1548.

[28] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, August 2010, pp. 283–288.

[29] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. of the International Symposium on Frontiers of Research on Speech and Music*, Mysore, India, January 2007.

[30] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550 – 563, March 2010.

[31] P. Smaragdis, "Relative pitch tracking of multiple arbitrary sounds," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3406–3413, May 2009.

[32] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, March 2010.

[33] M. Vinyes, "MTG MASS database,"
http://www.mtg.upf.edu/static/mass/resources, 2008.

[34] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music." *Journal of the Acoustic Society of America*, no. 111, pp. 1917–1930, April 2002.

[35] A. Klapuri, "A method for visualizing the pitch content of polyphonic music signals," in *Proc. of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, April 2009, pp. 615–620.

[36] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[37] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc.*

of IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, Nevada, USA, April 2008, pp. 2069 – 2072.

[38] K. Dressler, "Audio melody extraction for MIREX 2009," *extended abstract for the Music Information Retrieval Evaluation eXchange*, 2009.

[39] P. Cancela, E. López, and M. Rocamora, "Fan chirp transform for music representation," in *Proc. of the 13th International Conference on Digital Audio Effects*, Graz, Austria, September 6-10 2010, pp. 54–61.

[40] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *Proc. of IEEE International Conference on Multimedia and Expo*, Hannover, Germany, June 23-26 2008, pp. 1417–1420.

[41] MIREX, "Music Information Retrieval Evaluation eXchange," online: http://www.music-ir.org/mirex/2008/, September 2008.

[42] ——, "Music Information Retrieval Evaluation eXchange," online: http://www.music-ir.org/mirex/2009/, September 2009.

[43] SiSEC, "Professionally produced music recordings," Internet page: http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings, September 2010.

[44] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 19-24 2009, pp. 105–108.

[45] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, Paraty, Brazil, 15-18 March 2009, pp. 734–741.

[46] O. Gillet and G. Richard, "Transcription and Separation of Drum Signals From Polyphonic Music," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, March 2008.

Jean-Louis Durrieu was born on August 14th, 1982, in Saint-Denis, Reunion Island, France. He received the State Engineering degree and the Ph.D. degree, in the field of audio signal processing, from TELECOM ParisTech (formerly ENST), Paris, France, in 2006 and in 2010, respectively. He is currently a postdoctoral scientist in the Signal Processing Laboratory 5 (LTS5) at the EPFL.

His main research interests are statistical models for audio signals, with applications to language learning technologies, musical audio source separation and music information retrieval.

Gaël Richard received the State Engineering degree from TELECOM ParisTech (formerly ENST), Paris, France, in 1990, the Ph.D degree from LIMSI-CNRS, University of Paris-XI, in 1994 in speech synthesis and the *Habilitation à Diriger des Recherches* degree from the University of Paris XI in September 2001. After his Ph.D, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the speech processing group of Prof. J. Flanagan, where he explored innovative approaches for speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications, Bois d'Arcy, France, and for Philips Consumer Comunications, Montrouge, France. In particular, he was the project manager of several large-scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing of TELECOM ParisTech, where he is now full Professor in audio signal processing and Head of the Audio, Acoustics and Waves research group. He is co-author of over 80 papers, inventor in a number of patents and one of the experts of the European commission in the field of speech and audio signal processing. Pr. Richard is a senior member of IEEE and Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing.

Bertrand David (M'06) was born on March 12, 1967 in Paris, France. He received the M.Sc. degree from the University of Paris-Sud, in 1991, and the Agrégation, a competitive French examination for the recruitment of teachers, in the field of applied physics, from the École Normale Supérieure (ENS), Cachan. He received the Ph.D. degree from the University of Paris 6 in 1999, in the fields of musical acoustics and signal processing of musical signals.

He formerly taught in a graduate school in electrical engineering, computer science and communication. He also carried out industrial projects aiming at embarking a low complexity sound synthesizer. Since September 2001, he has worked as an Associate Professor with the Signal and Image Processing Department, TELECOM ParisTech (formerly ENST). His research interests include parametric methods for the analysis/synthesis of musical and mechanical signals, spectral parametrization and factorization, music information retrieval, and musical acoustics.